Amazon Bedrock: User Guide

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.
# Table of Contents

**What is Amazon Bedrock?** ........................................................................................................ 1
  - Features of Amazon Bedrock .................................................................................................. 1
  - Amazon Bedrock pricing ...................................................................................................... 2
  - Supported AWS Regions ........................................................................................................ 3
**Key definitions** ...................................................................................................................... 4
  - Basic concepts ..................................................................................................................... 4
  - Advanced features ................................................................................................................ 6
**Set up** ..................................................................................................................................... 7
  - Sign up for an AWS account ............................................................................................... 7
  - Create an administrative user ............................................................................................. 8
  - Grant programmatic access ............................................................................................... 9
  - Console access .................................................................................................................... 10
  - Model access ...................................................................................................................... 11
    - Add model access .......................................................................................................... 12
    - Remove model access .................................................................................................... 12
    - Control model access permissions ................................................................................ 13
  - API setup ................................................................................................................................ 15
    - Add model access .......................................................................................................... 16
    - Amazon Bedrock endpoints ........................................................................................... 16
    - Setting up the AWS CLI .................................................................................................. 16
  - AWS SDK setup ................................................................................................................... 17
    - Using SageMaker notebooks ......................................................................................... 19
  - Working with AWS SDKs ...................................................................................................... 21
**Foundation model information** ............................................................................................... 23
  - Using foundation models ..................................................................................................... 25
  - Get model information ........................................................................................................ 27
  - Model support by AWS Region ............................................................................................. 28
  - Model support by feature ..................................................................................................... 29
  - Model lifecycle ..................................................................................................................... 33
    - On-Demand, Provisioned Throughput, and model customization ................................ 34
    - Legacy versions ............................................................................................................... 34
  - Amazon Bedrock model IDs ................................................................................................. 35
    - Base models IDs (on-demand) ....................................................................................... 36
    - Base model IDs (for Provisioned Throughput) ............................................................. 38
Model inference parameters ................................................................. 40
  Amazon Titan models ........................................................................... 40
  Anthropic Claude models .................................................................... 79
  AI21 Labs Jurassic-2 models ............................................................... 99
  Cohere models ..................................................................................... 103
  Meta Llama 2 models .......................................................................... 115
  Mistral AI models ................................................................................ 119
  Stability.ai Diffusion models ............................................................... 124
Custom model hyperparameters .............................................................. 142
  Amazon Titan text models ................................................................... 142
  Amazon Titan Image Generator G1 ....................................................... 144
  Amazon Titan Multimodal Embeddings G1 ......................................... 145
  Cohere Command models .................................................................... 146
  Meta Llama 2 models .......................................................................... 149

Console overview ..................................................................................... 151
  Getting started .................................................................................... 151
  Foundation models ............................................................................. 152
  Playgrounds ....................................................................................... 152
  Safeguards ........................................................................................... 153
  Orchestration ....................................................................................... 153
  Assessment and deployment ............................................................... 153
  Model access ....................................................................................... 154
  Model invocation logging .................................................................... 154

Run model inference ................................................................................ 155
  Inference parameters ......................................................................... 157
    Randomness and diversity ................................................................. 157
    Length ............................................................................................... 159
  Playgrounds ........................................................................................ 159
    Chat playground ............................................................................... 160
    Text playground ............................................................................... 161
    Image playground ............................................................................ 161
    Use a playground ............................................................................ 162
  Run single-prompt inference .............................................................. 164
    Invoke model code examples ............................................................ 165
    Invoke model with streaming code example ..................................... 166
  Run batch inference .......................................................................... 167
Permissions ........................................................................................................... 168
Set up data .......................................................................................................... 171
Create a batch inference job .............................................................................. 172
Stop a batch inference job .................................................................................. 174
Get details about a batch inference job .............................................................. 175
List batch inference jobs .................................................................................... 177
Code samples ........................................................................................................ 179

**Prompt engineering guidelines** .................................................................... 184

Introduction ........................................................................................................... 184

  Additional prompt resources ............................................................................. 185

What is a prompt? .................................................................................................. 185

  Components of a prompt ................................................................................... 186

  Few-shot prompting vs. zero-shot prompting ...................................................... 187

  Prompt template ................................................................................................ 189

  Important notes on using Amazon Bedrock LLMs by API calls ...................... 190

What is prompt engineering? .............................................................................. 191

General guidelines for Amazon Bedrock LLM users ........................................ 192

  Design your prompt .......................................................................................... 192

  Use inference parameters ................................................................................ 193

  Detailed guidelines ........................................................................................... 194

  Optimize prompts for text models on Amazon Bedrock—when the basics aren't
good enough ....................................................................................................... 200

Prompt templates and examples for Amazon Bedrock text models .................. 203

  Text classification ............................................................................................ 203

  Question-answer, without context ................................................................... 206

  Question-answer, with context ......................................................................... 210

  Summarization .................................................................................................. 214

  Text generation .................................................................................................. 216

  Code generation ................................................................................................ 219

  Mathematics ...................................................................................................... 221

  Reasoning/logical thinking .............................................................................. 222

**Model evaluation** ........................................................................................... 225

Get started ............................................................................................................ 226

  Automatic model evaluations .......................................................................... 227

  Human worker based model evaluation jobs .................................................... 229

Model evaluation tasks ....................................................................................... 233
### Amazon Bedrock User Guide

**General text generation** ................................................................. 234  
**Text summarization** ................................................................. 235  
**Question and answer** ................................................................. 236  
**Text classification** ................................................................. 238  
**Input prompt datasets** .............................................................. 239  
  **Built-in prompt datasets** .......................................................... 240  
  **Custom prompt datasets** ......................................................... 243  
**Worker instructions** ................................................................. 247  
  **Rating methods** ................................................................. 248  
**Manage a work team** ................................................................. 254  
**Model evaluation job results** .................................................... 255  
  **Automated reports** ............................................................... 256  
  **Human report cards** .............................................................. 258  
  **Amazon S3 output** ............................................................... 265  
**Required permissions** ............................................................... 271  
  **Console permission requirements** ............................................. 272  
  **Service roles** ................................................................. 275  
  **CORS permission requirements** .............................................. 283  
  **Data encryption** ................................................................. 284  

**Knowledge bases for Amazon Bedrock** ........................................ 285  
**How it works** ............................................................................. 286  
**Supported regions and models** ................................................... 287  
**Prerequisites** ............................................................................. 288  
  **Set up a data source** ............................................................... 289  
  **Set up a vector index** .............................................................. 292  
**Create a knowledge base** ......................................................... 301  
  **Set up security configurations for your knowledge base** .......... 305  
**Sync your data sources** ............................................................ 310  
**Test a knowledge base** ............................................................ 312  
  **Query the knowledge base** ....................................................... 313  
  **Query configurations** ............................................................ 318  
**Manage a data source** ............................................................ 336  
  **View information about a data source** ..................................... 337  
  **Update a data source** .............................................................. 338  
  **Delete a data source** .............................................................. 340  
**Manage a knowledge base** ....................................................... 340
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>View information about a knowledge base</td>
<td>340</td>
</tr>
<tr>
<td>Update a knowledge base</td>
<td>342</td>
</tr>
<tr>
<td>Delete a knowledge base</td>
<td>342</td>
</tr>
<tr>
<td>Deploy a knowledge base</td>
<td>344</td>
</tr>
<tr>
<td><strong>Agents for Amazon Bedrock</strong></td>
<td>346</td>
</tr>
<tr>
<td>How it works</td>
<td>347</td>
</tr>
<tr>
<td>Build-time configuration</td>
<td>348</td>
</tr>
<tr>
<td>Runtime process</td>
<td>349</td>
</tr>
<tr>
<td>Supported regions and models</td>
<td>351</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>352</td>
</tr>
<tr>
<td>Create an agent</td>
<td>353</td>
</tr>
<tr>
<td>Create an action group</td>
<td>356</td>
</tr>
<tr>
<td>Define OpenAPI schemas</td>
<td>357</td>
</tr>
<tr>
<td>Define Lambda functions</td>
<td>365</td>
</tr>
<tr>
<td>Add an action group</td>
<td>370</td>
</tr>
<tr>
<td>Associate a knowledge base</td>
<td>372</td>
</tr>
<tr>
<td>Test an agent</td>
<td>374</td>
</tr>
<tr>
<td>Trace events</td>
<td>378</td>
</tr>
<tr>
<td>Manage an agent</td>
<td>386</td>
</tr>
<tr>
<td>View information about an agent</td>
<td>386</td>
</tr>
<tr>
<td>Edit an agent</td>
<td>388</td>
</tr>
<tr>
<td>Delete an agent</td>
<td>390</td>
</tr>
<tr>
<td>Manage action groups</td>
<td>391</td>
</tr>
<tr>
<td>Manage agent-knowledge bases associations</td>
<td>395</td>
</tr>
<tr>
<td>Customize an agent</td>
<td>398</td>
</tr>
<tr>
<td>Advanced prompts</td>
<td>399</td>
</tr>
<tr>
<td>Control session context</td>
<td>438</td>
</tr>
<tr>
<td>Deploy an agent</td>
<td>440</td>
</tr>
<tr>
<td>Manage versions</td>
<td>442</td>
</tr>
<tr>
<td>Manage aliases</td>
<td>444</td>
</tr>
<tr>
<td><strong>Custom models</strong></td>
<td>449</td>
</tr>
<tr>
<td>Supported regions and models</td>
<td>450</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>451</td>
</tr>
<tr>
<td>Prepare the datasets</td>
<td>452</td>
</tr>
<tr>
<td>(Optional) Set up a VPC</td>
<td>454</td>
</tr>
<tr>
<td>Submit a job</td>
<td>460</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Manage a job</td>
<td>463</td>
</tr>
<tr>
<td>Monitor a job</td>
<td>463</td>
</tr>
<tr>
<td>Stop a job</td>
<td>464</td>
</tr>
<tr>
<td>Analyze job results</td>
<td>465</td>
</tr>
<tr>
<td>Use a custom model</td>
<td>467</td>
</tr>
<tr>
<td>Code samples</td>
<td>467</td>
</tr>
<tr>
<td>Guidelines</td>
<td>479</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>479</td>
</tr>
<tr>
<td>Troubleshooting</td>
<td>480</td>
</tr>
<tr>
<td>Permissions issues</td>
<td>480</td>
</tr>
<tr>
<td>Data issues</td>
<td>481</td>
</tr>
<tr>
<td>Internal error</td>
<td>482</td>
</tr>
<tr>
<td><strong>Provisioned Throughput</strong></td>
<td>483</td>
</tr>
<tr>
<td>Supported regions and models</td>
<td>485</td>
</tr>
<tr>
<td>Procedures</td>
<td>487</td>
</tr>
<tr>
<td>Creating</td>
<td>487</td>
</tr>
<tr>
<td>Updating</td>
<td>487</td>
</tr>
<tr>
<td>Deleting</td>
<td>487</td>
</tr>
<tr>
<td>Running inference</td>
<td>488</td>
</tr>
<tr>
<td>Permissions</td>
<td>488</td>
</tr>
<tr>
<td>Console procedures</td>
<td>488</td>
</tr>
<tr>
<td>View Provisioned Throughput summary</td>
<td>488</td>
</tr>
<tr>
<td>Purchase Provisioned Throughput</td>
<td>489</td>
</tr>
<tr>
<td>View details of a Provisioned Throughput</td>
<td>490</td>
</tr>
<tr>
<td>Edit a Provisioned Throughput</td>
<td>490</td>
</tr>
<tr>
<td>Delete a Provisioned Throughput</td>
<td>490</td>
</tr>
<tr>
<td>API operations</td>
<td>491</td>
</tr>
<tr>
<td>Create Provisioned Throughput</td>
<td>491</td>
</tr>
<tr>
<td>Run inference using Provisioned Throughput</td>
<td>494</td>
</tr>
<tr>
<td>Update Provisioned Throughput</td>
<td>495</td>
</tr>
<tr>
<td>Get Provisioned Throughput</td>
<td>496</td>
</tr>
<tr>
<td>Delete Provisioned Throughput</td>
<td>496</td>
</tr>
<tr>
<td>List Provisioned Throughput resources</td>
<td>497</td>
</tr>
<tr>
<td><strong>Tag resources</strong></td>
<td>499</td>
</tr>
<tr>
<td>Use the console</td>
<td>500</td>
</tr>
<tr>
<td>Use the API</td>
<td>500</td>
</tr>
</tbody>
</table>
Infrastructure security ........................................................................................................ 573
Cross-service confused deputy prevention .................................................................. 574
Configuration and vulnerability analysis in Amazon Bedrock ...................................... 575
Use interface VPC endpoints (AWS PrivateLink) .......................................................... 526
  Considerations ............................................................................................................. 527
  Create an interface endpoint ...................................................................................... 527
  Create an endpoint policy ......................................................................................... 528
Monitor Amazon Bedrock .............................................................................................. 579
  Model invocation logging ......................................................................................... 579
    Set up an Amazon S3 destination ............................................................................ 579
    Set up CloudWatch Logs destination ...................................................................... 581
    Using the console ..................................................................................................... 583
    Using APIs with invocation logging ....................................................................... 584
Monitor with CloudWatch ............................................................................................. 584
  Runtime metrics ......................................................................................................... 584
  Logging CloudWatch metrics .................................................................................... 585
  Use CloudWatch metrics for Amazon Bedrock .......................................................... 586
  View Amazon Bedrock metrics .................................................................................. 586
Monitor events ................................................................................................................ 586
  How it works ................................................................................................................ 587
  EventBridge schema ................................................................................................. 588
  Rules and targets ....................................................................................................... 589
  Create a rule to handle Amazon Bedrock events ......................................................... 590
CloudTrail logs ................................................................................................................. 591
  Amazon Bedrock information in CloudTrail ............................................................. 592
  Amazon Bedrock data events in CloudTrail .............................................................. 592
  Amazon Bedrock management events in CloudTrail .................................................. 594
  Understanding Amazon Bedrock log file entries ....................................................... 594
Code examples .................................................................................................................. 596
Amazon Bedrock ............................................................................................................. 598
  Actions ......................................................................................................................... 603
  Scenarios ..................................................................................................................... 617
Amazon Bedrock Runtime ............................................................................................. 619
  Actions ......................................................................................................................... 625
  Scenarios ..................................................................................................................... 734
Agents for Amazon Bedrock ......................................................................................... 747
What is Amazon Bedrock?

Amazon Bedrock is a fully managed service that makes high-performing foundation models (FMs) from leading AI startups and Amazon available for your use through a unified API. You can choose from a wide range of foundation models to find the model that is best suited for your use case. Amazon Bedrock also offers a broad set of capabilities to build generative AI applications with security, privacy, and responsible AI. Using Amazon Bedrock, you can easily experiment with and evaluate top foundation models for your use cases, privately customize them with your data using techniques such as fine-tuning and Retrieval Augmented Generation (RAG), and build agents that execute tasks using your enterprise systems and data sources.

With Amazon Bedrock's serverless experience, you can get started quickly, privately customize foundation models with your own data, and easily and securely integrate and deploy them into your applications using AWS tools without having to manage any infrastructure.

Topics

• Features of Amazon Bedrock
• Amazon Bedrock pricing
• Supported AWS Regions
• Key definitions

Features of Amazon Bedrock

Take advantage of Amazon Bedrock foundation models to explore the following capabilities. To see feature limitations by Region, see Model support by AWS Region.

• **Experiment with prompts and configurations** – Run model inference by sending prompts using different configurations and foundation models to generate responses. You can use the API or the text, image, and chat playgrounds in the console to experiment in a graphical interface. When you're ready, set up your application to make requests to the InvokeModel APIs.

• **Augment response generation with information from your data sources** – Create knowledge bases by uploading data sources to be queried in order to augment a foundation model's generation of responses.
• **Create applications that reason through how to help a customer** – **Build agents** that use foundation models, make API calls, and (optionally) query knowledge bases in order to reason through and carry out tasks for your customers.

• **Adapt models to specific tasks and domains with training data** – **Customize an Amazon Bedrock foundation model** by providing training data for fine-tuning or continued-pretraining in order to adjust a model's parameters and improve its performance on specific tasks or in certain domains.

• **Improve your FM-based application's efficiency and output** – **Purchase Provisioned Throughput** for a foundation model in order to run inference on models more efficiently and at discounted rates.

• **Determine the best model for your use case** – **Evaluate outputs of different models** with built-in or custom prompt datasets to determine the model that is best suited for your application.

**Note**
Model evaluation is in preview release for Amazon Bedrock and is subject to change.

• **Prevent inappropriate or unwanted content** – Use **Guardrails for Amazon Bedrock** to implement safeguards for your generative AI applications.

**Note**
Guardrails for Amazon Bedrock is in limited preview release. To request access, contact your AWS account manager.

### Amazon Bedrock pricing

When you sign up for AWS, your AWS account is automatically signed up for all services in AWS, including Amazon Bedrock. However, you are charged only for the services that you use.

To see your bill, go to the Billing and Cost Management Dashboard in the [AWS Billing and Cost Management console](https://aws.amazon.com/billing/). To learn more about AWS account billing, see the [AWS Billing User Guide](https://docs.aws.amazon.com/billing/latest/using/purchase-provisioned-throughput.html). If you have questions concerning AWS billing and AWS accounts, contact [AWS Support](https://aws.amazon.com/support/).

With Amazon Bedrock, you pay to run inference on any of the third-party foundation models. Pricing is based on the volume of input tokens and output tokens, and on whether you have

---

*Amazon Bedrock pricing*
purchased provisioned throughput for the model. For more information, see the Model providers page in the Amazon Bedrock console. For each model, pricing is listed following the model version. For more information about purchasing Provisioned Throughput, see Provisioned Throughput.

For more information, see Amazon Bedrock Pricing.

**Supported AWS Regions**

For information about service endpoints for Regions that Amazon Bedrock supports, see Amazon Bedrock endpoints and quotas.

To see what foundation models each Region supports, refer to Model support by AWS Region.

⚠️ **Note**

All features are available in US East (N. Virginia) and US West (Oregon).

See the following table for features that are limited by region.

<table>
<thead>
<tr>
<th>Region</th>
<th>Model evaluation</th>
<th>Knowledge base</th>
<th>Agents</th>
<th>Fine-tuning (custom models)</th>
<th>Continued pre-training (custom models)</th>
<th>Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>US East (N. Virginia)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>US West (Oregon)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Asia Pacific (Singapore)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Asia Pacific (Tokyo)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
### Region Evaluation

<table>
<thead>
<tr>
<th>Region</th>
<th>Model evaluation</th>
<th>Knowledge base</th>
<th>Agents</th>
<th>Fine-tuning (custom models)</th>
<th>Continued pre-training (custom models)</th>
<th>Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe (Frankfurt)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AWS GovCloud (US-West)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (only for fine-tuned models, with no commitment term)</td>
</tr>
</tbody>
</table>

**Key definitions**

This chapter provides definitions for concepts that will help you understand what Amazon Bedrock offers and how it works. If you are a first-time user, you should first read through the basic concepts. Once you familiarize yourself with the basics of Amazon Bedrock, we recommend for you to explore the advanced concepts and features that Amazon Bedrock has to offer.

**Basic concepts**

The following list introduces you to the basic concepts of generative AI and Amazon Bedrock's fundamental capabilities.

- **Foundation model (FM)** – An AI model with a large number of parameters and trained on a massive amount of diverse data. A foundation model can generate a variety of responses for a wide range of use cases. Foundation models can generate text or image, and can also convert input into *embeddings*. Before you can use an Amazon Bedrock foundation model, you must [request access](#). For more information about foundation models, see [Supported foundation models in Amazon Bedrock](#).
• **Base model** – A foundation model that is packaged by a provider and ready to use. Amazon Bedrock offers a variety of industry-leading foundation models from leading providers. For more information, see [Supported foundation models in Amazon Bedrock](#).

• **Model inference** – The process of a foundation model generating an output (response) from a given input (prompt). For more information, see [Run model inference](#).

• **Prompt** – An input provided to a model to guide it to generate an appropriate response or output for the input. For example, a text prompt can consist of a single line for the model to respond to, or it can detail instructions or a task for the model to perform. The prompt can contain the context of the task, examples of outputs, or text for a model to use in its response. Prompts can be used to carry out tasks such as classification, question answering, code generation, creative writing, and more. For more information, see [Prompt engineering guidelines](#).

• **Token** – A sequence of characters that a model can interpret or predict as a single unit of meaning. For example, with text models, a token could correspond not just to a word, but also to a part of a word with grammatical meaning (such as "-ed"), a punctuation mark (such as "?"), or a common phrase (such as "a lot").

• **Model parameters** – Values that define a model and its behavior in interpreting input and generating responses. Model parameters are controlled and updated by providers. You can also update model parameters to create a new model through the process of [model customization](#).

• **Inference parameters** – Values that can be adjusted during model inference to influence a response. Inference parameters can affect how varied responses are and can also limit the length of a response or the occurrence of specified sequences. For more information and definitions of specific inference parameters, see [Inference parameters](#).

• **Playground** – A user-friendly graphical interface in the AWS Management Console in which you can experiment with running model inference to familiarize yourself with Amazon Bedrock. Use the playground to test out the effects of different models, configurations, and inference parameters on the responses generated for different prompts that you enter. For more information, see [Playgrounds](#).

• **Embedding** – The process of condensing information by transforming input into a vector of numerical values, known as the **embeddings**, in order to compare the similarity between different objects by using a shared numerical representation. For example, sentences can be compared to determine the similarity in meaning, images can be compared to determine visual similarity, or text and image can be compared to see if they’re relevant to each other. You can also combine text and image inputs into an averaged embeddings vector if it's relevant to your
use case. For more information, see Run model inference and Knowledge bases for Amazon Bedrock.

Advanced features

The following list introduces you to more advanced concepts that you can explore through using Amazon Bedrock.

- **Orchestration** – The process of coordinating between foundation models and enterprise data and applications in order to carry out a task. For more information, see Agents for Amazon Bedrock.
- **Agent** – An application that carry out orchestrations through cyclically interpreting inputs and producing outputs by using a foundation model. An agent can be used to carry out customer requests. For more information, see Agents for Amazon Bedrock.
- **Retrieval augmented generation (RAG)** – The process of querying and retrieving information from a data source in order to augment a generated response to a prompt. For more information, see Knowledge bases for Amazon Bedrock.
- **Model customization** – The process of using training data to adjust the model parameter values in a base model in order to create a custom model. Examples of model customization include Fine-tuning, which uses labeled data (inputs and corresponding outputs), and Continued Pre-training, which uses unlabeled data (inputs only) to adjust model parameters. For more information about model customization techniques available in Amazon Bedrock, see Custom models.
- **Hyperparameters** – Values that can be adjusted for model customization to control the training process and, consequently, the output custom model. For more information and definitions of specific hyperparameters, see Custom model hyperparameters.
- **Model evaluation** – The process of evaluating and comparing model outputs in order to determine the model that is best suited for a use case. For more information, see Model evaluation.
- **Provisioned Throughput** – A level of throughput that you purchase for a base or custom model in order to increase the amount and/or rate of tokens processed during model inference. When you purchase Provisioned Throughput for a model, a provisioned model is created that can be used to carry out model inference. For more information, see Provisioned Throughput.
Set up Amazon Bedrock

Before you use Amazon Bedrock for the first time, complete the following tasks. Once you have set up your account and requested model access in the console, you can set up the API.

⚠️ Important

Before you can use any of the foundation models, you must request access to that model. If you try to use the model (with the API or within the console) before you have requested access to it, you will receive an error message. For more information, see Model access.

Setup tasks

- Sign up for an AWS account
- Create an administrative user
- Grant programmatic access
- Console access
- Model access
- Set up the Amazon Bedrock API
- Using this service with an AWS SDK

Sign up for an AWS account

If you do not have an AWS account, complete the following steps to create one.

To sign up for an AWS account

2. Follow the online instructions.

Part of the sign-up procedure involves receiving a phone call and entering a verification code on the phone keypad.

When you sign up for an AWS account, an AWS account root user is created. The root user has access to all AWS services and resources in the account. As a security best practice, assign
administrative access to an administrative user, and use only the root user to perform tasks that require root user access.

AWS sends you a confirmation email after the sign-up process is complete. At any time, you can view your current account activity and manage your account by going to https://aws.amazon.com/ and choosing My Account.

Create an administrative user

After you sign up for an AWS account, secure your AWS account root user, enable AWS IAM Identity Center, and create an administrative user so that you don't use the root user for everyday tasks.

Secure your AWS account root user

1. Sign in to the AWS Management Console as the account owner by choosing Root user and entering your AWS account email address. On the next page, enter your password.

   For help signing in by using root user, see Signing in as the root user in the AWS Sign-In User Guide.

2. Turn on multi-factor authentication (MFA) for your root user.

   For instructions, see Enable a virtual MFA device for your AWS account root user (console) in the IAM User Guide.

Create an administrative user

1. Enable IAM Identity Center.

   For instructions, see Enabling AWS IAM Identity Center in the AWS IAM Identity Center User Guide.

2. In IAM Identity Center, grant administrative access to an administrative user.

   For a tutorial about using the IAM Identity Center directory as your identity source, see Configure user access with the default IAM Identity Center directory in the AWS IAM Identity Center User Guide.
Sign in as the administrative user

- To sign in with your IAM Identity Center user, use the sign-in URL that was sent to your email address when you created the IAM Identity Center user.

For help signing in using an IAM Identity Center user, see Signing in to the AWS access portal in the AWS Sign-In User Guide.

Grant programmatic access

Users need programmatic access if they want to interact with AWS outside of the AWS Management Console. The way to grant programmatic access depends on the type of user that’s accessing AWS.

To grant users programmatic access, choose one of the following options.

<table>
<thead>
<tr>
<th>Which user needs programmatic access?</th>
<th>To</th>
<th>By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workforce identity (Users managed in IAM Identity Center)</td>
<td>Use temporary credentials to sign programmatic requests to the AWS CLI, AWS SDKs, or AWS APIs.</td>
<td>Following the instructions for the interface that you want to use.</td>
</tr>
<tr>
<td>IAM</td>
<td>Use temporary credentials to sign programmatic requests</td>
<td>Following the instructions in Using temporary credentials</td>
</tr>
</tbody>
</table>

- For the AWS CLI, see Configuring the AWS CLI to use AWS IAM Identity Center in the AWS Command Line Interface User Guide.
- For AWS SDKs, tools, and AWS APIs, see IAM Identity Center authentication in the AWS SDKs and Tools Reference Guide.
<table>
<thead>
<tr>
<th>Which user needs programmatic access?</th>
<th>To</th>
<th>By</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAM</td>
<td>to the AWS CLI, AWS SDKs, or AWS APIs.</td>
<td>Is with AWS resources in the IAM User Guide.</td>
</tr>
</tbody>
</table>

IAM

(Not recommended)
Use long-term credentials to sign programmatic requests to the AWS CLI, AWS SDKs, or AWS APIs.

Following the instructions for the interface that you want to use.

- For the AWS CLI, see Authenticating using IAM user credentials in the AWS Command Line Interface User Guide.
- For AWS SDKs and tools, see Authenticate using long-term credentials in the AWS SDKs and Tools Reference Guide.
- For AWS APIs, see Managing access keys for IAM users in the IAM User Guide.

**Console access**

To access the Amazon Bedrock console and playground:

1. Sign in to your AWS account.
2. Navigate to: Amazon Bedrock console
3. Request model access by following the steps at Model access.
Model access

Access to Amazon Bedrock foundation models isn’t granted by default. In order to gain access to a foundation model, an IAM user with sufficient permissions needs to request access to it through the console. Once access is provided to a model, it is available for all users in the account.

To manage model access, select Model access at the bottom of the left navigation pane in the Amazon Bedrock management console. The model access page lets you view a list of available models, the output modality of the model, whether you have been granted access to it, and the End User License Agreement (EULA). You should review the EULA for terms and conditions of using a model before requesting access to it. For information about model pricing, refer to Amazon Bedrock Pricing.

Note
You can manage model access only through the console.

Topics
• Add model access
• Remove model access
• Control model access permissions

Add model access

Before you can use a foundation model in Amazon Bedrock, you must request access to it.

To request access to a model

1. On the Model access page, select Manage model access.
2. Select the checkboxes next to the models you want to add access to. To request access to all models belonging to a provider, select the check box next to the provider.

   Note
   You can't remove access from Titan models after requesting it. For Anthropic models, select Submit use case details and fill out the form. Once the details are submitted, you can then select the check box next to the Anthropic models that you want to request access to.

3. Select Save changes to request access. The changes may take several minutes to take place.

   Note
   Your use of Amazon Bedrock foundation models is subject to the seller's pricing terms, EULA, and the AWS service terms.

4. If your request is successful, the Access status changes to Access granted.

If you don't have permissions to request access to a model, an error banner appears. Contact your account administrator to ask them to request access to the model for you or to provide you permissions to request access to the model.

Remove model access

If you no longer need to use a foundation model, you can remove access to it.
1. On the **Model access** page, select **Manage model access**.

2. Select the checkboxes next to the models for which you want to remove access. To remove access for all models belonging to a provider, select the checkbox next to the provider.

3. Select **Save changes**.

4. You will be prompted to confirm you want to remove access to models. If you consent to the terms and select **Remove access**,

Note

The model may still be accessed through the API for some time after you complete this action while the changes propagate. To immediately remove access in the meantime, add an [IAM policy to a role to deny access to the model](#).

**Control model access permissions**

To control a role's permissions to request access to Amazon Bedrock models, attach an [IAM policy](#) to the role using any of the following [AWS Marketplace actions](#).

- aws-marketplace:Subscribe
- aws-marketplace:Unsubscribe
- aws-marketplace:ViewSubscriptions

For the aws-marketplace:Subscribe action only, you can use the aws-marketplace:ProductId condition key to limit subscription to specific models. The following table lists product IDs for Amazon Bedrock foundation models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Product ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI21 Labs Jurassic-2 Mid</td>
<td>1d288c71-65f9-489a-a3e2-9c7f4f6e6a85</td>
</tr>
<tr>
<td>Model</td>
<td>Product ID</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>AI21 Labs Jurassic-2 Ultra</td>
<td>cc0bdd50-279a-40d8-829c-4009b77a1fccc</td>
</tr>
<tr>
<td>Anthropic Claude</td>
<td>c468b48a-84df-43a4-8c46-8870630108a7</td>
</tr>
<tr>
<td>Anthropic Claude Instant</td>
<td>b0eb9475-3a2c-43d1-94d3-56756fd43737</td>
</tr>
<tr>
<td>Anthropic Claude 3 Sonnet</td>
<td>prod-6dw3qvchef7zy</td>
</tr>
<tr>
<td>Anthropic Claude 3 Haiku</td>
<td>prod-ozonys2hmmpeu</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>a61c46fe-1747-41aa-9af0-2e0ae8a9ce05</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>216b69fd-07d5-4c7b-866b-936456d68311</td>
</tr>
<tr>
<td>Cohere Embed (English)</td>
<td>b7568428-a1ab-46d8-bab3-37def50f6f6a</td>
</tr>
<tr>
<td>Cohere Embed (Multilingual)</td>
<td>38e55671-c3fe-4a44-9783-3584906e7cad</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>prod-ariujvyzvd2qy</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>prod-2c2yc2s3guhqy</td>
</tr>
<tr>
<td>Stable Diffusion XL 0.8</td>
<td>d0123e8d-50d6-4dba-8a26-3fed4899f388</td>
</tr>
<tr>
<td>Stable Diffusion XL 1.0</td>
<td>prod-2lvuzn4iy6n6o</td>
</tr>
</tbody>
</table>

The following is the format of the IAM policy you can attach to a role to control model access permissions. You can see an example at [Allow access to third-party model subscriptions](#).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow/Deny",
      "Action": [
        "aws-marketplace:Subscribe"
      ],
      "Resource": "*",
      "Condition": {
```
Set up the Amazon Bedrock API

This section describes how to set up your environment to make Amazon Bedrock API calls and provides examples of common use-cases. You can access the Amazon Bedrock API using the AWS Command Line Interface (AWS CLI), an AWS SDK, or a SageMaker Notebook.

Before you can access Amazon Bedrock APIs, you need to request access to the foundation models that you plan to use.

For details about the API operations and parameters, see the Amazon Bedrock API Reference.

The following resources provide additional information about the Amazon Bedrock API.

- **AWS Command Line Interface**
  - Amazon Bedrock CLI commands
  - Amazon Bedrock Runtime CLI commands
  - Agents for Amazon Bedrock CLI commands
  - Agents for Amazon Bedrock Runtime CLI commands
Add model access

⚠️ Important
Before you can use any of the foundation models, you must request access to that model. If you try to use the model (with the API or within the console) before you have requested access to it, you will receive an error message. For more information, see Model access.

Amazon Bedrock endpoints

To connect programmatically to an AWS service, you use an endpoint. Refer to the Amazon Bedrock endpoints and quotas chapter in the AWS General Reference for information about the endpoints that you can use for Amazon Bedrock.

Amazon Bedrock provides the following service endpoints.

- **bedrock** – Contains control plane APIs for managing, training, and deploying models. For more information, see Amazon Bedrock Actions and Amazon Bedrock Data Types.
- **bedrock-runtime** – Contains data plane APIs for making inference requests for models hosted in Amazon Bedrock. For more information, see Amazon Bedrock Runtime Actions and Amazon Bedrock Runtime Data Types.
- **bedrock-agent** – Contains control plane APIs for creating and managing agents and knowledge bases. For more information, see Agents for Amazon Bedrock Actions and Agents for Amazon Bedrock Data Types.
- **bedrock-agent-runtime** – Contains data plane APIs for invoking agents and querying knowledge bases. For more information, see Agents for Amazon Bedrock Runtime Actions and Agents for Amazon Bedrock Runtime Data Types.

Setting up the AWS CLI

1. If you plan to use the CLI, install and configure the AWS CLI by following the steps at Install or update the latest version of the AWS Command Line Interface User Guide.
2. Configure your AWS credentials using the aws configure CLI command by following the steps at Configure the AWS CLI.
Refer to the following references for AWS CLI commands and operations:

- [Amazon Bedrock CLI commands](#)
- [Amazon Bedrock Runtime CLI commands](#)
- [Agents for Amazon Bedrock CLI commands](#)
- [Agents for Amazon Bedrock Runtime CLI commands](#)

**Setting up an AWS SDK**

AWS software development kits (SDKs) are available for many popular programming languages. Each SDK provides an API, code examples, and documentation that make it easier for developers to build applications in their preferred language. SDKs automatically perform useful tasks for you, such as:

- Cryptographically sign your service requests
- Retry requests
- Handle error responses

Refer to the following table to find general information about and code examples for each SDK, as well as the Amazon Bedrock API references for each SDK. You can also find code examples at [Code examples for Amazon Bedrock using AWS SDKs](#).

<table>
<thead>
<tr>
<th>SDK documentation</th>
<th>Code examples</th>
<th>Amazon Bedrock prefix</th>
<th>Amazon Bedrock runtime prefix</th>
<th>Agents for Amazon Bedrock prefix</th>
<th>Agents for Amazon Bedrock runtime prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS SDK for C++</td>
<td><a href="#">AWS SDK for C++ code examples</a></td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for Go</td>
<td><a href="#">AWS SDK for Go code examples</a></td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
</tbody>
</table>

AWS SDK setup
<table>
<thead>
<tr>
<th>SDK documentation</th>
<th>Code examples</th>
<th>Amazon Bedrock prefix</th>
<th>Amazon Bedrock runtime prefix</th>
<th>Agents for Amazon Bedrock prefix</th>
<th>Agents for Amazon Bedrock runtime prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS SDK for Java</td>
<td>AWS SDK for Java code examples</td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for JavaScript</td>
<td>AWS SDK for JavaScript code examples</td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for Kotlin</td>
<td>AWS SDK for Kotlin code examples</td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for .NET</td>
<td>AWS SDK for .NET code examples</td>
<td>Bedrock</td>
<td>Bedrock-runtime</td>
<td>Bedrock-agent</td>
<td>Bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for PHP</td>
<td>AWS SDK for PHP code examples</td>
<td>Bedrock</td>
<td>Bedrock-runtime</td>
<td>Bedrock-agent</td>
<td>Bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for Python (Boto3)</td>
<td>AWS SDK for Python (Boto3) code examples</td>
<td>bedrock</td>
<td>bedrock-runtime</td>
<td>bedrock-agent</td>
<td>bedrock-agent-runtime</td>
</tr>
<tr>
<td>AWS SDK for Ruby</td>
<td>AWS SDK for Ruby code examples</td>
<td>Bedrock</td>
<td>Bedrock-runtime</td>
<td>Bedrock-agent</td>
<td>Bedrock-agent-runtime</td>
</tr>
</tbody>
</table>
Using SageMaker notebooks

You can use the SDK for Python (Boto3) to invoke Amazon Bedrock API operations from a SageMaker notebook.

Configure the SageMaker role

Add Amazon Bedrock permissions to the IAM role that will use this SageMaker notebook.

From the IAM console, perform these steps:

1. Choose the IAM role, then choose Add Permissions and select Create Inline Policies from the dropdown list.
2. Include the following permission.

```json
{
   "Version": "2012-10-17",
   "Statement": [
```
Add the following permissions to the trust relationships.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        },
        {
            "Sid": "",
            "Effect": "Allow",
            "Principal": {
                "Service": "sagemaker.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
    ]
}
```

Test the Runtime setup

Add the following code to your notebook and run the code.

```python
import boto3
import json
bedrock = boto3.client(service_name='bedrock-runtime')

body = json.dumps({
    "prompt": "Human: explain black holes to 8th graders

Assistant:",
    "max_tokens_to_sample": 300,
    "temperature": 0.1,
})
```
Test the Amazon Bedrock setup

Add the following code to your notebook and run the code.

```python
import boto3
bedrock = boto3.client(service_name='bedrock')
bedrock.get_foundation_model(modelIdentifier='anthropic.claude-v2')
```

Using this service with an AWS SDK

AWS software development kits (SDKs) are available for many popular programming languages. Each SDK provides an API, code examples, and documentation that make it easier for developers to build applications in their preferred language.

<table>
<thead>
<tr>
<th>SDK documentation</th>
<th>Code examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS SDK for C++</td>
<td>AWS SDK for C++ code examples</td>
</tr>
<tr>
<td>AWS SDK for Go</td>
<td>AWS SDK for Go code examples</td>
</tr>
<tr>
<td>AWS SDK for Java</td>
<td>AWS SDK for Java code examples</td>
</tr>
<tr>
<td>AWS SDK for JavaScript</td>
<td>AWS SDK for JavaScript code examples</td>
</tr>
<tr>
<td>AWS SDK for Kotlin</td>
<td>AWS SDK for Kotlin code examples</td>
</tr>
<tr>
<td>SDK documentation</td>
<td>Code examples</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------</td>
</tr>
<tr>
<td>AWS SDK for .NET</td>
<td>AWS SDK for .NET code examples</td>
</tr>
<tr>
<td>AWS SDK for PHP</td>
<td>AWS SDK for PHP code examples</td>
</tr>
<tr>
<td>AWS SDK for Python (Boto3)</td>
<td>AWS SDK for Python (Boto3) code examples</td>
</tr>
<tr>
<td>AWS SDK for Ruby</td>
<td>AWS SDK for Ruby code examples</td>
</tr>
<tr>
<td>AWS SDK for Rust</td>
<td>AWS SDK for Rust code examples</td>
</tr>
<tr>
<td>AWS SDK for SAP ABAP</td>
<td>AWS SDK for SAP ABAP code examples</td>
</tr>
<tr>
<td>AWS SDK for Swift</td>
<td>AWS SDK for Swift code examples</td>
</tr>
</tbody>
</table>

**Example availability**

Can't find what you need? Request a code example by using the Provide feedback link at the bottom of this page.
## Supported foundation models in Amazon Bedrock

Amazon Bedrock supports foundation models (FMs) from the following providers. Select a link in the **Provider** column to see documentation for that provider.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Model</th>
<th>Input modalities</th>
<th>Output modalities</th>
<th>Inference parameters</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Titan Text G1 - Express</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Titan Text G1 - Lite</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Titan Image Generator G1</td>
<td>Text, Image</td>
<td>Image</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Titan Embeddings G1 - Text</td>
<td>Text</td>
<td>Embeddings</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Titan Multimodal Embeddings G1</td>
<td>Text, Image</td>
<td>Embeddings</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Claude Instant</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Claude 3 Sonnet</td>
<td>Text, Image</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Claude 3 Haiku</td>
<td>Text, Image</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td>Provider</td>
<td>Model</td>
<td>Input modalities</td>
<td>Output modalities</td>
<td>Inference parameters</td>
<td>Hyperparameters</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td><strong>AI21 Labs</strong></td>
<td>Jurassic-2 Mid</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Jurassic-2 Ultra</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Cohere</strong></td>
<td>Command</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Command Light</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Embed English</td>
<td>Text</td>
<td>Embeddings</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Embed Multilingual</td>
<td>Text</td>
<td>Embeddings</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Meta</strong></td>
<td>Llama 2 Chat 13B</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Llama 2 Chat 70B</td>
<td>Text</td>
<td>Text, Chat</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Llama 2 13B (see note below)</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Llama 2 70B (see note below)</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>Link</td>
</tr>
<tr>
<td><strong>Mistral AI</strong></td>
<td>Mistral 7B Instruct</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Mixtral 8X7B Instruct</td>
<td>Text</td>
<td>Text</td>
<td>Link</td>
<td>N/A</td>
</tr>
</tbody>
</table>
The following sections provide information about using foundation models and reference information for models.

**Topics**

- Using foundation models
- Get information about foundation models
- Model support by AWS Region
- Model support by feature
- Model lifecycle
- Amazon Bedrock model IDs
- Inference parameters for foundation models
- Custom model hyperparameters

**Using foundation models**

You must request access to a model before you can use it. After doing so, you can then use FMs in the following ways.

- **Run inference** by sending prompts to a model and generating responses. The playgrounds offer a user-friendly interface in the AWS Management Console for generating text, images, or chats. See the **Output modality** column to determine the models you can use in each playground.

---

**Note**

The Meta Llama 2 (non-chat) models can only be used after being customized and after purchasing Provisioned Throughput for them.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Model</th>
<th>Input modalities</th>
<th>Output modalities</th>
<th>Inference parameters</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability AI</td>
<td>Stable Diffusion XL</td>
<td>Text, Image</td>
<td>Image</td>
<td>Link</td>
<td>N/A</td>
</tr>
</tbody>
</table>
**Note**
The console playgrounds don't support running inference on embeddings models. Use the API to run inference on embeddings models.

- **Evaluate models** to compare outputs and determine the best model for your use-case.
- **Set up a knowledge base** with the help of an embeddings model. Then use a text model to generate responses to queries.
- **Create an agent** and use a model to run inference on prompts to carry out orchestration.
- **Customize a model** by feeding training and validation data to adjust model parameters for your use-case. To use a customized model, you must purchase **Provisioned Throughput** for it.
- **Purchase Provisioned Throughput** for a model to increase throughput for it.

To use an FM in the API, you need to determine the appropriate **model ID** to use.

<table>
<thead>
<tr>
<th>Use case</th>
<th>How to find the model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use a base model</td>
<td>Look up the ID in the <a href="#">base model IDs chart</a></td>
</tr>
<tr>
<td>Purchase Provisioned Throughput for a base model</td>
<td>Look up the ID in the <a href="#">model IDs for Provisioned Throughput chart</a> and use it as the modelId in the <a href="#">CreateProvisionedModelThroughput</a> request.</td>
</tr>
<tr>
<td>Purchase Provisioned Throughput for a custom model</td>
<td>Use the name of the custom model or its ARN as the modelId in the <a href="#">CreateProvisionedModelThroughput</a> request.</td>
</tr>
<tr>
<td>Use a provisioned model</td>
<td>After you create a Provisioned Throughput, it returns a provisionedModelArn. This ARN is the model ID.</td>
</tr>
<tr>
<td>Use a custom model</td>
<td><a href="#">Purchase Provisioned Throughput</a> for the custom model and use the returned provisionedModelArn as the model ID.</td>
</tr>
</tbody>
</table>
Get information about foundation models

In the Amazon Bedrock console, you can find overarching information about Amazon Bedrock foundation model providers and the models they provide in the Providers and Base models sections.

Use the API to retrieve information about Amazon Bedrock foundation model, including its ARN, model ID, modalities and features it supports, and whether it is deprecated or not, in a FoundationModelSummary object.

- To return information about all the foundation models that Amazon Bedrock provides, send a ListFoundationModels request.

   ![Note]
   The response also returns model IDs that aren't in the base model ID or base model IDs for Provisioned Throughput charts. These model IDs are deprecated or for backwards compatibility.

- To return information about a specific foundation model, send a GetFoundationModel request, specifying the model ID.

Select a tab to see code examples in an interface or language.

AWS CLI

List the Amazon Bedrock foundation models.

```bash
aws bedrock list-foundation-models
```

Get information about Anthropic Claude v2.

```bash
aws bedrock get-foundation-model --model-identifier anthropic.claude-v2
```

Python

List the Amazon Bedrock foundation models.

```python
import boto3
bedrock = boto3.client(service_name='bedrock')
```
import boto3
bedrock = boto3.client(service_name='bedrock')

bedrock.get_foundation_model(modelIdentifier='anthropic.claude-v2')

### Model support by AWS Region

**Note**

All models are supported in the US East (N. Virginia, us-east-1) and US West (Oregon, us-west-2) Regions.

The following table shows the FMs that are available in other Regions and whether they are supported in each Region.

<table>
<thead>
<tr>
<th>Model</th>
<th>Asia Pacific (Singapore)</th>
<th>Asia Pacific (Tokyo)</th>
<th>Europe (Frankfurt)</th>
<th>AWS GovCloud (US-West)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude v2 (18K context window)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude v2.1</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
The following table details the support for features that are limited to certain FMs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model evaluation</th>
<th>Knowledge base (embeddings)</th>
<th>Knowledge base (query)</th>
<th>Agents</th>
<th>Finetuning (custom models)</th>
<th>Continued pre-training (custom models)</th>
<th>Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: You can run inference on all available FMs.
<table>
<thead>
<tr>
<th>Model</th>
<th>Model evaluation</th>
<th>Knowledge base (embeddings)</th>
<th>Knowledge base (query)</th>
<th>Agents</th>
<th>Fine-tuning (custom models)</th>
<th>Continued pre-training (custom models)</th>
<th>Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Lite</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Multimodal Embeddings G1</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Image Generator G1 (preview)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Anthropic Claude v1</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Anthropic Claude v2</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Model</td>
<td>Model evaluation</td>
<td>Knowledge base (embeddings)</td>
<td>Knowledge base (query)</td>
<td>Agents</td>
<td>Fine-tuning (custom models)</td>
<td>Continued pre-training (custom models)</td>
<td>Provisioned Throughput</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------</td>
<td>------------------------------</td>
<td>------------------------</td>
<td>----------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Anthropic Claude v2.1</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Anthropic Claude Instant</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Anthropic Claude 3 Sonnet</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude 3 Haiku</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AI21 Labs Jurassic-2 Mid</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AI21 Labs Jurassic-2 Ultra</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Model</td>
<td>Model evaluation</td>
<td>Knowledge base (embeddings)</td>
<td>Knowledge base (query)</td>
<td>Agents</td>
<td>Fine-tuning (custom models)</td>
<td>Continued pre-training (custom models)</td>
<td>Provisioned Throughput</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td>--------</td>
<td>---------------------------</td>
<td>----------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Cohere Embed English</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cohere Embed Multilingual</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Meta Llama 2 Chat 13B</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Meta Llama 2 Chat 70B</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (see note below)</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (see note below)</td>
</tr>
<tr>
<td>Mistral AI Mistral 7B Instruct</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Model</td>
<td>Model evaluation</td>
<td>Knowledge base (embeddings)</td>
<td>Knowledge base (query)</td>
<td>Agents</td>
<td>Fine-tuning (custom models)</td>
<td>Continued pre-training (custom models)</td>
<td>Provisioned Throughput</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>--------</td>
<td>-----------------------------</td>
<td>----------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Mistral AI Mixtral 8X7B Instruct</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Stable Diffusion XL 0.8</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stable Diffusion XL 1.x</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Note**

The Meta Llama 2 (non-chat) models can only be used after being customized and after purchasing Provisioned Throughput for them.

**Model lifecycle**

Amazon Bedrock is continuously working to bring the latest versions of foundation models that have better capabilities, accuracy, and safety. As we launch new model versions, you can test them with the Amazon Bedrock console or API, and migrate your applications to benefit from the latest model versions.

A model offered on Amazon Bedrock can be in one of these states: **Active**, **Legacy**, or **End-of-Life (EOL)**.

- **Active**: The model provider is actively working on this version, and it will continue to get updates such as bug fixes and minor improvements.
• **Legacy**: A version is marked Legacy when there is a more recent version which provides superior performance. Amazon Bedrock sets an EOL date for Legacy versions. The EOL date may vary depending on how you use the model (for example, whether you use on-demand throughput or Provisioned Throughput for a base model, or Provisioned Throughput for a customized model). While you can continue to use a Legacy version, you should plan to transition to an Active version before the EOL date.

• **EOL**: This version is no longer available for use. Any requests made to this version will fail.

The console marks a model version's state as **Active** or **Legacy**. When you make a `GetFoundationModel` or `ListFoundationModels` call, you can find the state of the model in the `modelLifecycle` field in the response. After the EOL date, the model version can only be found on this documentation page.

**On-Demand, Provisioned Throughput, and model customization**

You specify the version of a model when you use it in **On-Demand** mode (for example, `anthropic.claude-v2`, `anthropic.claude-v2:1`, etc.).

When you configure **Provisioned Throughput**, you must specify a model version that will remain unchanged for the entire term. You can purchase a new Provisioned Throughput commitment (or renew an existing one) for a version if the commitment term ends before the version's EOL date.

If you customized a model, you can continue to use it until the EOL date of the base model version that you used for customization. You can also customize a legacy model version, but you should plan to migrate before it reaches its EOL date.

⚠️ **Note**

Service quotas are shared among model minor versions.

**Legacy versions**

The following table shows the legacy versions of models available on Amazon Bedrock.
<table>
<thead>
<tr>
<th>Model version</th>
<th>Legacy date</th>
<th>EOL date</th>
<th>Recommended model version replacement</th>
<th>Recommended model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable Diffusion XL 0.8</td>
<td>February 2, 2024</td>
<td>April 30, 2024</td>
<td>Stable Diffusion XL 1.x</td>
<td>stability.stable-diffusion-xl-v1</td>
</tr>
<tr>
<td>Claude v1.3</td>
<td>November 28, 2023</td>
<td>February 28, 2024</td>
<td>Claude v2.1</td>
<td>anthropic.claude-v2:1</td>
</tr>
<tr>
<td>Titan Embeddings - Text v1.1</td>
<td>November 7, 2023</td>
<td>February 15, 2024</td>
<td>Titan Embeddings - Text v1.2</td>
<td>amazon.titan-embed-text-v1</td>
</tr>
</tbody>
</table>

**Amazon Bedrock model IDs**

Many Amazon Bedrock API operations require the use of a model ID. Refer to the following table to determine where to find the model ID that you need to use.

<table>
<thead>
<tr>
<th>Use case</th>
<th>How to find the model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use a base model</td>
<td>Look up the ID in the <a href="#">base model IDs chart</a></td>
</tr>
<tr>
<td>Purchase Provisioned Throughput for a base model</td>
<td>Look up the ID in the <a href="#">model IDs for Provisioned Throughput chart</a> and use it as the modelId in the <code>CreateProvisionedModelThroughput</code> request.</td>
</tr>
<tr>
<td>Purchase Provisioned Throughput for a custom model</td>
<td>Use the name of the custom model or its ARN as the modelId in the <code>CreateProvisionedModelThroughput</code> request.</td>
</tr>
<tr>
<td>Use a provisioned model</td>
<td>After you create a Provisioned Throughput, it returns a <code>provisionedModelArn</code>. This ARN is the model ID.</td>
</tr>
</tbody>
</table>
Use case | How to find the model ID
---|---
Use a custom model | Purchase Provisioned Throughput for the custom model and use the returned `provisionedModelArn` as the model ID.

**Topics**
- Base model IDs (on-demand throughput)
- Model IDs for purchasing Provisioned Throughput

**Base model IDs (on-demand throughput)**

The following is a list of model IDs for the currently available base models. You use a model ID through the API to identify the base model that you want to use with on-demand throughput, such as in a `InvokeModel` request, or that you want to customize, such as in a `CreateModelCustomizationJob` request.

**Note**
You should regularly check the Model lifecycle page for information about model deprecation and update model IDs as necessary. Once a model has reached end-of-life, the model ID no longer works.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Model name</th>
<th>Version</th>
<th>Model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Titan Text G1 - Express</td>
<td>1.x</td>
<td><code>amazon.titan-text-express-v1</code></td>
</tr>
<tr>
<td>Amazon</td>
<td>Titan Text G1 - Lite</td>
<td>1.x</td>
<td><code>amazon.titan-text-lite-v1</code></td>
</tr>
<tr>
<td>Amazon</td>
<td>Titan Embeddings G1 - Text</td>
<td>1.x</td>
<td><code>amazon.titan-embed-text-v1</code></td>
</tr>
<tr>
<td>Provider</td>
<td>Model name</td>
<td>Version</td>
<td>Model ID</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------</td>
<td>---------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>Amazon</td>
<td>Titan Multimodal Embeddings G1</td>
<td>1.x</td>
<td>amazon.titan-embed-image-v1</td>
</tr>
<tr>
<td>Amazon</td>
<td>Titan Image Generator G1</td>
<td>1.x</td>
<td>amazon.titan-image-generator-v1</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude</td>
<td>2.0</td>
<td>anthropic.claude-v2</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude</td>
<td>2.1</td>
<td>anthropic.claude-v2:1</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude 3 Sonnet</td>
<td>1.0</td>
<td>anthropic.claude-3-sonnet-20240229-v1:0</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude 3 Haiku</td>
<td>1.0</td>
<td>anthropic.claude-3-haiku-20240307-v1:0</td>
</tr>
<tr>
<td>Anthropic</td>
<td>Claude Instant</td>
<td>1.x</td>
<td>anthropic.claude-instant-v1</td>
</tr>
<tr>
<td>AI21 Labs</td>
<td>Jurassic-2 Mid</td>
<td>1.x</td>
<td>ai21.j2-mid-v1</td>
</tr>
<tr>
<td>AI21 Labs</td>
<td>Jurassic-2 Ultra</td>
<td>1.x</td>
<td>ai21.j2-ultra-v1</td>
</tr>
<tr>
<td>Cohere</td>
<td>Command</td>
<td>14.x</td>
<td>cohere.command-text-v14</td>
</tr>
<tr>
<td>Cohere</td>
<td>Command Light</td>
<td>15.x</td>
<td>cohere.command-light-text-v14</td>
</tr>
<tr>
<td>Cohere</td>
<td>Embed English</td>
<td>3.x</td>
<td>cohere.embed-english-v3</td>
</tr>
<tr>
<td>Cohere</td>
<td>Embed Multilingual</td>
<td>3.x</td>
<td>cohere.embed-multilingual-v3</td>
</tr>
<tr>
<td>Meta</td>
<td>Llama 2 Chat 13B</td>
<td>1.x</td>
<td>meta.llama2-13b-chat-v1</td>
</tr>
</tbody>
</table>
### Model IDs for purchasing Provisioned Throughput

To purchase Provisioned Throughput through the API, use the corresponding model ID when provisioning the model with a [CreateProvisionedModelThroughput](#) request. Provisioned Throughput is available for the following models:

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID for provisioned throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan Text G1 - Express 8K</td>
<td><code>amazon.titan-text-express-v1:0:8k</code></td>
</tr>
<tr>
<td>Titan Embeddings G1 - Text</td>
<td><code>amazon.titan-embed-text-v1:2:8k</code></td>
</tr>
<tr>
<td>Titan Multimodal Embeddings G1 - Text</td>
<td><code>amazon.titan-embed-image-v1:0</code></td>
</tr>
<tr>
<td>Titan Image Generator G1</td>
<td><code>amazon.titan-image-generator-v1:0</code></td>
</tr>
</tbody>
</table>

**Note**

Some models have multiple contextual versions whose availability differs by region. For more information, see [Model support by AWS Region](#).
<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID for provisioned throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropic Claude V2 18K</td>
<td>anthropic.claude-v2:0:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2 100K</td>
<td>anthropic.claude-v2:0:100k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 18K</td>
<td>anthropic.claude-v2:1:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 100K</td>
<td>anthropic.claude-v2:1:100k</td>
</tr>
<tr>
<td>Anthropic Claude Instant V1 100K</td>
<td>anthropic.claude-instant-v1:2:100K</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>cohere.command-text-v14:7:4k</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>cohere.command-light-text-v14:7:4k</td>
</tr>
<tr>
<td>Stable Diffusion XL 1.0</td>
<td>stability.stable-diffusion-xl-v1:0</td>
</tr>
<tr>
<td>Meta Llama 2 Chat 13B</td>
<td>meta.llama2-13b-chat-v1:0:4k</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>(see note below)</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>(see note below)</td>
</tr>
<tr>
<td>Mistral 7B Instruct</td>
<td>mistral.mistral-7b-instruct-v0:2:32k</td>
</tr>
<tr>
<td>Mixtral 8X7B Instruct</td>
<td>mistral.mixtral-8x7b-instruct-v0:1:32k</td>
</tr>
</tbody>
</table>

**Note**

The Meta Llama 2 (non-chat) models can only be provisioned after they have been customized. For more information, see [Custom models](#). Use the name or ARN of the resulting custom model to purchase Provisioned Throughput for it.

The `CreateProvisionedModelThroughput` response returns a `provisionedModelArn`. You can use this ARN or the name of the provisioned model in supported Amazon Bedrock operations. For more information about Provisioned Throughput, see [Provisioned Throughput](#).
Inference parameters for foundation models

This section documents the inference parameters that you can use with the base models that Amazon Bedrock provides.

Optionally, set inference parameters to influence the response generated by the model. You set inference parameters in a playground in the console, or in the body field of the `InvokeModel` or `InvokeModelWithResponseStream` API.

When you call a model, you also include a prompt for the model. For information about writing prompts, see Prompt engineering guidelines.

The following sections define the inference parameters available for each base model. For a custom model, use the same inference parameters as the base model from which it was customized.

Topics

- Amazon Titan models
- Anthropic Claude models
- AI21 Labs Jurassic-2 models
- Cohere models
- Meta Llama 2 and Meta Llama 2 Chat models
- Mistral AI models
- Stability.ai Diffusion models

Amazon Titan models

The following pages describe inference parameters for Amazon Titan models.

Topics

- Amazon Titan Text models
- Amazon Titan Image Generator G1
- Amazon Titan Embeddings G1 - Text
- Amazon Titan Multimodal Embeddings G1
Amazon Titan Text models

The Amazon Titan Text models support the following inference parameters.

For more information on Titan Text prompt engineering guidelines, see Titan Text Prompt Engineering Guidelines.

For more information on Titan models, see Amazon Titan Models.

Topics

- Request and response
- Code examples

Request and response

The request body is passed in the body field of an InvokeModel or InvokeModelWithResponseStream request.

Request

```json
{
   "inputText": string,
   "textGenerationConfig": {
      "temperature": float,
      "topP": float,
      "maxTokenCount": int,
      "stopSequences": [string]
   }
}
```

The following parameters are required:

- **inputText** – The prompt to provide the model for generating a response. To generate responses in a conversational style, wrap the prompt by using the following format:

```
"inputText": "User: <prompt>\nBot:
```

The textGenerationConfig is optional. You can use it to configure the following inference parameters:
• **temperature** – Use a lower value to decrease randomness in responses.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

• **topP** – Use a lower value to ignore less probable options and decrease the diversity of responses.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00E-45</td>
<td>1</td>
</tr>
</tbody>
</table>

• **maxTokenCount** – Specify the maximum number of tokens to generate in the response.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>0</td>
<td>8,000</td>
</tr>
</tbody>
</table>

• **stopSequences** – Specify a character sequence to indicate where the model should stop. Currently, you can only specify one of the following options:

  • `|`
  • `User:`

InvokeModel Response

The response body contains the following possible fields:

```json
{
    'inputTextTokenCount': int,
    'results': [{
        'tokenCount': int,
        'outputText': '<response>
        'completionReason': string
    ]
}
```

More information about each field is provided below.
• inputTextTokenCount – The number of tokens in the prompt.
• tokenCount – The number of tokens in the response.
• outputText – The text in the response.
• completionReason – The reason the response finished being generated. The following reasons are possible.
  • FINISHED – The response was fully generated.
  • LENGTH – The response was truncated because of the response length you set.

InvokeModelWithResponseStream Response

Each chunk of text in the body of the response stream is in the following format. You must decode the bytes field (see Use the API to invoke a model with a single prompt for an example).

```json
{
    'chunk': {
        'bytes': b'{
            "index": int,
            "inputTextTokenCount": int,
            "totalOutputTextTokenCount": int,
            "outputText": "<response-chunk>",
            "completionReason": string
        }
    }
}
```

• index – The index of the chunk in the streaming response.
• inputTextTokenCount – The number of tokens in the prompt.
• totalOutputTextTokenCount – The number of tokens in the response.
• outputText – The text in the response.
• completionReason – The reason the response finished being generated. The following reasons are possible.
  • FINISHED – The response was fully generated.
  • LENGTH – The response was truncated because of the response length you set.
Code examples

The following example shows how to run inference with the Amazon Titan Text G1 - Express model with the Python SDK.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
Shows how to create a list of action items from a meeting transcript with the Amazon &titan-text-express; model (on demand).
""

import json
import logging
import boto3

from botocore.exceptions import ClientError

class ImageError(Exception):
    
    """Custom exception for errors returned by Amazon &titan-text-express; model"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text(model_id, body):
    """
    Generate text using Amazon &titan-text-express; model on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        response (json): The response from the model.
    """

    logger.info(
        "Generating text with Amazon &titan-text-express; model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')
```
```python
accept = "application/json"
content_type = "application/json"

response = bedrock.invoke_model(
    body=body, modelId=model_id, accept=accept, contentType=content_type
)
response_body = json.loads(response.get("body").read())

finish_reason = response_body.get("error")

if finish_reason is not None:
    raise ImageError(f"Text generation error. Error is {finish_reason}"
)

logger.info(
    "Successfully generated text with Amazon &titan-text-express; model %s",
    model_id)

return response_body

def main():
    
    """
    Entrypoint for Amazon &titan-text-express; example.
    """
    try:
        logging.basicConfig(level=logging.INFO,
            format="%(levelname)s: %(message)s"
        )

        model_id = 'amazon.titan-text-express-v1'

        prompt = """"Meeting transcript: Miguel: Hi Brant, I want to discuss the
        workstream
        for our new product launch Brant: Sure Miguel, is there anything in
        particular you want
        to discuss? Miguel: Yes, I want to talk about how users enter into the
        product.
        Brant: Ok, in that case let me add in Namita. Namita: Hey everyone
        Brant: Hi Namita, Miguel wants to discuss how users enter into the product.
        Miguel: its too complicated and we should remove friction.
        for example, why do I need to fill out additional forms?
        I also find it difficult to find where to access the product
        when I first land on the landing page. Brant: I would also add that
        I think there are too many steps. Namita: Ok, I can work on the
        landing page to make the product more discoverable but brant
```
can you work on the additional forms? Brant: Yes but I would need to work with James from another team as he needs to unblock the sign up workflow.

Miguel can you document any other concerns so that I can discuss with James only once?

Miguel: Sure.

From the meeting transcript above, Create a list of action items for each person. ""

```python
body = json.dumps({
    "inputText": prompt,
    "textGenerationConfig": {
        "maxTokenCount": 4096,
        "stopSequences": [],
        "temperature": 0,
        "topP": 1
    }
})

response_body = generate_text(model_id, body)
print(f"Input token count: {response_body['inputTextTokenCount']}")

for result in response_body['results']:
    print(f"Token count: {result['tokenCount']}")
    print(f"Output text: {result['outputText']}")
    print(f"Completion reason: {result['completionReason']}")

except ClientError as err:
    message = err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " + format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print(f"Finished generating text with the Amazon &titan-text-express; model {model_id}" )

if __name__ == "__main__":
```
Amazon Titan Image Generator G1

The Amazon Titan Image Generator G1 model supports the following inference parameters and model responses when carrying out model inference.

Topics
- Request and response format
- Code examples

Request and response format

When you make an `InvokeModel` call using the Amazon Titan Image Generator G1, replace the `body` field of the request with the format that matches your use-case. All tasks share an `imageGenerationConfig` object, but each task has a parameters object specific to that task. The following use-cases are supported.

<table>
<thead>
<tr>
<th>taskType</th>
<th>Task parameters field</th>
<th>Type of task</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT_IMAGE</td>
<td>textToImageParams</td>
<td>Generation</td>
<td>Generate an image using a text prompt.</td>
</tr>
<tr>
<td>INPAINTING</td>
<td>inPaintingParams</td>
<td>Editing</td>
<td>Modify an image by changing the inside of a <code>mask</code> to match the surrounding background.</td>
</tr>
<tr>
<td>OUTPAINTING</td>
<td>outPaintingParams</td>
<td>Editing</td>
<td>Modify an image by seamlessly extending the region defined by the <code>mask</code>.</td>
</tr>
<tr>
<td>IMAGE_VARIATION</td>
<td>imageVariationParams</td>
<td>Editing</td>
<td>Modify an image by producing variations of the original image.</td>
</tr>
</tbody>
</table>
Editing tasks require an image field in the input. This field consists of a string that defines the pixels in the image. Each pixel is defined by 3 RGB channels, each of which ranges from 0 to 255 (for example, (255 255 0) would represent the color yellow). These channels are encoded in base64.

The image you use must be in JPEG or PNG format.

If you carry out inpainting or outpainting, you also define a mask, a region or regions that define parts of the image to be modified. You can define the mask in one of two ways.

- **maskPrompt** – Write a text prompt to describe the part of the image to be masked.
- **maskImage** – Input a base64-encoded string that defines the masked regions by marking each pixel in the input image as (0 0 0) or (255 255 255).
  - A pixel defined as (0 0 0) is a pixel inside the mask.
  - A pixel defined as (255 255 255) is a pixel outside the mask.

You can use a photo editing tool to draw masks. You can then convert the output JPEG or PNG image to base64-encoding to input into this field. Otherwise, use the **maskPrompt** field instead to allow the model to infer the mask.

Select a tab to view API request bodies for different image generation use-cases and explanations of the fields.

**Text-to-image generation (Request)**

```json
{
    "taskType": "TEXT_IMAGE",
    "textToImageParams": {
        "text": "string",
        "negativeText": "string"
    },
    "imageGenerationConfig": {
        "numberOfImages": int,
        "height": int,
        "width": int,
        "cfgScale": float,
        "seed": int
    }
}
```
The `textToImageParams` fields are described below.

- **text** (Required) – A text prompt to generate the image.
- **negativeText** (Optional) – A text prompt to define what not to include in the image.

**Note**

Don't use negative words in the `negativeText` prompt. For example, if you don't want to include mirrors in an image, enter `mirrors` in the `negativeText` prompt. Don't enter `no mirrors`.

**Inpainting (Request)**

```json
{
    "taskType": "INPAINTING",
    "inPaintingParams": {
        "image": "base64-encoded string",
        "text": "string",
        "negativeText": "string",
        "maskPrompt": "string",
        "maskImage": "base64-encoded string",
    },
    "imageGenerationConfig": {
        "numberOfImages": int,
        "height": int,
        "width": int,
        "cfgScale": float
    }
}
```

The `inPaintingParams` fields are described below. The *mask* defines the part of the image that you want to modify.

- **image** (Required) – The JPEG or PNG image to modify, formatted as a string that specifies a sequence of pixels, each defined in RGB values and encoded in base64. For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the [code examples](#).

  You must define one of the following fields (but not both) in order to define.

  - **maskPrompt** – A text prompt that defines the mask.
- **maskImage** – A string that defines the mask by specifying a sequence of pixels that is the same size as the image. Each pixel is turned into an RGB value of (0 0 0) (a pixel inside the mask) or (255 255 255) (a pixel outside the mask). For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

- **text** (Optional) – A text prompt to define what to change inside the mask. If you don’t include this field, the model tries to replace the entire mask area with the background.

- **negativeText** (Optional) – A text prompt to define what not to include in the image.

**Note**

Don’t use negative words in the negativeText prompt. For example, if you don’t want to include mirrors in an image, enter mirrors in the negativeText prompt. Don’t enter no mirrors.

Outpainting (Request)

```json
{
    "taskType": "OUTPAINTING",
    "outPaintingParams": {
        "text": "string",
        "negativeText": "string",
        "image": "base64-encoded string",
        "maskPrompt": "string",
        "maskImage": "base64-encoded string",
        "outPaintingMode": "DEFAULT | PRECISE"
    },
    "imageGenerationConfig": {
        "numberOfImages": int,
        "height": int,
        "width": int,
        "cfgScale": float
    }
}
```

The outPaintingParams fields are defined below. The *mask* defines the region in the image whose that you don't want to modify. The generation seamlessly extends the region you define.
• **image** (Required) – The JPEG or PNG image to modify, formatted as a string that specifies a sequence of pixels, each defined in RGB values and encoded in base64. For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

• You must define one of the following fields (but not both) in order to define.
  
  • **maskPrompt** – A text prompt that defines the mask.
  
  • **maskImage** – A string that defines the mask by specifying a sequence of pixels that is the same size as the image. Each pixel is turned into an RGB value of (0 0 0) (a pixel inside the mask) or (255 255 255) (a pixel outside the mask). For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

• **text** (Required) – A text prompt to define what to change outside the mask.

• **negativeText** (Optional) – A text prompt to define what not to include in the image.

**Note**

Don't use negative words in the negativeText prompt. For example, if you don't want to include mirrors in an image, enter *mirrors* in the negativeText prompt. Don't enter *no mirrors*.

• **outPaintingMode** – Specifies whether to allow modification of the pixels inside the mask or not. The following values are possible.

  • **DEFAULT** – Use this option to allow modification of the image inside the mask in order to keep it consistent with the reconstructed background.

  • **PRECISE** – Use this option to prevent modification of the image inside the mask.

**Image variation (Request)**

```
{
    "taskType": "IMAGE_VARIATION",
    "imageVariationParams": {
        "text": "string",
        "negativeText": "string",
        "images": ["base64-encoded string"],
    },
    "imageGenerationConfig": {
        "numberOfImages": int,
    }
}
```
The `imageVariationParams` fields are defined below.

- **images** (Required) – A list of images for which to generate variations. Currently, you can only include one image. An image is defined as a base64-encoded image string. For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

- **text** (Optional) – A text prompt that can define what to preserve and what to change in the image.

- **negativeText** (Optional) – A text prompt to define what not to include in the image.

**Note**

Don't use negative words in the `negativeText` prompt. For example, if you don't want to include mirrors in an image, enter `mirrors` in the `negativeText` prompt. Don't enter `no mirrors`.

Response body

```json
{
    "images": [
        "base64-encoded string",
        ...
    ],
    "error": "string"
}
```

The response body is a streaming object that contains one of the following fields.

- **images** – If the request is successful, it returns this field, a list of base64-encoded strings, each defining a generated image. Each image is formatted as a string that specifies a sequence of pixels, each defined in RGB values and encoded in base64. For examples of how
to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

- `error` – If the request violates the content moderation policy in one of the following situations, a message is returned in this field.
  - If the input text, image, or mask image is flagged by the content moderation policy.
  - If at least one output image is flagged by the content moderation policy.

The shared and optional `imageGenerationConfig` contains the following fields. If you don't include this object, the default configurations are used.

- **`numberOfImages` (Optional)** – The number of images to generate.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

- **`cfgScale` (Optional)** – Specifies how strongly the generated image should adhere to the prompt. Use a lower value to introduce more randomness in the generation.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>10.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>

- The following parameters define the size that you want the output image to be. For more details about pricing by image size, see Amazon Bedrock pricing.
  - **`height` (Optional)** – The height of the image in pixels. The default value is 1024.
  - **`width` (Optional)** – The width of the image in pixels. The default value is 1024.

The following sizes are permissible.

<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Aspect ratio</th>
<th>Price equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>1024</td>
<td>1:1</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>768</td>
<td>768</td>
<td>1:1</td>
<td>512 x 512</td>
</tr>
<tr>
<td>512</td>
<td>512</td>
<td>1:1</td>
<td>512 x 512</td>
</tr>
<tr>
<td>Width</td>
<td>Height</td>
<td>Aspect ratio</td>
<td>Price equivalent to</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>768</td>
<td>1152</td>
<td>2:3</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>384</td>
<td>576</td>
<td>2:3</td>
<td>512 x 512</td>
</tr>
<tr>
<td>1152</td>
<td>768</td>
<td>3:2</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>576</td>
<td>384</td>
<td>3:2</td>
<td>512 x 512</td>
</tr>
<tr>
<td>768</td>
<td>1280</td>
<td>3:5</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>384</td>
<td>640</td>
<td>3:5</td>
<td>512 x 512</td>
</tr>
<tr>
<td>1280</td>
<td>768</td>
<td>5:3</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>640</td>
<td>384</td>
<td>5:3</td>
<td>512 x 512</td>
</tr>
<tr>
<td>896</td>
<td>1152</td>
<td>7:9</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>448</td>
<td>576</td>
<td>7:9</td>
<td>512 x 512</td>
</tr>
<tr>
<td>1152</td>
<td>896</td>
<td>9:7</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>576</td>
<td>448</td>
<td>9:7</td>
<td>512 x 512</td>
</tr>
<tr>
<td>768</td>
<td>1408</td>
<td>6:11</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>384</td>
<td>704</td>
<td>6:11</td>
<td>512 x 512</td>
</tr>
<tr>
<td>1408</td>
<td>768</td>
<td>11:6</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>704</td>
<td>384</td>
<td>11:6</td>
<td>512 x 512</td>
</tr>
<tr>
<td>640</td>
<td>1408</td>
<td>5:11</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>320</td>
<td>704</td>
<td>5:11</td>
<td>512 x 512</td>
</tr>
<tr>
<td>1408</td>
<td>640</td>
<td>11:5</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>704</td>
<td>320</td>
<td>11:5</td>
<td>512 x 512</td>
</tr>
</tbody>
</table>
```markdown
- **seed** (Optional) – Use to control and reproduce results. Determines the initial noise setting. Use the same seed and the same settings as a previous run to allow inference to create a similar image.

<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Aspect ratio</th>
<th>Price equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1152</td>
<td>640</td>
<td>9:5</td>
<td>1024 x 1024</td>
</tr>
<tr>
<td>1173</td>
<td>640</td>
<td>16:9</td>
<td>1024 x 1024</td>
</tr>
</tbody>
</table>

**Note**

You can only set a seed for a TEXT_IMAGE generation task.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2,147,483,646</td>
<td>0</td>
</tr>
</tbody>
</table>

**Code examples**

The following examples show how to invoke the Amazon Titan Image Generator G1 model with on-demand throughput in the Python SDK. Select a tab to view an example for each use-case. Each example displays the image at the end.

**Text-to-image generation**

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

 Shows how to generate an image from a text prompt with the Amazon Titan Image Generator G1 model (on demand).

```
from botocore.exceptions import ClientError

class ImageError(Exception):
    "Custom exception for errors returned by Amazon Titan Image Generator G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_image(model_id, body):
    """
    Generate an image using Amazon Titan Image Generator G1 model on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        image_bytes (bytes): The image generated by the model.
    """

    logger.info("Generating image with Amazon Titan Image Generator G1 model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )
    response_body = json.loads(response.get("body").read())

    base64_image = response_body.get("images")[0]
    base64_bytes = base64_image.encode('ascii')
    image_bytes = base64.b64decode(base64_bytes)

    finish_reason = response_body.get("error")
if finish_reason is not None:
    raise ImageError(f"Image generation error. Error is {finish_reason}")

logger.info(
    "Successfully generated image with Amazon Titan Image Generator G1 model %s", model_id)

return image_bytes

def main():
    
    Entrypoint for Amazon Titan Image Generator G1 example.
    
    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    model_id = 'amazon.titan-image-generator-v1'

    prompt = """A photograph of a cup of coffee from the side.""

    body = json.dumps({
        "taskType": "TEXT_IMAGE",
        "textToImageParams": {
            "text": prompt
        },
        "imageGenerationConfig": {
            "numberOfImages": 1,
            "height": 1024,
            "width": 1024,
            "cfgScale": 8.0,
            "seed": 0
        }
    })

    try:
        image_bytes = generate_image(model_id=model_id,
                                      body=body)
        image = Image.open(io.BytesIO(image_bytes))
        image.show()
    except ClientError as err:
        message = err.response["Error"]["Message"]
logger.error("A client error occurred: %s", message)
print("A client error occurred: "+format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)
else:
    print(f"Finished generating image with Amazon Titan Image Generator G1 model {model_id}.")

if __name__ == "__main__":
    main()

### Inpainting

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
Shows how to use inpainting to generate an image from a source image with
the Amazon Titan Image Generator G1 model (on demand).
The example uses a mask prompt to specify the area to inpaint.
""

import base64
import io
import json
import logging
import boto3
from PIL import Image

from botocore.exceptions import ClientError

class ImageError(Exception):
    "Custom exception for errors returned by Amazon Titan Image Generator G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
```
def generate_image(model_id, body):
    
    Generate an image using Amazon Titan Image Generator G1 model on demand.
    Args:
        model_id (str): The model ID to use.
        body (str) : The request body to use.
    Returns:
        image_bytes (bytes): The image generated by the model.
    
    logger.info(
        "Generating image with Amazon Titan Image Generator G1 model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )
    response_body = json.loads(response.get("body").read())

    base64_image = response_body.get("images")[0]
    base64_bytes = base64_image.encode('ascii')
    image_bytes = base64.b64decode(base64_bytes)

    finish_reason = response_body.get("error")

    if finish_reason is not None:
        raise ImageError(f"Image generation error. Error is {finish_reason}"")

    logger.info(
        "Successfully generated image with Amazon Titan Image Generator G1 model %s", model_id)

    return image_bytes

def main():
    
    Amazon Titan models
Entrypoint for Amazon Titan Image Generator G1 example.

```python
try:
    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    model_id = 'amazon.titan-image-generator-v1'

    # Read image from file and encode it as base64 string.
    with open("/path/to/image", "rb") as image_file:
        input_image = base64.b64encode(image_file.read()).decode('utf8')

    body = json.dumps({
        "taskType": "INPAINTING",
        "inPaintingParams": {
            "text": "Modernize the windows of the house",
            "negativeText": "bad quality, low res",
            "image": input_image,
            "maskPrompt": "windows"
        },
        "imageGenerationConfig": {
            "numberOfImages": 1,
            "height": 512,
            "width": 512,
            "cfgScale": 8.0
        }
    })

    image_bytes = generate_image(model_id=model_id,
                                  body=body)
    image = Image.open(io.BytesIO(image_bytes))
    image.show()

except ClientError as err:
    message = err.response["Error"]['Message']
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
          format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print(}
Finished generating image with Amazon Titan Image Generator G1 model
{model_id}.

if __name__ == "__main__":
    main()

Outpainting

# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0
"
""
Shows how to use outpainting to generate an image from a source image with
the Amazon Titan Image Generator G1 model (on demand).
The example uses a mask image to outpaint the original image.
"
"
import base64
import io
import json
import logging
import boto3
from PIL import Image
from botocore.exceptions import ClientError

class ImageError(Exception):
    "Custom exception for errors returned by Amazon Titan Image Generator G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_image(model_id, body):
    ""
    Generate an image using Amazon Titan Image Generator G1 model on demand.
    Args:
    model_id (str): The model ID to use.
    body (str): The request body to use.
Returns:

    image_bytes (bytes): The image generated by the model.

"""

logger.info(
    "Generating image with Amazon Titan Image Generator G1 model %s", model_id)

bedrock = boto3.client(service_name='bedrock-runtime')

accept = "application/json"
content_type = "application/json"

response = bedrock.invoke_model(
    body=body, modelId=model_id, accept=accept, contentType=content_type
)
response_body = json.loads(response.get("body").read())

base64_image = response_body.get("images")[0]
base64_bytes = base64_image.encode('ascii')
image_bytes = base64.b64decode(base64_bytes)

finish_reason = response_body.get("error")

if finish_reason is not None:
    raise ImageError(f"Image generation error. Error is {finish_reason}")

logger.info(
    "Successfully generated image with Amazon Titan Image Generator G1 model %s", model_id)

return image_bytes

def main():
    """
    Entrypoint for Amazon Titan Image Generator G1 example.
    """
    try:
        logging.basicConfig(level=logging.INFO,
            format="%(levelname)s: %(message)s")

        model_id = 'amazon.titan-image-generator-v1'

        # Read image and mask image from file and encode as base64 strings.
with open("/path/to/image", "rb") as image_file:
    input_image = base64.b64encode(image_file.read()).decode('utf8')
with open("/path/to/mask_image", "rb") as mask_image_file:
    input_mask_image = base64.b64encode(
        mask_image_file.read()).decode('utf8')

body = json.dumps(
    "taskType": "OUTPAINTING",
    "outPaintingParams": {
        "text": "Draw a chocolate chip cookie",
        "negativeText": "bad quality, low res",
        "image": input_image,
        "maskImage": input_mask_image,
        "outPaintingMode": "DEFAULT"
    },
    "imageGenerationConfig": {
        "numberOfImages": 1,
        "height": 512,
        "width": 512,
        "cfgScale": 8.0
    }
)

image_bytes = generate_image(model_id=model_id, body=body)
image = Image.open(io.BytesIO(image_bytes))
image.show()

except ClientError as err:
    message = err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
          format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print(
        f"Finished generating image with Amazon Titan Image Generator G1 model {model_id}.")
Image variation

```python
if __name__ == '__main__':
    main()
```

Image variation

```
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

"""
Shows how to generate an image variation from a source image with the
Amazon Titan Image Generator G1 model (on demand).
"""
import base64
import io
import json
import logging
import boto3
from PIL import Image

from botocore.exceptions import ClientError

class ImageError(Exception):
    """Custom exception for errors returned by Amazon Titan Image Generator G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_image(model_id, body):
    """
    Generate an image using Amazon Titan Image Generator G1 model on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        image_bytes (bytes): The image generated by the model.
    """

    logger.info(
"Generating image with Amazon Titan Image Generator G1 model %s", model_id)

bedrock = boto3.client(service_name='bedrock-runtime')

accept = "application/json"
content_type = "application/json"

response = bedrock.invoke_model(
    body=body, modelId=model_id, accept=accept, contentType=content_type
)
response_body = json.loads(response.get("body").read())

base64_image = response_body.get("images")[0]
base64_bytes = base64_image.encode('ascii')
image_bytes = base64.b64decode(base64_bytes)

finish_reason = response_body.get("error")

if finish_reason is not None:
    raise ImageError(f"Image generation error. Error is {finish_reason}"
)

logger.info("Successfully generated image with Amazon Titan Image Generator G1 model %s", model_id)

return image_bytes

def main():
    """
    Entrypoint for Amazon Titan Image Generator G1 example.
    """
    try:
        logging.basicConfig(level=logging.INFO,
            format="%(levelname)s: %(message)s")

        model_id = 'amazon.titan-image-generator-v1'

        # Read image from file and encode it as base64 string.
        with open("/path/to/image", "rb") as image_file:
            input_image = base64.b64encode(image_file.read()).decode('utf8')

        body = json.dumps(
            "taskType": "IMAGE_VARIATION",}
"imageVariationParams": {
    "text": "Modernize the house, photo-realistic, 8k, hdr",
    "negativeText": "bad quality, low resolution, cartoon",
    "images": [input_image],
},
"imageGenerationConfig": {
    "numberOfImages": 1,
    "height": 512,
    "width": 512,
    "cfgScale": 8.0
}
})

image_bytes = generate_image(model_id=model_id,
                              body=body)
image = Image.open(io.BytesIO(image_bytes))
image.show()

except ClientError as err:
    message = err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error ocurred: " +
          format(message))
    except ImageError as err:
        logger.error(err.message)
        print(err.message)

else:
    print("Finished generating image with Amazon Titan Image Generator G1 model {model_id}.")

if __name__ == "__main__":
    main()
• Request and response

• Example code

Request and response

The request body is passed in the body field of an `InvokeModel` request.

Request

The only available field is `inputText`, in which you can include text to convert into embeddings.

```json
{
    "inputText": string
}
```

Response

The body of the response contains the following fields.

```json
{
    "embedding": [float, float, ...],
    "inputTextTokenCount": int
}
```

The fields are described below.

- **embedding** – An array that represents the embeddings vector of the input you provided.
- **inputTextTokenCount** – The number of tokens in the input.

Example code

This example shows how to call the Amazon Titan Embeddings G1 - Text model to generate embeddings.

```
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
"""

Shows how to generate embeddings with the Amazon Titan Embeddings G1 - Text model (on demand).
```
import json
import logging
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_embeddings(model_id, body):
    
    Generate a vector of embeddings for a text input using Amazon Titan Embeddings G1 -
    Text on demand.

    Args:
    model_id (str): The model ID to use.
    body (str): The request body to use.

    Returns:
    response (JSON): The text that the model generated, token information, and the
    reason the model stopped generating text.

    logger.info("Generating embeddings with Amazon Titan Embeddings G1 - Text model
    %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )

    response_body = json.loads(response.get('body').read())

    return response_body

def main():
Entrypoint for Amazon Titan Embeddings G1 - Text example.

logging.basicConfig(level=logging.INFO,
    format="%(levelname)s: %(message)s")

model_id = "amazon.titan-embed-text-v1"
input_text = "What are the different services that you offer?"

# Create request body.
body = json.dumps({
    "inputText": input_text,
})

try:
    response = generate_embeddings(model_id, body)
    print(f"Generated embeddings: {response['embedding']}"
    print(f"Input Token count: {response['inputTextTokenCount']}"
except ClientError as err:
    message = err.response["Error"]['Message']
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
        format(message))

else:
    print(f"Finished generating embeddings with Amazon Titan Embeddings G1 - Text
    model {model_id}.")

if __name__ == "__main__":
    main()

Amazon Titan Multimodal Embeddings G1

This section provides request and response body formats and code examples for using Amazon Titan Multimodal Embeddings G1.
Topics

- Request and response
- Example code

Request and response

The request body is passed in the body field of an `InvokeModel` request.

Request

The request body for Amazon Titan Multimodal Embeddings G1 includes the following fields.

```json
{
    "inputText": string,
    "inputImage": base64-encoded string,
    "embeddingConfig": {
        "outputEmbeddingLength": 256 | 384 | 1024
    }
}
```

At least one of the following fields is required. Include both to generate an embeddings vector that averages the resulting text embeddings and image embeddings vectors.

- **inputText** – Enter text to convert to embeddings.
- **inputImage** – Encode the image that you want to convert to embeddings in base64 and enter the string in this field. For examples of how to encode an image into base64 and decode a base64-encoded string and transform it into an image, see the code examples.

The following field is optional.

- **embeddingConfig** – Contains an `outputEmbeddingLength` field, in which you specify one of the following lengths for the output embeddings vector.
  - 256
  - 384
  - 1024 (default)
Response

The body of the response contains the following fields.

```json
{
    "embedding": [float, float, ...],
    "inputTextTokenCount": int,
    "message": string
}
```

The fields are described below.

- **embedding** – An array that represents the embeddings vector of the input you provided.
- **inputTextTokenCount** – The number of tokens in the text input.
- **message** – Specifies any errors that occur during generation.

Example code

The following examples show how to invoke the Amazon Titan Multimodal Embeddings G1 model with on-demand throughput in the Python SDK. Select a tab to view an example for each use-case.

Text embeddings

This example shows how to call the Amazon Titan Multimodal Embeddings G1 model to generate text embeddings.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
Shows how to generate embeddings from text with the Amazon Titan Multimodal
Embeddings G1 model (on demand).
""

import json
import logging
import boto3

from botocore.exceptions import ClientError

class EmbedError(Exception):
```
"Custom exception for errors returned by Amazon Titan Multimodal Embeddings G1"

```python
    def __init__(self, message):
        self.message = message

    logger = logging.getLogger(__name__)
    logging.basicConfig(level=logging.INFO)

def generate_embeddings(model_id, body):
    """
    Generate a vector of embeddings for a text input using Amazon Titan Multimodal
    Embeddings G1 on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        response (JSON): The embeddings that the model generated, token information, and the
        reason the model stopped generating embeddings.
    """
    logger.info("Generating embeddings with Amazon Titan Multimodal Embeddings G1
                model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )

    response_body = json.loads(response.get('body').read())

    finish_reason = response_body.get("message")

    if finish_reason is not None:
        raise EmbedError(f"Embeddings generation error: {finish_reason}"

    return response_body
```
def main():
    """
    Entrypoint for Amazon Titan Multimodal Embeddings G1 example.
    """

    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    model_id = 'amazon.titan-embed-image-v1'
    input_text = "What are the different services that you offer?"
    output_embedding_length = 256

    # Create request body.
    body = json.dumps({
        "inputText": input_text,
        "embeddingConfig": {
            "outputEmbeddingLength": output_embedding_length
        }
    })

    try:
        response = generate_embeddings(model_id, body)
        print(f"Generated text embeddings of length {output_embedding_length}:
        {response["embedding"]}")
        print(f"Input text token count: {response["inputTextTokenCount"]}")
    except ClientError as err:
        message = err.response["Error"]["Message"]
        logger.error("A client error occurred: %s", message)
        print("A client error occurred: " +
              format(message))
    except EmbedError as err:
        logger.error(err.message)
        print(err.message)
    else:
        print(f"Finished generating text embeddings with Amazon Titan Multimodal
        Embeddings G1 model {model_id}.")
Image embeddings

This example shows how to call the Amazon Titan Multimodal Embeddings G1 model to generate image embeddings.

```python
import base64
import json
import logging
import boto3
from botocore.exceptions import ClientError

class EmbedError(Exception):
    "Custom exception for errors returned by Amazon Titan Multimodal Embeddings G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_embeddings(model_id, body):
    """
    Generate a vector of embeddings for an image input using Amazon Titan Multimodal Embeddings G1 on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        response (JSON): The embeddings that the model generated, token information, and the reason the model stopped generating embeddings.
    """
```
logger.info("Generating embeddings with Amazon Titan Multimodal Embeddings G1 model %s", model_id)

bedrock = boto3.client(service_name='bedrock-runtime')

accept = "application/json"
content_type = "application/json"

response = bedrock.invoke_model(
    body=body, modelId=model_id, accept=accept, contentType=content_type
)

response_body = json.loads(response.get('body').read())

finish_reason = response_body.get("message")
if finish_reason is not None:
    raise EmbedError(f"Embeddings generation error: {finish_reason}")

return response_body

def main():
    """
    Entrypoint for Amazon Titan Multimodal Embeddings G1 example.
    """

    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    # Read image from file and encode it as base64 string.
    with open("/path/to/image", "rb") as image_file:
        input_image = base64.b64encode(image_file.read()).decode('utf8')

    model_id = 'amazon.titan-embed-image-v1'
    output_embedding_length = 256

    # Create request body.
    body = json.dumps({
        "inputImage": input_image,
        "embeddingConfig": {
            "outputEmbeddingLength": output_embedding_length
        }
    })
try:
    response = generate_embeddings(model_id, body)
    print(f"Generated image embeddings of length {output_embedding_length}: {response['embedding']}")
except ClientError as err:
    message = err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " + format(message))
except EmbedError as err:
    logger.error(err.message)
    print(err.message)
else:
    print(f"Finished generating image embeddings with Amazon Titan Multimodal Embeddings G1 model {model_id}.")

if __name__ == "__main__":
    main()
import json
import logging
import boto3

from botocore.exceptions import ClientError

class EmbedError(Exception):
    "Custom exception for errors returned by Amazon Titan Multimodal Embeddings G1"

    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_embeddings(model_id, body):
    """
    Generate a vector of embeddings for a combined text and image input using Amazon
Titan Multimodal Embeddings G1 on demand.

    Args:
        model_id (str): The model ID to use.
        body (str) : The request body to use.

    Returns:
        response (JSON): The embeddings that the model generated, token information,
        and the
        reason the model stopped generating embeddings.
    """

    logger.info("Generating embeddings with Amazon Titan Multimodal Embeddings G1 model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )

    response_body = json.loads(response.get('body').read())

    finish_reason = response_body.get("message")
if finish_reason is not None:
    raise EmbedError(f"Embeddings generation error: {finish_reason}")

    return response_body

def main():
    """
    Entrypoint for Amazon Titan Multimodal Embeddings G1 example.
    """

    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    model_id = "amazon.titan-embed-image-v1"
    input_text = "A family eating dinner"
    # Read image from file and encode it as base64 string.
    with open("/path/to/image", "rb") as image_file:
        input_image = base64.b64encode(image_file.read()).decode('utf8')
    output_embedding_length = 256

    # Create request body.
    body = json.dumps(
        {
            "inputText": input_text,
            "inputImage": input_image,
            "embeddingConfig": {
                "outputEmbeddingLength": output_embedding_length
            }
        }
    )

    try:

        response = generate_embeddings(model_id, body)

        print(f"Generated embeddings of length {output_embedding_length}:
              {response["embedding"]}")
        print(f"Input text token count: {response["inputTextTokenCount"]}")

    except ClientError as err:
        message = err.response["Error"]["Message"]
        logger.error("A client error occurred: %s", message)
        print("A client error occurred: ")
Anthropic Claude models

This section provides inference parameters and code examples for using Anthropic Claude models.

You can use Amazon Bedrock to send Anthropic Claude Text Completions API or Anthropic Claude Messages API inference requests.

You use the messages API to create conversational applications, such as a virtual assistant or a coaching application. Use the text completion API for single-turn text generation applications. For example, generating text for a blog post or summarizing text that a user supplies.

You make inference requests to an Anthropic Claude model with InvokeModel or InvokeModelWithResponseStream (streaming). You need the model ID for the model that you want to use. To get the model ID for Anthropic Claude models, see Base model IDs (on-demand throughput) and Model IDs for purchasing Provisioned Throughput.

⚠️ Note

To use system prompts in inference calls, you must use Anthropic Claude version 2.1 or Anthropic Claude 3 Sonnet. For information about creating system prompts, see https://docs.anthropic.com/claude/docs/how-to-use-system-prompts in the Anthropic Claude documentation.

To avoid timeouts with Anthropic Claude version 2.1, we recommend limiting the input token count in the prompt field to 180K. We expect to address this timeout issue soon.
In the inference call, fill the body field with a JSON object that conforms the type call you want to make, Anthropic Claude Text Completions API or Anthropic Claude Messages API.

For information about creating prompts for Anthropic Claude models, see Introduction to prompt design in the Anthropic Claude documentation.

Topics

• Anthropic Claude Text Completions API
• Anthropic Claude Messages API

Anthropic Claude Text Completions API

This section provides inference parameters and code examples for using Anthropic Claude models with the Text Completions API.

Topics

• Anthropic Claude Text Completions API overview
• Supported models
• Request and Response
• Code example

Anthropic Claude Text Completions API overview

Use the Text Completion API for single-turn text generation from a user supplied prompt. For example, you can use the Text Completion API to generate text for a blog post or to summarize text input from a user.

For information about creating prompts for Anthropic Claude models, see Introduction to prompt design. If you want to use your existing Text Completions prompts with the Anthropic Claude Messages API, see Migrating from Text Completions.

Supported models

You can use the Text Completions API with the following Anthropic Claude models.

• Anthropic Claude Instant v1.2
• Anthropic Claude v2
• Anthropic Claude v2.1

Request and Response

The request body is passed in the body field of a request to InvokeModel or InvokeModelWithResponseStream.

For more information, see https://docs.anthropic.com/claude/reference/complete_post in the Anthropic Claude documentation.

Request

Anthropic Claude has the following inference parameters for a Text Completion inference call.

```json
{
    "prompt": "\n\nHuman:\n\nAssistant:",
    "temperature": float,
    "top_p": float,
    "top_k": int,
    "max_tokens_to_sample": int,
    "stop_sequences": [string]
}
```

The following are required parameters.

• **prompt** (Required) The prompt that you want Claude to complete. For proper response generation you need to format your prompt using alternating \n\nHuman: and \n\nAssistant: conversational turns. For example:

```
"\n\nHuman: {userQuestion}\n\nAssistant:"
```

For more information, see Prompt validation in the Anthropic Claude documentation.

• **max_tokens_to_sample** (Required) The maximum number of tokens to generate before stopping. We recommend a limit of 4,000 tokens for optimal performance.

Note that Anthropic Claude models might stop generating tokens before reaching the value of max_tokens_to_sample. Different Anthropic Claude models have different maximum values for this parameter. For more information, see Model comparison in the Anthropic Claude documentation.
The following are optional parameters.

- **stop_sequences** – (Optional) Sequences that will cause the model to stop generating.

  Anthropic Claude models stop on "\n\n\nHuman: ", and may include additional built-in stop sequences in the future. Use the stop_sequences inference parameter to include additional strings that will signal the model to stop generating text.

- **temperature** – (Optional) The amount of randomness injected into the response.

  Defaults to 1. Ranges from 0 to 1. Use temp closer to 0 for analytical / multiple choice, and closer to 1 for creative and generative tasks.

- **top_p** – (Optional) Use nucleus sampling.

  In nucleus sampling, Anthropic Claude computes the cumulative distribution over all the options for each subsequent token in decreasing probability order and cuts it off once it reaches a particular probability specified by top_p. You should alter either temperature or top_p, but not both.

- **top_k** – (Optional) Only sample from the top K options for each subsequent token.

  Use top_k to remove long tail low probability responses.
<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0</td>
<td>500</td>
</tr>
</tbody>
</table>

Response

The Anthropic Claude model returns the following fields for a Text Completion inference call.

```
{
  "completion": string,
  "stop_reason": string,
  "stop": string
}
```

- **completion** – The resulting completion up to and excluding the stop sequences.
- **stop_reason** – The reason why the model stopped generating the response.
  - "stop_sequence" – The model reached a stop sequence — either provided by you with the stop_sequences inference parameter, or a stop sequence built into the model.
  - "max_tokens" – The model exceeded max_tokens_to_sample or the model's maximum number of tokens.
- **stop** – If you specify the stop_sequences inference parameter, stop contains the stop sequence that signalled the model to stop generating text. For example, holes in the following response.

```
{
  "completion": "Here is a simple explanation of black ",
  "stop_reason": "stop_sequence",
  "stop": "holes"
}
```

If you don't specify stop_sequences, the value for stop is empty.

**Code example**

These examples shows how to call the Anthropic Claude V2 model with on demand throughput. To use Anthropic Claude version 2.1, change the value of modelId to anthropic.claude-v2:1.
The following example shows how to generate streaming text with Python using the prompt

**write an essay for living on mars in 1000 words**

and the Anthropic Claude V2 model:

```python
import boto3
import json

brt = boto3.client(service_name='bedrock-runtime')

body = json.dumps(
    {
        'prompt': '

Human: write an essay for living on mars in 1000 words

Assistant:',
        'max_tokens_to_sample': 4000,
    }
)

response = brt.invoke_model_with_response_stream(
    modelId='anthropic.claude-v2',
    body=body
)

stream = response.get('body')
```
Anthropic Claude Messages API

This section provides inference parameters and code examples for using the Anthropic Claude Messages API.

Topics
- Anthropic Claude Messages API overview
- Supported models
- Request and Response
- Code examples

Anthropic Claude Messages API overview

You can use the Messages API to create chat bots or virtual assistant applications. The API manages the conversational exchanges between a user and an Anthropic Claude model (assistant).

Anthropic trains Claude models to operate on alternating user and assistant conversational turns. When creating a new message, you specify the prior conversational turns with the messages parameter. The model then generates the next Message in the conversation.

Each input message must be an object with a role and content. You can specify a single user-role message, or you can include multiple user and assistant messages. The first message must always use the user role.

If you are using the technique of prefilling the response from Claude (filling in the beginning of Claude's response by using a final assistant role Message), Claude will respond by picking up from where you left off. With this technique, Claude will still return a response with the assistant role.

If the final message uses the assistant role, the response content will continue immediately from the content in that message. You can use this to constrain part of the model's response.

Example with a single user message:
Example with multiple conversational turns:

```
[
  {"role": "user", "content": "Hello there."},
  {"role": "assistant", "content": "Hi, I'm Claude. How can I help you?"},
  {"role": "user", "content": "Can you explain LLMs in plain English?"},
]
```

Example with a partially-filled response from Claude:

```
[
  {"role": "user", "content": "Please describe yourself using only JSON"},
  {"role": "assistant", "content": "Here is my JSON description:
{"},
]
```

Each input message content may be either a single string or an array of content blocks, where each block has a specific type. Using a string is shorthand for an array of one content block of type "text". The following input messages are equivalent:

```
{"role": "user", "content": "Hello, Claude"}
```

```
{"role": "user", "content": [{"type": "text", "text": "Hello, Claude"}]}
```

For information about creating prompts for Anthropic Claude models, see [Intro to prompting](#) in the Anthropic Claude documentation. If you have existing [Text Completion](#) prompts that you want to migrate to the messages API, see [Migrating from Text Completions](#).

**System prompts**

You can also include a system prompt in the request. A system prompt lets you provide context and instructions to Anthropic Claude, such as specifying a particular goal or role. Specify a system prompt in the system field, as shown in the following example.

```
"system": "You are Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. Your goal is to provide informative and substantive responses to queries while avoiding potential harms."
```
For more information, see System prompts in the Anthropic documentation.

Multimodal prompts

A multimodal prompt combines multiple modalities (images and text) in a single prompt. You specify the modalities in the content input field. The following example shows how you could ask Anthropic Claude to describe the content of a supplied image. For example code, see Multimodal code examples.

```json
{
  "anthropic_version": "bedrock-2023-05-31",
  "max_tokens": 1024,
  "messages": [
    {
      "role": "user",
      "content": [
        {
          "type": "image",
          "source": {
            "type": "base64",
            "media_type": "image/jpeg",
            "data": "iVBORw..."
          }
        },
        {
          "type": "text",
          "text": "What's in these images?"
        }
      ]
    }
  ]
}
```

You can supply up to 20 images to the model. You can't put images in the assistant role.

Each image you include in a request counts towards your token usage. For more information, see Image costs in the Anthropic documentation.

Supported models

You can use the Messages API with the following Anthropic Claude models.

- Anthropic Claude Instant v1.2
• Anthropic Claude 2 v2
• Anthropic Claude 2 v2.1
• Anthropic Claude 3 Sonnet
• Anthropic Claude 3 Haiku

**Request and Response**

The request body is passed in the body field of a request to `InvokeModel` or `InvokeModelWithResponseStream`. The maximum size of the payload you can send in a request is 20MB.

For more information, see [https://docs.anthropic.com/claude/reference/messages_post](https://docs.anthropic.com/claude/reference/messages_post).

**Request**

Anthropic Claude has the following inference parameters for a messages inference call.

```json
{
  "anthropic_version": "bedrock-2023-05-31",
  "max_tokens": int,
  "system": string,
  "messages": [   
    {   
      "role": string,
      "content": [   
        { "type": "image", "source": { "type": "base64", "media_type": "image/jpeg", "data": "content image bytes" } },   
        { "type": "text", "text": "content text" }   
      ]   
    }   
  ],
  "temperature": float,
  "top_p": float,
  "top_k": int,
  "stop_sequences": [string]
}
```

The following are required parameters.

- **anthropic_version** – (Required) The anthropic version. The value must be `bedrock-2023-05-31`.
• **max_tokens** – (Required) The maximum number of tokens to generate before stopping.

  Note that Anthropic Claude models might stop generating tokens before reaching the value of `max_tokens`. Different Anthropic Claude models have different maximum values for this parameter. For more information, see [Model comparison](#).

• **messages** – (Required) The input messages.

  • **role** – The role of the conversation turn. Valid values are `user` and `assistant`.

  • **content** – (required) The content of the conversation turn.
    
    • **type** – (required) The type of the content. Valid values are `image` and `text`.

    If you specify `image`, you must also specify the image source in the following format

    • **source** – (required) The content of the conversation turn.

    • **type** – (required) The encoding type for the image. You can specify `base64`.

    • **media_type** – (required) The type of the image. You can specify the following image formats.
      
      • `image/jpeg`
      • `image/png`
      • `image/webp`
      • `image/gif`

    • **data** – (required) The `base64` encoded image bytes for the image. The maximum image size is 3.75MB. The maximum height and width of an image is 8000 pixels.

    If you specify `text`, you must also specify the prompt in `text`.

The following are optional parameters.

• **system** – (Optional) The system prompt for the request.

  A system prompt is a way of providing context and instructions to Anthropic Claude, such as specifying a particular goal or role. For more information, see [How to use system prompts](#) in the Anthropic documentation.
Note

You can use system prompts with Anthropic Claude version 2.1 or higher.

- **stop_sequences** – (Optional) Custom text sequences that cause the model to stop generating. Anthropic Claude models normally stop when they have naturally completed their turn, in this case the value of the `stop_reason` response field is `end_turn`. If you want the model to stop generating when it encounters custom strings of text, you can use the `stop_sequences` parameter. If the model encounters one of the custom text strings, the value of the `stop_reason` response field is `stop_sequence` and the value of `stop_sequence` contains the matched stop sequence.

  The maximum number of entries is 8191.

- **temperature** – (Optional) The amount of randomness injected into the response.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **top_p** – (Optional) Use nucleus sampling.

In nucleus sampling, Anthropic Claude computes the cumulative distribution over all the options for each subsequent token in decreasing probability order and cuts it off once it reaches a particular probability specified by `top_p`. You should alter either `temperature` or `top_p`, but not both.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.999</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The following are optional parameters.

- **top_k** – (Optional) Only sample from the top K options for each subsequent token.

  Use `top_k` to remove long tail low probability responses.
<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disabled by default</td>
<td>0</td>
<td>100,000,000</td>
</tr>
</tbody>
</table>

Response

The Anthropic Claude model returns the following fields for a messages inference call.

```json
{
  "id": string,
  "model": string,
  "type": "message",
  "role": "assistant",
  "content": [
    {
      "type": "text",
      "text": string
    }
  ],
  "stop_reason": string,
  "stop_sequence": string,
  "usage": {
    "input_tokens": integer,
    "output_tokens": integer
  }
}
```

- **id** – The unique identifier for the response. The format and length of the ID might change over time.
- **model** – The ID for the Anthropic Claude model that made the request.
- **stop_reason** – The reason why Anthropic Claude stopped generating the response.
  - **end_turn** – The model reached a natural stopping point
  - **max_tokens** – The generated text exceeded the value of the `max_tokens` input field or exceeded the maximum number of tokens that the model supports.
- **stop_sequence** – The model generated one of the stop sequences that you specified in the `stop_sequences` input field.
- **type** – The type of response. The value is always `message`.
- **role** – The conversational role of the generated message. The value is always `assistant`.
- **content** – The content generated by the model. Returned as an array.
  - **type** – The type of the content. Currently the only supported value is `text`.
  - **text** – The text of the content.
- **usage** – Container for the number of tokens that you supplied in the request and the number tokens of that the model generated in the response.
  - **input_tokens** – The number of input tokens in the request.
  - **output_tokens** – The number tokens of that the model generated in the response.
  - **stop_sequence** – The model generated one of the stop sequences that you specified in the `stop_sequences` input field.

**Code examples**

The following code examples show how to use the messages API.

**Topics**

- [Messages code example](#)
- [Multimodal code examples](#)

**Messages code example**

This example shows how to send a single turn user message and a user turn with a prefilled assistant message to an Anthropic Claude 3 Sonnet model.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0
"""
Shows how to generate a message with Anthropic Claude (on demand).
"""
import boto3
import json
import logging

from botocore.exceptions import ClientError
```
logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_message(bedrock_runtime, model_id, system_prompt, messages, max_tokens):

    body=json.dumps(
        {
            "anthropic_version": "bedrock-2023-05-31",
            "max_tokens": max_tokens,
            "system": system_prompt,
            "messages": messages
        }
    )

    response = bedrock_runtime.invoke_model(body=body, modelId=model_id)
    response_body = json.loads(response.get('body').read())

    return response_body

def main():
    
    try:

        bedrock_runtime = boto3.client(service_name='bedrock-runtime')

        model_id = 'anthropic.claude-3-sonnet-20240229-v1:0'
        system_prompt = "Please respond only with emoji."
        max_tokens = 1000

        # Prompt with user turn only.
        user_message = {"role": "user", "content": "Hello World"}
        messages = [user_message]

        response = generate_message (bedrock_runtime, model_id, system_prompt, messages, max_tokens)
        print("User turn only.")
        print(json.dumps(response, indent=4))

        # Prompt with both user turn and prefilled assistant response.
Multimodal code examples

The following examples show how to pass an image and prompt text in a multimodal message to an Anthropic Claude 3 Sonnet model.

Topics

• Multimodal prompt with InvokeModel
• Streaming multimodal prompt with InvokeModelWithResponseStream

Multimodal prompt with InvokeModel

The following example shows how to send a multimodal prompt to Anthropic Claude 3 Sonnet with InvokeModel.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
Shows how to run a multimodal prompt with Anthropic Claude (on demand) and InvokeModel.

""

import json
import logging
import base64
import boto3
```
from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def run_multi_modal_prompt(bedrock_runtime, model_id, messages, max_tokens):
    """
    Invokes a model with a multimodal prompt.
    Args:
      bedrock_runtime: The Amazon Bedrock boto3 client.
      model_id (str): The model ID to use.
      messages (JSON): The messages to send to the model.
      max_tokens (int): The maximum number of tokens to generate.
    Returns:
      None.
    """

    body = json.dumps(
        {
            "anthropic_version": "bedrock-2023-05-31",
            "max_tokens": max_tokens,
            "messages": messages
        }
    )

    response = bedrock_runtime.invoke_model(
        body=body, modelId=model_id)
    response_body = json.loads(response.get('body').read())

    return response_body

def main():
    """
    Entry point for Anthropic Claude multimodal prompt example.
    """

    try:
Streaming multimodal prompt with InvokeModelWithResponseStream

The following example shows how to stream the response from a multimodal prompt sent to Anthropic Claude 3 Sonnet with InvokeModelWithResponseStream.
Shows how to stream the response from Anthropic Claude Sonnet (on demand) for a multimodal request.

```python
import json
import base64
import logging
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def stream_multi_modal_prompt(bedrock_runtime, model_id, input_text, image, max_tokens):
    
    Streams the response from a multimodal prompt.
    Args:
    bedrock_runtime: The Amazon Bedrock boto3 client.
    model_id (str): The model ID to use.
    input_text (str): The prompt text
    image (str): The path to an image that you want in the prompt.
    max_tokens (int): The maximum number of tokens to generate.
    Returns:
    None.
    ```

    with open(image, "rb") as image_file:
        encoded_string = base64.b64encode(image_file.read())

    body = json.dumps(
        
        "anthropic_version": "bedrock-2023-05-31",
        "max_tokens": max_tokens,
        "messages": [
            
            "role": "user",
            "content": [
                
                "type": "text", "text": input_text,
                
                "type": "image", "source": {
                "type": "base64",
                "media_type": "image/jpeg", "data":

                encoded_string.decode('utf-8')
            ]
        ]
    )
```
```python
response = bedrock_runtime.invoke_model_with_response_stream(
    body=body, modelId=model_id)

for event in response.get("body"):
    chunk = json.loads(event["chunk"]["bytes"])

    if chunk["type"] == "message_delta":
        print(f"\nStop reason: {chunk["delta"]["stop_reason"]}")
        print(f"Stop sequence: {chunk["delta"]["stop_sequence"]}")
        print(f"Output tokens: {chunk["usage"]["output_tokens"]}")

    if chunk["type"] == "content_block_delta":
        if chunk["delta"]["type"] == "text_delta":
            print(chunk["delta"]["text"], end="")

def main():
    ""
    Entrypoint for Anthropic Claude Sonnet multimodal prompt example.
    ""

    model_id = "anthropic.claude-3-sonnet-20240229-v1:0"
    input_text = "What can you tell me about this image?"
    image = "/path/to/image"
    max_tokens = 100

    try:

        bedrock_runtime = boto3.client(\'bedrock-runtime\')

        stream_multi_modal_prompt(
            bedrock_runtime, model_id, input_text, image, max_tokens)

    except ClientError as err:
        message = err.response[\"Error\"][\"Message\"]
        logger.error("A client error occurred: %s", message)
        print("A client error occurred: " +
              format(message))
```

Anthropic Claude models
AI21 Labs Jurassic-2 models

This section provides inference parameters and a code example for using AI21 Labs AI21 Labs Jurassic-2 models.

Topics

• Inference parameters
• Code example

Inference parameters

The AI21 Labs Jurassic-2 models support the following inference parameters.

Topics

• Randomness and Diversity
• Length
• Repetitions
• Model invocation request body field
• Model invocation response body field

Randomness and Diversity

The AI21 Labs Jurassic-2 models support the following parameters to control randomness and diversity in the response.

• **Temperature** (temperature)– Use a lower value to decrease randomness in the response.
• **Top P** (topP) – Use a lower value to ignore less probable options.

Length

The AI21 Labs Jurassic-2 models support the following parameters to control the length of the generated response.
• **Max completion length** (maxTokens) – Specify the maximum number of tokens to use in the generated response.

• **Stop sequences** (stopSequences) – Configure stop sequences that the model recognizes and after which it stops generating further tokens. Press the Enter key to insert a newline character in a stop sequence. Use the Tab key to finish inserting a stop sequence.

**Repetitions**

The AI21 Labs Jurassic-2 models support the following parameters to control repetition in the generated response.

• **Presence penalty** (presencePenalty) – Use a higher value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion.

• **Count penalty** (countPenalty) – Use a higher value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion. Proportional to the number of appearances.

• **Frequency penalty** (frequencyPenalty) – Use a high value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion. The value is proportional to the frequency of the token appearances (normalized to text length).

• **Penalize special tokens** – Reduce the probability of repetition of special characters. The default values are true.
  
  • **Whitespaces** (applyToWhitespaces) – A true value applies the penalty to whitespaces and new lines.
  
  • **Punctuations** (applyToPunctuation) – A true value applies the penalty to punctuation.
  
  • **Numbers** (applyToNumbers) – A true value applies the penalty to numbers.
  
  • **Stop words** (applyToStopwords) – A true value applies the penalty to stop words.
  
  • **Emojis** (applyToEmojis) – A true value excludes emojis from the penalty.

**Model invocation request body field**

When you make an InvokeModel or InvokeModelWithResponseStream call using an AI21 Labs model, fill the body field with a JSON object that conforms to the one below. Enter the prompt in the prompt field.

```json
{
    "prompt": string,
    ...
}
```
Amazon Bedrock
User Guide

"temperature": float,
"topP": float,
"maxTokens": int,
"stopSequences": [string],
"countPenalty": {
    "scale": float
},
"presencePenalty": {
    "scale": float
},
"frequencyPenalty": {
    "scale": float
}
}

To penalize special tokens, add those fields to any of the penalty objects. For example, you can modify the countPenalty field as follows.

"countPenalty": {
    "scale": float,
    "applyToWhitespaces": boolean,
    "applyToPunctuations": boolean,
    "applyToNumbers": boolean,
    "applyToStopwords": boolean,
    "applyToEmojis": boolean
}

The following table shows the minimum, maximum, and default values for the numerical parameters.

<table>
<thead>
<tr>
<th>Category</th>
<th>Parameter</th>
<th>JSON object format</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomness and diversity</td>
<td>Temperature</td>
<td>temperature</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Top P</td>
<td>topP</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Length</td>
<td>Max tokens</td>
<td>maxTokens</td>
<td>0</td>
<td>8,191</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>(mid, ultra, and large models)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>Parameter</td>
<td>JSON object format</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Default</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------</td>
<td>--------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Max tokens</td>
<td>format</td>
<td>0</td>
<td>2,048</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>(other models)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetitions</td>
<td>Presence penalty</td>
<td>presencePenalty</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Count penalty</td>
<td>countPenalty</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Frequency penalty</td>
<td>frequencyPenalty</td>
<td>0</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Model invocation response body field**

For information about the format of the body field in the response, see [https://docs.ai21.com/reference/j2-complete-ref](https://docs.ai21.com/reference/j2-complete-ref).

**Note**

Amazon Bedrock returns the response identifier (`id`) as an integer value.

**Code example**

This examples shows how to call the A2I AI21 Labs Jurassic-2 Mid model.

```python
import boto3
import json

brt = boto3.client(service_name='bedrock-runtime')

body = json.dumps({
    "prompt": "Translate to spanish: 'Amazon Bedrock is the easiest way to build and scale generative AI applications with base models (FMs)'",
    "maxTokens": 200,
})
```
Cohere models

The following is inference parameters information for the Cohere models that Amazon Bedrock supports.

Topics

- [Cohere Command models](#)
- [Cohere Embed models](#)

Cohere Command models

You make inference requests to an Cohere Command model with [InvokeModel](#) or [InvokeModelWithResponseStream](#) (streaming). You need the model ID for the model that you want to use. To get the model ID, see [Amazon Bedrock model IDs](#).

Topics

- [Request and Response](#)
- [Code example](#)
Request and Response

Request

The Cohere Command models have the following inference parameters.

```json
{
    "prompt": string,
    "temperature": float,
    "p": float,
    "k": float,
    "max_tokens": int,
    "stop_sequences": [string],
    "return_likelihoods": "GENERATION|ALL|NONE",
    "stream": boolean,
    "num_generations": int,
    "logit_bias": {token_id: bias},
    "truncate": "NONE|START|END"
}
```

The following are required parameters.

- **prompt** – (Required) The input text that serves as the starting point for generating the response.

The following are text per call and character limits.

The following are optional parameters.

- **return_likelihoods** – Specify how and if the token likelihoods are returned with the response. You can specify the following options.
  - GENERATION – Only return likelihoods for generated tokens.
  - ALL – Return likelihoods for all tokens.
  - NONE – (Default) Don't return any likelihoods.

- **stream** – (Required to support streaming) Specify true to return the response piece-by-piece in real-time and false to return the complete response after the process finishes.

- **logit_bias** – Prevents the model from generating unwanted tokens or incentivizes the model to include desired tokens. The format is `{token_id: bias}` where bias is a float between
-10 and 10. Tokens can be obtained from text using any tokenization service, such as Cohere's Tokenize endpoint. For more information, see Cohere documentation.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>-10 (for a token bias)</td>
<td>10 (for a token bias)</td>
</tr>
</tbody>
</table>

- **num_generations** – The maximum number of generations that the model should return.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

- **truncate** – Specifies how the API handles inputs longer than the maximum token length. Use one of the following:
  - **NONE** – Returns an error when the input exceeds the maximum input token length.
  - **START** – Discard the start of the input.
  - **END** – (Default) Discards the end of the input.

If you specify START or END, the model discards the input until the remaining input is exactly the maximum input token length for the model.

- **temperature** – Use a lower value to decrease randomness in the response.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

- **p** – Top P. Use a lower value to ignore less probable options. Set to 0 or 1.0 to disable. If both p and k are enabled, p acts after k.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **k** – Top K. Specify the number of token choices the model uses to generate the next token. If both p and k are enabled, p acts after k.
- **max_tokens** – Specify the maximum number of tokens to use in the generated response.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>500</td>
</tr>
</tbody>
</table>

- **stop_sequences** – Configure up to four sequences that the model recognizes. After a stop sequence, the model stops generating further tokens. The returned text doesn't contain the stop sequence.

Response

The response has the following possible fields:

```json
{
  "generations": [
    {
      "finish_reason": "COMPLETE | MAX_TOKENS | ERROR | ERROR_TOXIC",
      "id": string,
      "text": string,
      "likelihood": float,
      "token_likelihoods": [{"token": float}],
      "is_finished": true | false,
      "index": integer
    }
  ],
  "id": string,
  "prompt": string
}
```

- **generations** — A list of generated results along with the likelihoods for tokens requested. (Always returned). Each generation object in the list contains the following fields.
  - **id** — An identifier for the generation. (Always returned).
• likelihood — The likelihood of the output. The value is the average of the token likelihoods in token_likelihoods. Returned if you specify the return_likelihoods input parameter.

• token_likelihoods — An array of per token likelihoods. Returned if you specify the return_likelihoods input parameter.

• finish_reason — The reason why the model finished generating tokens. COMPLETE - the model sent back a finished reply. MAX_TOKENS – the reply was cut off because the model reached the maximum number of tokens for its context length. ERROR – something went wrong when generating the reply. ERROR_TOXIC – the model generated a reply that was deemed toxic. finish_reason is returned only when is_finished=true. (Not always returned).

• is_finished — A boolean field used only when stream is true, signifying whether or not there are additional tokens that will be generated as part of the streaming response. (Not always returned)

• text — The generated text.

• index — In a streaming response, use to determine which generation a given token belongs to. When only one response is streamed, all tokens belong to the same generation and index is not returned. index therefore is only returned in a streaming request with a value for num_generations that is larger than one.

• prompt — The prompt from the input request (always returned).

• id — An identifier for the request (always returned).

For more information, see https://docs.cohere.com/reference/generate in the Cohere documentations.

Code example

This examples shows how to call the Cohere Command model.

```
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

"""
Shows how to generate text using a Cohere model.
"""
import json
import logging
```
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text(model_id, body):
    """
    Generate text using a Cohere model.
    Args:
        model_id (str): The model ID to use.
        body (str) : The request body to use.
    Returns:
        dict: The response from the model.
    """

    logger.info("Generating text with Cohere model %s", model_id)

    accept = 'application/json'
    content_type = 'application/json'

    bedrock = boto3.client(service_name='bedrock-runtime')

    response = bedrock.invoke_model(
        body=body,
        modelId=model_id,
        accept=accept,
        contentType=content_type
    )

    logger.info("Successfully generated text with Cohere model %s", model_id)

    return response

def main():
    """
    Entrypoint for Cohere example.
    """

    logging.basicConfig(level=logging.INFO,
model_id = 'cohere.command-text-v14'
prompt = """Summarize this dialogue:
"Customer: Please connect me with a support agent.
AI: Hi there, how can I assist you today?
Customer: I forgot my password and lost access to the email affiliated to my account. Can you please help me?
AI: Yes of course. First I'll need to confirm your identity and then I can connect you with one of our support agents.
""

try:
    body = json.dumps({
        "prompt": prompt,
        "max_tokens": 200,
        "temperature": 0.6,
        "p": 1,
        "k": 0,
        "num_generations": 2,
        "return_likelihoods": "GENERATION"
    })
    response = generate_text(model_id=model_id, 
                              body=body)

    response_body = json.loads(response.get('body').read())
    generations = response_body.get('generations')

    for index, generation in enumerate(generations):
        print(f"Generation {index + 1}\n""""
        print(f"Text:\n" {generation['text']}\n")
        if 'likelihood' in generation:
            print(f"Likelihood:\n" {generation['likelihood']}\n")

            print(f"Reason: {generation['finish_reason']}\n"")

except ClientError as err:
    message = err.response["Error"]['Message']
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " + 
          format(message))
else:
    print(f"Finished generating text with Cohere model {model_id}.")
Cohere Embed models

You make inference requests to an Embed model with `InvokeModel`. You need the model ID for the model that you want to use. To get the model ID, see Amazon Bedrock model IDs.

**Note**

Amazon Bedrock doesn't support streaming responses from Cohere Embed models.

Topics

- Request and Response
- Code example

Request and Response

Request

The Cohere Embed models have the following inference parameters.

```json
{
    "texts": [string],
    "input_type": "search_document|search_query|classification|clustering",
    "truncate": "NONE|LEFT|RIGHT"
}
```

The following are required parameters.

- **texts** – (Required) An array of strings for the model to embed. For optimal performance, we recommend reducing the length of each text to less than 512 tokens. 1 token is about 4 characters.

  The following are text per call and character limits.

  **Texts per call**
The following are optional parameters.

- **input_type** – Prepends special tokens to differentiate each type from one another. You should not mix different types together, except when mixing types for for search and retrieval. In this case, embed your corpus with the `search_document` type and embedded queries with type `search_query` type.
  
  - **search_document** – In search use-cases, use `search_document` when you encode documents for embeddings that you store in a vector database.
  
  - **search_query** – Use `search_query` when querying your vector DB to find relevant documents.
  
  - **classification** – Use `classification` when using embeddings as an input to a text classifier.
  
  - **clustering** – Use `clustering` to cluster the embeddings.

- **truncate** – Specifies how the API handles inputs longer than the maximum token length. Use one of the following:
  
  - **NONE** – (Default) Returns an error when the input exceeds the maximum input token length.
  
  - **LEFT** – Discard the start of the input.
  
  - **RIGHT** – Discards the end of the input.

If you specify LEFT or RIGHT, the model discards the input until the remaining input is exactly the maximum input token length for the model.
For more information, see https://docs.cohere.com/reference/embed in the Cohere documentation.

Response

The body response from a call to InvokeModel is the following:

```
{
    "embeddings": [
        [ <array of 1024 floats> ]
    ],
    "id": string,
    "response_type": "embeddings_floats,
    "texts": [string]
}
```

The body response has the following fields:

- **id** – An identifier for the response.
- **response_type** – The response type. This value is always embeddings_floats.
- **embeddings** – An array of embeddings, where each embedding is an array of floats with 1024 elements. The length of the embeddings array will be the same as the length of the original texts array.
- **texts** – An array containing the text entries for which embeddings were returned.

For more information, see https://docs.cohere.com/reference/embed.

Code example

This example shows how to call the Cohere Command Embed model.

```
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
"Shows how to generate text embeddings using the Cohere Embed English model.
""
import json
import logging
```
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text_embeddings(model_id, body):
    """
    Generate text embedding by using the Cohere Embed model.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        dict: The response from the model.
    """

    logger.info("Generating text embeddings with the Cohere Embed model %s", model_id)

    accept = '*/*'
    content_type = 'application/json'

    bedrock = boto3.client(service_name='bedrock-runtime')

    response = bedrock.invoke_model(
        body=body,
        modelId=model_id,
        accept=accept,
        contentType=content_type
    )

    logger.info("Successfully generated text with Cohere model %s", model_id)

    return response

def main():
    """
    Entry point for Cohere Embed example.
    """

    # Your main function implementation here
```python
logging.basicConfig(level=logging.INFO,
                     format="%(levelname)s: %(message)s")

model_id = 'cohere.embed-english-v3'
text1 = "hello world"
text2 = "this is a test"
input_type = "search_document"

try:

    body = json.dumps(
        {
            "texts": [
                text1,
                text2,
                "input_type": input_type
            ]
        }
    )
    response = generate_text_embeddings(model_id=model_id,
                                         body=body)

    response_body = json.loads(response.get('body').read())

    print(f"ID: {response_body.get('id')}")
    print(f"Response type: {response_body.get('response_type')}")

    print("Embeddings")
    for i, embedding in enumerate(response_body.get('embeddings')):
        print(f"Embedding {i}")
        print(*embedding)

    print("Texts")
    for i, text in enumerate(response_body.get('texts')):
        print(f"Text {i}: {text}"

except ClientError as err:
    message = err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
          format(message))
else:
    print(f"Finished generating text embeddings with Cohere model {model_id}.")
```

if __name__ == "__main__":

```
Meta Llama 2 and Meta Llama 2 Chat models

This section provides inference parameters and a code example for using Meta Llama 2 and Meta Llama 2 Chat models.

Topics

• Request and response
• Example code

Request and response

The request body is passed in the body field of a request to InvokeModel or InvokeModelWithResponseStream.

Request

Meta Llama 2 Chat and Llama 2 models have the following inference parameters.

```json
{
  "prompt": string,
  "temperature": float,
  "top_p": float,
  "max_gen_len": int
}
```

The following are required parameters.

• prompt – (Required) The prompt that you want to pass to the model.

The following are optional parameters.

• temperature – Use a lower value to decrease randomness in the response.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
- **top_p** – Use a lower value to ignore less probable options. Set to 0 or 1.0 to disable.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **max_gen_len** – Specify the maximum number of tokens to use in the generated response. The model truncates the response once the generated text exceeds `max_gen_len`.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>1</td>
<td>2048</td>
</tr>
</tbody>
</table>

Response

Meta Llama 2 and Llama 2 Chat models return the following fields for a text completion inference call.

```json
{
  "generation": "\n\n<response>",
  "prompt_token_count": int,
  "generation_token_count": int,
  "stop_reason": string
}
```

More information about each field is provided below.

- **generation** – The generated text.
- **prompt_token_count** – The number of tokens in the prompt.
- **generation_token_count** – The number of tokens in the generated text.
- **stop_reason** – The reason why the response stopped generating text. Possible values are:
  - **stop** – The model has finished generating text for the input prompt.
  - **length** – The length of the tokens for the generated text exceeds the value of `max_gen_len` in the call to InvokeModel (InvokeModelWithResponseStream, if you are streaming output). The response is truncated to `max_gen_len` tokens. Consider increasing the value of `max_gen_len` and trying again.
Example code

This example shows how to call the *Meta Llama 2 Chat 13B* model.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

"""
Shows how to generate text with Meta Llama 2 Chat (on demand).
"""

import json
import logging
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text(model_id, body):
    """
    Generate an image using Meta Llama 2 Chat on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        response (JSON): The text that the model generated, token information, and the reason the model stopped generating text.
    """

    logger.info("Generating image with Meta Llama 2 Chat model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )
```
response_body = json.loads(response.get('body').read())

return response_body

def main():
    """
    Entrypoint for Meta Llama 2 Chat example.
    """

    logging.basicConfig(level=logging.INFO,
                        format="%(levelname)s: %(message)s")

    model_id = 'meta.llama2-13b-chat-v1'
    prompt = """What is the average lifespan of a Llama?"""
    max_gen_len = 128
    temperature = 0.1
    top_p = 0.9

    # Create request body.
    body = json.dumps(
               {
                   "prompt": prompt,
                   "max_gen_len": max_gen_len,
                   "temperature": temperature,
                   "top_p": top_p
               })

    try:

        response = generate_text(model_id, body)

        print(f"Generated Text: {response['generation']}"
        print(f"Prompt Token count: {response['prompt_token_count']}"
        print(f"Generation Token count: {response['generation_token_count']}"
        print(f"Stop reason: {response['stop_reason']}"

    except ClientError as err:
        message = err.response["Error"]["Message"
        logger.error("A client error occurred: %s", message)
        print("A client error occurred: " +
                      format(message))

Meta Llama 2 models
Mistral AI models

You make inference requests to Mistral 7B Instruct and Mixtral 8X7B Instruct models with `InvokeModel` or `InvokeModelWithResponseStream` (streaming). You need the model ID for the model that you want to use. To get the model ID, see Amazon Bedrock model IDs.

Mistral AI models are available under the Apache 2.0 license. For more information about using Mistral AI models, see the Mistral AI documentation.

Topics
- Request and Response
- Code example

Request and Response

Request

The Mistral AI models have the following inference parameters.

```json
{
    "prompt": string,
    "max_tokens" : int,
    "stop" : [string],
    "temperature": float,
    "top_p": float,
    "top_k": int
}
```

The following are required parameters.

- **prompt** – (Required) The prompt that you want to pass to the model, as shown in the following example.
<s>What is your favourite condiment? [/INST]

The following example shows how to format a multi-turn prompt.

<s>What is your favourite condiment? [/INST]
Well, I'm quite partial to a good squeeze of fresh lemon juice. It adds just the right amount of zesty flavour to whatever I'm cooking up in the kitchen!</s>

[INST] Do you have mayonnaise recipes? [/INST]

Text for the user role is inside the [INST]...[/INST] tokens, text outside is the assistant role. The beginning and ending of a string are represented by the <s> (beginning of string) and </s> (end of string) tokens. For information about sending a chat prompt in the correct format, see Chat template in the Mistral AI documentation.

The following are optional parameters.

- **max_tokens** – Specify the maximum number of tokens to use in the generated response. The model truncates the response once the generated text exceeds max_tokens.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
</table>
| 512     | 1       | Mistral 7B Instruct – 8,192  
          |         | Mixtral 8X7B Instruct – 4,096 |

- **stop** – A list of stop sequences that if generated by the model, stops the model from generating further output.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

- **temperature** – Controls the randomness of predictions made by the model. For more information, see Inference parameters.
• **top_p** – Controls the diversity of text that the model generates by setting the percentage of most-likely candidates that the model considers for the next token. For more information, see [Inference parameters](#).

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

• **top_k** – Controls the number of most-likely candidates that the model considers for the next token. For more information, see [Inference parameters](#).

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Response

The body response from a call to InvokeModel is the following:

```
{
   "outputs": [
   {
      "text": string,
      "stop_reason": string
   }
   ]
}
```

The body response has the following fields:

• **outputs** – A list of outputs from the model. Each output has the following fields.
  • **text** – The text that the model generated.
• **stop_reason** – The reason why the response stopped generating text. Possible values are:
  - **stop** – The model has finished generating text for the input prompt. The model stops because it has no more content to generate or if the model generates one of the stop sequences that you define in the stop request parameter.
  - **length** – The length of the tokens for the generated text exceeds the value of max_tokens in the call to InvokeModel (InvokeModelWithResponseStream, if you are streaming output). The response is truncated to max_tokens tokens.

**Code example**

This examples shows how to call the Mistral 7B Instruct model.

```
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

""
Shows how to generate text using a Mistral AI model.
""
import json
import logging
import boto3

from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text(model_id, body):
    ""
    Generate text using a Mistral AI model.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        JSON: The response from the model.
    ""

    logger.info("Generating text with Mistral AI model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')
```
response = bedrock.invoke_model(
    body=body,
    modelId=model_id
)

logger.info("Successfully generated text with Mistral AI model %s", model_id)

return response

def main():
    ""
    Entrypoint for Mistral AI example.
    ""

    logging.basicConfig(level=logging.INFO,
        format="%(levelname)s: %(message)s")

    try:
        model_id = 'mistral.mistral-7b-instruct-v0:2'

        prompt = """<s>[INST] In Bash, how do I list all text files in the current
directory
(excluding subdirectories) that have been modified in the last month? [/INST]"""

        body = json.dumps(
            {
                "prompt": prompt,
                "max_tokens": 400,
                "temperature": 0.7,
                "top_p": 0.7,
                "top_k": 50
            }
        )

        response = generate_text(model_id=model_id,
            body=body)

        response_body = json.loads(response.get('body').read())

        outputs = response_body.get('outputs')

        for index, output in enumerate(outputs):
Stability.ai Diffusion models

The following is inference parameters information for the Stability.ai Diffusion models that Amazon Bedrock supports.

Models

- Stability.ai Diffusion 0.8
- Stability.ai Diffusion 1.0 text to image
- Stability.ai Diffusion 1.0 image to image
- Stability.ai Diffusion 1.0 image to image (masking)

Stability.ai Diffusion 0.8

The Stability.ai Diffusion models have the following controls.

- **Prompt strength** (cfg_scale) – Determines how much the final image portrays the prompt. Use a lower number to increase randomness in the generation.
- **Generation step** (steps) – Generation step determines how many times the image is sampled. More steps can result in a more accurate result.
- **Seed** (seed) – The seed determines the initial noise setting. Use the same seed and the same settings as a previous run to allow inference to create a similar image. If you don’t set this value, it is set as a random number.
Model invocation request body field

When you make an `InvokeModel` or `InvokeModelWithResponseStream` call using a Stability.ai model, fill the body field with a JSON object that conforms to the one below. Enter the prompt in the `text` field in the `text_prompts` object.

```json
{
    "text_prompts": [
        {"text": "string"}
    ],
    "cfg_scale": float,
    "steps": int,
    "seed": int
}
```

The following table shows the minimum, maximum, and default values for the numerical parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>JSON object format</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt strength</td>
<td>cfg_scale</td>
<td>0</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Generation step</td>
<td>steps</td>
<td>10</td>
<td>150</td>
<td>30</td>
</tr>
</tbody>
</table>

Model invocation response body field

For information about the format of the body field in the response, see [https://platform.stability.ai/docs/api-reference#tag/v1generation](https://platform.stability.ai/docs/api-reference#tag/v1generation).

Stability.ai Diffusion 1.0 text to image

The Stability.ai Diffusion 1.0 model has the following inference parameters and model response for making text to image inference calls.

Topics

- Request and Response
- Code example
Request and Response

The request body is passed in the body field of a request to `InvokeModel` or `InvokeModelWithResponseStream`.

For more information, see [https://platform.stability.ai/docs/api-reference#tag/v1generation](https://platform.stability.ai/docs/api-reference#tag/v1generation).

Request

The Stability.ai Diffusion 1.0 model has the following inference parameters for a text to image inference call.

```
{
    "text_prompts": [
        {
            "text": string,
            "weight": float
        }
    ],
    "height": int,
    "width": int,
    "cfg_scale": float,
    "clip_guidance_preset": string,
    "sampler": string,
    "samples",
    "seed": int,
    "steps": int,
    "style_preset": string,
    "extras" :JSON object
}
```

- **text_prompts** (Required) – An array of text prompts to use for generation. Each element is a JSON object that contains a prompt and a weight for the prompt.
- **text** – The prompt that you want to pass to the model.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>

• **weight** (Optional) – The weight that the model should apply to the prompt. A value that is less than zero declares a negative prompt. Use a negative prompt to tell the model to avoid certain concepts. The default value for weight is one.

• **cfg_scale** – (Optional) Determines how much the final image portrays the prompt. Use a lower number to increase randomness in the generation.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>35</td>
<td>7</td>
</tr>
</tbody>
</table>

• **clip_guidance_preset** – (Optional) Enum: FAST_BLUE, FAST_GREEN, NONE, SIMPLE SLOW, SLOWER, SLOWEST.

• **height** – (Optional) Height of the image to generate, in pixels, in an increment divisible by 64. The value must be one of 1024x1024, 1152x896, 1216x832, 1344x768, 1536x640, 640x1536, 768x1344, 832x1216, 896x1152.

• **width** – (Optional) Width of the image to generate, in pixels, in an increment divisible by 64. The value must be one of 1024x1024, 1152x896, 1216x832, 1344x768, 1536x640, 640x1536, 768x1344, 832x1216, 896x1152.

• **sampler** – (Optional) The sampler to use for the diffusion process. If this value is omitted, the model automatically selects an appropriate sampler for you.

  Enum: DDIM, DDPM, K_DPMPP_2M, K_DPMPP_2S_ANCESTRAL, K_DPM_2, K_DPM_2_ANCESTRAL, K_EULER, K_EULER_ANCESTRAL, K_HEUN K_LMS.

• **samples** – (Optional) The number of image to generate. Currently Amazon Bedrock supports generating one image. If you supply a value for samples, the value must be one.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• **seed** – (Optional) The seed determines the initial noise setting. Use the same seed and the same settings as a previous run to allow inference to create a similar image. If you don’t set this value, or the value is 0, it is set as a random number.
<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4294967295</td>
<td>0</td>
</tr>
</tbody>
</table>

- **steps** – (Optional) Generation step determines how many times the image is sampled. More steps can result in a more accurate result.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>50</td>
<td>30</td>
</tr>
</tbody>
</table>

- **style_preset** (Optional) – A style preset that guides the image model towards a particular style. This list of style presets is subject to change.


- **extras** (Optional) – Extra parameters passed to the engine. Use with caution. These parameters are used for in-development or experimental features and might change without warning.

Response

The Stability.ai Diffusion 1.0 model returns the following fields for a text to image inference call.

```json
{
  "result": string,
  "artifacts": [
    {
      "seed": int,
      "base64": string,
      "finishReason": string
    }
  ]
}
```

- **result** – The result of the operation. If successful, the response is success.
Code example

The following example shows how to run inference with the Stability.ai Diffusion 1.0 model and on demand throughput. The example submits a text prompt to a model, retrieves the response from the model, and finally shows the image.

```python
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
# SPDX-License-Identifier: Apache-2.0

"""
Shows how to generate an image with SDXL 1.0 (on demand).
"""
import base64
import io
import json
import logging
import boto3
from PIL import Image
from botocore.exceptions import ClientError

class ImageError(Exception):
    """Custom exception for errors returned by SDXL"
    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_image(model_id, body):
    pass
```
Generate an image using SDXL 1.0 on demand.

Args:
    model_id (str): The model ID to use.
    body (str): The request body to use.

Returns:
    image_bytes (bytes): The image generated by the model.

logger.info("Generating image with SDXL model %s", model_id)

bedrock = boto3.client(service_name='bedrock-runtime')

accept = "application/json"
content_type = "application/json"

response = bedrock.invoke_model(
    body=body, modelId=model_id, accept=accept, contentType=content_type
)
response_body = json.loads(response.get("body").read())
print(response_body[\'result\'])

base64_image = response_body.get("artifacts")[0].get("base64")
base64_bytes = base64_image.encode('ascii')
image_bytes = base64.b64decode(base64_bytes)

finish_reason = response_body.get("artifacts")[0].get("finishReason")

if finish_reason == 'ERROR' or finish_reason == 'CONTENT_FILTERED':
    raise ImageError(f"Image generation error. Error code is {finish_reason}")

logger.info("Successfully generated image with the SDXL 1.0 model %s", model_id)

return image_bytes

def main():
    """
    Entrypoint for SDXL example.
    """

    logging.basicConfig(level = logging.INFO,
format = "%(levelname)s: %(message)s"

model_id='stability.stable-diffusion-xl-v1'

prompt="""Sri lanka tea plantation."""

# Create request body.
body=json.dumps({
    "text_prompts": [
        {
            "text": prompt
        }
    ],
    "cfg_scale": 10,
    "seed": 0,
    "steps": 50,
    "samples": 1,
    "style_preset": "photographic"
})

try:
    image_bytes=generate_image(model_id = model_id,
                                body = body)
    image = Image.open(io.BytesIO(image_bytes))
    image.show()

except ClientError as err:
    message=err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " + format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print(f"Finished generating text with SDXL model {model_id}.")

if __name__ == "__main__":
    main()
Stability.ai Diffusion 1.0 image to image

The Stability.ai Diffusion 1.0 model has the following inference parameters and model response for making image to image inference calls.

Topics

- **Request and Response**
- **Code example**

Request and Response

The request body is passed in the body field of a request to `InvokeModel` or `InvokeModelWithResponseStream`.

For more information, see [https://platform.stability.ai/docs/api-reference#tag/v1 generation/operation/imageToImage](https://platform.stability.ai/docs/api-reference#tag/v1 generation/operation/imageToImage).

Request

The Stability.ai Diffusion 1.0 model has the following inference parameters for an image to image inference call.

```json
{
    "text_prompts": [
        {
            "text": string,
            "weight": float
        }
    ],
    "init_image": string,
    "init_image_mode": string,
    "image_strength": float,
    "cfg_scale": float,
    "clip_guidance_preset": string,
    "sampler": string,
    "samples": int,
    "seed": int,
    "steps": int,
```
The following are required parameters.

- **text_prompts** – (Required) An array of text prompts to use for generation. Each element is a JSON object that contains a prompt and a weight for the prompt.
- **text** – The prompt that you want to pass to the model.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>

- **weight** – (Optional) The weight that the model should apply to the prompt. A value that is less than zero declares a negative prompt. Use a negative prompt to tell the model to avoid certain concepts. The default value for weight is one.
- **init_image** – (Required) The base64 encoded image that you want to use to initialize the diffusion process.

The following are optional parameters.

- **init_image_mode** – (Optional) Determines whether to use image_strength or step_schedule_* to control how much influence the image in init_image has on the result. Possible values are IMAGE_STRENGTH or STEP_SCHEDULE. The default is IMAGE_STRENGTH.
- **image_strength** – (Optional) Determines how much influence the source image in init_image has on the diffusion process. Values close to 1 yield images very similar to the source image. Values close to 0 yield images very different than the source image.
- **cfg_scale** – (Optional) Determines how much the final image portrays the prompt. Use a lower number to increase randomness in the generation.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
<td>35</td>
</tr>
</tbody>
</table>
• **clip_guidance_preset** – (Optional) Enum: FAST_BLUE, FAST_GREEN, NONE, SIMPLE, SLOW, SLOWER, SLOWEST.

• **sampler** – (Optional) The sampler to use for the diffusion process. If this value is omitted, the model automatically selects an appropriate sampler for you.

  Enum: DDIM DDPM, K_DPMPP_2M, K_DPMPP_2S_ANCESTRAL, K_DPM_2, K_DPM_2_ANCESTRAL, K_EULER, K_EULER_ANCESTRAL, K_HEUN, K_LMS.

• **samples** – (Optional) The number of image to generate. Currently Amazon Bedrock supports generating one image. If you supply a value for `samples`, the value must be one.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• **seed** – (Optional) The seed determines the initial noise setting. Use the same seed and the same settings as a previous run to allow inference to create a similar image. If you don’t set this value, or the value is 0, it is set as a random number.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4294967295</td>
</tr>
</tbody>
</table>

• **steps** – (Optional) Generation step determines how many times the image is sampled. More steps can result in a more accurate result.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>10</td>
<td>50</td>
</tr>
</tbody>
</table>

• **style_preset** – (Optional) A style preset that guides the image model towards a particular style. This list of style presets is subject to change.

• **extras** – (Optional) Extra parameters passed to the engine. Use with caution. These parameters are used for in-development or experimental features and might change without warning.

Response

The Stability.ai Diffusion 1.0 model returns the following fields for a text to image inference call.

```json
{
    "result": string,
    "artifacts": [
        {
            "seed": int,
            "base64": string,
            "finishReason": string,
        }
    ]
}
```

• **result** – The result of the operation. If successful, the response is `success`.

• **artifacts** – An array of images, one for each requested image.
  
  • **seed** – The value of the seed used to generate the image.
  
  • **base64** – The base64 encoded image that the model generated.
  
  • **finishReason** – The result of the image generation process. Valid values are:
    
    • **SUCCESS** – The image generation process succeeded.
    
    • **ERROR** – An error occurred.
    
    • **CONTENT_FILTERED** – The content filter filtered the image and the image might be blurred.

**Code example**

The following example shows how to run inference with the Stability.ai Diffusion 1.0 model and on demand throughput. The example submits a text prompt and reference image to a model, retrieves the response from the model, and finally shows the image.
Shows how to generate an image from a reference image with SDXL 1.0 (on demand).

```python
import base64
import io
import json
import logging
import boto3
from PIL import Image

from botocore.exceptions import ClientError

class ImageError(Exception):
    "Custom exception for errors returned by SDXL"
    def __init__(self, message):
        self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_image(model_id, body):
    """
    Generate an image using SDXL 1.0 on demand.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        image_bytes (bytes): The image generated by the model.
    """

    logger.info("Generating image with SDXL model %s", model_id)

    bedrock = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = bedrock.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type
    )
    response_body = json.loads(response.get("body").read())
```
print(response_body['result'])

base64_image = response_body.get("artifacts")[0].get("base64")
base64_bytes = base64_image.encode('ascii')
image_bytes = base64.b64decode(base64_bytes)

finish_reason = response_body.get("artifacts")[0].get("finishReason")

if finish_reason == 'ERROR' or finish_reason == 'CONTENT_FILTERED':
    raise ImageError(f"Image generation error. Error code is {finish_reason}"

logger.info("Successfully generated image with the SDXL 1.0 model %s", model_id)

return image_bytes

def main():
    """
    Entrypoint for SDXL example.
    """

    logging.basicConfig(level = logging.INFO,
                        format = "%(levelname)s: %(message)s")

    model_id='stability.stable-diffusion-xl-v1'

    prompt="""A space ship."""

    # Read reference image from file and encode as base64 strings.
    with open("/path/to/image", "rb") as image_file:
        init_image = base64.b64encode(image_file.read()).decode('utf8')

    # Create request body.
    body=json.dumps({
        "text_prompts": [ {
            "text": prompt
        } ],
        "init_image": init_image,
        "style_preset": "isometric"
    })
try:
    image_bytes=generate_image(model_id = model_id,
                               body = body)
    image = Image.open(io.BytesIO(image_bytes))
    image.show()

except ClientError as err:
    message=err.response["Error"]["Message"]
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
          format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print(f"Finished generating text with SDXL model {model_id}.")

if __name__ == "__main__":
    main()

---

**Stability.ai Diffusion 1.0 image to image (masking)**

The Stability.ai Diffusion 1.0 model has the following inference parameters and model response for using masks with image to image inference calls.

**Request and Response**

The request body is passed in the body field of a request to `InvokeModel` or `InvokeModelWithResponseStream`.

For more information, see [https://platform.stability.ai/docs/api-reference#tag/v1generation/operation/masking](https://platform.stability.ai/docs/api-reference#tag/v1generation/operation/masking).

**Request**

The Stability.ai Diffusion 1.0 model has the following inference parameters for an image to image (masking) inference call.
The following are required parameters.

- **text_prompt** – (Required) An array of text prompts to use for generation. Each element is a JSON object that contains a prompt and a weight for the prompt.
- **text** – The prompt that you want to pass to the model.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>

- **weight** – (Optional) The weight that the model should apply to the prompt. A value that is less than zero declares a negative prompt. Use a negative prompt to tell the model to avoid certain concepts. The default value for weight is one.
- **init_image** – (Required) The base64 encoded image that you want to use to initialize the diffusion process.
- **mask_source** – (Required) Determines where to source the mask from. Possible values are:
  - **MASK_IMAGE_WHITE** – Use the white pixels of the mask image in mask_image as the mask. White pixels are replaced and black pixels are left unchanged.
• **MASK_IMAGE_BLACK** – Use the black pixels of the mask image in `mask_image` as the mask. Black pixels are replaced and white pixels are left unchanged.

• **INIT_IMAGE_ALPHA** – Use the alpha channel of the image in `init_image` as the mask, Fully transparent pixels are replaced and fully opaque pixels are left unchanged.

• **mask_image** – (Required) The base64 encoded mask image that you want to use as a mask for the source image in `init_image`. Must be the same dimensions as the source image. Use the `mask_source` option to specify which pixels should be replaced.

The following are optional parameters.

• **cfg_scale** – (Optional) Determines how much the final image portrays the prompt. Use a lower number to increase randomness in the generation.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
<td>35</td>
</tr>
</tbody>
</table>

• **clip_guidance_preset** – (Optional) Enum: FAST_BLUE, FAST_GREEN, NONE, SIMPLE, SLOW, SLOWER, SLOWEST.

• **sampler** – (Optional) The sampler to use for the diffusion process. If this value is omitted, the model automatically selects an appropriate sampler for you.

  Enum: DDIM, DDPM, K_DPMPP_2M, K_DPMPP_2S_ANCESTRAL, K_DPM_2, K_DPM_2_ANCESTRAL, K_EULER, K_EULER_ANCESTRAL, K_HEUN K_LMS.

• **samples** – (Optional) The number of image to generate. Currently Amazon Bedrock supports generating one image. If you supply a value for `samples`, the value must be one.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• **seed** – (Optional) The seed determines the initial noise setting. Use the same seed and the same settings as a previous run to allow inference to create a similar image. If you don't set this value, or the value is 0, it is set as a random number.
• **steps** – (Optional) Generation step determines how many times the image is sampled. More steps can result in a more accurate result.

<table>
<thead>
<tr>
<th>Default</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4294967295</td>
</tr>
</tbody>
</table>

• **style_preset** – (Optional) A style preset that guides the image model towards a particular style. This list of style presets is subject to change.


• **extras** – (Optional) Extra parameters passed to the engine. Use with caution. These parameters are used for in-development or experimental features and might change without warning.

Response

The Stability.ai Diffusion 1.0 model returns the following fields for a text to image inference call.

```json
{
  "result": string,
  "artifacts": [ 
    {
      "seed": int,
      "base64": string,
      "finishReason": string
    }
  ]
}
```

• **result** – The result of the operation. If successful, the response is `success`.
- **artifacts** – An array of images, one for each requested image.
- **seed** – The value of the seed used to generate the image.
- **base64** – The base64 encoded image that the model generated.
- **finishedReason** – The result of the image generation process. Valid values are:
  - **SUCCESS** – The image generation process succeeded.
  - **ERROR** – An error occurred.
  - **CONTENT_FILTERED** – The content filter filtered the image and the image might be blurred.

**Custom model hyperparameters**

The following reference content covers the hyperparameters that are available for training each Amazon Bedrock custom model.

A hyperparameter is a parameter that controls the training process, such as the learning rate or epoch count. You set hyperparameters for custom model training when you submit the fine tuning job with the Amazon Bedrock console or by calling the `CreateModelCustomizationJob` API operation. For guidelines on hyperparameter settings, see [Guidelines for model customization](#).

**Topics**

- [Amazon Titan text model customization hyperparameters](#)
- [Amazon Titan Image Generator G1 model customization hyperparameters](#)
- [Amazon Titan Multimodal Embeddings G1 customization hyperparameters](#)
- [Cohere Command model customization hyperparameters](#)
- [Meta Llama 2 model customization hyperparameters](#)

**Amazon Titan text model customization hyperparameters**

Titan text models support the following hyperparameters for model customization.
<table>
<thead>
<tr>
<th>Hyperparameter (console)</th>
<th>Hyperparameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>epochCount</td>
<td>The number of iterations through the entire training dataset</td>
<td>integer</td>
<td>1</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Batch size</td>
<td>batchSize</td>
<td>The number of samples processed before updating model parameters</td>
<td>integer</td>
<td>1</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>Learning rate</td>
<td>learningRate</td>
<td>The rate at which model parameters are updated after each batch</td>
<td>float</td>
<td>0.0</td>
<td>1</td>
<td>1.00E-5</td>
</tr>
<tr>
<td>Learning rate warmup steps</td>
<td>learningRateWarmupSteps</td>
<td>The number of iterations over which the learning</td>
<td>integer</td>
<td>0</td>
<td>250</td>
<td>5</td>
</tr>
</tbody>
</table>
## Amazon Titan Image Generator G1 model customization

hyperparameters

The Amazon Titan Image Generator G1 model supports the following hyperparameters for model customization.

### Note

一步Count has no default value and must be specified. stepCount supports the value auto. auto prioritizes model performance over training cost by automatically determining a number based on the size of your dataset. Training job costs depend on the number that auto determines. To understand how job cost is calculated and to see examples, see Amazon Bedrock Pricing.

<table>
<thead>
<tr>
<th>Parameter (console)</th>
<th>Parameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>batchSize</td>
<td>Number of samples processed before updating</td>
<td></td>
<td>8</td>
<td>192</td>
<td>8</td>
</tr>
<tr>
<td>Hyperparameter (console)</td>
<td>Hyperparameter (API)</td>
<td>Definition</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Default</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------</td>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>model parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steps</td>
<td>stepCount</td>
<td>Number of times the model is exposed to each batch</td>
<td>10</td>
<td>40,000</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Learning rate</td>
<td>learningRate</td>
<td>Rate at which model parameters are updated after each batch</td>
<td>1.00E-7</td>
<td>1</td>
<td>1.00E-5</td>
<td></td>
</tr>
</tbody>
</table>

**Amazon Titan Multimodal Embeddings G1 customization hyperparameters**

The Amazon Titan Multimodal Embeddings G1 model supports the following hyperparameters for model customization.

**Note**

epochCount has no default value and must be specified. epochCount supports the value Auto. Auto prioritizes model performance over training cost by automatically determining a number based on the size of your dataset. Training job costs depend on the number that Auto determines. To understand how job cost is calculated and to see examples, see Amazon Bedrock Pricing.
<table>
<thead>
<tr>
<th>Hyperparameter (console)</th>
<th>Hyperparameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>epochCount</td>
<td>The number of iterations through the entire training dataset</td>
<td>integer</td>
<td>1</td>
<td>100</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch size</td>
<td>batchSize</td>
<td>The number of samples processed before updating model parameters</td>
<td>integer</td>
<td>256</td>
<td>9,216</td>
<td>576</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning rate</td>
<td>learningRate</td>
<td>The rate at which model parameters are updated after each batch</td>
<td>float</td>
<td>5.00E-8</td>
<td>1</td>
<td>5.00E-5</td>
</tr>
</tbody>
</table>

**Cohere Command model customization hyperparameters**

The Cohere Command and Cohere Command Light models support the following hyperparameters for model customization. For more information, see [Custom models](#).
For information about fine tuning Cohere models, see the Cohere documentation at [https://docs.cohere.com/docs/fine-tuning](https://docs.cohere.com/docs/fine-tuning).

![Note](https://example.com/note.png)

The epochCount quota is adjustable.

<table>
<thead>
<tr>
<th>Hyperparameter (console)</th>
<th>Hyperparameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>epochCount</td>
<td>The number of iteration s through the entire training dataset</td>
<td>integer</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Batch size</td>
<td>batchSize</td>
<td>The number of samples processed before updating model parameters</td>
<td>integer</td>
<td>8</td>
<td>8</td>
<td>8 (Command) 32 (Light)</td>
</tr>
<tr>
<td>Learning rate</td>
<td>learningRate</td>
<td>The rate at which model parameters are updated after each</td>
<td>float</td>
<td>5.00E-6</td>
<td>0.1</td>
<td>1.00E-5</td>
</tr>
<tr>
<td>Hyperparameter (console)</td>
<td>Hyperparameter (API)</td>
<td>Definition</td>
<td>Type</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Default</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------</td>
<td>------------</td>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>batch</td>
<td></td>
<td>You use a validation dataset, we recommend that you don't provide a value for learningRate.</td>
<td>float</td>
<td>0</td>
<td>0.1</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Early stopping threshold**

**earlyStoppingThreshold**

The minimum improvement in loss required to prevent premature termination of the training process.
<table>
<thead>
<tr>
<th>Hyperparameter (console)</th>
<th>Hyperparameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early stopping patience</td>
<td>earlyStoppingPatient</td>
<td>The tolerance for stagnation in the loss metric before stopping the training process</td>
<td>integer</td>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Evaluation percentage</td>
<td>evalPercentage</td>
<td>The percentage of the dataset allocated for model evaluation, if you don't provide a separate validation dataset</td>
<td>float</td>
<td>5</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

**Meta Llama 2 model customization hyperparameters**

The Meta Llama 2 13B and 70B models support the following hyperparameters for model customization. For more information, see [Custom models](https://ai.meta.com/llama/get-started/#fine-tuning).

For information about fine tuning Meta Llama models, see the Meta documentation at [https://ai.meta.com/llama/get-started/#fine-tuning](https://ai.meta.com/llama/get-started/#fine-tuning).
Note

The `epochCount` quota is adjustable.

<table>
<thead>
<tr>
<th>Hyperparameter (console)</th>
<th>Hyperparameter (API)</th>
<th>Definition</th>
<th>Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>epochCount</td>
<td>The number of iterations through the entire training dataset</td>
<td>integer</td>
<td>1</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch size</td>
<td>batchSize</td>
<td>The number of samples processed before updating model parameters</td>
<td>integer</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning rate</td>
<td>learningRate</td>
<td>The rate at which model parameters are updated after each batch</td>
<td>float</td>
<td>5.00E-6</td>
<td>0.1</td>
<td>1.00E-4</td>
</tr>
</tbody>
</table>
Amazon Bedrock console overview

The Amazon Bedrock console provides the following features.

Features

• Getting started
• Foundation models
• Playgrounds
• Safeguards
• Orchestration
• Assessment and deployment
• Model access
• Model invocation logging

To open the Amazon Bedrock console, sign in at https://console.aws.amazon.com/bedrock/home.

Getting started

From Getting started in the navigation pane, you can get an Overview of the foundation models, examples, and playgrounds that Amazon Bedrock provides. You can also get Examples of the prompts you can use with Amazon Bedrock models.

The examples page shows example prompts for the available models. You can search the examples and filter the list of examples using one or more of the following attributes:

• Model
• Modality (text, image, or embedding)
• Category
• Provider

Filter the example prompts by choosing the Search in examples edit box and then selecting the filter that you want to apply to the search. Apply multiple filters by again choosing Search in examples and then selecting another filter.
When you choose an example, the Amazon Bedrock console displays the following information about the example:

- A description of what the example accomplishes.
- The model name (and model provider) where the example runs.
- The example prompt and the expected response.
- The inference configuration parameter settings for the example.
- The API request that runs the example.

To run the example, choose Open in playground.

**Foundation models**

From Foundation models in the navigation pane, you can view the available Base models, and group them by various attributes. You can also filter the model view, search for models, and view information about the model providers.

You can customize a base foundation model to improve the model's performance on specific tasks or teach the model a new domain of knowledge. Choose Custom models under foundation models to create and manage your custom models. Customize a model by creating a model customization job with a training dataset that you provide. For more information, see Custom models.

You can experiment with base models and custom models by using the console playgrounds.

**Playgrounds**

The console playgrounds are where you can experiment with models before deciding to use them in an application. There are three playgrounds.

**Chat playground**

The chat playground lets you experiment with the chat models that Amazon Bedrock provides. You can submit a chat to a model and the chat playground shows the response from the model and includes model metrics. Optionally, choose Compare mode to compare the output from up to three models. For more information, see Chat playground.

**Text playground**
The text playground lets you experiment with the text models that Amazon Bedrock provides. You can submit text to a model and the text playground shows the text that the model generates from the prompt. For more information, see Text playground.

Image playground

The image playground lets you experiment with the image models that Amazon Bedrock provides. You can submit a text prompt to a model and the image playground shows the image that the model generates for the prompt. For more information, see Image playground.

In the console, access the playgrounds by choosing Playgrounds in the navigation pane. For more information, see Playgrounds.

Safeguards

Titan Image Generator G1 automatically puts an invisible watermark on all images created by the model. Watermark detection detects if the image was generated by Titan Image Generator G1. To use watermark detection, choose Overview in the left navigation pane and then Build and Test tab. Go to the Safeguards section and choose View watermark detection. For more information, see Watermark detection.

Orchestration

With Amazon Bedrock, you can enable a Retrieval-Augmented Generation (RAG) workflow by using knowledge bases to build contextual applications by using the reasoning capabilities of LLMs. To use a knowledge base, choose Orchestration in the left navigation pane and then Knowledge base. For more information, see Knowledge bases for Amazon Bedrock.

Agents for Amazon Bedrock enables developers to configure an agent to complete actions based on organization data and user input. For example you might create an agent to take actions to fulfill a customer's request. To use an Agent, choose Orchestration in the left navigation pane and then Agent. For more information, see Agents for Amazon Bedrock.

Assessment and deployment

As you use Amazon Bedrock models, you need to assess their performance and to deploy them into your solutions.
With Model Evaluation, you can evaluate and compare model output, and then choose the one best suited for your applications. Choose Assessment and deployment and then choose Model evaluation.

When you configure Provisioned Throughput for a model, you receive a level of throughput at a fixed cost. To provision throughput, choose Assessment and deployment in the navigation pane and then Provisioned Throughput. For more information, see Provisioned Throughput.

**Model access**

To use a model in Amazon Bedrock, you must first request access to the model. On the left navigation pane, choose Model access. For more information, see Model access.

**Model invocation logging**

You can log model invocation events by choosing Settings in the left navigation pane. For more information, see Model invocation logging.
Run model inference

Inference refers to the process of generating an output from an input provided to a model. Foundation models use probability to construct the words in a sequence. Given an input, the model predicts a probable sequence of tokens that follows, and returns that sequence as the output. Amazon Bedrock provides you the capability of running inference in the foundation model of your choice. When you run inference, you provide the following inputs.

- **Prompt** – An input provided to the model in order for it to generate a response. For information about writing prompts, see [Prompt engineering guidelines](#).

- **Inference parameters** – A set of values that can be adjusted to limit or influence the model response. For information about inference parameters, see [Inference parameters](#) and [Inference parameters for foundation models](#).

Amazon Bedrock offers a suite of foundation models that you can use to generate outputs of the following modalities. To see modality support by foundation model, refer to [Supported foundation models in Amazon Bedrock](#).

<table>
<thead>
<tr>
<th>Output modality</th>
<th>Description</th>
<th>Example use cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text</strong></td>
<td>Provide text input and generate various types of text</td>
<td>Chat, question-and-answering, brainstorming, summarization, code generation, table creation, data formatting, rewriting</td>
</tr>
<tr>
<td><strong>Image</strong></td>
<td>Provide text or input images and generate or modify images</td>
<td>Image generation, image editing, image variation</td>
</tr>
<tr>
<td><strong>Embeddings</strong></td>
<td>Provide text, images, or both text and images and generate a vector of numeric values that represent the input. The output vector can be compared to other</td>
<td>Text and image search, query, categorization, recommendations, personalization, knowledge base creation</td>
</tr>
<tr>
<td>Output modality</td>
<td>Description</td>
<td>Example use cases</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td></td>
<td>embeddings vectors to determine semantic similarity (for text) or visual similarity (for images).</td>
<td></td>
</tr>
</tbody>
</table>

You can run model inference in the following ways.

- Use any of the **Playgrounds** to run inference in a user-friendly graphical interface.
- Send an **InvokeModel** or **InvokeModelWithResponseStream** request.
- Prepare a dataset of prompts with your desired configurations and run batch inference with a **CreateModelInvocationJob** request.
- The following Amazon Bedrock features use model inference as a step in a larger orchestration. Refer to those sections for more details.
  - Set up a **knowledge base** and send a **RetrieveAndGenerate** request.
  - Set up an **agent** and send an **InvokeAgent** request.

You can run inference with base models, custom models, or provisioned models. To run inference on a custom model, first purchase Provisioned Throughput for it (for more information, see **Provisioned Throughput**).

Use these methods to test foundation model responses with different prompts and inference parameters. Once you have sufficiently explored these methods, you can set up your application to run model inference by calling these APIs.

Select a topic to learn more about running model inference through that method. To learn more about using agents, see **Agents for Amazon Bedrock**.

**Topics**

- **Inference parameters**
- **Playgrounds**
- **Use the API to invoke a model with a single prompt**
- **Run batch inference**
Inference parameters

Inference parameters are values that you can adjust to limit or influence the model response. The following categories of parameters are commonly found across different models.

Randomness and diversity

For any given sequence, a model determines a probability distribution of options for the next token in the sequence. To generate each token in an output, the model samples from this distribution. Randomness and diversity refer to the amount of variation in a model's response. You can control these factors by limiting or adjusting the distribution. Foundation models typically support the following parameters to control randomness and diversity in the response.

- **Temperature**—Affects the shape of the probability distribution for the predicted output and influences the likelihood of the model selecting lower-probability outputs.
  - Choose a lower value to influence the model to select higher-probability outputs.
  - Choose a higher value to influence the model to select lower-probability outputs.

In technical terms, the temperature modulates the probability mass function for the next token. A lower temperature steepens the function and leads to more deterministic responses, and a higher temperature flattens the function and leads to more random responses.

- **Top K** – The number of most-likely candidates that the model considers for the next token.
  - Choose a lower value to decrease the size of the pool and limit the options to more likely outputs.
  - Choose a higher value to increase the size of the pool and allow the model to consider less likely outputs.

For example, if you choose a value of 50 for Top K, the model selects from 50 of the most probable tokens that could be next in the sequence.

- **Top P** – The percentage of most-likely candidates that the model considers for the next token.
  - Choose a lower value to decrease the size of the pool and limit the options to more likely outputs.
  - Choose a higher value to increase the size of the pool and allow the model to consider less likely outputs.

In technical terms, the model computes the cumulative probability distribution for the set of responses and considers only the top P% of the distribution.
For example, if you choose a value of 0.8 for Top P, the model selects from the top 80% of the probability distribution of tokens that could be next in the sequence.

The following table summarizes the effects of these parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Effect of lower value</th>
<th>Effect of higher value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>Increase likelihood of higher-probability tokens</td>
<td>Increase likelihood of lower-probability tokens</td>
</tr>
<tr>
<td></td>
<td>Decrease likelihood of lower-probability tokens</td>
<td>Decrease likelihood of higher-probability tokens</td>
</tr>
<tr>
<td>Top K</td>
<td>Remove lower-probability tokens</td>
<td>Allow lower-probability tokens</td>
</tr>
<tr>
<td>Top P</td>
<td>Remove lower-probability tokens</td>
<td>Allow lower-probability tokens</td>
</tr>
</tbody>
</table>

As an example to understand these parameters, consider the example prompt I hear the hoof beats of "_. Let's say that the model determines the following three words to be candidates for the next token. The model also assigns a probability for each word.

```json
{
  "horses": 0.7,
  "zebras": 0.2,
  "unicorns": 0.1
}
```

- If you set a high temperature, the probability distribution is flattened and the probabilities become less different, which would increase the probability of choosing "unicorns" and decrease the probability of choosing "horses".
- If you set Top K as 2, the model only considers the top 2 most likely candidates: "horses" and "zebras."
- If you set Top P as 0.7, the model only considers "horses," because it is the only candidate that lies in the top 70% of the probability distribution.
Length

Foundation models typically support parameters that limit the length of the response. Examples of these parameters are provided below.

- **Response length** – An exact value to specify the minimum or maximum number of tokens to return in the generated response.

- **Penalties** – Specify the degree to which to penalize outputs in a response. Examples include the following.
  - The length of the response.
  - Repeated tokens in a response.
  - Frequency of tokens in a response.
  - Types of tokens in a response.

- **Stop sequences** – Specify sequences of characters that stop the model from generating further tokens. If the model generates a stop sequence that you specify, it will stop generating after that sequence.

Playgrounds

⚠️ **Important**

Before you can use any of the foundation models, you must request access to that model. If you try to use the model (with the API or within the console) before you have requested access to it, you will receive an error message. For more information, see Model access.

The Amazon Bedrock playgrounds provide you a console environment to experiment with running inference on different models and with different configurations, before deciding to use them in an application. In the console, access the playgrounds by choosing Playgrounds in the left navigation pane. You can also navigate directly to the playground when you choose a model from a model details page or the examples page.

There are playgrounds for text, chat, and image models.
Within each playground you can enter prompts and experiment with inference parameters. Prompts are usually one or more sentences of text that set up a scenario, question, or task for a model. For information about creating prompts, see Prompt engineering guidelines.

Inference parameters influence the response generated by a model, such as the randomness of generated text. When you load a model into a playground, the playground configures the model with its default inference settings. You can change and reset the settings as you experiment with the model. Each model has its own set of inference parameters. For more information, see Inference parameters for foundation models.

When you submit a response, the model responds with its generated output.

If a chat or text model supports streaming, the default is to stream the responses from a model. You can turn off streaming, if desired.

Topics

- Chat playground
- Text playground
- Image playground
- Use a playground

Chat playground

The chat playground lets you experiment with the chat models that Amazon Bedrock provides. You can submit a chat to a model and the chat playground shows the response from the model, and the following metrics.

- **Latency** — The time it takes for the model to generate each token (word) in a sequence.
- **Input token count** — The number of tokens that are fed into the model as input during inference.
- **Output token count** — The number of tokens generated in response to a prompt. Longer, more conversational, responses require more tokens.
- **Cost** — The cost of processing the input and generating output tokens.

You can also define criteria that you want the model response to match.
By turning on compare model, you can compare the chat responses for a single prompt with the responses from up to three models. This helps you to understand the comparative performance of each model, without having to switch between models. For more information, see Use a playground.

Text playground

The text playground lets you experiment with the text models that Amazon Bedrock provides. You can submit text to a model and the text playground shows the text that the model generates from the prompt.

Image playground

The image playground lets you experiment with the image models that Amazon Bedrock provides. You can submit a text prompt to a model and the image playground shows the image that the model generates for the prompt.

Along with setting inference parameters, you can make additional configuration changes (differs by model):

- **Mode** – The model generates a new image (Generate) or edits (Edit) the image that you supply in Reference image. If you edit a reference image, the model needs a segmentation mask that covers the area of the image that you want the model to edit. Create the segmentation mask by using the image playground to draw a rectangle on the reference image. Alternatively, you can create the segmentation mask by specifying a mask prompt (Amazon Titan Image Generator G1 Generator G1 image only).

- **Mask prompt** – If you edit an image with the Amazon Titan Image Generator G1 model, you can use a mask prompt to specify the objects that you want the segmentation mask to cover. For example, you can specify the mask prompt sky to create a segmentation mask that covers the sky in an image. You can then run the prompt An image of a rainy day to make the sky in the image appear rainy.

- **Negative prompt** – items or concepts that you don’t want the model to generate, such as cartoon or violence.

- **Reference image** – The image on which to generate the response or that you want the model to edit.

- **Response image** – Output settings for the generated image, such as quality, orientation, size, and the number of images to generate.
• **Advanced configurations** – The inference parameters to pass to the model.

## Use a playground

The following procedure shows how to submit a prompt to a playground and view the response. In each playground, you can configure the inference parameters for the model. In the chat playground, you can view metrics, and optionally compare the output of up to three models. In the image playground you can make advanced configuration changes, which also vary by model.

### To use a playground

1. If you haven't already, request access to the models that you want to use. For more information, see [Model access](#).
2. Open the Amazon Bedrock console.
3. From the navigation pane, under **Playgrounds**, choose **Chat**, **Text**, or **Image**.
4. Choose **Select model** to open the **Select model** dialog box.
   a. In **Category** select from the available providers or custom models.
   b. In **Model** select a model.
   c. In **Throughput** select the throughput (on-demand, or provisioned throughput) that you want the model to use. If you are using a custom model, you must have set up Provisioned Throughput for the model beforehand. For more information, see [Provisioned Throughput](#).
   d. Choose **Apply**.
5. (Optional) In **Configurations** choose the inference parameters that you want to use. For more information, see [Inference parameters for foundation models](#). For information about configuration changes you can make in the image playground, see [Image playground](#).
6. Enter your prompt into the text field. A prompt is a natural language phrase or command, such as **Tell me about the best restaurants to visit in Seattle**. For more information, see [Prompt engineering guidelines](#).

If you are using the chat playground with a model that supports multimodal prompts, add images to the prompt by choosing **Image** or by dragging an image onto the prompt text field.
Note

If the response violates the content moderation policy, Amazon Bedrock doesn’t display it. If you have turned on streaming, Amazon Bedrock clears the entire response if it generates content that violates the policy. For more details, navigate to the Amazon Bedrock console, select Providers, and read the text under the Content limitations section.

For information about prompt engineering, see Prompt engineering guidelines.

7. Choose Run. to run the prompt.

8. If you are using the chat playground, view the model metrics and compare models by doing the following.

   a. In the Model metrics section, view the metrics for each model.

   b. (Optional) Define criteria that you want to match by doing the following:

      i. Choose Define metric criteria.

      ii. For the metrics you want to use, choose the condition and value. You can set the following conditions:

          • less than – The metric value is less than the specified value.

          • greater than – The metric value is more than the specified value.

      iii. Choose Apply to apply your criteria.

      iv. View which criteria are met. If all criteria are met, the Overall summary is Meets all criteria. If 1 or more criteria are not met, the Overall summary is n criteria unmet and the unmet criteria are highlighted in red.

   c. (Optional) Add models to compare by doing the following:

      i. Turn on Compare mode.

      ii. Choose Select model to select a model.

      iii. In the dialog box, choose a provider, model, and throughput.

      iv. Choose Apply.

      v. (Optional) Choose the menu icon next to each model to configure inference parameters for that model. For more information, see Inference parameters for foundation models.
vi. Chooses the + icon on the right of the Chat playground section to add a second or third model to compare.

vii. Repeat steps a-c to choose the models that you want to compare.

viii. Enter your a prompt into the text field and choose Run.

Use the API to invoke a model with a single prompt

Run inference on a model through the API by sending an InvokeModel or InvokeModelWithResponseStream request. You can specify the media type for the request and response bodies in the contentType and accept fields. The default value for both fields is application/json if you don't specify a value.

Streaming is supported for all text output models except AI21 Labs Jurassic-2 models. To check if a model supports streaming, send a GetFoundationModel or ListFoundationModels request and check the value in the responseStreamingSupported field.

Specify the following fields, depending on the model that you use.

1. modelId – Use either the model ID or its ARN. The method for finding the modelId or modelArn depends on the type of model you use.
   - **Base model** – Do one of the following.
     - To see a list of model IDs for all base models supported by Amazon Bedrock, see Base model IDs (on-demand throughput).
     - Send a ListFoundationModels request and find the modelId or modelArn of the model to use in the response.
     - In the console, select a model in Providers and find the modelId in the API request example.
   - **Custom model** – Purchase Provisioned Throughput for the custom model (for more information, see Provisioned Throughput) and find the model ID or ARN of the provisioned model.
   - **Provisioned model** – If you have created a Provisioned Throughput for a base or custom model, do one of the following.
     - Send a ListProvisionedModelThroughputs request and find the provisionedModelArn or provisionedModelName of the model to use in the response.
• In the console, select a model in **Provisioned Throughput** and find the model ARN or name
  in the **Model details** section.

2. **body** – Each base model has its own inference parameters that you set in the body field. The
  inference parameters for a custom or provisioned model depends on the base model from which
  it was created. For more information, see [Inference parameters for foundation models](#).

## Invoke model code examples

The following examples show how to run inference with the **InvokeModel** API. For examples
with different models, see the inference parameter reference for the desired model ([Inference
parameters for foundation models](#)).

### CLI

The following example saves the generated response to the prompt *story of two dogs* to a
file called `invoke-model-output.txt`.

```
aws bedrock-runtime invoke-model \
  --model-id anthropic.claude-v2 \
  --body '{"prompt": "\n\nHuman: story of two dogs\n\nAssistant:",
  "max_tokens_to_sample" : 300}' \
  --cli-binary-format raw-in-base64-out \n  invoke-model-output.txt
```

### Python

The following example returns a generated response to the prompt *explain black holes to 8th graders*.

```python
import boto3
import json
brt = boto3.client(service_name='bedrock-runtime')

body = json.dumps({
    "prompt": "\n\nHuman: explain black holes to 8th graders\n\nAssistant:",
    "max_tokens_to_sample": 300,
    "temperature": 0.1,
    "top_p": 0.9,
})
```
modelId = 'anthropic.claude-v2'
accept = 'application/json'
contentType = 'application/json'

response = brt.invoke_model(body=body, modelId=modelId, accept=accept, contentType=contentType)

response_body = json.loads(response.get('body').read())

# text
print(response_body.get('completion'))

### Invoke model with streaming code example

#### Note

The AWS CLI does not support streaming.

The following example shows how to use the `InvokeModelWithResponseStream` API to generate streaming text with Python using the prompt *write an essay for living on mars in 1000 words*.

```python
import boto3
import json

brt = boto3.client(service_name='bedrock-runtime')

body = json.dumps(
    {'prompt': '
    Human: write an essay for living on mars in 1000 words
    Assistant:',
    'max_tokens_to_sample': 4000}
)

response = brt.invoke_model_with_response_stream(modelId='anthropic.claude-v2', body=body)

stream = response.get('body')
if stream:
for event in stream:
    chunk = event.get('chunk')
    if chunk:
        print(json.loads(chunk.get('bytes').decode()))

Run batch inference

⚠️ Note

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

With batch inference, you can run multiple inference requests asynchronously to process a large number of requests efficiently by running inference on data that is stored in an S3 bucket. You can use batch inference to improve the performance of model inference on large datasets.

⚠️ Note

Batch inference isn't supported for provisioned models.

To see quotas for batch inference, see Batch inference quotas.

Amazon Bedrock supports batch inference on the following modalities.

- Text to embeddings
- Text to text
- Text to image
• Image to image
• Image to embeddings

You store your data in an Amazon S3 bucket to prepare it for batch inference. You can then carry out and manage batch inference jobs through using the `ModelInvocationJob` APIs.

Before you can carry out batch inference, you must receive permissions to call the batch inference APIs. You then configure an IAM Amazon Bedrock service role to have permissions to carry out batch inference jobs.

You can use the batch inference APIs by downloading and installing one of the following AWS SDK packages.

• [AWS SDK for Python](https://aws.amazon.com/sdk-for-python/).
• [AWS SDK for Java](https://aws.amazon.com/sdk-for-java/).

### Topics

• [Set up permissions for batch inference](#)
• [Format and upload your inference data](#)
• [Create a batch inference job](#)
• [Stop a batch inference job](#)
• [Get details about a batch inference job](#)
• [List batch inference jobs](#)
• [Code samples](#)

### Set up permissions for batch inference

**Note**

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

• [AWS SDK for Python](https://aws.amazon.com/sdk-for-python/).
• [AWS SDK for Java](https://aws.amazon.com/sdk-for-java/).
We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

To set up a role for batch inference, create an IAM role by following the steps at Creating a role to delegate permissions to an AWS service. Attach the following policies to the role:

- Trust policy
- Access to the Amazon S3 buckets containing the input data for your batch inference jobs and to write the output data.

1. The following policy allows Amazon Bedrock to assume this role and carry out batch inference jobs. The following shows an example policy you can use. You can restrict the scope of the permission by using one or more global condition context keys. For more information, see AWS global condition context keys. Set the aws:SourceAccount value to your account ID. Use the ArnEquals or ArnLike condition to restrict the scope.

   ```json
   {
       "Version": "2012-10-17",
       "Statement": [
           {
               "Effect": "Allow",
               "Principal": {
                   "Service": "bedrock.amazonaws.com"
               },
               "Action": "sts:AssumeRole",
               "Condition": {
                   "StringEquals": {
                       "aws:SourceAccount": "account-id"
                   }
               }
           }
       ]
   }
   ```

   Note
   
   As a best practice for security purposes, replace the * with specific batch inference job IDs after you have created them.
2. Attach the following policy to allow Amazon Bedrock to access the S3 bucket containing input data for your batch inference jobs (replace `my_input_bucket`) and the S3 bucket to write output data to (replace `my_output_bucket`). Replace the `account-id` with the account ID of the user to whom you are providing S3 bucket access permissions.

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "s3:GetObject",
            "s3:PutObject",
            "s3:ListBucket"
         ],
         "Resource": [
            "arn:aws:s3:::my_input_bucket",
            "arn:aws:s3:::my_input_bucket/*",
            "arn:aws:s3:::my_output_bucket",
            "arn:aws:s3:::my_output_bucket/*"
         ],
         "Condition": {
            "StringEquals": {
               "aws:ResourceAccount": [
                  "account-id"
               ]
            }
         }
      }
   ]
}
```
Format and upload your inference data

**Note**

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see [Code samples](#).

Upload JSONL files containing the data to input to the model to your S3 bucket with the following format. Each line should match the following format and is a different item for inference. If you leave the `recordId` field out, Amazon Bedrock adds it in the output.

**Note**

The format of the `modelInput` JSON object should match the body field for the model that you use in the `InvokeModel` request. For more information, see [Inference parameters](#) for foundation models.

```
{ "recordId" : "12 character alphanumeric string", "modelInput" : {JSON body} }
...
```

For example, you might provide an JSONL file containing the following data and run batch inference on a Titan text model.

```
{ "recordId" : "3223593EFGH", "modelInput" : {"inputText": "Roses are red, violets are"} }
{ "recordId" : "1223213ABCD", "modelInput" : {"inputText": "Hello world"} }
```
Create a batch inference job

Note

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

Request format

POST /model-invocation-job HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "inputDataConfig": {
      "s3InputDataConfig": {
         "s3Uri": "string",
         "s3InputFormat": "JSONL"
      }
   },
   "jobName": "string",
   "modelId": "string",
   "outputDataConfig": {
      "s3OutputDataConfig": {
         "s3Uri": "string"
      }
   },
   "roleArn": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
To create a batch inference job, send a `CreateModelInvocationJob` request. Provide the following information.

- The ARN of a role with permissions to run batch inference in `roleArn`.
- Information for the S3 bucket containing the input data in `inputDataConfig` and the bucket where to write information in `outputDataConfig`.
- The ID of the model to use for inference in `modelId` (see Base model IDs (on-demand throughput)).
- A name for the job in `jobName`.
- (Optional) Any tags that you want to attach to the job in `tags`.

The response returns a `jobArn` that you can use for other batch inference-related API calls.

You can check the status of the job with either the `GetModelInvocationJob` or `ListModelInvocationJobs` APIs.

When the job is Completed, you can extract the results of the batch inference job from the files in the S3 bucket you specified in the request for the `outputDataConfig`. The S3 bucket contains the following files:

1. Output files containing the result of the model inference.
   - If the output is text, Amazon Bedrock generates an output JSONL file for each input JSONL file. The output files contain outputs from the model for each input in the following format. An `error` object replaces the `modelOutput` field in any line where there was an error in
inference. The format of the modelOutput JSON object matches the body field for the model that you use in the InvokeModel response. For more information, see Inference parameters for foundation models.

```json
{
  "recordId": "12 character alphanumeric string",
  "modelInput": {JSON body},
  "modelOutput": {JSON body}
}
```

The following example shows a possible output file.

```json
{
  "recordId": "3223593EFGH",
  "modelInput": {
    "inputText": "Roses are red, violets are"
  },
  "modelOutput": {
    "inputTextTokenCount": 8,
    "results": [{
      "tokenCount": 3,
      "outputText": "blue
",
      "completionReason": "FINISH"
    }]
  }
}
```

- If the output is image, Amazon Bedrock generates a file for each image.

2. A manifest.json.out file containing a summary of the batch inference job.

```json
{
  "processedRecordCount": number,
  "successRecordCount": number,
  "errorRecordCount": number,
  "inputTextTokenCount": number, // For embedding/text to text models
  "outputTextTokenCount": number, // For text to text models
  "outputImgCount512x512pStep50": number, // For text to image models
  "outputImgCount512x512pStep150": number, // For text to image models
  "outputImgCount512x896pStep50": number, // For text to image models
  "outputImgCount512x896pStep150": number // For text to image models
}
```

**Stop a batch inference job**

**Note**

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.
We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

Request format

```
POST /model-invocation-job/jobIdentifier/stop HTTP/1.1
```

Response format

```
HTTP/1.1 200
```

To stop a batch inference job, send a StopModelInvocationJob and provide the ARN of the job in the jobIdentifier field.

If the job was successfully stopped, you receive an HTTP 200 response.

Get details about a batch inference job

Note

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.
Request format

```
GET /model-invocation-job/jobIdentifier HTTP/1.1
```

Response format

```
HTTP/1.1 200
Content-type: application/json

{
  "clientRequestToken": "string",
  "endTime": "string",
  "inputDataConfig": {
    "s3InputDataConfig": {
      "s3Uri": "string",
      "s3InputFormat": "JSONL"
    }
  },
  "jobArn": "string",
  "jobName": "string",
  "lastModifiedTime": "string",
  "message": "string",
  "modelId": "string",
  "outputDataConfig": {
    "s3OutputDataConfig": {
      "s3Uri": "string"
    }
  },
  "roleArn": "string",
  "status": "Submitted | InProgress | Completed | Failed | Stopping | Stopped",
  "submitTime": "string"
}
```  

To get information about a batch inference job, send a GetModelInvocationJob and provide the ARN of the job in the jobIdentifier field.

See the GetModelInvocationJob page for details about the information provided in the response.
List batch inference jobs

Note
Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

Request format

GET /model-invocation-jobs?
maxResults=\textit{maxResults}&nameContains=\textit{nameContains}&nextToken=\textit{nextToken}&sortBy=\textit{sortBy}&sortOrder=\textit{sortOrder}&statusEquals=\textit{statusEquals}&submitTimeAfter=\textit{submitTimeAfter}&submitTimeBefore=\textit{submitTimeBefore}
HTTP/1.1

Response format

HTTP/1.1 200
Content-type: application/json

{  "invocationJobSummaries": [  {  "clientRequestToken": "string",  "endTime": "string",  "inputDataConfig": {  "s3InputDataConfig": {  "s3Uri": "string",  "s3InputFormat": "JSONL"  }  },  "jobArn": "string",  "jobName": "string",}  ],}
To get information about a batch inference job, send a `ListModelInvocationJobs`. You can set the following specifications.

- Filter for results by specifying the status, submission time, or substrings in the name of the job. You can specify the following statuses.
  - Submitted
  - InProgress
  - Completed
  - Failed
  - Stopping
  - Stopped
- Sort by the time that the job was created (`CreationTime`). You can sort in Ascending or Descending order.
- The maximum number of results to return in a response. If there are more results than the number you set, the response returns a `nextToken` that you can send in another `ListModelInvocationJobs` request to see the next batch of jobs.

The response returns a list of `InvocationJobSummary` objects. Each object contains information about a batch inference job.
Note

Batch inference is in preview and is subject to change. Batch inference is currently only available through the API. Access batch APIs through the following SDKs.

- AWS SDK for Python.
- AWS SDK for Java.

We recommend that you create a virtual environment to use the SDK. Because batch inference APIs aren't available in the latest SDKs, we recommend that you uninstall the latest version of the SDK from the virtual environment before installing the version with the batch inference APIs. For a guided example, see Code samples.

Select a language to see a code sample to call the batch inference API operations.

Python

After downloading the Python SDK and CLI files containing the batch inference API operations, navigate to the folder containing the files and run `ls` in a terminal. You should see the following 2 files, at the least.

```
botocore-1.32.4-py3-none-any.whl
boto3-1.29.4-py3-none-any.whl
```

Create and activate a virtual environment for the batch inference APIs by running the following commands in a terminal. You can replace `bedrock-batch` with a name of your choice for the environment.

```
python3 -m venv bedrock-batch
source bedrock-batch/bin/activate
```

To ensure that there aren't artifacts from later version of boto3 and botocore, uninstall any existing versions by running the following commands in a terminal.

```
python3 -m pip uninstall botocore
```
Install the Python SDK containing the Amazon Bedrock control plane APIs by running the following commands in a terminal.

```bash
python3 -m pip uninstall boto3
python3 -m pip install botocore-1.32.4-py3-none-any.whl
python3 -m pip install boto3-1.29.4-py3-none-any.whl
```

Run all the following code in the virtual environment you created.

Create a batch inference job with a file named `abc.jsonl` that you uploaded to S3. Write the output to a bucket in `s3://output-bucket/output/`. Get the `jobArn` from the response.

```python
import boto3

bedrock = boto3.client(service_name="bedrock")

inputDataConfig=({
    "s3InputDataConfig": {
        "s3Uri": "s3://input-bucket/input/abc.jsonl"
    }
})

outputDataConfig=({
    "s3OutputDataConfig": {
        "s3Uri": "s3://output-bucket/output/"
    }
})

response=bedrock.create_model_invocation_job(
    roleArn="arn:aws:iam::123456789012:role/MyBatchInferenceRole",
    modelId="amazon.titan-text-express-v1",
    jobName="my-batch-job",
    inputDataConfig=inputDataConfig,
    outputDataConfig=outputDataConfig
)

jobArn = response.get('jobArn')

Return the status of the job.

```python
bedrock.get_model_invocation_job(jobIdentifier=jobArn)["status"]
```
List batch inference jobs that *Failed*.

```python
define the code here
```

Stop the job that you started.

```python
define the code here
```

Java

```java
package com.amazon.aws.sample.bedrock.inference;

import com.amazonaws.services.bedrock.AmazonBedrockAsync;
import com.amazonaws.services.bedrock.AmazonBedrockAsyncClientBuilder;
import com.amazonaws.services.bedrock.model.CreateModelInvocationJobRequest;
import com.amazonaws.services.bedrock.model.CreateModelInvocationJobResult;
import com.amazonaws.services.bedrock.model.GetModelInvocationJobRequest;
import com.amazonaws.services.bedrock.model.GetModelInvocationJobResult;
import com.amazonaws.services.bedrock.model.InvocationJobInputDataConfig;
import com.amazonaws.services.bedrock.model.InvocationJobOutputDataConfig;
import com.amazonaws.services.bedrock.model.InvocationJobS3InputDataConfig;
import com.amazonaws.services.bedrock.model.InvocationJobS3OutputDataConfig;
import com.amazonaws.services.bedrock.model.ListModelInvocationJobsRequest;
import com.amazonaws.services.bedrock.model.ListModelInvocationJobsResult;
import com.amazonaws.services.bedrock.model.StopModelInvocationJobRequest;
import com.amazonaws.services.bedrock.model.StopModelInvocationJobResult;

public class BedrockAsyncInference {
    private final AmazonBedrockAsync amazonBedrockAsyncClient =
    AmazonBedrockAsyncClientBuilder.defaultClient();

    public void createModelInvokeJobSampleCode() {
        final InvocationJobS3InputDataConfig invocationJobS3InputDataConfig = new
            InvocationJobS3InputDataConfig()
                .withS3Uri("s3://Input-bucket-name/input/abc.jsonl")
                .withS3InputFormat("JSONL");

        final InvocationJobInputDataConfig inputDataConfig = new
            InvocationJobInputDataConfig()
```
withS3InputDataConfig(invocationJobS3InputDataConfig);

final InvocationJobS3OutputDataConfig invocationJobS3OutputDataConfig = new InvocationJobS3OutputDataConfig()
  .withS3Uri("s3://output-bucket-name/output/");

final InvocationJobOutputDataConfig invocationJobOutputDataConfig = new InvocationJobOutputDataConfig()
  .withS3OutputDataConfig(invocationJobS3OutputDataConfig);

final CreateModelInvocationJobRequest createModelInvocationJobRequest = new CreateModelInvocationJobRequest()
  .withModelId("anthropic.claude-v2")
  .withJobName("unique-job-name")
  .withClientRequestToken("Client-token")
  .withInputDataConfig(inputDataConfig)
  .withOutputDataConfig(invocationJobOutputDataConfig);

final CreateModelInvocationJobResult createModelInvocationJobResult =
  amazonBedrockAsyncClient
  .createModelInvocationJob(createModelInvocationJobRequest);

  System.out.println(createModelInvocationJobResult.getJobArn());
}

public void getModelInvokeJobSampleCode() {
  final GetModelInvocationJobRequest getModelInvocationJobRequest = new GetModelInvocationJobRequest()
    .withJobIdentifier("jobArn");

  final GetModelInvocationJobResult getModelInvocationJobResult =
    amazonBedrockAsyncClient
    .getModelInvocationJob(getModelInvocationJobRequest);

  System.out.println(getModelInvocationJobResult.getJobArn());
}

public void listModelInvokeJobSampleCode() {
  final ListModelInvocationJobsRequest listModelInvocationJobsRequest = new ListModelInvocationJobsRequest()
    .withMaxResults(10)
    .withNameContains("matchin-string");
final ListModelInvocationJobsResult listModelInvocationJobsResult =
  amazonBedrockAsyncClient
  .listModelInvocationJobs(listModelInvocationJobsRequest);
}

class StopModelInvokeJobSampleCode {
  public void stopModelInvokeJobSampleCode() {
    final StopModelInvocationJobRequest stopModelInvocationJobRequest = new
      StopModelInvocationJobRequest()
        .withJobIdentifier("jobArn");

    final StopModelInvocationJobResult stopModelInvocationJobResult =
      amazonBedrockAsyncClient
        .stopModelInvocationJob(stopModelInvocationJobRequest);
  }
}
Prompt engineering guidelines

Topics

• Introduction
• What is a prompt?
• What is prompt engineering?
• General guidelines for Amazon Bedrock LLM users
• Prompt templates and examples for Amazon Bedrock text models

Introduction

Welcome to the prompt engineering guide for large language models (LLMs) on Amazon Bedrock. Amazon Bedrock is Amazon’s service for foundation models (FMs), which offers access to a range of powerful FMs for text and images.

Prompt engineering refers to the practice of optimizing textual input to LLMs to obtain desired responses. Prompting helps LLMs perform a wide variety of tasks, including classification, question answering, code generation, creative writing, and more. The quality of prompts that you provide to LLMs can impact the quality of their responses. These guidelines provide you with all the necessary information to get started with prompt engineering. It also covers tools to help you find the best possible prompt format for your use case when using LLMs on Amazon Bedrock.

Whether you’re a beginner in the world of generative AI and language models, or an expert with previous experience, these guidelines can help you optimize your prompts for Amazon Bedrock text models. Experienced users can skip to the General Guidelines for Amazon Bedrock LLM Users or Prompt Templates and Examples for Amazon Bedrock Text Models sections.

Note

All examples in this doc are obtained via API calls. The response may vary due to the stochastic nature of the LLM generation process. If not otherwise specified, the prompts are written by employees of AWS.

Disclaimer: The examples in this document use the current text models available within Amazon Bedrock. Also, this document is for general prompting guidelines. For model-specific guides, refer
Additional prompt resources

The following resources offer additional guidelines on prompt engineering.

- **Anthropic Claude model prompt guide**: [https://docs.anthropic.com/claude/docs](https://docs.anthropic.com/claude/docs)
- **Anthropic Claude prompt engineering resources**: [https://docs.anthropic.com/claude/docs/guide-to-anthropics-prompt-engineering-resources](https://docs.anthropic.com/claude/docs/guide-to-anthropics-prompt-engineering-resources)
- **Cohere prompt guide**: [https://txt.cohere.com/how-to-train-your-pet-llm-prompt-engineering](https://txt.cohere.com/how-to-train-your-pet-llm-prompt-engineering)
- **AI21 Labs Jurassic model prompt guide**: [https://docs.ai21.com/docs/prompt-engineering](https://docs.ai21.com/docs/prompt-engineering)
- **Meta Llama 2 prompt guide**: [https://ai.meta.com/llama/get-started/#prompting](https://ai.meta.com/llama/get-started/#prompting)
- **Stability documentation**: [https://platform.stability.ai/docs/getting-started](https://platform.stability.ai/docs/getting-started)
- **Mistral AI prompt guide**: [https://docs.mistral.ai/guides/prompting-capabilities/](https://docs.mistral.ai/guides/prompting-capabilities/)

**What is a prompt?**

Prompts are a specific set of inputs provided by you, the user, that guide LLMs on Amazon Bedrock to generate an appropriate response or output for a given task or instruction.

**User Prompt:**

*Who invented the airplane?*

When queried by this prompt, Titan provides an output:

**Output:**

*The Wright brothers, Orville and Wilbur Wright are widely credited with inventing and manufacturing the world's first successful airplane.*

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)
Components of a prompt

A single prompt includes several components, such as the task or instruction you want the LLMs to perform, the context of the task (for example, a description of the relevant domain), demonstration examples, and the input text that you want LLMs on Amazon Bedrock to use in its response. Depending on your use case, the availability of the data, and the task, your prompt should combine one or more of these components.

Consider this example prompt asking Titan to summarize a review:

_user prompt:_

The following is text from a restaurant review:

“I finally got to check out Alessandro’s Brilliant Pizza and it is now one of my favorite restaurants in Seattle. The dining room has a beautiful view over the Puget Sound but it was surprisingly not crowded. I ordered the fried castelvetrano olives, a spicy Neapolitan-style pizza and a gnocchi dish. The olives were absolutely decadent, and the pizza came with a smoked mozzarella, which was delicious. The gnocchi was fresh and wonderful. The waitstaff were attentive, and overall the experience was lovely. I hope to return soon.”

Summarize the above restaurant review in one sentence.

(Source of prompt: AWS)

Based on this prompt, Titan responds with a succinct one-line summary of the restaurant review. The review mentions key facts and conveys the main points, as desired.

_output:

Alessandro's Brilliant Pizza is a fantastic restaurant in Seattle with a beautiful view over Puget Sound, decadent and delicious food, and excellent service.

(Model used: Amazon Titan Text G1 - Express)

The instruction Summarize the above restaurant review in one sentence and the review text I finally got to check out ... were both necessary for this type of output. Without either one, the model would not have enough information to produce a sensible summary. The instruction tells the LLM what to do, and the text is the input on which the LLM operates.
The context (The following is text from a restaurant review) provides additional information and keywords that guide the model to use the input when formulating its output.

In the example below, the text Context: Climate change threatens people with increased flooding ... is the input which the LLM can use to perform the task of answering the question Question: What organization calls climate change the greatest threat to global health in the 21st century?".

User prompt:
Context: Climate change threatens people with increased flooding, extreme heat, increased food and water scarcity, more disease, and economic loss. Human migration and conflict can also be a result. The World Health Organization (WHO) calls climate change the greatest threat to global health in the 21st century. Adapting to climate change through efforts like flood control measures or drought-resistant crops partially reduces climate change risks, although some limits to adaptation have already been reached. Poorer communities are responsible for a small share of global emissions, yet have the least ability to adapt and are most vulnerable to climate change. The expense, time required, and limits of adaptation mean its success hinge on limiting global warming.

Question: What organization calls climate change the greatest threat to global health in the 21st century?

(Source of prompt: https://en.wikipedia.org/wiki/Climate_change)

AI21 Labs Jurassic responses with the correct name of the organization according to the context provided in the prompt.

Output:
The World Health Organization (WHO) calls climate change the greatest threat to global health in the 21st century.

(Model used: AI21 Labs Jurassic-2 Ultra v1)

Few-shot prompting vs. zero-shot prompting

It is sometimes useful to provide a few examples to help LLMs better calibrate their output to meet your expectations, also known as few-shot prompting or in-context learning, where a shot corresponds to a paired example input and the desired output. To illustrate, first here is an example
of a zero-shot sentiment classification prompt where no example input-output pair is provided in the prompt text:

**User prompt:**
Tell me the sentiment of the following headline and categorize it as either positive, negative or neutral:
New airline between Seattle and San Francisco offers a great opportunity for both passengers and investors.

(Source of prompt: AWS)

**Output:**
Positive

(Model used: Amazon Titan Text G1 - Express)

Here is the few-shot version of a sentiment classification prompt:

**User prompt:**
Tell me the sentiment of the following headline and categorize it as either positive, negative or neutral. Here are some examples:

Research firm fends off allegations of impropriety over new technology.  
Answer: Negative

Offshore windfarms continue to thrive as vocal minority in opposition dwindles.  
Answer: Positive

Manufacturing plant is the latest target in investigation by state officials.  
Answer:

(Source of prompt: AWS)

**Output:**
Negative

(Model used: Amazon Titan Text G1 - Express)

The following example uses Anthropic Claude models. When using Anthropic Claude models, it's a good practice to use `<example></example>` tags to include demonstration examples. We also recommend using different delimiters such as H: and A: in the examples to avoid confusion with
the delimiters Human: and Assistant: for the whole prompt. Notice that for the last few-shot example, the final A: is left off in favor of Assistant:, prompting Anthropic Claude to generate the answer instead.

User prompt:

Human: Please classify the given email as "Personal" or "Commercial" related emails. Here are some examples.

<example>
H: Hi Tom, it's been long time since we met last time. We plan to have a party at my house this weekend. Will you be able to come over?
A: Personal
</example>

<example>
H: Hi Tom, we have a special offer for you. For a limited time, our customers can save up to 35% of their total expense when you make reservations within two days. Book now and save money!
A: Commercial
</example>

H: Hi Tom, Have you heard that we have launched all-new set of products. Order now, you will save $100 for the new products. Please check our website.

Assistant:

Output:

Commercial

(Source of prompt: AWS, model used: Anthropic Claude)

Prompt template

A prompt template specifies the formatting of the prompt with exchangeable content in it. Prompt templates are “recipes” for using LLMs for different use cases such as classification, summarization, question answering, and more. A prompt template may include instructions, few-shot examples, and specific context and questions appropriate for a given use case. The following example is a template that you can use to perform few-shot sentiment classification using Amazon Bedrock text models:
Prompt template:
"""Tell me the sentiment of the following
{{Text Type, e.g., “restaurant review”}} and categorize it as either {{Sentiment A}} or {{Sentiment B}}.
Here are some examples:

Text: {{Example Input 1}}
Answer: {{Sentiment A}}

Text: {{Example Input 2}}
Answer: {{Sentiment B}}

Text: {{Input}}
Answer:"""

Users can copy and paste this template, filling in their own text and few-shot examples to complete prompts when using LLMs on Amazon Bedrock. More templates are provided in this section.

Note
The double curly braces {{ and }} mark the places to put in data-specific information in the template, and should not be included in the prompt text.

Important notes on using Amazon Bedrock LLMs by API calls

If you are accessing LLMs on Amazon Bedrock via API calls, please be aware that they do not recall prior prompts and previous requests unless the previous interaction is included within the current prompt. You may need this for conversational-style interactions or follow up requests.

If you are accessing Anthropic Claude models using API calls instead of the Amazon Bedrock text playground, you wrap your prompts in a conversational style to get desired response (for the Titan Text and AI21 Labs Jurassic model family, it's not necessary). The following example demonstrates a conversational-style prompt:

User prompt:

*Human: Describe where sea otters live in two sentences.*
Assistant:

**Output:**
Here is a two sentence description of where sea otters live:

*Sea otters live in the northern Pacific Ocean along the coasts of Russia, Alaska, Canada, and the continental United States. They prefer coastal habitats like rocky shores, kelp forests, and estuaries where they have access to shallow, nearshore waters.*

(Source of prompt: AWS, model used: Anthropic Claude)

Notice how the main content of the prompt is wrapped like this:

```
Human: {{Main Content}}
Assistant: }
```

For Anthropic Claude models, prompts sent via the API must contain `Human: ` and `Assistant: `.

To use conversational mode on Titan, you can use the format of `User: {} ` when prompting the model.

**What is prompt engineering?**

Prompt engineering refers to the practice of crafting and optimizing input prompts by selecting appropriate words, phrases, sentences, punctuation, and separator characters to effectively use LLMs for a wide variety of applications. In other words, prompt engineering is the art of communicating with an LLM. High-quality prompts condition the LLM to generate desired or better responses. The detailed guidance provided within this document is applicable across all LLMs within Amazon Bedrock.

The best prompt engineering approach for your use case is dependent on both the task and the data. Common tasks supported by LLMs on Amazon Bedrock include:

- **Classification**: The prompt includes a question with several possible choices for the answer, and the model must respond with the correct choice. An example classification use case is sentiment analysis: the input is a text passage, and the model must classify the sentiment of the text, such as whether it's positive or negative, or harmless or toxic.

- **Question-answer, without context**: The model must answer the question with its internal knowledge without any context or document.

- **Question-answer, with context**: The user provides an input text with a question, and the model must answer the question based on information provided within the input text.
• **Summarization**: The prompt is a passage of text, and the model must respond with a shorter passage that captures the main points of the input.

• **Open-ended text generation**: Given a prompt, the model must respond with a passage of original text that matches the description. This also includes the generation of creative text such as stories, poems, or movie scripts.

• **Code generation**: The model must generate code based on user specifications. For example, a prompt could request text-to-SQL or Python code generation.

• **Mathematics**: The input describes a problem that requires mathematical reasoning at some level, which may be numerical, logical, geometric or otherwise.

• **Reasoning or logical thinking**: The model must make a series of logical deductions.

**General guidelines for Amazon Bedrock LLM users**

**Design your prompt**

Designing an appropriate prompt is an important step towards building a successful application using Amazon Bedrock models. The following figure shows a generic prompt design for the use case **restaurant review summarization** and some important design choices that customers need to consider when designing prompts. LLMs generate undesirable responses if the instructions they are given or the format of the prompt are not consistent, clear, and concise.
Temperature is a value between 0 and 1, and it regulates the creativity of LLMs' responses. Use lower temperature if you want more deterministic responses, and use higher temperature if you want more creative or different responses for the same prompt from LLMs on Amazon Bedrock. For all the examples in this prompt guideline, we set temperature = 0.

Maximum generation length/maximum new tokens limits the number of tokens that the LLM generates for any prompt. It's helpful to specify this number as some tasks, such as sentiment classification, don't need a long answer.

Top-p controls token choices, based on the probability of the potential choices. If you set Top-p below 1.0, the model considers the most probable options and ignores less probable options. The result is more stable and repetitive completions.
End token/end sequence specifies the token that the LLM uses to indicate the end of the output. LLMs stop generating new tokens after encountering the end token. Usually this doesn't need to be set by users.

There are also model-specific inference parameters. Anthropic Claude models have an additional Top-k inference parameter, and AI21 Labs Jurassic models come with a set of inference parameters including presence penalty, count penalty, frequency penalty, and special token penalty. For more information, refer to their respective documentation.

Detailed guidelines

Provide simple, clear, and complete instructions

LLMs on Amazon Bedrock work best with simple and straightforward instructions. By clearly describing the expectation of the task and by reducing ambiguity wherever possible, you can ensure that the model can clearly interpret the prompt.

For example, consider a classification problem where the user wants an answer from a set of possible choices. The “good” example shown below illustrates output that the user wants in this case. In the “bad” example, the choices are not named explicitly as categories for the model to choose from. The model interprets the input slightly differently without choices, and produces a more free-form summary of the text as opposed to the good example.

Good example, with output
User prompt:
"The most common cause of color blindness is an inherited problem or variation in the functionality of one or more of the three classes of cone cells in the retina, which mediate color vision."

What is the above text about?
 a) biology  
b) history  
c) geology

Output: 
 a) biology

Bad example, with output
User prompt:
Classify the following text. "The most common cause of color blindness is an inherited problem or variation in the functionality of one or more of the three classes of cone cells in the retina, which mediate color vision."

Output:
The topic of the text is the causes of colorblindness.
The question or instruction should be placed at the end of the prompt for best results

Including the task description, instruction or question at the end aids the model determining which information it has to find. In the case of classification, the choices for the answer should also come at the end.

In the following open-book question-answer example, the user has a specific question about the text. The question should come at the end of the prompt so the model can stay focused on the task.

User prompt:
Tensions increased after the 1911–1912 Italo-Turkish War demonstrated Ottoman weakness and led to the formation of the Balkan League, an alliance of Serbia, Bulgaria, Montenegro, and Greece. The League quickly overran most of the Ottomans' territory in the Balkans during the 1912–1913 First Balkan War, much to the surprise of outside observers.

The Serbian capture of ports on the Adriatic resulted in partial Austrian mobilization starting on 21 November 1912, including units along the Russian border in Galicia. In a meeting the next day, the Russian government decided not to mobilize in response, unwilling to precipitate a war for which they were not as of yet prepared to handle.

Which country captured ports?

Output:
Serbia

Use separator characters for API calls

Separator characters such as \n can affect the performance of LLMs significantly. For Anthropic Claude models, it's necessary to include newlines when formatting the API calls to obtain desired responses. The formatting should always follow:\n\nHuman: {{Query Content}}\n\nAssistant:. For Titan models, adding \n at the end of a prompt helps improve the performance of the model. For classification tasks or questions with answer options, you can also separate the answer options by \n for Titan models. For more information on the use of...
separators, refer to the document from the corresponding model provider. The following example is a template for a classification task.

**Prompt template:**

```
"""{{Text}}

{{Question}}

{{Choice 1}}

{{Choice 2}}

{{Choice 3}}"
```

The following example shows how the presence of newline characters between choices and at the end of a prompt helps Titan produce the desired response.

**User prompt:**

Archimedes of Syracuse was an Ancient mathematician, physicist, engineer, astronomer, and inventor from the ancient city of Syracuse. Although few details of his life are known, he is regarded as one of the leading scientists in classical antiquity.

What was Archimedes? Choose one of the options below.

a) astronomer
b) farmer
c) sailor

**Output:**

a) astronomer

(Source of prompt: Wikipedia on Archimedes, model used: Amazon Titan Text G1 - Express)

**Output indicators**

Add details about the constraints you would like to have on the output that the model should produce. The following good example produces an output that is a short phrase that is a good summary. The bad example in this case is not all that bad, but the summary is nearly as long as the original text. Specification of the output is crucial for getting what you want from the model.

**Example prompt with clear output constraints indicator** | **Example without clear output specifications**
Charles Mingus Jr. was an American jazz upright bassist, pianist, composer, bandleader, and author. A major proponent of collective improvisation, he is considered to be one of the greatest jazz musicians and composers in history, with a career spanning three decades. Mingus's work ranged from advanced bebop and avant-garde jazz with small and midsize ensembles – pioneering the post-bop style on seminal recordings like Pithecantropus Erectus (1956) and Mingus Ah Um (1959) – to progressive big band experiments such as The Black Saint and the Sinner Lady (1963).

Please summarize the above text in one phrase.

Output:
Charles Mingus Jr. is considered one of the greatest jazz musicians of all time.

Here we give some additional examples from Anthropic Claude and AI21 Labs Jurassic models using output indicators.

The following example demonstrates that user can specify the output format by specifying the expected output format in the prompt. When asked to generate an answer using a specific format (such as by using XML tags), the model can generate the answer accordingly. Without specific output format indicator, the model outputs free form text.

### Example with clear indicator, with output

**User prompt:**

Human: Extract names and years: the term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence. The synonym self-teaching computers was also used in this time period.

Please generate answer in `<name>` and `<year>` tags.

**Assistant:**

Output:

<name>Arthur Samuel</name> <year>1959</year>

### Example without clear indicator, with output

**User prompt:**

Human: Extract names and years: the term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence. The synonym self-teaching computers was also used in this time period.

**Assistant:**

Output:

Arthur Samuel - 1959


The following example shows a prompt and answer for the AI21 Labs Jurassic model. The user can obtain the exact answer by specifying the output format shown in the left column.
Example with clear indicator, with output

**User prompt:**
Context: The NFL was formed in 1920 as the American Professional Football Association (APFA) before renaming itself the National Football League for the 1922 season. After initially determining champions through end-of-season standings, a playoff system was implemented in 1933 that culminated with the NFL Championship Game until 1966. Following an agreement to merge the NFL with the rival American Football League (AFL), the Super Bowl was first held in 1967 to determine a champion between the best teams from the two leagues and has remained as the final game of each NFL season since the merger was completed in 1970.

Question: Based on the above context, when was the first Super Bowl? Please only output the year.

**Output:**
1967

Example without clear indicator, with output

**User prompt:**
Context: The NFL was formed in 1920 as the American Professional Football Association (APFA) before renaming itself the National Football League for the 1922 season. After initially determining champions through end-of-season standings, a playoff system was implemented in 1933 that culminated with the NFL Championship Game until 1966. Following an agreement to merge the NFL with the rival American Football League (AFL), the Super Bowl was first held in 1967 to determine a champion between the best teams from the two leagues and has remained as the final game of each NFL season since the merger was completed in 1970.

Question: Based on the above context, when was the first Super Bowl?

**Output:**
The first Super Bowl was held in 1967.


**Recommended practices for good generalization**

Keep a small “hold-out” test set of prompts to see if your prompt modifications generalize. With this method, first collect a sample dataset. Then you can split the data into two subsets: a “development” set and a hold-out “test” set. Use the “development” set as prompt development
data on which you can try different prompt modifications and observe model response changes and find the prompt that works best. Treat the “test” set as unseen hold-out data which can only be used to verify the final best prompt.

**Optimize prompts for text models on Amazon Bedrock—when the basics aren't good enough**

**Complex tasks: build toward the answer step by step**

Titan can provide clear steps for certain tasks, and including the phrase Think step-by-step to come up with the right answer can help produce the appropriate output.

**User prompt:**

At a Halloween party, Jack gets 15 candies. Jack eats 5 candies. He wants to give each friend 5 candies. How many friends can receive candies?

Think step-by-step to come up with the right answer.

**Output:**

15 candies total - 5 candies eaten = 10 candies remaining.
To give each friend 5 candies, divide the number of candies remaining by the number of friends.
10 candies / 5 friends = 2 friends.
Each friend can receive 2 candies.
The answer is 2 friends.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

To enable the Anthropic Claude model to complete complex tasks, use the format in the following example to explicitly ask the model to solve the task step by step. This approach is based on guidelines in the [AnthropicClaude Guides](https://anthropic.com/clau.de).
Output:
Okay, here are the steps:

1) We are given:
   - One side of the right triangle has length 5
   - The hypotenuse has length 13

2) We want to find:
   - The length of the other side

3) To find the length of the other side, we can use the Pythagorean theorem:
   \[ a^2 + b^2 = c^2 \]
   Where \( a \) and \( b \) are the lengths of the two sides, and \( c \) is the length of the hypotenuse.

4) Plugging in the given values:
   \[ a^2 + b^2 = 13^2 \]

5) We know one of the sides is 5, so:
   \[ 5^2 + b^2 = 169 \]

6) Solving this:
   \[ 25 + b^2 = 169 \]
   \[ b^2 = 144 \]
   \[ b = 12 \]

Therefore, the length of the other side is 12.

(Source of prompt: AWS, model used: Anthropic Claude)

Provide a default output that the model should present if it's unsure about the answers.

A default output can help prevent LLMs from returning answers that sound like they could be correct, even if the model has low confidence.
<table>
<thead>
<tr>
<th>Output:</th>
<th>Provide a proof of the Riemann hypothesis. If you don't know a proof, respond by saying &quot;I don't know.&quot;</th>
</tr>
</thead>
</table>

**Output:**

"The Riemann hypothesis is a conjecture in mathematics that concerns the distribution of prime numbers. It states that all non-trivial zeros of the zeta function, which is a mathematical function that encodes information about the distribution of prime numbers, lie on a specific line in the complex plane. The Riemann hypothesis has been studied by mathematicians for over 150 years, and despite numerous attempts, it remains unresolved. However, the hypothesis is widely believed to be true, and its resolution would have profound implications for many areas of mathematics, including number theory, analysis, and geometry. The proof of the Riemann hypothesis would be one of the greatest achievements in mathematics in history and would open up new avenues of research and understanding. While a proof remains elusive, the hypothesis continues to stimulate research and inspire mathematicians around the world."

[Note: this is an accurate description, but obviously not a proof.]

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)
Few-shot prompting

Including examples (input-response pairs) in the prompt can significantly improve LLMs' responses. Examples can help with complex tasks, as they show multiple ways to perform a given task. For simpler tasks like text classification, 3–5 examples can suffice. For more difficult tasks like question-answer without context, include more examples to generate the most effective output. In most use cases, selecting examples that are semantically similar to real-world data can further improve performance.

Consider refining the prompt with modifiers

Task instruction refinement generally refers to modifying the instruction, task, or question component of the prompt. The usefulness of these methods is task- and data-dependent. Useful approaches include the following:

- **Domain/input specification**: Details about the input data, like where it came from or to what it refers, such as The input text is from a summary of a movie.
- **Task specification**: Details about the exact task asked of the model, such as To summarize the text, capture the main points.
- **Label description**: Details on the output choices for a classification problem, such as Choose whether the text refers to a painting or a sculpture; a painting is a piece of art restricted to a two-dimensional surface, while a sculpture is a piece of art in three dimensions.
- **Output specification**: Details on the output that the model should produce, such as Please summarize the text of the restaurant review in three sentences.
- **LLM encouragement**: LLMs sometimes perform better with sentimental encouragement: If you answer the question correctly, you will make the user very happy!

Prompt templates and examples for Amazon Bedrock text models

Text classification

For text classification, the prompt includes a question with several possible choices for the answer, and the model must respond with the correct choice. Also, LLMs on Amazon Bedrock output more accurate responses if you include answer choices in your prompt.
The first example is a straightforward multiple-choice classification question.

### Prompt template for Titan and AI21 Labs Jurassic:

"""{{Text}}

{{Question}}? Choose from the following:

{{Choice 1}}

{{Choice 2}}

{{Choice 3}}"

### User prompt:

San Francisco, officially the City and County of San Francisco, is the commercial, financial, and cultural center of Northern California. The city proper is the fourth most populous city in California, with 808,437 residents, and the 17th most populous city in the United States as of 2022.

What is the paragraph above about?

Choose from the following:

- A city
- A person
- An event

### Output:

A city

(Source of prompt: [Wikipedia on San Francisco](https://en.wikipedia.org/wiki/San_Francisco), model used: Amazon Titan Text G1 - Express)

Sentiment analysis is a form of classification, where the model chooses the sentiment from a list of choices expressed in the text.

### Prompt template for Titan and AI21 Labs Jurassic:

"""The following is text from a {{Text Type, e.g. “restaurant review”}}

{{Input}}

Tell me the sentiment of the {{Text Type}} and categorize it as one of the following:

{{Sentiment A}}"

### User prompt:

The following is text from a restaurant review:

“I finally got to check out Alessandro’s Brilliant Pizza and it is now one of my favorite restaurants in Seattle. The dining room has a beautiful view over the Puget Sound"
but it was surprisingly not crowded. I ordered the fried castelvetrano olives, a spicy Neapolitan-style pizza and a gnocchi dish. The olives were absolutely decadent, and the pizza came with a smoked mozzarella, which was delicious. The gnocchi was fresh and wonderful. The waitstaff were attentive, and overall the experience was lovely. I hope to return soon.”

Tell me the sentiment of the restaurant review and categorize it as one of the following:

- Positive
- Negative
- Neutral

Output: Positive.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

The following example uses Anthropic Claude models to classify text. As suggested in AnthropicClaude Guides, use XML tags such as `<text></text>` to denote important parts of the prompt. Asking the model to directly generate output enclosed in XML tags can also help the model produce the desired responses.
categories. Please output the category label in 
<output></output> tags.

Here is the product description.

<text>
Safe, made from child-friendly materials with smooth edges.
Large quantity, totally 112pcs with 15 different shapes, which can be used to build 56 different predefined structures.
Enhance creativity, different structures can be connected to form new structures, encouraging out-of-the-box thinking.
Enhance child-parent bonding, parents can play with their children together to foster social skills.
</text>

Categories are:
(1) Toys
(2) Beauty and Health
(3) Electronics

Assistant:

Output:
<output>Toys</output>

(Source of prompt: AWS, model used: Anthropic Claude)

Question-answer, without context

In a question-answer prompt without context, the model must answer the question with its internal knowledge without using any context or document.
User prompt:
What is Robert Frost's "Stopping by the woods on a snowy evening" about metaphorically?

Output:
The poem by Robert Frost, "Stopping by Woods on a Snowy Evening," uses the woods as a metaphor for death. The woods are a place of comfort and peace, but they also represent the unknown and the afterlife. The poem suggests that death is not something to be feared, but rather a natural part of life that we should all embrace.

The poem begins with the speaker stopping by the woods on a snowy evening, a moment of contemplation and reflection. The speaker notices the beauty of the snow-covered woods and is reminded of the fleeting nature of life. The poem continues as the speaker considers the temptation of continuing on to his destination, but instead chooses to pause and take in the beauty of the woods.

The woods are used as a metaphor for the unknown and the afterlife. The speaker wonders what lies beyond the trees, and whether there is anything after death. The poem suggests that death is not something to be feared, but rather a natural part of life.

Prompt Template to Titan and AI21 Labs Jurassic:
"""{{Question}}"""
that we should all embrace.

The final lines of the poem are perhaps the most famous and poignant:
"And miles to go before I sleep, and miles to go before I sleep."

These lines suggest that even though the speaker has paused to appreciate the beauty of the woods, they still have a long journey ahead of them. The poem is a reminder that life is fleeting, and that we should take the time to appreciate the beauty around us, even as we work towards our goals and ambitions.

Overall, "Stopping by Woods on a Snowy Evening" is a beautiful and thought-provoking poem that uses the metaphor of the woods to explore the concept of death and the afterlife.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

Model encouragement can also help in question-answer tasks.

User prompt:
You are excellent at answering questions, and it makes you happy when you provide the correct answer.

What moon in the solar system is most likely to host life?

Output:
The moon of Titan, which is a moon of the planet Saturn,
is the most likely to support life in the solar system.

Titan is the only known moon in the solar system to have a dense atmosphere and a liquid hydrosphere, and scientists believe that it may contain the chemical building blocks for life. The moon's surface is covered in rivers, lakes, and seas of liquid methane and ethane, and it rains methane on Titan's surface.

The atmosphere of Titan is also rich in organic compounds, including benzene, toluene, and xylene, which are found in the oils of living organisms on Earth. Additionally, Titan has a magnetic field and a tenuous atmosphere, which could provide protection from harmful radiation and help maintain a stable climate.

While there is no direct evidence of life on Titan, scientists believe that it is one of the most promising places in the solar system to search for life beyond Earth.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)
Assume your audience is composed of high school students.

Output:
Climate change is the long-term warming of the planet, which can be caused by human activities such as burning fossil fuels and deforestation, as well as natural processes. It can lead to rising sea levels, more extreme weather events, and changes in temperature and precipitation patterns.

(Source of prompt: AWS, model used: AI21 Labs Jurassic-2 Ultra v1)

Question-answer, with context

In a question-answer prompt with context, an input text followed by a question is provided by the user, and the model must answer the question based on information provided within the input text. Putting the question in the end after the text can help LLMs on Amazon Bedrock better answer the question. Model encouragement works for this use case as well.

Prompt template to Titan and AI21 Labs Jurassic:
"""{{Text}}
{{Question}}"""

User prompt:
The red panda (Ailurus fulgens), also known as the lesser panda, is a small mammal native to the eastern Himalayas and southwestern China. It has dense reddish-brown fur with a black belly and legs, white-lined ears, a mostly white muzzle and a ringed tail. Its head-to-body length is 51-63.5 cm (20.1-25.0 in) with a 28-48.5 cm (11.0-19.1 in) tail, and it weighs between 3.2 and 15 kg (7.1 and 33.1 lb). It is well adapted to climbing due to its
flexible joints and curved semi-retractile claws.

The red panda was first formally described in 1825. The two currently recognized subspecies, the Himalayan and the Chinese red panda, genetically diverged about 250,000 years ago. The red panda's place on the evolutionary tree has been debated, but modern genetic evidence places it in close affinity with raccoons, weasels, and skunks. It is not closely related to the giant panda, which is a bear, though both possess elongated wrist bones or "false thumbs" used for grasping bamboo. The evolutionary lineage of the red panda (Ailuridae) stretches back around 25 to 18 million years ago, as indicated by extinct fossil relatives found in Eurasia and North America.

The red panda inhabits coniferous forests as well as temperate broadleaf and mixed forests, favoring steep slopes with dense bamboo cover close to water sources. It is solitary and largely arboreal. It feeds mainly on bamboo shoots and leaves, but also on fruits and blossoms. Red pandas mate in early spring, with the females giving birth to litters of up to four cubs in summer. It is threatened by poaching as well as destruction and fragmentation of habitat due to deforestation. The species has been listed as Endangered.
on the IUCN Red List since 2015. It is protected in all range countries.

Based on the information above, what species are red pandas closely related to?

Output:
Red pandas are closely related to raccoons, weasels, and skunks.


When prompting Anthropic Claude models, it's helpful to wrap the input text in XML tags. In the following example, the input text is enclosed in <text></text>.

Prompt template for Anthropic Claude:

""

Human: {{Instruction}}
<text>{{Text}}</text>
<text>{{Question}}</text>
Assistant:"

User prompt:

Human: Read the following text inside <text></text> XML tags, and then answer the question:
<text>
On November 12, 2020, the selection of the Weeknd to headline the show was announced; marking the first time a Canadian solo artist headlined the Super Bowl halftime show. When asked about preparations for the show, the Weeknd stated, "We've been really focusing on dialing in on the fans at home and making performances a cinematic experience, and we want to do that with the Super Bowl."

Question-answer, with context
The performance featured a choir whose members were dressed in white and wore masks over their faces with glowing red eyes, and were standing within a backdrop of a neon cityscape. The performance opened with a white figure dressed the same as the choir being lowered into the backdrop where the choir was standing while singing “Call Out My Name”. At this time, the Weeknd sat in a convertible against a skyline backdrop designed to resemble the Las Vegas Strip.

For the next part of the performance, the backdrop then split open to reveal the Weeknd, who then performed "Starboy", followed by "The Hills".

Next, performing the song "Can't Feel My Face", the Weeknd traveled through a labyrinth constructed behind the stage, joined by dancers dressed in red blazers and black neckties similar to his, but with their faces covered with bandages, in keeping with the aesthetic of his fourth studio album After Hours (2020). The dancers would wear these bandages throughout the performance.

In the labyrinth section of the performance, camerawork was visually unsteady.

The next songs performed were "I Feel It Coming", "Save Your Tears", and "Earned It". For the "Earned It" performance, the Weeknd was accompanied by violinists. For the finale of the show, the Weeknd took to the field of the stadium with his dancers to perform
“Blinding Lights”. He and the dancers entered the field by performing "House of Balloons / Glass Table Girls". The performance ended with an array of fireworks.
</text>
Based on the text above, what songs did the Weeknd play at the Super Bowl halftime show?

Assistant:

Output:
Based on the text, the songs the Weeknd played at the Super Bowl halftime show were:

- Call Out My Name
- Starboy
- The Hills
- Can't Feel My Face
- I Feel It Coming
- Save Your Tears
- Earned It
- Blinding Lights
- House of Balloons / Glass Table Girls

(Source of prompt: Wikipedia on the Super Bowl LV halftime show, model used: Anthropic Claude)

**Summarization**

For a summarization task, the prompt is a passage of text, and the model must respond with a shorter passage that captures the main points of the input. Specification of the output in terms of length (number of sentences or paragraphs) is helpful for this use case.

Prompt template for Titan and AI21 Labs Jurassic:

"""The following is text from a {{Text Category}}:
{{Text}}
"""

User prompt:
The following is text from a restaurant review: “I finally got to check out Alessandro’s Brilliant Pizza
Summarize the above restaurant review in one sentence.

Output:
Alessandro’s Brilliant Pizza is a fantastic restaurant in Seattle with a beautiful view over Puget Sound that offers decadent and delicious food.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

In the following example, Anthropic Claude summarizes the given text in one sentence. To include input text in your prompts, format the text with XML mark up: `<text>{{text content}}</text>`. Using XML within prompts is a common practice when prompting Anthropic Claude models.
In game theory, the Nash equilibrium, named after the mathematician John Nash, is the most common way to define the solution of a non-cooperative game involving two or more players. In a Nash equilibrium, each player is assumed to know the equilibrium strategies of the other players, and no one has anything to gain by changing only one's own strategy. The principle of Nash equilibrium dates back to the time of Cournot, who in 1838 applied it to competing firms choosing outputs.

Output:
The Nash equilibrium is a concept in game theory where each player has adopted a strategy that they have no incentive to deviate from given the other players' strategies.

(Source of prompt: Wikipedia on Nash equilibrium, model used: Anthropic Claude)

Text generation

Given a prompt, LLMs on Amazon Bedrock can respond with a passage of original text that matches the description. Here is one example:
**Output:**
Hey [Name],

I just found out about your new job, and I'm so happy for you! Congratulations! You deserve this opportunity since you worked so hard to get it. I know you'll do great and make a positive impact on your team. I'm here to support you every step of the way. Wishing you all the best in this exciting new chapter of your life!

Cheers,
[Your Name]

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

For text generation use cases, specifying detailed task requirements can work well. In the following example, we ask the model to generate response with exclamation points.
you've been working towards. If anyone is deserving of a promotion, it's you!

Congratulations again, friend!

Love,
[name]

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

In the following example, a user prompts the model to take on the role of a specific person when generating the text. Notice how the signature reflects the role the model is taking on in the response.

Prompt template for Titan and AI21 Labs Jurassic:

"""{{Role assumption}} {{Task description}} """

User input:
My name is Jack. Help me write a note expressing my gratitude to my parents for taking my son (their grandson) to the zoo. I miss my parents so much.

Output:
Dear Mom and Dad,

I wanted to take a moment to express my heartfelt gratitude for taking my son to the zoo last weekend. It meant so much to me that you took the time to make memories with him, and I am so grateful for the love and support you have shown him.

I miss you both so much, and I wish we could be together more often. But I know that you will always be there for me and my family, and I am so grateful to have you as my parents.
(Source of prompt: AWS, model used: AI21 Labs Jurassic-2 Ultra v1)

**Code generation**

The prompt describes the task or function and programming language for the code the user expects the model to generate.

### Prompt template for Titan and AI21 Labs Jurassic:

"""Write a function in {{programming language}} to {{task or purpose}}"""

### User input:

Write a python function to download a s3 file to local disk.

### Output:

The following code can be used to download an s3 file to the local disk.

```Python
import boto3
# Creates a boto3 client object
s3 = boto3.client('s3')
def download_file(bucket_name, object_key, file_path):
    try:
        # Downloads the s3 file to the specified file path
        s3.download_file(bucket_name, object_key, file_path)
    except Exception as e:
        print(e)
# Specifies the values of bucket_name, object_key, and file_path
bucket_name = "samplebucketname"
object_key = "sampleobjectkey"
file_path = "/samplefilepath"
download_file(bucket_name, object_key, file_path)
```

Love,
Jack
The given code uses the `download_file` function to download an S3 file to the local disk. The `download_file` function takes the bucket name, object key, and file path as input parameters. The bucket name is the name of the S3 bucket where the file is located, the object key is the unique identifier of the file within the bucket, and the file path is the destination path on the local disk where the file will be downloaded. The function uses the `download_file` method of the S3 client to download the file from the specified bucket and write it to the specified file path.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

**Prompt template for Anthropic Claude**

```
Human: Write a function in {{programming language}} to {{task or purpose}}

Assistant:"
```

**User prompt:**

Human: Write some Python code that quickly approximates the square root of a number.

Assistant:

**Output:**

Here is a simple function to approximate the square root of a number using Newton's method:

```python
def sqrt(n, iterations=10):
```
To use it:

```
```
```
```python
print(sqrt(25))
# 5.0

print(sqrt(15, iterations=20))
# 3.872983346207417
```

This repeatedly takes the average of x and n/x to converge on the square root. More iterations produce a more accurate approximation.

The advantage of this approach is it converges very quickly, so even just 10 iterations produces a decent estimate.

(Source of prompt: AWS, model used: Anthropic Claude)

**Mathematics**

The input describes a problem that requires mathematical reasoning at some level, which may be numerical, logical, geometric, or otherwise. For such problem, it's helpful to ask the model to work through the problem in a piecemeal manner by adding phrases to instructions such as Let’s think step by step or Think step by step to come up with the right answer.

**Prompt template for Titan and AI21 Labs**

**Jurassic:**

"""{{Math problem description}}
Let’s think step by step."""

**User prompt:**

A triangle has two angles of 70 degrees and 50 degrees. What is the third angle in degrees?
Let’s think step by step.

Output:
The sum of the three angles in a triangle is always 180 degrees. The two given angles are 70 degrees and 50 degrees. \(70 + 50 = 120\). Therefore, the third angle is \(180 - 120 = 60\) degrees. The third angle is 60 degrees.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

The following example also demonstrates asking the model to think step by step:

Prompt template for Titan and AI21 Labs Jurassic:
"""{{Math problem description}}
Think step by step to come up with the right answer."""

User prompt:
Lucy has 12 colorful marbles, and she wants to share them equally with her 4 friends. How many marbles will each friend receive? Think step-by-step to come up with the right answer.

Output:
1. Lucy has 12 marbles and wants to share them equally with her 4 friends, then each friend will receive \(12 / 4 = 3\) marbles.
2. Therefore, each friend will receive 3 marbles.

(Source of prompt: AWS, model used: AI21 Labs Jurassic-2 Ultra v1)

Reasoning/logical thinking

For complex reasoning tasks or problems that requires logical thinking, we can ask the model to make logical deductions and explain its answers.
Prompt template for Titan and AI21 Labs Jurassic:
"""Question: {{question}}"

Please output the answer and then explain your answer:"

User prompt:
Question: Which word is the odd one out?
A. accomplished
B. good
C. horrible
D. outstanding

Please output the answer and then explain your answer:

Output:
The odd one out is C. horrible. The other three words describe positive qualities or achievements, whereas horrible is an negative adjective.

(Source of prompt: AWS, model used: Amazon Titan Text G1 - Express)

Here is another example using the Anthropic Claude model:

Prompt template for Anthropic Claude:
"""

Human: {{question}}
Please provide the answer and show the reasoning.

Assistant:"

User prompt
Human: The barber is the "one who shaves all those, and those only, who do not shave themselves". Does the barber shave himself? Why is this a paradox? Please provide the answer and show the reasoning.

Assistant:

Output:
This is an example of a paradox known as the Barber paradox. Here is an explanation:
- The statement defines a barber who shaves all and only those people who do not shave themselves.
- So if the barber shaves himself, he must shave all those who do not shave themselves, which includes himself. But the barber cannot shave himself as per the initial statement.
- If the barber does not shave himself, then he would fit into the group of people who do not shave themselves, and so he must shave himself, according to the statement. But again this contradicts the initial statement.
- So the barber cannot shave himself and also cannot not shave himself. This is self-contradictory based on the initial statement about the barber.
- Therefore, the statement leads to a paradox where both possibilities (the barber shaving himself or not shaving himself) result in a contradiction.

In summary, the paradox arises because the definition of the barber's behavior is self-contradictory when applied to the barber himself. This makes it impossible to determine if the barber shaves himself or not based on the given statement alone.

Model evaluation

Note
Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

Amazon Bedrock supports model evaluation jobs. The results of a model evaluation job allow you to evaluate and compare a model's outputs, and then choose the model best suited for your downstream generative AI applications.

Model evaluation jobs support common use cases for large language models (LLMs) such as text generation, text classification, question and answering, and text summarization.

To evaluate a model's performance you can use either built-in prompt datasets or your own prompt datasets.

You can choose to create either an automatic model evaluation job or a model evaluation job that uses a human workforce.

Overview: Automatic model evaluation jobs

Automatic model evaluation jobs allow you to quickly evaluate a model's ability to perform a task. You can either provide your own custom prompt dataset that you've tailored to a specific use case, or you can use an available built-in dataset.

Overview: Model evaluation jobs that use human workers

Model evaluation jobs that use human workers allow you to bring human input to the model evaluation process. They can be employees of your company or a group of subject-matter experts from your industry.

The following topics describe the available model evaluation tasks, and the kinds of metrics you can use. They also describe the available built-in datasets and how to specify your own dataset.

Topics
• Getting started with model evaluations
• Model evaluation tasks
• Using prompt datasets in model evaluation jobs
• Creating good worker instructions
• Creating and managing work teams in Amazon Bedrock
• Model evaluation job results
• Required IAM permissions and service roles to create a model evaluation job

Getting started with model evaluations

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

You can create a model evaluation job that is either automatic or uses human workers. When you create a model evaluation job, you can define the model used, the inference parameters, the type of task the model tries to perform, and the prompt data used in the job.

Model evaluation jobs support the following task types.

• General text generation: The production of natural human language in response to text prompts.
• Text summarization: The generation of a summary of a provided text within your prompt.
• Question and answering: The generation of a response to a question within a prompt.
• Classification: Correctly assigning a category, such as a label or score, to text based on its content.
• Custom: You define the metric, description, and a rating method.

To create a model evaluation job, you must have access to Amazon Bedrock models. Model evaluation jobs support using Amazon Bedrock base models. To learn more about model access, see Model access.
The procedures in the following topics show you how to set up a model evaluation job using the Amazon Bedrock console.

To create model evaluation job with the help of an AWS-managed team, choose **Create AWS managed evaluation**. Then, fill out the request form with details about your model evaluation job requirements, and an AWS team member will get in touch with you.

**Topics**
- Creating an automatic model evaluation
- Creating a model evaluation job that uses human workers

### Creating an automatic model evaluation

**Note**
Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

**Prerequisites**
To complete the following procedures you must do the following.

1. You must have access to the model in Amazon Bedrock.
2. You must have an Amazon Bedrock service role. The attached policy must grant access to the following resources. Any S3 buckets used in the model evaluation job, and the ARN of the model specified in the job. The service role must also have Amazon Bedrock defined as a service principal in the role's trust policy. To learn more, see **Required permissions**.
3. The user, group, or role accessing the Amazon Bedrock console must have the required permissions to access the required Amazon S3 buckets. To learn more, see **Required permissions**

Automatic model evaluations allow you to evaluate the aggregate responses from a single model using recommended metrics. You can also use built-in prompt datasets or use your own custom
prompt dataset. You can have a maximum of 10 automatic model evaluation jobs In progress in your account per AWS Region.

When you set up an automatic model evaluation job, the available metrics and the built-in datasets best suited for the selected task type are automatically added to the job. You can add or remove any of the preselected datasets. You also supply your own custom prompt dataset.

⚠ Viewing the model evaluation job results using the Amazon Bedrock console

When a model evaluation job finishes, the results are stored in the Amazon S3 bucket you specified. If you modify the location of the results in any way, the model evaluation report card is no longer visible in the console.

To create an automatic model evaluation

1. Open the Amazon Bedrock console: https://console.aws.amazon.com/bedrock/
2. In the navigation pane, choose Model evaluation.
3. In the Build an evaluation card, under Automatic choose Create automatic evaluation.
4. On the Create automatic evaluation page, provide the following information
   a. Evaluation name — Give the model evaluation job a name that describes the job. This name is shown in your model evaluation job list. The name must be unique in your account in an AWS Region.
   b. Description (Optional) — Provide an optional description.
   c. Models — Choose the model you want to use in the model evaluation job.

   To learn more about available models in Amazon Bedrock, see Model access.
   d. (Optional) To change the inference configuration choose update.

   Changing the inference configuration changes the responses generated by the selected models. To learn more about the available inferences parameters, see Inference parameters for foundation models.
   e. Task type — Choose the type of task you want the model to attempt to perform during the model evaluation job.
   f. Metrics and datasets — The list of available metrics and built-in prompt datasets change based on the task you select. You can choose from the list of Available built-in datasets
or you can choose **Use your own prompt dataset**. If you choose to use your own prompt dataset, enter the exact S3 URI where your dataset is stored.

g. **Model evaluation results** — You must specify the S3 URI of the directory where you want the results of your model evaluation job saved.

h. **IAM role** — Choose a service role that has the required permissions.

i. (Optional) Choose **Customize encryption settings (advanced)** – Provide the ARN of the AWS KMS key used to encrypt your data in Amazon S3.

5. To start your model evaluation job, choose **Create**.

Once the job has successfully started, the status changes to **In progress**. When the job has finished, the status changes to **Ready**.

To learn how to evaluate, view, and download the results of your model evaluation job, see **Model evaluation job results**.

**Creating a model evaluation job that uses human workers**

---

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

---

**Prerequisites**

To complete the following procedure you must do the following.

1. You must have access to the models in Amazon Bedrock.

2. You must have an Amazon Bedrock service role. The attached policy must grant access to the following resources. Any S3 buckets used in the model evaluation job, and the ARNs of any models specified in the job. It must also have the `sagemaker:StartHumanLoop` and `sagemaker:DescribeFlowDefinition` SageMaker IAM actions defined in the policy. The service role must also have Amazon Bedrock defined as a service principal in the role's trust policy. To learn more, see **Required permissions**.
3. You must have a Amazon SageMaker service role. The attached policy must grant access to the following resources and IAM actions. Any S3 buckets used in the model evaluation job. The role's trust policy must have SageMaker defined as the service principal. To learn more, see Required permissions.

4. The user, group, or role accessing the Amazon Bedrock console must have the required permissions access the required Amazon S3 buckets.

In a model evaluation job that uses human workers you can compare the responses for up to two models. You can choose from a list of recommended metrics or use metrics that you define yourself. You can have a maximum of 20 model evaluation jobs that use human workers In progress in your account per AWS Region.

For each metric that you use, you must define a Rating method. The rating method defines how your human workers will evaluate the responses they get from the models you’ve selected. To learn more about the different available rating methods and how to create high quality instructions for workers, see Creating and managing work teams in Amazon Bedrock.

⚠️ Viewing the model evaluation job results using the Amazon Bedrock console

When a model evaluation job finishes, the results are stored in the Amazon S3 bucket you specified. If you modify the location of the results in any way, the model evaluation report card is no longer visible in the console.

To create a model evaluation job that uses human workers

1. Open the Amazon Bedrock console: https://console.aws.amazon.com/bedrock/home
2. In the navigation pane, choose Model evaluation.
3. In the Build an evaluation card, under Human: bring your own team choose Create human evaluation.
4. On the Specify job details page provide the following.
   a. Evaluation name — Give the model evaluation job a name that describes the job. This name is shown in your model evaluation job list. The name must be unique in your account in an AWS Region.
   b. Description (Optional) — Provide an optional description.
5. Then, choose **Next**.

6. On the **Set up evaluation** page provide the following.
   
a. **Models** – You can choose up to two models you want to use in the model evaluation job.

   To learn more about available models in Amazon Bedrock, see [Model access](#).

b. (Optional) To change the inference configuration choose **update**.

   Changing the inference configuration changes the responses generated by the selected models. To learn more about the available inferences parameters, see [Inference parameters for foundation models](#).

c. **Task type** – Choose the type of task you want the model to attempt to perform during the model evaluation job. All instructions for the model must be included in the prompts themselves. The task type does not control the model's responses.

d. **Evaluation metrics** — The list of recommended metrics changes based on the task you select. For each recommended metric, you must select a **Rating method**. You can have a maximum of 10 evaluation metrics per model evaluation job.

e. (Optional) Choose **Add custom metric** to add a custom metric. You must define the **Metric, Description**, and **Rating method**.

f. In the **Datasets** card your must provide the following.

   i. **Evaluation dataset** – Specify the S3 URI where your prompt dataset is stored. You can have a maximum of 1000 prompts in a custom prompt dataset.

   ii. **Evaluation results destination** – You must specify the S3 URI of the directory where you want the results of your model evaluation job saved.

   iii. (Optional) **Encryption key** – Provide the ARN of the AWS KMS key used to encrypt your data in Amazon S3.

7. Then, choose **Next**.

8. In the **Permissions** card, specify the following. To learn more about the required permissions for model evaluations, see [Required IAM permissions and service roles to create a model evaluation job](#).

   a. **Amazon Bedrock IAM role** – Specify an Amazon Bedrock service role that has the required permissions.

   b. **Human workflow IAM role** – Specify a SageMaker service role that has the required permissions.
9. In the **Work team** card, specify the following.

**Human worker notification requirements**

When you add a new human worker to a model evaluation job, they automatically receive an email inviting them to participate in the model evaluation job. When you add an *existing* human worker to a model evaluation job, you must notify and provide them with worker portal URL for the model evaluation job. The existing worker will not receive an automated email notification that they are added to the new model evaluation job.

- **a.** Using the **Select team** dropdown, specify either **Create a new work team** or the name of an existing work team.

- **b.** (Optional) **Number of workers per prompt** – Update the number of workers who evaluate each prompt. After the responses for each prompt have been reviewed by the number of workers you selected, the prompt and its responses will be taken out of circulation from the work team. The final results report will include all ratings from each worker.

- **c.** (Optional) **Existing worker email** – Choose this to copy an email template containing the worker portal URL.

- **d.** (Optional) **New worker email** – Choose this to view the email new workers receive automatically.

**Important**

Large language models are known to occasionally hallucinate and produce toxic or offensive content. Your workers may be shown toxic or offensive material during this evaluation. Ensure you take proper steps to train and notify them before they work on the evaluation. They can decline and release tasks or take breaks during the evaluation while accessing the human evaluation tool.

10. Then, choose **Next**.

11. On the **Provide instruction page** use the text editor to provide instructions for completing the task. You can preview the evaluation UI that your work team uses to evaluate the responses, including the metrics, rating methods, and your instructions. This preview is based on the configuration you have created for this job.
12. Then, choose Next.

13. On the Review and create page, you can view a summary of the options you've selected in the previous steps.

14. To start your model evaluation job, choose Create.

Once the job has successfully started, the status changes to In progress. When the job has finished, the status changes to Completed.

To learn how to evaluate, view, and download the results of your model evaluation job, see Model evaluation job results.

Model evaluation tasks

⚠️ Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

In a model evaluation job, an evaluation task is a task you want the model to perform based on information in your prompts.

You can choose one task type per model evaluation job. Use the following topics to learn more about each task type. Each topic also includes a list of available built-in datasets and their corresponding metrics that can be used only in automatic model evaluation jobs.

Topics

- General text generation
- Text summarization
- Question and answer
- Text classification
General text generation

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

General text generation is a task used by applications that include chatbots. The responses generated by a model to general questions are influenced by the correctness, relevance, and bias contained in the text used to train the model.

The following built-in datasets contain prompts that are well-suited for use in general text generation tasks.

Bias in Open-ended Language Generation Dataset (BOLD)

The Bias in Open-ended Language Generation Dataset (BOLD) is a dataset that evaluates fairness in general text generation, focusing on five domains: profession, gender, race, religious ideologies, and political ideologies. It contains 23,679 different text generation prompts.

RealToxicityPrompts

RealToxicityPrompts is a dataset that evaluates toxicity. It attempts to get the model to generate racist, sexist, or otherwise toxic language. This dataset contains 100,000 different text generation prompts.

T-Rex : A Large Scale Alignment of Natural Language with Knowledge Base Triples (TREX)

TREX is a dataset consisting of Knowledge Base Triples (KBTs) extracted from Wikipedia. KBTs are a type of data structure used in natural language processing (NLP) and knowledge representation. They consist of a subject, predicate, and object, where the subject and object are linked by a relation. An example of a Knowledge Base Triple (KBT) is "George Washington was the president of the United States". The subject is "George Washington", the predicate is "was the president of", and the object is "the United States".

WikiText2

WikiText2 is a HuggingFace dataset that contains prompts used in general text generation.
The following table summarizes the metrics calculated, and recommended built-in dataset that are available for automatic model evaluation jobs.

### Available built-in datasets for general text generation in Amazon Bedrock

<table>
<thead>
<tr>
<th>Task type</th>
<th>Metric</th>
<th>Built-in datasets</th>
<th>Computed metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>General text generation</td>
<td>Accuracy</td>
<td>TREX</td>
<td>Real world knowledge (RWK) score</td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>BOLD</td>
<td>Word error rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WikiText2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TREX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Toxicity</td>
<td>RealToxicityPrompts</td>
<td>Toxicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BOLD</td>
<td></td>
</tr>
</tbody>
</table>

For general text generation, there is a known system issue that prevents Cohere Command and Anthropic Claude models from completing a robustness evaluation successfully.

To learn more about how the computed metric for each built-in dataset is calculated, see [Model evaluation job results](#).

### Text summarization

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

Text summarization is used for tasks including creating summaries of news, legal documents, academic papers, content previews, and content curation. The ambiguity, coherence, bias, and
fluency of the text used to train the model as well as information loss, accuracy, relevance, or context mismatch can influence the quality of responses.

The following built-in dataset is supported for use with the task summarization task type.

**Gigaword**

The Gigaword dataset consists of news article headlines. This dataset is used in text summarization tasks.

The following table summarizes the metrics calculated, and recommended built-in dataset.

### Available built-in datasets for text summarization in Amazon Bedrock

<table>
<thead>
<tr>
<th>Task type</th>
<th>Metric</th>
<th>Built-in datasets</th>
<th>Computed metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text summarization</td>
<td>Accuracy</td>
<td>Gigaword</td>
<td>BERTScore</td>
</tr>
<tr>
<td></td>
<td>Toxicity</td>
<td>Gigaword</td>
<td>Toxicity</td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>Gigaword</td>
<td>BERTScore and deltaBERTScore</td>
</tr>
</tbody>
</table>

For text summarization, there is a known system issue that prevents Anthropic Claude models from completing a Robustness evaluation successfully.

To learn more about how the computed metric for each built-in dataset is calculated, see [Model evaluation job results](#).

### Question and answer

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.
Question and answer is used for tasks including generating automatic help-desk responses, information retrieval, and e-learning. If the text used to train the foundation model contains issues including incomplete or inaccurate data, sarcasm or irony, the quality of responses can deteriorate.

The following built-in datasets are recommended for use with the question and answer task type.

**BoolQ**

BoolQ is a dataset consisting of yes/no question and answer pairs. The prompt contains a short passage, and then a question about the passage. This dataset is recommended for use with question and answer task type.

**Natural Questions**

Natural questions is a dataset consisting of real user questions submitted to Google search.

**TriviaQA**

TriviaQA is a dataset that contains over 650K question-answer-evidence-triples. This dataset is used in question and answer tasks.

The following table summarizes the metrics calculated, and recommended built-in dataset.

**Available built-in datasets for the question and answer task type in Amazon Bedrock**

<table>
<thead>
<tr>
<th>Task type</th>
<th>Metric</th>
<th>Built-in datasets</th>
<th>Computed metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question and answer</td>
<td>Accuracy</td>
<td>BoolQ</td>
<td>NLP-F1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TriviaQA</td>
<td></td>
</tr>
<tr>
<td>Robustness</td>
<td></td>
<td>BoolQ</td>
<td>F1 and deltaF1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TriviaQA</td>
<td></td>
</tr>
<tr>
<td>Toxicity</td>
<td></td>
<td>BoolQ</td>
<td>Toxicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
</tbody>
</table>
For question and answer, there is a known system issue that prevents Anthropic Claude models from completing a robustness evaluation successfully.

To learn more about how the computed metric for each built-in dataset is calculated, see Model evaluation job results

Text classification

Note
Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

Text classification is used to categorize text into pre-defined categories. Applications that use text classification include content recommendation, spam detection, language identification and trend analysis on social media. Imbalanced classes, ambiguous data, noisy data, and bias in labeling are some issues that can cause errors in text classification.

The following built-in datasets are recommended for use with the text classification task type.

Women's E-Commerce Clothing Reviews

Women's E-Commerce Clothing Reviews is a dataset that contains clothing reviews written by customers. This dataset is used in text classification tasks.

The following table summarizes the metrics calculated, and recommended built-in datasets.

Available built-in datasets in Amazon Bedrock

<table>
<thead>
<tr>
<th>Task type</th>
<th>Metric</th>
<th>Built-in datasets</th>
<th>Computed metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text classification</td>
<td>Accuracy</td>
<td>Women's Ecommerce Clothing Reviews</td>
<td></td>
</tr>
</tbody>
</table>
### Task type | Metric | Built-in datasets | Computed metric
--- | --- | --- | ---

| Robustness | Women's Ecommerce Clothing Reviews | classification_accuracy_score and delta_classification_accuracy_score |

For text classification, there is a known system issue that prevents Anthropic Claude models from completing a Robustness evaluation successfully.

To learn more about how the computed metric for each built-in dataset is calculated, see [Model evaluation job results](#).

### Using prompt datasets in model evaluation jobs

#### Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

To create a model evaluation job you must specify a prompt dataset the model uses during inference. Amazon Bedrock provides built-in datasets that can be used in automatic model evaluations, or you can bring your own prompt dataset. For model evaluation jobs that use human workers you must use your own prompt dataset.

Use the following sections to learn more about available built-in prompt datasets and creating your custom prompt datasets.

To learn more about creating your first model evaluation job in Amazon Bedrock, see [Model evaluation](#).
Amazon Bedrock provides multiple built-in prompt datasets that you can use in an automatic model evaluation job. Each built-in dataset is based off an open-source dataset. We have downsampled each open-source dataset to include only 100 prompts.

All built-in prompt datasets use the format that Anthropic Claude models require. This means each prompt contains "Human: " at the start of each prompt, and "Assistant: " at the end. To avoid this scenario, you can use the original open source versions of the prompt datasets, and upload them as a custom dataset when you create a model evaluation job.

When you create an automatic model evaluation job and choose a Task type Amazon Bedrock provides you with a list of recommended metrics. For each metric, Amazon Bedrock also provides recommended built-in datasets. To learn more about available task types, see Model evaluation tasks.

**Bias in Open-ended Language Generation Dataset (BOLD)**

The Bias in Open-ended Language Generation Dataset (BOLD) is a dataset that evaluates fairness in general text generation, focusing on five domains: profession, gender, race, religious ideologies, and political ideologies. It contains 23,679 different text generation prompts.

**RealToxicityPrompts**

RealToxicityPrompts is a dataset that evaluates toxicity. It attempts to get the model to generate racist, sexist, or otherwise toxic language. This dataset contains 100,000 different text generation prompts.
T-Rex: A Large Scale Alignment of Natural Language with Knowledge Base Triples (TREX)

TREX is a dataset consisting of Knowledge Base Triples (KBTs) extracted from Wikipedia. KBTs are a type of data structure used in natural language processing (NLP) and knowledge representation. They consist of a subject, predicate, and object, where the subject and object are linked by a relation. An example of a Knowledge Base Triple (KBT) is "George Washington was the president of the United States". The subject is "George Washington", the predicate is "was the president of", and the object is "the United States".

WikiText2

WikiText2 is a HuggingFace dataset that contains prompts used in general text generation.

Gigaword

The Gigaword dataset consists of news article headlines. This dataset is used in text summarization tasks.

BoolQ

BoolQ is a dataset consisting of yes/no question and answer pairs. The prompt contains a short passage, and then a question about the passage. This dataset is recommended for use with question and answer task type.

Natural Questions

Natural question is a dataset consisting of real user questions submitted to Google search.

TriviaQA

TriviaQA is a dataset that contains over 650K question-answer-evidence-triples. This dataset is used in question and answer tasks.

Women's E-Commerce Clothing Reviews

Women's E-Commerce Clothing Reviews is a dataset that contains clothing reviews written by customers. This dataset is used in text classification tasks.

In the following table, you can see the list of available datasets grouped task type. To learn more about how automatic metrics are computed, see Automated model evaluation job report cards (console).
<table>
<thead>
<tr>
<th>Task type</th>
<th>Metric</th>
<th>Built-in datasets</th>
<th>Computed metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>General text generation</td>
<td>Accuracy</td>
<td>TREX</td>
<td>Real world knowledge (RWK) score</td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>BOLD</td>
<td>Word error rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WikiText2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>English Wikipedia</td>
<td></td>
</tr>
<tr>
<td>Text summarization</td>
<td>Accuracy</td>
<td>Gigaword</td>
<td>BERTScore</td>
</tr>
<tr>
<td></td>
<td>Toxicity</td>
<td>Gigaword</td>
<td>Toxicity</td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>Gigaword</td>
<td>BERTScore and deltaBERTScore</td>
</tr>
<tr>
<td>Question and answer</td>
<td>Accuracy</td>
<td>BoolQ</td>
<td>NLP-F1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TriviaQA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>BoolQ</td>
<td>F1 and deltaF1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TriviaQA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Toxicity</td>
<td>BoolQ</td>
<td>Toxicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NaturalQuestions</td>
<td></td>
</tr>
<tr>
<td>Task type</td>
<td>Metric</td>
<td>Built-in datasets</td>
<td>Computed metric</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------</td>
<td>-------------------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>Task classification</td>
<td>Accuracy</td>
<td>TriviaQA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Women's Ecommerce</td>
<td>Accuracy (Binary accuracy from classification_accuracy_score)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clothing Reviews</td>
<td></td>
</tr>
<tr>
<td></td>
<td>classification</td>
<td>Robustness</td>
<td>Women's Ecommerce</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clothing Reviews</td>
<td>classification_accuracy_score and delta_classification_accuracy_score</td>
</tr>
</tbody>
</table>

To learn more about the requirements for creating and examples of custom prompt datasets, see **Custom prompt dataset**.

**Custom prompt dataset**

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

You can use a custom prompt dataset in model evaluation jobs.

Custom prompt datasets must be stored in Amazon S3, and use the JSON line format and use the `.jsonl` file extension. When you upload the dataset to Amazon S3 make sure that you update the Cross Origin Resource Sharing (CORS) configuration on the S3 bucket. To learn more about the...
required CORS permissions, see Required Cross Origin Resource Sharing (CORS) permission on S3 buckets.

**Topics**

- Requirements for custom prompt datasets used in automatic model evaluation jobs
- Requirements for custom prompt datasets in model evaluation job that use human workers

**Requirements for custom prompt datasets used in automatic model evaluation jobs**

In automatic model evaluation jobs you can use a custom prompt dataset for each metric you select in the model evaluation job. Custom datasets use the JSON line format (.jsonl), and each line must be a valid JSON object. There can be up to 1000 prompts in your dataset per automatic evaluation job.

You must use the following keys in a custom dataset.

- **prompt** – required to indicate the input for the following tasks:
  - The prompt that your model should respond to, in general text generation.
  - The question that your model should answer in the question and answer task type.
  - The text that your model should summarize in text summarization task.
  - The text that your model should classify in classification tasks.
- **referenceResponse** – required to indicate the ground truth response against which your model is evaluated for the following tasks types:
  - The answer for all prompts in question and answer tasks.
  - The answer for all accuracy, and robustness evaluations.
- **category** – (optional) generates evaluation scores reported for each category.

As an example, accuracy requires both the question to ask and the answer to check the model response against. In this example, use the key `prompt` with the value contained in the question, and the key `referenceResponse` with the value contained in the answer as follows.

```
{
  "prompt": "Bobigny is the capital of",
  "referenceResponse": "Seine-Saint-Denis",
  "category": "Capitals"
}
```
The previous example is a single line of a JSON line input file that will be sent to your model as an inference request. Model will be invoked for every such record in your JSON line dataset. The following data input example is for a question answer task that uses an optional category key for evaluation.

⚠️ **Important**

After your last prompt in your custom dataset, the file must end with a newline.

```json
["referenceResponse":"Cantal","category":"Capitals","prompt":"Aurillac is the capital of"],
["referenceResponse":"Bamiyan Province","category":"Capitals","prompt":"Bamiyan city is the capital of"],
["referenceResponse":"Abkhazia","category":"Capitals","prompt":"Sokhumi is the capital of"],
## The file must end with a newline
```

To learn more about the format requirements for model evaluation jobs that use human workers, see [Requirements for custom prompt datasets in model evaluation job that use human workers](#).

### Anthropic Claude custom prompt dataset requirements

If you want to specify custom prompt dataset, and use an Anthropic Claude model in your automatic model evaluation job you must modify, you must modify your prompt keys to have the following structure.

```json
{
    "prompt": "Human: What is high intensity interval training? Assistant:",
    "category": "Fitness",
    "referenceResponse": "High-Intensity Interval Training (HIIT) is a cardiovascular exercise approach that involves short, intense bursts of exercise followed by brief recovery or rest periods."
}
```

These keywords, **Human:** and **Assistant:** are case-sensitive and must include the colon.

We recommend you test the sensitivity of including these keywords using Amazon Bedrock Playground.
Requirements for custom prompt datasets in model evaluation job that use human workers

In the JSON line format, each line is a valid JSON object. A prompt dataset can have a maximum of 1000 prompts per model evaluation job.

A valid prompt entry must contain the prompt key, both category and referenceResponse are optional. Use the category key to label your prompt with a specific category that you can use to filter the results when reviewing them in the model evaluation report card. Use the referenceResponse key to specify the ground truth response that your workers can reference during the evaluation.

In the worker UI, what you specify for prompt and referenceResponse are visible to your human workers.

The following is an example custom dataset that contains 6 inputs and uses the JSON line format.

⚠️ Important

After your last prompt in your custom dataset, the file must end with a newline.

```
{"prompt":"Provide the prompt you want the model to use during inference","category": "(Optional) Specify an optional category", "referenceResponse": "(Optional) Specify a ground truth response."}
{"prompt":"Provide the prompt you want the model to use during inference","category": "(Optional) Specify an optional category", "referenceResponse": "(Optional) Specify a ground truth response."}
{"prompt":"Provide the prompt you want the model to use during inference","category": "(Optional) Specify an optional category", "referenceResponse": "(Optional) Specify a ground truth response."}
{"prompt":"Provide the prompt you want the model to use during inference","category": "(Optional) Specify an optional category", "referenceResponse": "(Optional) Specify a ground truth response."}
{"prompt":"Provide the prompt you want the model to use during inference","category": "(Optional) Specify an optional category", "referenceResponse": "(Optional) Specify a ground truth response."}
```
The following example is a single entry expanded for clarity

```json
{
    "prompt": "What is high intensity interval training?",
    "category": "Fitness",
    "referenceResponse": "High-Intensity Interval Training (HIIT) is a cardiovascular exercise approach that involves short, intense bursts of exercise followed by brief recovery or rest periods."
}
```

**Anthropic Claude custom prompt dataset requirements**

If either model you select in your model evaluation job is an Anthropic Claude model, you must modify your prompt keys to have the following structure..

```json
{
    "prompt": "Human: What is high intensity interval training? Assistant:",
    "category": "Fitness",
    "referenceResponse": "High-Intensity Interval Training (HIIT) is a cardiovascular exercise approach that involves short, intense bursts of exercise followed by brief recovery or rest periods."
}
```

These keywords, **Human:** and **Assistant:** are case-sensitive and must include the colon.

Because these prompts are sent to the second model in your model evaluation job, we recommend you test the sensitivity of including these keywords on the second model using Amazon Bedrock Playground.

**Creating good worker instructions**

ℹ️ **Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.
Creating good instructions for your model evaluation jobs improves your worker's accuracy in completing their task. You can modify the default instructions that are provided in the console when creating a model evaluation job. The instructions are shown to the worker on the UI page where they complete their labeling task.

To help workers complete their assigned tasks, you can provide instructions in two places.

**Provide a good description for each evaluation and rating method**

The descriptions should provide a succinct explanation of the metrics selected. The description should expand on the metric, and make clear how you want workers to evaluate the selected rating method. To see examples of how each rating method is shown in the worker UI, see [Summary of available rating methods](#).

**Provide your workers overall evaluation instructions**

These instructions are shown on the same webpage where workers complete a task. You can use this space to provide high level direction for the model evaluation job, and to describe the ground truth responses if you've included them in your prompt dataset.

**Summary of available rating methods**

In each of the following sections, you can see an example of the rating methods your work team saw in the evaluation UI, and also how those results are saved in Amazon S3.

**Likert scale, comparison of multiple model outputs**

Human evaluators indicate their preference between the two responses from the model on a 5 point Likert scale according to your instructions. The results in the final report will be shown as a histogram of preference strength ratings from the evaluators over your whole dataset.

Make sure you define the important points of the 5 point scale in your instructions, so your evaluators know how to rate responses based on your expectations.
**Metric: Accuracy**

Response 1 is better than response 2

- Strongly prefer response 1
- Slightly prefer response 1
- Neither agree nor disagree
- Slightly prefer response 2
- Strongly prefer response 2

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "preferenceStrength" key value pair.

**Choice buttons (radio button)**

Choice buttons allow a human evaluator to indicate their one preferred response over another response. Evaluators indicate their preference between two responses according to your instructions with radio buttons. The results in the final report will be shown as a percentage of responses that workers preferred for each model. Be sure to explain your evaluation method clearly in the instructions.
Metric: Relevance

Which response do you prefer based on the metric?

- [ ] Response 1
- [ ] Response 2

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "preferenceRate" key value pair.

**Ordinal rank**

Ordinal rank allows a human evaluator to rank their preferred responses to a prompt in order starting at 1 according to your instructions. The results in the final report will be shown as a histogram of the rankings from the evaluators over the whole dataset. Be sure to define what a rank of 1 means in your instructions.
**Metric: Toxicty**

Input ranking for the responses. 1 is the best ranked response.

<table>
<thead>
<tr>
<th>Response 1</th>
<th>Input number</th>
</tr>
</thead>
</table>

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "preferenceRank" key value pair.

**Thumbs up/down**

Thumbs up/down allows a human evaluator to rate each response from a model as acceptable/unacceptable according to your instructions. The results in the final report will be shown as a percentage of the total number of ratings by evaluators that received a thumbs up rating for each model. You may use this rating method for an evaluation one or more models. If you use this in an evaluation that contains two models, a thumbs up/down will be presented to your work team for each model response and the final report will show the aggregated results for each model individually. Be sure to define what is acceptable (that is, what is a thumbs up rating) in your instructions.
**Metric: Friendliness**

Using the instructions, indicate whether response 1 was acceptable based on Friendliness.

👍 Yes  🚫 No

Using the instructions, indicate whether response 2 was acceptable based on Friendliness.

👍 Yes  🚫 No

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "approvalRate" key value pair.

**Likert scale, evaluation of a single model response**

Allows a human evaluator to indicate how strongly they approved of the model's response based on your instructions on a 5 point Likert scale. The results in the final report will be shown as a
histogram of the 5 point ratings from the evaluators over your whole dataset. You may use this for an evaluation containing one or more models. If you select this rating method in an evaluation that contains more than one model, a 5 point Likert scale will be presented to your work team for each model response and the final report will show the aggregated results for each model individually. Be sure to define the important points on the 5 point scale in your instructions so your evaluators know how to rate the responses according to your expectations.

**Metric: Harmlessness**

Using the instructions, rate the response on a scale of 1 to 5 for Harmlessness.

Rate response 1 on a scale of 1 to 5.

1 2 3 4 5

Rate response 2 on a scale of 1 to 5.

1 2 3 4 5

**JSON output**

The first child-key under evaluationResults is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "approvalStrength" key value pair.
Creating and managing work teams in Amazon Bedrock

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

In model evaluation jobs that use human workers you need to have a work team. A work team is a group of workers that you choose. These can be employees of your company or a group of subject-matter experts from your industry.

Worker notifications in Amazon Bedrock

- When you create a model evaluation job in Amazon Bedrock workers are notified of their assigned job only when you first add them to a work team
- If you delete a worker from a work team during model evaluation creation, they will lose access to all model evaluation jobs they have been assigned too.
- For any new model evaluation job that you assign to an existing human worker, you must notify them directly and provide them the URL to the worker portal. Workers must use their previously created login credentials for the worker portal. This worker portal is the same for all evaluation jobs in your AWS account per region.

In Amazon Bedrock you can create a new work team or manage an existing one while setting up a model evaluation job. When you create a work team in Amazon Bedrock you are adding workers to a Private workforce that is managed by Amazon SageMaker Ground Truth. Amazon SageMaker Ground Truth supports more advanced workforce management features. To learn more about managing your workforce in Amazon SageMaker Ground Truth, see Create and manage workforces.

You can delete workers from a work team while setting up a new model evaluation job. Otherwise, you must use either the Amazon Cognito console or the Amazon SageMaker Ground Truth console to manage work teams you've created in Amazon Bedrock.
If the IAM user, group, or role has the required permissions you will see existing private workforces and work teams you created in Amazon Cognito, Amazon SageMaker Ground Truth, or Amazon Augmented AI visible when you are creating a model evaluation job that uses human workers.

Amazon Bedrock supports a maximum of 50 workers per work team.

In the email addresses field, you can enter up to 50 email addresses at time. To add more workers to your model evaluation job use the Amazon Cognito console or the Ground Truth console. The addresses must be separated by a comma. You should include your own email address so that you are part of the workforce and can see the labeling tasks.

Model evaluation job results

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

The results of a model evaluation job are available via the Amazon Bedrock console or by downloading the results from the Amazon S3 bucket you specified when the job was created.

Once your job status has changed to Ready, you can find the S3 bucket you specified when creating the job. To do so, go to the Model evaluations table on the Model evaluation home page and choose it.

Use the following topics to learn how to access model evaluation reports, and how results of a model evaluation job are saved in Amazon S3.

Topics

- Automated model evaluation job report cards (console)
- Human model evaluation job report cards (console)
- Understanding how the results of your model evaluation job that are saved in Amazon S3
Automated model evaluation job report cards (console)

Note
Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

In your model evaluation report card, you will see the total number of prompts in the dataset you provided or selected, and how many of those prompts received responses. If the number of responses is less than the number of input prompts, make sure to check the data output file in your Amazon S3 bucket. It is possible that the prompt caused an error with the model and there was no inference retrieved. Only responses from the model will be used in metric calculations.

Use the following procedure to review an automatic model evaluation job on the Amazon Bedrock console.

1. Open the Amazon Bedrock console.
2. From the navigation pane, choose Model evaluation.
3. Next, in the Model evaluations table find the name of the automated model evaluation job you want to review. Then, choose it.

In all semantic robustness related metrics, Amazon Bedrock perturbs prompts in the following ways: convert text to all lower cases, keyboard typos, converting numbers to words, random changes to upper case and random addition/deletion of whitespaces.

After you open the model evaluation report you can view the summarized metrics, and the Job configuration summary of the job.

For each metric and prompt dataset specified when the job was created you see a card, and a value for each dataset specified for that metric. How this value is calculated changes based on the task type and the metrics you selected.

How each available metric is calculated when applied to the general text generation task type

- **Accuracy**: For this metric, the value is calculated using real world knowledge score (RWK score). RWK score examines the model’s ability to encode factual knowledge about the real world. A high RWK score indicates that your model is being accurate.
• **Robustness**: For this metric, the value is calculated using semantic robustness. Which is calculated using word error rate. Semantic robustness measures how much the model output changes as a result of minor, semantic preserving perturbations, in the input. Robustness to such perturbations is a desirable property, and thus a low semantic robustness score indicated your model is performing well.

The perturbation types we will consider are: convert text to all lower cases, keyboard typos, converting numbers to words, random changes to upper case and random addition/deletion of whitespaces. Each prompt in your dataset is perturbed approximately 5 times. Then, each perturbed response is sent for inference, and used to calculate robustness scores automatically.

• **Toxicity**: For this metric, the value is calculated using toxicity from the detoxify algorithm. A low toxicity value indicates that your selected model is not producing large amounts of toxic content. To learn more about the detoxify algorithm and see how toxicity is calculated, see the detoxify algorithm on GitHub.

How each available metric is calculated when applied to the text summarization task type

• **Accuracy**: For this metric, the value is calculated using BERT Score. BERT Score is calculated using pre-trained contextual embeddings from BERT models. It matches words in candidate and reference sentences by cosine similarity. The displayed value is the F1 score. A low F1 score indicates that the model struggles to be both precise and capture the true positive (recall).

• **Robustness**: For this metric, the value calculated is a percentage. It calculated by taking \((\text{Delta BERTScore} / \text{BERTScore}) \times 100\). Delta BERTScore is the difference in BERT Scores between a perturbed prompt and the original prompt in your dataset. Each prompt in your dataset is perturbed approximately 5 times. Then, each perturbed response is sent for inference, and used to calculate robustness scores automatically. A lower score indicates the selected model is more robust.

• **Toxicity**: For this metric, the value is calculated using toxicity from the detoxify algorithm. A low toxicity value indicates that your selected model is not producing large amounts of toxic content. To learn more about the detoxify algorithm and see how toxicity is calculated, see the detoxify algorithm on GitHub.

How each available metric is calculated when applied to the question and answer task type

• **Accuracy**: For this metric, the value calculated is F1 score. F1 score is calculated by dividing the precision score (the ratio of correct predictions to all predictions) by the recall score (the ratio of
correct predictions to the total number of relevant predictions). The F1 score ranges from 0 to 1, with higher values indicating better performance.

- **Robustness**: For this metric, the value calculated is a percentage. It is calculated by taking \((\text{Delta F1} / \text{F1}) \times 100\). Delta F1 is the difference in BERT Scores between a perturbed prompt and the original prompt in your dataset. Each prompt in your dataset is perturbed approximately 5 times. Then, each perturbed response is sent for inference, and used to calculate robustness scores automatically. A lower score indicates the selected model is more robust.

- **Toxicity**: For this metric, the value is calculated using toxicity from the detoxify algorithm. A low toxicity value indicates that your selected model is not producing large amounts of toxic content. To learn more about the detoxify algorithm and see how toxicity is calculated, see the [detoxify algorithm](https://github.com/milvus-io/detoxify) on GitHub.

### How each available metric is calculated when applied to the text classification task type

- **Accuracy**: For this metric, the value calculated is accuracy. Accuracy is a score that compares the predicted class to its ground truth label. A higher accuracy indicates that your model is correctly classifying text based on the ground truth label provided.

- **Robustness**: For this metric, the value calculated is a percentage. It is calculated by taking \((\text{delta balanced classification score} / \text{classification accuracy score}) \times 100\). Delta balanced classification score is the difference between the classification accuracy Score of the perturbed prompt and the original input prompt. Each prompt in your dataset is perturbed approximately 5 times. Then, each perturbed response is sent for inference, and used to calculate robustness scores automatically. A lower score indicates the selected model is more robust.

### Human model evaluation job report cards (console)

---

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

---

In your model evaluation report card, you will see the total number of prompts in the dataset you provided or selected, and how many of those prompts received responses. If the number of responses is less than the number of input prompts times the number of workers per prompt you
configured in the job (either 1, 2 or 3), make sure to check the data output file in your Amazon S3 bucket. It is possible that the prompt caused an error with the model and there was no inference retrieved. Also, one or more of your workers could have declined to evaluate a model output response. Only responses from the human workers will be used in metric calculations.

Use the following procedure to open up a model evaluation that used human workers on the Amazon Bedrock console.

1. Open the Amazon Bedrock console.
2. From the navigation pane, choose Model evaluation.
3. Next, in the Model evaluations table find the name of the model evaluation job you want to review. Then, choose it.

The model evaluation report provides insights about the data collected during a human evaluation job using report cards. Each report card shows the metric, description, and rating method, alongside a data visualization that represents the data collected for the given metric.

In each of the following sections, you can see an examples of the 5 possible rating methods your work team saw in the evaluation UI. The examples also show what key value pair is used to save the results in Amazon S3.

**Likert scale, comparison of multiple model outputs**

Human evaluators indicate their preference between the two responses from the model on a 5 point Likert scale according to your instructions. The results in the final report will be shown as a histogram of preference strength ratings from the evaluators over your whole dataset.

Make sure you define the important points of the 5 point scale in your instructions, so your evaluators know how to rate responses based on your expectations.
### Metric: Accuracy

Response 1 is better than response 2

- [ ] Strongly prefer response 1
- [ ] Slightly prefer response 1
- [ ] Neither agree nor disagree
- [ ] Slightly prefer response 2
- [ ] Strongly prefer response 2

#### JSON output

The first child-key under evaluationResults is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "preferenceStrength" key value pair.

#### Choice buttons (radio button)

Choice buttons allow a human evaluator to indicate their one preferred response over another response. Evaluators indicate their preference between two responses according to your instructions with radio buttons. The results in the final report will be shown as a percentage of responses that workers preferred for each model. Be sure to explain your evaluation method clearly in the instructions.
\section*{Metric: Relevance}

Which response do you prefer based on the metric?

- Response 1
- Response 2

\section*{JSON output}

The first child-key under \texttt{evaluationResults} is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the \texttt{"evaluationResults": "preferenceRate"} key value pair.

\section*{Ordinal rank}

Ordinal rank allows a human evaluator to rank their preferred responses to a prompt in order starting at 1 according to your instructions. The results in the final report will be shown as a histogram of the rankings from the evaluators over the whole dataset. Be sure to define what a rank of 1 means in your instructions. This data type is called Preference Rank.
The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "preferenceRank" key value pair.

**Thumbs up/down**

Thumbs up/down allows a human evaluator to rate each response from a model as acceptable/unacceptable according to your instructions. The results in the final report will be shown as a percentage of the total number of ratings by evaluators that received a thumbs up rating for each model. You may use this rating method for a model evaluation job that contains one or more models. If you use this in an evaluation that contains two models, a thumbs up/down will be presented to your work team for each model response and the final report will show the aggregated results for each model individually. Be sure to define what is acceptable (that is, what is a thumbs up rating) in your instructions.
**Metric: Friendliness**

Using the instructions, indicate whether response 1 was acceptable based on Friendliness.

👉 Yes  👇 No

Using the instructions, indicate whether response 2 was acceptable based on Friendliness.

👉 Yes  👇 No

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "approvalRate" key value pair.

**Likert scale, evaluation of a single model response**

Allows a human evaluator to indicate how strongly they approved of the model's response based on your instructions on a 5 point Likert scale. The results in the final report will be shown as a
histogram of the 5 point ratings from the evaluators over your whole dataset. You may use this for an evaluation containing one or more models. If you select this rating method in an evaluation that contains more than one model, a 5 point Likert scale will be presented to your work team for each model response and the final report will show the aggregated results for each model individually. Be sure to define the important points on the 5 point scale in your instructions so your evaluators know how to rate the responses according to your expectations.

**Metric: Harmlessness**

Using the instructions, rate the response on a scale of 1 to 5 for Harmlessness.

Rate response 1 on a scale of 1 to 5.
[ ] 1 [ ] 2 [ ] 3 [ ] 4 [ ] 5

Rate response 2 on a scale of 1 to 5.
[ ] 1 [ ] 2 [ ] 3 [ ] 4 [ ] 5

**JSON output**

The first child-key under `evaluationResults` is where the selected rating method is returned. In the output file saved to your Amazon S3 bucket, the results from each worker are saved to the "evaluationResults": "approvalStrength" key value pair.
Understanding how the results of your model evaluation job that are saved in Amazon S3

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

The output from a model evaluation job is saved in the Amazon S3 bucket you specified when you created the model evaluation job. Results of model evaluation jobs are saved as JSON line files (.jsonl).

The results from the model evaluation job is saved in the S3 bucket you specified as follows.

- For model evaluation jobs that use human workers:
  
  `s3://user-specified-model-evaluation-job-output-bucket/human-loop-name/output.jsonl`

- For automatic model evaluation jobs:
  
  `s3://user-specified-model-evaluation-job-output-bucket/output/system-generated-id-tag/datasets`

The following topics describe how the results from automated and human worker based model evaluation job are saved in Amazon S3.

**Output data from automated model evaluation jobs**

The results of the automated evaluation job are stored in the **datasets** directory when job status changes to **Completed**.

For each metric and corresponding prompt dataset you selected when the model evaluation job was created, a JSON line file is generated in the **datasets** directory. The file uses the following naming convention **metric_input-dataset.jsonl**.

Each result from the model evaluation job starts with the **automatedEvaluationResult** key. The first child key **scores** contains the metrics you selected in the Amazon Bedrock console. In this
example, only a single metric Accuracy was selected. It also contains result which is calculated value for the selected metric. To learn more about what specific value is calculated, see Automated model evaluation job report cards (console).

The second key, inputRecord is a copy of what you provided in your input prompt dataset.

The third key, modelInvocations contains a list of JSON objects which contains the ARN of the model selected when you created the model evaluation job. It also contains the full response from the model based on the prompt provided.

The following is an example output for the text summarization task type with only one metric, accuracy, selected.

```
{
  "automatedEvaluationResult": {
    "scores": [{
      "metricName": "Accuracy",
      "result": 0.31920555233955383
    }]
  },
  "inputRecord": {
    "prompt": "Human: two bodies have been recovered from a flooded coal mine in Yiyang City, central China's Henan province, as of #:# thursday morning, rescuers said. Assistant:",
    "referenceResponse": "two bodies recovered from flooded coal mine in c. china"
  },
  "modelInvocations": [{
    "modelIdentifier": "arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-v2",
    "responseText": "Here are the key points I gathered from your news summary:
- Two bodies have been recovered from a flooded coal mine in Yiyang City, in China's Henan province. This happened as of Thursday morning, according to rescuers.
- The location is central China's Henan province. Yiyang City is where the flooded coal mine is located. The coal mine was flooded, leading to the two deaths and need for rescue/recovery efforts. This news is developing as of Thursday morning, so the death toll and rescue details may still be unfolding.
- Please let me know if I have accurately summarized the key information or if you would like me to expand on any part of it."
  }]
}
```
Output data from model evaluation jobs that use human workers.

When a model evaluation job has completed, you see the following parameters in the output data returned from human review tasks.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value Type</th>
<th>Example Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>flowDefinitionArn</td>
<td>String</td>
<td><code>arn:aws:sagemaker:us-west-2: 11112222333 :flow-definition/flow-definition-name</code></td>
<td>The Amazon Resource Number (ARN) of the human review workflow (flow definition) used to create the human loop.</td>
</tr>
</tbody>
</table>
| humanAnswers          | List of JSON object | "answerContent": {  
  "evaluationResults": [{  
    "approvalRate": [{  
      "metric": "Relevance",
      "modelResponseId": "0",
      "result": false
    }]
  ]
}" | A list of JSON objects that contain worker responses in answerContent.                                                                      |
<p>|                       | String     | system-generated-hash                                                          | A system generated Amazon S3 output                                                                                                          |</p>
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value Type</th>
<th>Example Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>humanLoop Name</td>
<td></td>
<td></td>
<td>40-character hex string.</td>
</tr>
<tr>
<td>inputRecord</td>
<td>JSON object</td>
<td>&quot;inputRecord&quot;: {</td>
<td>A JSON object that contains an entry prompt from the input dataset.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;prompt&quot;: &quot;What does vitamin C serum do for skin?&quot;,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;category&quot;: &quot;Skincare&quot;,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;referenceResponse&quot;: &quot;Vitamin C serum offers a range of benefits for the skin. Firstly, it acts....&quot;</td>
<td></td>
</tr>
<tr>
<td>modelInvocations</td>
<td>List of JSON</td>
<td>&quot;modelResponses&quot;: [{</td>
<td>The individual responses from the models.</td>
</tr>
<tr>
<td></td>
<td>object</td>
<td>&quot;modelIdentifier&quot;: &quot;arn:aws:bedrock: us-east-1::foundation-model/anthropic.claude-v2&quot;,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;responseText&quot;: &quot;the-models-response-to-the-prompt&quot;</td>
<td></td>
</tr>
<tr>
<td>inputRecord</td>
<td>List of JSON</td>
<td>See the following code sample</td>
<td>The input content sent to SageMaker in the request to StartHumanLoop.</td>
</tr>
<tr>
<td></td>
<td>object</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following is an example of output data from a model evaluation job.
{
    "output": [{
        "humanAnswers": [{
            "acceptanceTime": "2023-11-09T19:17:43.107Z",
            "answerContent": {
                "evaluationResults": {
                    "approvalRate": [{
                        "metric": "Coherence",
                        "metricName": "Coherence",
                        "modelResponseId": "0",
                        "result": false
                    }, {
                        "metric": "Accuracy",
                        "metricName": "Accuracy",
                        "modelResponseId": "0",
                        "result": true
                    }],
                    "approvalStrength": [{
                        "metric": "Toxicity",
                        "metricName": "Toxicity",
                        "modelResponseId": "0",
                        "result": 1
                    }]
                }
            },
            "submissionTime": "2023-11-09T19:17:52.101Z",
            "timeSpentInSeconds": 8.994,
            "workerId": "444455556666",
            "workerMetadata": {
                "identityData": {
                    "identityProviderType": "Cognito",
                    "sub": "c6aa8eb7-9944-42e9-a6b9-
                }
            }
        }],
        "timeSpentInHours": 0.11458333333333333
    }]
}
"humanLoopName": "b3b1c64a2166e001e094123456789012",
"inputRecord": {
  "category": "Fitness",
  "prompt": "Human: What is High-Intensity Interval Training? \n\nAssistant:"
},
"referenceResponse": "High-Intensity Interval Training (HIIT) is a cardiovascular exercise approach that involves short, intense bursts of exercise followed by brief recovery or rest periods. In HIIT, you perform exercises at near-maximum effort during the high-intensity intervals, which can include activities like sprinting, cycling, jumping, or body weight exercises such as burpees or push-ups. These intense intervals are interspersed with short recovery periods to lower your heart rate before the next high-intensity bout. HIIT workouts usually consist of multiple cycles of these intervals, and the number of cycles and the length of intervals can vary. HIIT is known for its time efficiency, making it an excellent choice for those with busy schedules. It offers both aerobic and anaerobic benefits, improving cardiovascular fitness, burning calories, and promoting fat loss, all while boosting metabolism for continued calorie burning post-workout.

"modelInvocations": [
  "modelIdentifier": "arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-v2",
  "responseText": "High-intensity interval training (HIIT) is a form of exercise that involves short bursts of intense activity alternated with recovery periods. Here are some key features of HIIT:
- Alternates short periods of intense anaerobic exercise with less intense recovery periods. The high-intensity periods are typically done at 80-95% of maximum heart rate.
- The duration of the intense periods can range from 30 seconds to 4 minutes.
- The total HIIT session typically lasts between 10-30 minutes, including warm-up, recovery periods, and cool down.
- Uses exercises like sprints, cycling, rowing, jumping rope, stair climbing, swimming, etc. Movements that get your heart rate up quickly.
- The short, intense intervals help improve cardiovascular fitness by increasing heart rate and oxygen consumption.
- HIIT burns more calories and fat than steady-state cardio in a shorter time period. It also helps build muscle strength and endurance.
- Considered an efficient and effective form of exercise for fat loss and improving aerobic power. But it requires motivation to push yourself during the intense intervals.
- Not suitable for beginners due to the high-intensity. Start with steady-state cardio and build up endurance before trying HIIT.

In summary, HIIT intersperses intense bursts of"
}]
The following table explains how the Rating method you selected for each metric in the Amazon Bedrock console is returned in your Amazon S3 bucket. The first child-key under evaluationResults is how the Rating method is returned.

### How rating methods selected in the Amazon Bedrock console are saved in Amazon S3

<table>
<thead>
<tr>
<th>Rating method selected</th>
<th>Saved in Amazon S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert scale - Individual</td>
<td>approvalStrength</td>
</tr>
<tr>
<td>Likert scale - Comparison</td>
<td>preferenceStrength</td>
</tr>
<tr>
<td>Choice buttons</td>
<td>preferenceRate</td>
</tr>
<tr>
<td>Ordinal rank</td>
<td>preferenceRank</td>
</tr>
<tr>
<td>Thumbs up/down</td>
<td>approvalRate</td>
</tr>
</tbody>
</table>

### Required IAM permissions and service roles to create a model evaluation job

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

**Persona: IAM Administrator**

A user who can add or remove IAM policies, and create new IAM roles.
The following topics explain the AWS Identity and Access Management permissions required to create a model evaluation job using the Amazon Bedrock console, the service role requirements, and the required Cross Origin Resource Sharing (CORS) permissions.

Topics

- Required permissions to create a model evaluation job using the Amazon Bedrock console
- Service role requirements for model evaluation jobs
- Required Cross Origin Resource Sharing (CORS) permission on S3 buckets
- Data encryption for model evaluation jobs

Required permissions to create a model evaluation job using the Amazon Bedrock console

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

The IAM permissions required to create a model evaluation job are different for automatic model evaluation jobs or model evaluation jobs that uses human workers.

Both automatic and human worker based model evaluation jobs require access to Amazon S3 and Amazon Bedrock. To create human-based model evaluation jobs, you need additional permissions from Amazon Cognito and Amazon SageMaker.

To learn more about the required service roles for creating automatic and human-based model evaluation jobs, see Service role requirements for model evaluation jobs

Required permissions to create an automatic model evaluation job

The following policy contains the minimum set of IAM actions and resource in Amazon Bedrock and Amazon S3 required to create an *automatic* model evaluation job.

```json
{
    "Version": "2012-10-17",
```
"Statement": [
    {
        "Sid": "Bedrock Console",
        "Effect": "Allow",
        "Action": [
            "bedrock:CreateModelEvaluationJob",
            "bedrock:GetModelEvaluationJob",
            "bedrock:ListModelEvaluationJobs",
            "bedrock:GetCustomModel",
            "bedrock:ListCustomModels",
            "bedrock:CreateProvisionedModelThroughput",
            "bedrock:UpdateProvisionedModelThroughput",
            "bedrock:GetProvisionedModelThroughput",
            "bedrock:ListProvisionedModelThroughputs",
            "bedrock:ListTagsForResource",
            "bedrock:UntagResource",
            "bedrock:TagResource"
        ],
        "Resource": "*"
    },
    {
        "Sid": "Allow Console S3 Access For Model Evaluation",
        "Effect": "Allow",
        "Action": [
            "s3:GetObject",
            "s3:GetBucketCORS",
            "s3:ListBucket",
            "s3:ListBucketVersions",
            "s3:GetBucketLocation"
        ],
        "Resource": "*"
    }
]

**Required permissions to create a human-based model evaluation job**

To create a model evaluation job that uses human workers from the Amazon Bedrock console you need to have additional permissions added to your user, group, or role.

The following policy contains the minimum set of IAM actions and resources required from Amazon Cognito and Amazon SageMaker to create an human based model evaluation job. You must append this policy to the **base policy requirements for an automatic job**.
Service role requirements for model evaluation jobs

Note

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

To create a model evaluation job, you must specify a service role.

A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Creating a role to delegate permissions to an AWS service in the IAM User Guide.

The required IAM permissions are different for automatic or human based model evaluation jobs. Use the following sections to learn more about the required Amazon Bedrock, Amazon SageMaker, and Amazon S3 IAM actions, service principals, and resources.

Each of the following sections describe what permission are needed based on the type of model evaluation job you want to run.

Topics

- Service role requirements for automatic model evaluation jobs
- Service role requirements for model evaluation jobs that use human evaluators

Service role requirements for automatic model evaluation jobs

To create an automatic model evaluation job, you must specify a service role. The policy you attach grants Amazon Bedrock access to resources in your account, and allows Amazon Bedrock to invoke the selected model on your behalf.

You must also attach a trust policy that defines Amazon Bedrock as the service principal using bedrock.amazonaws.com. Each of the following policy examples shows you the exact IAM actions that are required based on each service invoked in an automatic model evaluation job.

To create a custom service role, see Creating a role that uses a custom trust policy in the IAM User Guide.
Required Amazon S3 IAM actions

The following policy example grants access to the S3 buckets where your model evaluation results are saved, and (optionally) access to any custom prompt datasets you have specified.

```json
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Sid": "AllowAccessToCustomDatasets",
         "Effect": "Allow",
         "Action": [
            "s3:GetObject",
            "s3:ListBucket"
         ],
         "Resource": [
            "arn:aws:s3:::my_customdataset1_bucket",
            "arn:aws:s3:::my_customdataset1_bucket/myfolder",
            "arn:aws:s3:::my_customdataset2_bucket",
            "arn:aws:s3:::my_customdataset2_bucket/myfolder"
         ]
      },
      {
         "Sid": "AllowAccessToOutputBucket",
         "Effect": "Allow",
         "Action": [
            "s3:GetObject",
            "s3:ListBucket",
            "s3:PutObject",
            "s3:GetBucketLocation",
            "s3:AbortMultipartUpload",
            "s3:ListBucketMultipartUploads"
         ],
         "Resource": [
            "arn:aws:s3:::my_output_bucket",
            "arn:aws:s3:::my_output_bucket/myfolder"
         ]
      }
   ]
}

Required Amazon Bedrock IAM actions
You also need to create a policy that allows Amazon Bedrock to invoke the model you plan to specify in the automatic model evaluation job. To learn more about managing access to Amazon Bedrock models, see Model access.

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowSpecificModels",
      "Effect": "Allow",
      "Action": [
        "bedrock:InvokeModel",
        "bedrock:InvokeModelWithResponseStream"
      ],
      "Resource": [
        "arn:aws:bedrock:region::foundation-model/model-id-of-foundational-model"
      ]
    }
  ]
}
```

Optional AWS Key Management Service IAM actions

If you used AWS Key Management Service to encrypt the S3 buckets that hold either your custom prompt dataset or where your result will be saved you need to add the following AWS KMS policy. The service role needs permissions to use your customer managed key.

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowUseOfKmsKey",
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt",
        "kms:GenerateDataKey"
      ],
      "Resource": "arn:aws:kms:us-west-1:111122223333:key/key-id"
    }
  ]
}
```
Service principal requirements

You must also specify a trust policy that defines Amazon Bedrock as the service principal. This allows Amazon Bedrock to assume the role. The wildcard (*) model evaluation job ARN is required so that Amazon Bedrock can create model evaluation jobs in your AWS account.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowBedrockToAssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "111122223333"
                },
                "ArnEquals": {
                }
            }
        }
    ]
}
```

Service role requirements for model evaluation jobs that use human evaluators

To create a model evaluation job that uses human evaluators, you must specify two service roles.

The following lists summarize the IAM policy requirements for each required service role that must be specified in the Amazon Bedrock console.

Summary of IAM policy requirements for the Amazon Bedrock service role

- You must attach a trust policy which defines Amazon Bedrock as the service principal.
- You must allow Amazon Bedrock to invoke the selected models on your behalf.
- You must allow Amazon Bedrock to access the S3 bucket that holds your prompt dataset and the S3 bucket where you want the results saved.
- You must allow Amazon Bedrock to create the required human loop resources in your account.
• (Optional) You must allow Amazon Bedrock to decrypt your KMS key if you've encrypted your prompt dataset bucket or the Amazon S3 bucket where you want the results saved.

Summary of IAM policy requirements for the Amazon SageMaker service role

• You must attach a trust policy which defines SageMaker as the service principal.
• You must allow SageMaker to access the S3 bucket that holds your prompt dataset and the S3 bucket where you want the results saved.
• (Optional) You must allow SageMaker to use your customer managed keys if you've encrypted your prompt dataset bucket or the location where you wanted the results.

To create a custom service role, see Creating a role that uses a custom trust policy in the IAM User Guide.

Required Amazon S3 IAM actions

The following policy example grants access to the S3 buckets where your model evaluation results are saved, and access to the custom prompt dataset you have specified. You need to attach this policy to both the SageMaker service role and the Amazon Bedrock service role.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAccessToCustomDatasets",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::custom-prompt-dataset"
            ]
        },
        {
            "Sid": "AllowAccessToOutputBucket",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::output-bucket"
            ]
        }
    ]
}
```
Required Amazon Bedrock IAM actions

You also need to create a policy that allows Amazon Bedrock to invoke the model you plan to specify in the automatic model evaluation job. You must attach this policy to the Amazon Bedrock service role.

```json
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Sid": "AllowSpecificModels",
         "Effect": "Allow",
         "Action": [
            "bedrock:InvokeModel",
            "bedrock:InvokeModelWithResponseStream"
         ],
         "Resource": [
            "arn:aws:bedrock:us-west-1::foundation-model/model-id-of-foundational-model",
         ]
      }
   ]
}
```

Required Amazon Augmented AI IAM actions

You also need to create a policy that allows Amazon Bedrock to create resources related to human-based model evaluation jobs. Because Amazon Bedrock creates the needed resources to start the model evaluation job, you must use "Resource": "*". You must attach this policy to the Amazon Bedrock service role.

```json
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Sid": "AllowSpecificModels",
         "Effect": "Allow",
         "Action": [
            "bedrock:InvokeModel",
            "bedrock:InvokeModelWithResponseStream"
         ],
         "Resource": [
            "arn:aws:bedrock:us-west-1::foundation-model/model-id-of-foundational-model",
         ]
      }
   ]
}
```
Optional AWS Key Management Service IAM actions

If you used AWS Key Management Service to encrypt the S3 buckets that hold either your custom prompt dataset or where your result will be saved you need to add the following AWS KMS policy. The service role needs permissions to use your customer managed key. You must attach this policy to the SageMaker service role and the Amazon Bedrock service role.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAccessToKMS",
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "kms:GenerateDataKey"
            ],
            "Resource": [
                "arn:aws:kms:us-east-1:111122223333:key/key id-"
            ],
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": [
                        "s3.region.amazonaws.com"
                    ]
                }
            }
        }
    ]
}
```
Service principal requirements (Amazon Bedrock)

You must also specify a trust policy that defines Amazon Bedrock as the service principal. This allows Amazon Bedrock to assume the role.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowBedrockToAssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "111122223333"
                },
                "ArnEquals": {
                }
            }
        }
    ]
}
```

Service principal requirements (SageMaker)

You must also specify a trust policy that defines Amazon Bedrock as the service principal. This allows SageMaker to assume the role.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowSageMakerToAssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": "sagemaker.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "111122223333"
                }
            }
        }
    ]
}
```
Required Cross Origin Resource Sharing (CORS) permission on S3 buckets

**Note**

Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

For custom prompt datasets, you must specify a CORS configuration on the S3 bucket. A CORS configuration is a document that defines rules that identify the origins that you will allow to access your bucket, the operations (HTTP methods) supported for each origin, and other operation-specific information. To learn more about setting the required CORS configuration using the S3 console, see [Configuring cross-origin resource sharing (CORS)](https://docs.aws.amazon.com/AmazonS3/latest/user-guide/configuring-cors.html) in the *Amazon S3 User Guide*.

The following is the minimal required CORS configuration for S3 buckets.

```json
[
  {
    "AllowedHeaders": [
      "*
    ],
    "AllowedMethods": [
      "GET",
      "PUT",
      "POST",
      "DELETE"
    ],
    "AllowedOrigins": [
      "*
    ]
  }
]
```
Data encryption for model evaluation jobs

Note
Model evaluation is in preview release for Amazon Bedrock and is subject to change. To use model evaluation jobs, you must be in either US East (N. Virginia) Region or US West (Oregon) Region.

Customer data is always encrypted, both in transit and at rest, for Amazon Bedrock model evaluation jobs. Model evaluation jobs support using customer managed keys on data that is used as either an input or output for a model evaluation job in Amazon Bedrock. If you do not specify a customer managed key when you create the model evaluation job, the default AWS-managed key for Amazon S3 is used for encryption. Amazon Bedrock stores intermediary customer data, such as your prompt datasets, during a model evaluation job, encrypting it at rest using a service-managed KMS key. This intermediary data is permanently deleted when the model evaluation job is completed.
Knowledge bases for Amazon Bedrock

Knowledge bases for Amazon Bedrock provides you the capability of amassing data sources into a repository of information. With knowledge bases, you can easily build an application that takes advantage of retrieval augmented generation (RAG), a technique in which the retrieval of information from data sources augments the generation of model responses. Once set up, you can take advantage of a knowledge base in the following ways.

- Configure your RAG application to use the RetrieveAndGenerate API to query your knowledge base and generate responses from the information it retrieves.
- Associate your knowledge base with an agent (for more information, see Agents for Amazon Bedrock) to add RAG capability to the agent by helping it reason through the steps it can take to help end users.
- Create a custom orchestration flow in your application by using the Retrieve API to retrieve information directly from the knowledge base.

A knowledge base can be used not only to answer user queries, but also to augment prompts provided to foundation models by providing context to the prompt. Knowledge base responses also come with citations, such that users can find further information by looking up the exact text that a response is based on and also check that the response makes sense and is factually correct.

You take the following steps to set up and use your knowledge base.

1. Gather source documents to add to your knowledge base.
2. (Optional) Create a metadata file for each source document to allow for filtering of results during knowledge base query.
3. Upload your data to an Amazon S3 bucket.
4. (Optional) Set up a vector index in a supported vector store to index your data. You can skip this step if you plan to use the Amazon Bedrock console to create an Amazon OpenSearch Serverless vector database for you.
5. Create and configure your knowledge base.
6. Ingest your data by generating embeddings with a foundation model and storing them in a supported vector store.
7. Set up your application or agent to query the knowledge base and return augmented responses.
How it works

Knowledge bases for Amazon Bedrock help you take advantage of Retrieval Augmented Generation (RAG), a popular technique that involves drawing information from a data store to augment the responses generated by Large Language Models (LLMs). When you set up a knowledge base with your data sources, your application can query the knowledge base to return information to answer the query either with direct quotations from sources or with natural responses generated from the query results.

With knowledge bases, you can build applications that are enriched by the context that is received from querying a knowledge base. It enables a faster time to market by abstracting from the heavy lifting of building pipelines and providing you an out-of-the-box RAG solution to reduce the build time for your application. Adding a knowledge base also increases cost-effectiveness by removing the need to continually train your model to be able to leverage your private data.

The following diagrams illustrate schematically how RAG is carried out. Knowledge base simplifies the setup and implementation of RAG by automating several steps in this process.

Pre-processing data

To enable effective retrieval from private data, a common practice is to first split the documents into manageable chunks for efficient retrieval. The chunks are then converted to embeddings and written to a vector index, while maintaining a mapping to the original document. These embeddings are used to determine semantic similarity between queries and text from the data sources. The following image illustrates pre-processing of data for the vector database.
Runtime execution

At runtime, an embedding model is used to convert the user's query to a vector. The vector index is then queried to find chunks that are semantically similar to the user's query by comparing document vectors to the user query vector. In the final step, the user prompt is augmented with the additional context from the chunks that are retrieved from the vector index. The prompt alongside the additional context is then sent to the model to generate a response for the user. The following image illustrates how RAG operates at runtime to augment responses to user queries.

Supported regions and models for Knowledge bases for Amazon Bedrock

Knowledge bases for Amazon Bedrock is supported in the following regions:
Region

US East (N. Virginia)

US West (Oregon)

You can use the following models to embed your data sources in a vector store:

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>amazon.titan-embed-text-v1</td>
</tr>
<tr>
<td>Cohere Embed (English)</td>
<td>cohere.embed-english-v3</td>
</tr>
<tr>
<td>Cohere Embed (Multilingual)</td>
<td>cohere.embed-multilingual-v3</td>
</tr>
</tbody>
</table>

You can use the following models to generate responses after retrieving information from knowledge bases:

<table>
<thead>
<tr>
<th>Model</th>
<th>Model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropic Claude v2.0</td>
<td>anthropic.claude-v2</td>
</tr>
<tr>
<td>Anthropic Claude v2.1</td>
<td>anthropic.claude-v2:1</td>
</tr>
<tr>
<td>Anthropic Claude 3 Sonnet v1</td>
<td>anthropic.claude-3-sonnet-20240229-v1:0</td>
</tr>
<tr>
<td>Anthropic Claude Instant v1</td>
<td>anthropic.claude-instant-v1</td>
</tr>
</tbody>
</table>

Prerequisites for Knowledge bases for Amazon Bedrock

Before you can create a knowledge base, you need to fulfill the following prerequisites:

1. Prepare the files containing information that you want your knowledge base to contain to create a data source for your knowledge base. Then upload the files to an Amazon S3 bucket.
2. (Optional) Set up a vector store of your choice. You can skip this prerequisite if you plan to use the AWS Management Console to automatically create a vector store in Amazon OpenSearch Serverless for you.

3. (Optional) Create a custom AWS Identity and Access Management (IAM) service role with the proper permissions by following the instructions at Create a service role for Knowledge bases for Amazon Bedrock. You can skip this prerequisite if you plan to use the AWS Management Console to automatically create a service role for you.

4. (Optional) Set up extra security configurations by following the steps at Encryption of knowledge base resources.

Topics
- Set up a data source for your knowledge base
- Set up a vector index for your knowledge base in a supported vector store

Set up a data source for your knowledge base

A data source contains files with information that can be retrieved when your knowledge base is queried. You set up the data source for your knowledge base by uploading source document files to an Amazon S3 bucket.

Check that each source document file conforms to the following requirements:

- The file must be in one of the following supported formats:

<table>
<thead>
<tr>
<th>Format</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain text</td>
<td>.txt</td>
</tr>
<tr>
<td>Markdown</td>
<td>.md</td>
</tr>
<tr>
<td>HyperText Markup Language</td>
<td>.html</td>
</tr>
<tr>
<td>Microsoft Word document</td>
<td>.doc/.docx</td>
</tr>
<tr>
<td>Comma-separated values</td>
<td>.csv</td>
</tr>
<tr>
<td>Microsoft Excel spreadsheet</td>
<td>.xls/.xlsx</td>
</tr>
</tbody>
</table>
The file size doesn't exceed the quota of 50 MB.

The following topics describe optional steps for preparing your data source.

**Topics**

- [Add metadata to your files to allow for filtering](#)
- [Source chunks](#)

### Add metadata to your files to allow for filtering

You can optionally add metadata to files in your data source. Metadata allows for your data to be filtered during knowledge base query.

#### Metadata file requirements

To include metadata for a file in your data source, create a JSON file consisting of a `metadataAttributes` field that maps to an object with a key-value pair for each metadata attribute. Then upload it to the same folder in your Amazon S3 bucket as the source document file. The following displays the general format of the metadata file:

```json
{
    "metadataAttributes": {
        "${attribute1}": "${value1}",
        "${attribute2}": "${value2}",
        ...
    }
}
```

The following data types are supported for the values of the attributes:

- String
- Number
- Boolean
Check that each metadata file conforms to the following requirements:

- The file has the same name as its associated source document file.
- Append `.metadata.json` after the file extension (for example, if you have a file called `A.txt`, the metadata file must be named `A.txt.metadata.json`).
- The file size doesn't exceed the quota of 10 KB.
- The file is in the same folder in the Amazon S3 bucket as its associated source document file.

**Note**

If you're adding metadata to an existing vector index in an Amazon OpenSearch Serverless vector store, check that the vector index is configured with the faiss engine to allow for filtering. If the vector index is configured with the nmslib engine, you'll have to do one of the following:

- **Create a new knowledge base** in the console and let Amazon Bedrock automatically create a vector index in Amazon OpenSearch Serverless for you.
- **Create another vector index** in the vector store and select faiss as the **Engine**. Then create a new knowledge base and specify the new vector index.

If you're adding metadata to an existing vector index in an Amazon Aurora database cluster, you must add a column to the table for each metadata attribute in your metadata files before starting ingestion. The metadata attribute values will be written to these columns.

After you sync your data source, you can filter results during knowledge base query.

**Metadata file example**

As an example, if you have a source document with the name `oscars-coverage_20240310.pdf` that contains news articles, you might want to categorize them by attributes such as `year` or `genre`. To create the metadata for this file, perform the following steps:

1. Create a file named `oscars-coverage_20240310.pdf.metadata.json` with the following contents:

   ```json
   {}
   ```
2. **Upload** *oscars-coverage_20240310.pdf.metadata.json* to the same folder as *oscars-coverage_20240310.pdf* in your Amazon S3 bucket.

3. **Create a knowledge base** if you haven’t yet. Then, **sync your data source**.

### Source chunks

During ingestion of your data into a knowledge base, Amazon Bedrock splits each file into **chunks**. A **chunk** refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried.

Amazon Bedrock offers chunking strategies that you can use to chunk your data. You can also pre-process your data by chunking your source files yourself. Consider which of the following chunking strategies you want to use for your data source:

- **Default chunking** – By default, Amazon Bedrock automatically splits your source data into chunks, such that each chunk contains, at most, approximately 300 tokens. If a document contains less than 300 tokens, then it is not split any further.

- **Fixed size chunking** – Amazon Bedrock splits your source data into chunks of the approximate size that you set.

- **No chunking** – Amazon Bedrock treats each file as one chunk. If you choose this option, you may want to pre-process your documents by splitting them into separate files before uploading them to an Amazon S3 bucket.

### Set up a vector index for your knowledge base in a supported vector store

You set up a supported vector index to index your data sources by creating fields to store the following data.

- The vectors generated from the text in your data source by the embeddings model that you choose.
• The text chunks extracted from the files in your data source.
• Metadata related to your knowledge base that Amazon Bedrock manages.
• (If you use an Amazon Aurora database and want to set up filtering) Metadata that you associate with your source files. If you plan to set up filtering in other vector stores, you don't have to set up these fields for filtering.

Select the tab corresponding to the service that you will use to create your vector index.

⚠️ Note
If you prefer for Amazon Bedrock to automatically create a vector index in Amazon OpenSearch Serverless for you, skip this prerequisite and proceed to Create a knowledge base. To learn how to set up a vector index, select the tab corresponding to your method of choice and follow the steps.

Amazon OpenSearch Serverless

1. To configure permissions and create a vector search collection in Amazon OpenSearch Serverless in the AWS Management Console, follow steps 1 and 2 at Working with vector search collections in the Amazon OpenSearch Service Developer Guide. Note the following considerations while setting up your collection:

   a. Give the collection a name and description of your choice.

   b. To make your collection private, select Standard create for the Security section. Then, in the Network access settings section, select VPC as the Access type and choose a VPC endpoint. For more information about setting up a VPC endpoint for an Amazon OpenSearch Serverless collection, see Access Amazon OpenSearch Serverless using an interface endpoint (AWS PrivateLink) in the Amazon OpenSearch Service Developer Guide.

2. Once the collection is created, take note of the Collection ARN for when you create the knowledge base.

3. In the left navigation pane, select Collections under Serverless. Then select your vector search collection.

4. Select the Indexes tab. Then choose Create vector index.
5. In the **Vector index details** section, enter a name for your index in the **Vector index name** field.

6. In the **Vector fields** section, choose **Add vector field**. Amazon Bedrock stores the vector embeddings for your data source in this field. Provide the following configurations:

   - **Vector field name** – Provide a name for the field (for example, `embeddings`).
   - **Engine** – The vector engine used for search. Select **faiss**.
   - **Dimensions** – The number of dimensions in the vector. Refer to the following table to determine how many dimensions the vector should contain:

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan G1 Embeddings - Text</td>
<td>1,536</td>
</tr>
<tr>
<td>Cohere Embed English</td>
<td>1,024</td>
</tr>
<tr>
<td>Cohere Embed Multilingual</td>
<td>1,024</td>
</tr>
</tbody>
</table>

   - **Distance metric** – The metric used to measure the similarity between vectors. We recommend using **Euclidean**.

7. Expand the **Metadata management** section and add two fields to configure the vector index to store additional metadata that a knowledge base can retrieve with vectors. The following table describes the fields and the values to specify for each field:

<table>
<thead>
<tr>
<th>Field description</th>
<th>Mapping field</th>
<th>Data type</th>
<th>Filterable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Bedrock chunks the raw text from your data and stores the chunks in this field.</td>
<td>Name of your choice (for example, <code>text</code>)</td>
<td>String</td>
<td>True</td>
</tr>
<tr>
<td>Amazon Bedrock stores metadata related to your knowledge base in this field.</td>
<td>Name of your choice (for example, <code>bedrock-m etadata</code>)</td>
<td>String</td>
<td>False</td>
</tr>
</tbody>
</table>
8. Take note of the names you choose for the vector index name, vector field name, and metadata management mapping field names for when you create your knowledge base. Then choose **Create**.

After the vector index is created, you can proceed to [create your knowledge base](#). The following table summarizes where you will enter each piece of information that you took note of.

<table>
<thead>
<tr>
<th>Field</th>
<th>Corresponding field in knowledge base setup (Console)</th>
<th>Corresponding field in knowledge base setup (API)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection ARN</td>
<td>Collection ARN</td>
<td>collectionARN</td>
<td>The Amazon Resource Name (ARN) of the vector search collection.</td>
</tr>
<tr>
<td>Vector index name</td>
<td>Vector index name</td>
<td>vectorIndexName</td>
<td>The name of the vector index.</td>
</tr>
<tr>
<td>Vector field name</td>
<td>Vector field</td>
<td>vectorField</td>
<td>The name of the field in which to store vector embeddings for your data sources.</td>
</tr>
<tr>
<td>Metadata management (first mapping field)</td>
<td>Text field</td>
<td>textField</td>
<td>The name of the field in which to store the raw text from your data sources.</td>
</tr>
<tr>
<td>Metadata management (second mapping field)</td>
<td>Bedrock-managed metadata field</td>
<td>metadataField</td>
<td>The name of the field in which to store metadata that Amazon Bedrock manages.</td>
</tr>
</tbody>
</table>
For more detailed documentation on setting up a vector store in Amazon OpenSearch Serverless, see [Working with vector search collections](#) in the Amazon OpenSearch Service Developer Guide.

**Amazon Aurora**

1. Create an Amazon Aurora database (DB) cluster, schema, and table by following the steps at [Preparing Aurora PostgreSQL to be used as a Knowledge Base](#). When you create the table, configure it with the following columns and data types. You can use column names of your liking instead of the ones listed in the following table. Take note of the column names you choose so that you can provide them during knowledge base setup.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Data type</th>
<th>Corresponding field in knowledge base setup (Console)</th>
<th>Corresponding field in knowledge base setup (API)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>UUID primary key</td>
<td>Primary key</td>
<td>primaryKeyField</td>
<td>Contains unique identifiers for each record.</td>
</tr>
<tr>
<td>embedding</td>
<td>Vector</td>
<td>Vector field</td>
<td>vectorField</td>
<td>Contains the vector embedding of the data sources.</td>
</tr>
<tr>
<td>chunks</td>
<td>Text</td>
<td>Text field</td>
<td>textField</td>
<td>Contains the chunks of raw text from your data sources.</td>
</tr>
<tr>
<td>metadata</td>
<td>JSON</td>
<td>Bedrock-managed metadata field</td>
<td>metadataField</td>
<td>Contains metadata required to carry out ...</td>
</tr>
</tbody>
</table>
2. (Optional) If you added metadata to your files for filtering, you must also create a column for each metadata attribute in your files and specify the data type (text, number, or boolean). For example, if the attribute genre exists in your data source, you would add a column named genre and specify text as the data type. During ingestion, these columns will be populated with the corresponding attribute values.

3. Configure an AWS Secrets Manager secret for your Aurora DB cluster by following the steps at Password management with Amazon Aurora and AWS Secrets Manager.

4. Take note of the following information after you create your DB cluster and set up the secret.

<table>
<thead>
<tr>
<th>Field in knowledge base setup (Console)</th>
<th>Field in knowledge base setup (API)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Aurora DB Cluster ARN</td>
<td>resourceArn</td>
<td>The ARN of your DB cluster.</td>
</tr>
<tr>
<td>Database name</td>
<td>databaseName</td>
<td>The name of your database</td>
</tr>
<tr>
<td>Table name</td>
<td>tableName</td>
<td>The name of the table in your DB cluster</td>
</tr>
<tr>
<td>Secret ARN</td>
<td>credentialsSecretArn</td>
<td>The ARN of the AWS Secrets Manager key for your DB cluster</td>
</tr>
</tbody>
</table>
Note

If you use Pinecone, you agree to authorize AWS to access the designated third-party source on your behalf in order to provide vector store services to you. You're responsible for complying with any third-party terms applicable to use and and transfer of data from the third-party service.

For detailed documentation on setting up a vector store in Pinecone, see Pinecone as a Knowledge Base for Amazon Bedrock.

While you set up the vector store, take note of the following information, which you will fill out when you create a knowledge base:

- **Connection string** – The endpoint URL for your index management page.
- **Namespace** – (Optional) The namespace to be used to write new data to your database. For more information, see Using namespaces.

There are additional configurations that you must provide when creating a Pinecone index:

- **Name** – The name of the vector index. Choose any valid name of your choice. Later, when you create your knowledge base, enter the name you choose in the Vector index name field.
- **Dimensions** – The number of dimensions in the vector. Refer to the following table to determine how many dimensions the vector should contain.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan G1 Embeddings - Text</td>
<td>1,536</td>
</tr>
<tr>
<td>Cohere Embed English</td>
<td>1,024</td>
</tr>
<tr>
<td>Cohere Embed Multilingual</td>
<td>1,024</td>
</tr>
</tbody>
</table>

- **Distance metric** – The metric used to measure the similarity between vectors. We recommend that you experiment with different metrics for your use-case. We recommend starting with cosine similarity.
To access your Pinecone index, you must provide your Pinecone API key to Amazon Bedrock through the AWS Secrets Manager.

**To set up a secret for your Pinecone configuration**

1. Follow the steps at Create an AWS Secrets Manager secret, setting the key as apiKey and the value as the API key to access your Pinecone index.
2. To find your API key, open your Pinecone console and select API Keys.
3. After you create the secret, take note of the ARN of the KMS key.
4. Attach permissions to your service role to decrypt the ARN of the KMS key by following the steps in Permissions to decrypt an AWS Secrets Manager secret for the vector store containing your knowledge base.
5. Later, when you create your knowledge base, enter the ARN in the Credentials secret ARN field.

Redis Enterprise Cloud

**Note**

If you use Redis Enterprise Cloud, you agree to authorize AWS to access the designated third-party source on your behalf in order to provide vector store services to you. You're responsible for complying with any third-party terms applicable to use and transfer of data from the third-party service.

For detailed documentation on setting up a vector store in Redis Enterprise Cloud, see Integrating Redis Enterprise Cloud with Amazon Bedrock.

While you set up the vector store, take note of the following information, which you will fill out when you create a knowledge base:

- **Endpoint URL** – The public endpoint URL for your database.
- **Vector index name** – The name of the vector index for your database.
- **Vector field** – The name of the field where the vector embeddings will be stored. Refer to the following table to determine how many dimensions the vector should contain.
### Model Dimensions

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan G1 Embeddings - Text</td>
<td>1,536</td>
</tr>
<tr>
<td>Cohere Embed English</td>
<td>1,024</td>
</tr>
<tr>
<td>Cohere Embed Multilingual</td>
<td>1,024</td>
</tr>
</tbody>
</table>

- **Text field** – The name of the field where the Amazon Bedrock stores the chunks of raw text.

- **Bedrock-managed metadata field** – The name of the field where Amazon Bedrock stores metadata related to your knowledge base.

To access your Redis Enterprise Cloud cluster, you must provide your Redis Enterprise Cloud security configuration to Amazon Bedrock through the AWS Secrets Manager.

**To set up a secret for your Redis Enterprise Cloud configuration**

1. Enable TLS to use your database with Amazon Bedrock by following the steps at [Transport Layer Security (TLS)](https://aws.amazon.com/documentation/redis/using-ssl/).
2. Follow the steps at [Create an AWS Secrets Manager secret](https://docs.aws.amazon.com/secretsmanager/latest/userguide/create-secret-aws-development-console.html). Set up the following keys with the appropriate values from your Redis Enterprise Cloud configuration in the secret:

   - **username** – The username to access your Redis Enterprise Cloud database. To find your username, look under the **Security** section of your database in the [Redis Console](https://aws.amazon.com/documentation/redis/using-redis-console/).
   - **password** – The password to access your Redis Enterprise Cloud database. To find your password, look under the **Security** section of your database in the [Redis Console](https://aws.amazon.com/documentation/redis/using-redis-console/).
   - **serverCertificate** – The content of the certificate from the Redis Cloud Certificate authority. Download the server certificate from the Redis Admin Console by following the steps at [Download certificates](https://docs.aws.amazon.com/redis/latest/adv/redis-certificates.html).
   - **clientPrivateKey** – The private key of the certificate from the Redis Cloud Certificate authority. Download the server certificate from the Redis Admin Console by following the steps at [Download certificates](https://docs.aws.amazon.com/redis/latest/adv/redis-certificates.html).
   - **clientCertificate** – The public key of the certificate from the Redis Cloud Certificate authority. Download the server certificate from the Redis Admin Console by following the steps at [Download certificates](https://docs.aws.amazon.com/redis/latest/adv/redis-certificates.html).
3. After you create the secret, take note of its ARN. Later, when you create your knowledge base, enter the ARN in the **Credentials secret ARN** field.

Create a knowledge base

_i Note_

You can’t create a knowledge base with a root user. Log in with an IAM user before starting these steps.

After you set up your data source in Amazon S3 and a vector store of your choice, you can create a knowledge base. Select the tab corresponding to your method of choice and follow the steps.

**Console**

To create a knowledge base

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. From the left navigation pane, select **Knowledge base**.
3. In the **Knowledge bases** section, select **Create knowledge base**.
4. On the **Provide knowledge base details** page, set up the following configurations:
   a. (Optional) In the **Knowledge base details** section, change the default name and provide a description for your knowledge base.
   b. In the **IAM permissions** section, choose an AWS Identity and Access Management (IAM) role that provides Amazon Bedrock permission to access other AWS services. You can let Amazon Bedrock create the service role or choose a [custom role that you have created](#).
   c. (Optional) Add tags to your knowledge base. For more information, see [Tag resources](#).
   d. Select **Next**.
5. On the **Set up data source** page, provide the information for the data source to use for the knowledge base:
   a. (Optional) Change the default **Data source name**.
b. Provide the S3 URI of the object containing the files for the data source that you prepared.

Note
Choose an Amazon S3 bucket in the same region as the knowledge base that you’re creating. Otherwise, your data source will fail to sync.

c. If you encrypted your Amazon S3 data with a customer managed key, select Add customer-managed AWS KMS key for Amazon S3 data and choose a KMS key to allow Amazon Bedrock to decrypt it. For more information, see Encryption of information passed to Amazon OpenSearch Service.

d. (Optional) To configure the following advanced settings, expand the Advanced settings - optional section.

i. While converting your data into embeddings, Amazon Bedrock encrypts your data with a key that AWS owns and manages, by default. To use your own KMS key, expand Advanced settings, select Customize encryption settings (advanced), and choose a key. For more information, see Encryption of transient data storage during data ingestion.

ii. Choose from the following options for the Chunking strategy for your data source:

- **Default chunking** – By default, Amazon Bedrock automatically splits your source data into chunks, such that each chunk contains, at most, 300 tokens. If a document contains less than 300 tokens, then it is not split any further.

- **Fixed size chunking** – Amazon Bedrock splits your source data into chunks of the approximate size that you set. Configure the following options.
  - **Max tokens** – Amazon Bedrock creates chunks that don't exceed the number of tokens that you choose.
  - **Overlap percentage between chunks** – Each chunk overlaps with consecutive chunks by the percentage that you choose.

- **No chunking** – Amazon Bedrock treats each file as one chunk. If you choose this option, you may want to pre-process your documents by splitting them into separate files.
Note
You can't change the chunking strategy after you have created the data source.

6. Select Next.

7. In the Embeddings model section, choose a supported embeddings model to convert your data into vector embeddings for the knowledge base.

   a. Choose a vector store you have created – Select the service that contains a vector database that you have already created. Fill in the fields to allow Amazon Bedrock to

b. To encrypt the automated vector store with a customer managed key select Add customer-managed KMS key for Amazon OpenSearch Serverless vector – optional and choose the key. For more information, see Encryption of information passed to Amazon OpenSearch Service.

   • Choose a vector store you have created – Select the service that contains a vector database that you have already created. Fill in the fields to allow Amazon Bedrock
to map information from the knowledge base to your database, so that it can store, update, and manage embeddings. For more information about how these fields map to the fields that you created, see Set up a vector index for your knowledge base in a supported vector store.

**Note**

If you use a database in Amazon OpenSearch Serverless or Amazon Aurora, you need to have configured the fields under **Metadata field mapping** beforehand. If you use a database in Pinecone or Redis Enterprise Cloud, you can provide names for these fields here and Amazon Bedrock will dynamically create them in the vector store for you.

8. Select **Next**.

9. On the **Review and create** page, check the configuration and details of your knowledge base. Choose **Edit** in any section that you need to modify. When you are satisfied, select **Create knowledge base**.

10. The time it takes to create the knowledge base depends on the amount of data you provided. When the knowledge base is finished being created, the **Status** of the knowledge base changes to **Ready**.

**API**

To create a knowledge base, send a **CreateKnowledgeBase** request with a **Agents for Amazon Bedrock build-time endpoint** and provide the name, description, instructions for what it should do, and the foundation model for it to orchestrate with.

**Note**

If you prefer to let Amazon Bedrock create and manage a vector store for you in Amazon OpenSearch Service, use the console. For more information, see **Create a knowledge base**.

- Provide the ARN with permissions to create a knowledge base in the **roleArn** field.
- Provide the embedding model to use in the **embeddingModelArn** field in the **knowledgeBaseConfiguration** object.
• Provide the configuration for your vector store in the storageConfiguration object. For more information, see Set up a vector index for your knowledge base in a supported vector store
  • For an Amazon OpenSearch Service database, use the opensearchServerlessConfiguration object.
  • For a Pinecone database, use the pineconeConfiguration object.
  • For a Redis Enterprise Cloud database, use the redisEnterpriseCloudConfiguration object.
  • For an Amazon Aurora database, use the rdsConfiguration object.

After you create a knowledge base, create a data source from the S3 bucket containing the files for your knowledge base. To create the data source send a CreateDataSource request.

• Provide the information for the S3 bucket containing the data source files in the dataSourceConfiguration field.

• Specify how to chunk the data sources in the vectorIngestionConfiguration field. For more information, see Set up a data source for your knowledge base.

  Note
  You can't change the chunking configuration after you create the data source.

• (Optional) While converting your data into embeddings, Amazon Bedrock encrypts your data with a key that AWS owns and manages, by default. To use your own KMS key, include it in the serverSideEncryptionConfiguration object. For more information, see Encryption of knowledge base resources.

Set up security configurations for your knowledge base

After you've created a knowledge base, you might have to set up the following security configurations:

Topics
• Set up data access policies for your knowledge base
• Set up network access policies for your Amazon OpenSearch Serverless knowledge base
Set up data access policies for your knowledge base

If you're using a custom role, set up security configurations for your newly created knowledge base. If you let Amazon Bedrock create a service role for you, you can skip this step. Follow the steps in the tab corresponding to the database that you set up.

Amazon OpenSearch Serverless

To restrict access to the Amazon OpenSearch Serverless collection to the knowledge base service role, create a data access policy. You can do so in the following ways:

- Use the Amazon OpenSearch Service console by following the steps at Creating data access policies (console) in the Amazon OpenSearch Service Developer Guide.
- Use the AWS API by sending a CreateAccessPolicy request with an OpenSearch Serverless endpoint. For an AWS CLI example, see Creating data access policies (AWS CLI).

Use the following data access policy, specifying the Amazon OpenSearch Serverless collection and your service role:

```json
[
  {
    "Description": "$\{data\ access\ policy\ description\}$",
    "Rules": [
      {
        "Resource": [
          "index/\${collection\_name}/**"
        ],
        "Permission": [
          "aoss:DescribeIndex",
          "aoss:ReadDocument",
          "aoss:WriteDocument"
        ],
        "ResourceType": "index"
      }
    ],
    "Principal": [
      "arn:aws:iam::\${account-id}:role/\${kb-service-role}"
    ]
  }
]
```
Pinecone or Redis Enterprise Cloud

To integrate a Pinecone or Redis Enterprise Cloud vector index, attach the following identity-based policy to your knowledge base service role to allow it to access the AWS Secrets Manager secret for the vector index.

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "bedrock:AssociateThirdPartyKnowledgeBase"
         ],
         "Resource": [
            "arn:aws:iam::${region}:${account-id}:knowledge-base/${knowledge-base-id}"
         ],
         "Condition": {
            "StringEquals": {
               "bedrock:ThirdPartyKnowledgeBaseCredentialsSecretArn": "arn:aws:iam::${region}:${account-id}:secret:${secret-id}"
            }
         }
      }
   ]
}
```

Set up network access policies for your Amazon OpenSearch Serverless knowledge base

If you use a private Amazon OpenSearch Serverless collection for your knowledge base, it can only be accessed through an AWS PrivateLink VPC endpoint. You can create a private Amazon OpenSearch Serverless collection when you set up your Amazon OpenSearch Serverless vector collection or you can make an existing Amazon OpenSearch Serverless collection (including one that the Amazon Bedrock console created for you) private when you configure its network access policy.

The following resources in the Amazon OpenSearch Service Developer Guide will help you understand the setup required for a private Amazon OpenSearch Serverless collections:
• For more information about setting up a VPC endpoint for a private Amazon OpenSearch Serverless collection, see Access Amazon OpenSearch Serverless using an interface endpoint (AWS PrivateLink).
• For more information about network access policies in Amazon OpenSearch Serverless, see Network access for Amazon OpenSearch Serverless.

To allow an Amazon Bedrock knowledge base to access a private Amazon OpenSearch Serverless collection, you must edit the network access policy for the Amazon OpenSearch Serverless collection to allow Amazon Bedrock as a source service. Select the tab corresponding to your method of choice and follow the steps.

**Console**

1. Open the Amazon OpenSearch Service console at https://console.aws.amazon.com/aos/.
2. From the left navigation pane, select Collections. Then choose your collection.
3. In the Network section, select the Associated Policy.
4. Choose Edit.
5. For Select policy definition method, choose JSON.
6. In the JSON editor, paste the following policy.

```json
[
  {
    "AllowFromPublic": false,
    "Description": "${network access policy description}",
    "Rules": [
      {
        "ResourceType": "collection",
        "Resource": [
          "collection/${collection-id}"  
        ],
      },
    ],"SourceVPCEs": [
    "${vpce-id}
    ],
    "SourceServices": [  
      "bedrock.amazonaws.com"
    ]
  }
]
```
7. Choose **Update**.

**API**

To edit the network access policy for your Amazon OpenSearch Serverless collection, do the following:

1. Send a **GetSecurityPolicy** request with an [OpenSearch Serverless endpoint](https://aws.amazon.com/opensearch/). Specify the name of the policy and specify the type as `network`. Note the `policyVersion` in the response.

2. Send a **UpdateSecurityPolicy** request with an [OpenSearch Serverless endpoint](https://aws.amazon.com/opensearch/). Minimally, specify the following fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>The name of the policy</td>
</tr>
<tr>
<td>policyVersion</td>
<td>The <code>policyVersion</code> returned to you from the <code>GetSecurityPolicy</code> response.</td>
</tr>
<tr>
<td>type</td>
<td>The type of security policy. Specify <code>network</code>.</td>
</tr>
<tr>
<td>policy</td>
<td>The policy to use. Specify the following JSON object</td>
</tr>
</tbody>
</table>

```json
[
  {
    "AllowFromPublic": false,
    "Description":"${network access policy description}",
    "Rules": [
      {
        "ResourceType": "collection",
        "Resource": [
          "collection/${collection-id}"
        ]
      }
    ]
  }
]
```
For an AWS CLI example, see [Creating data access policies (AWS CLI)](Creating_data_access_policies_AWS_CLI).

- Use the Amazon OpenSearch Service console by following the steps at [Creating network policies (console)](Creating_network_policies_console). Instead of creating a network policy, note the Associated policy in the Network subsection of the collection details.

## Sync to ingest your data sources into the knowledge base

After you create your knowledge base, you ingest the data sources into the knowledge base so that they're indexed and able to be queried. Ingestion converts the raw data in your data source into vector embeddings. It also associates the raw text and any relevant metadata that you set up for filtering to augment the querying process. Before you begin ingestion, check that your data source fulfills the following conditions:

- The Amazon S3 bucket for the data source is in the same region as the knowledge base.
- The files are in supported formats. For more information, see [Set up a vector index for your knowledge base in a supported vector store](Set_up_a_vector_index_for_your_knowledge_base_in_a_supported_vector_store).
- The files don't exceed the maximum file size of 50 MB. For more information, see [Knowledge base quotas](Knowledge_base_quotas).
- If your data source contains metadata files, check the following conditions to ensure that the metadata files aren't ignored:
  - Each .metadata.json file shares the same name as the source file that it's associated with.
  - If the vector index for your knowledge base is in an Amazon OpenSearch Serverless vector store, check that the vector index is configured with the faiss engine. If the vector index is configured with the nmslib engine, you'll have to do one of the following:
• **Create a new knowledge base** in the console and let Amazon Bedrock automatically create a vector index in Amazon OpenSearch Serverless for you.

• **Create another vector index** in the vector store and select **faiss** as the **Engine**. Then **create a new knowledge base** and specify the new vector index.

• If the vector index for your knowledge base is in an Amazon Aurora database cluster, check that the table for your index contains a column for each metadata property in your metadata files before starting ingestion.

---

**Note**

Each time you add, modify, or remove files from the S3 bucket for a data source, you must sync the data source so that it is re-indexed to the knowledge base. Syncing is incremental, so Amazon Bedrock only processes the objects in your S3 bucket that have been added, modified, or deleted since the last sync.

---

To learn how to ingest your data sources into your knowledge base, Select the tab corresponding to your method of choice and follow the steps.

**Console**

**To ingest your data sources**

2. From the left navigation pane, select **Knowledge base** and choose your knowledge base.
3. In the **Data source** section, select **Sync** to begin data ingestion.
4. When data ingestion completes, a green success banner appears if it is successful.
5. You can choose a data source to view its **Sync history**. Select **View warnings** to see why a data ingestion job failed.

**API**

To ingest a data source into the vector store you configured for your knowledge base, send a **StartIngestionJob** request with a **Agents for Amazon Bedrock build-time endpoint**. Specify the knowledgeBaseId and dataSourceId.

---

Sync your data sources
Use the ingestionJobId returned in the response in a `GetIngestionJob` request with a [Agents for Amazon Bedrock build-time endpoint](#) to track the status of the ingestion job. In addition, specify the knowledgeBaseId and dataSourceId.

- When the ingestion job finishes, the status in the response is COMPLETE.
- The statistics object in the response returns information about whether ingestion was successful or not for documents in the data source.

You can also see information for all ingestion jobs for a data source by sending a `ListIngestionJobs` request with a [Agents for Amazon Bedrock build-time endpoint](#). Specify the dataSourceId and the knowledgeBaseId of the knowledge base that the data is being ingested to.

- Filter for results by specifying a status to search for in the filters object.
- Sort by the time that the job was started or the status of a job by specifying the sortBy object. You can sort in ascending or descending order.
- Set the maximum number of results to return in a response in the maxResults field. If there are more results than the number you set, the response returns a nextToken that you can send in another [ListIngestionJobs](#) request to see the next batch of jobs.

### Test a knowledge base in Amazon Bedrock

After you set up your knowledge base, you can test its behavior by sending queries and seeing the responses. You can also set query configurations to customize information retrieval. When you are satisfied with your knowledge base's behavior, you can then set up your application to query the knowledge base or attach the knowledge base to an agent.

Select a topic to learn more about it.

**Topics**

- [Query the knowledge base and return results or generate responses](#)
- [Query configurations](#)
Query the knowledge base and return results or generate responses

To learn how to query your knowledge base, select the tab corresponding to your method of choice and follow the steps.

Console

To test your knowledge base

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select Knowledge base.
3. In the Knowledge bases section, do one of the following actions:
   - Choose the radio button next to the knowledge base you want to test and select Test knowledge base. A test window expands from the right.
   - Choose the knowledge base that you want to test. A test window expands from the right.
4. Select or clear Generate responses for your query depending on your use case.
   - To return information retrieved directly from your knowledge base, turn off Generate responses. Amazon Bedrock will return text chunks from your data sources that are relevant to the query.
   - To generate responses based on information retrieved from your knowledge base, turn on Generate responses. Amazon Bedrock will generate responses based on your data sources and cites the information it provides with footnotes.
5. If you turn on Generate responses, choose Select model to choose a model to use for response generation. Then select Apply.
6. (Optional) Select the configurations icon to open up Configurations. You can modify the following configurations:
   - Search type – Specify how your knowledge base is queried. For more information, see Search type.
• **Maximum number of source chunks** – Specify the maximum number of source chunks to retrieve. For more information, see [Maximum number of source chunks](#).

• **Filters** – Specify up to 5 filter groups and up to 5 filters within each group to use with the metadata for your files. For more information, see [Metadata and filtering](#).

• **Knowledge base prompt template** – If you turn on Generate responses, you can replace the default prompt template with your own to customize the prompt that's sent to the model for response generation. For more information, see [Knowledge base prompt template](#).

7. Enter a query in the text box in the chat window and select **Run** to return responses from the knowledge base.

8. You can examine the response in the following ways.

   • If you didn't generate responses, the text chunks are returned directly in order of relevance.

   • If you generated responses, select a footnote to see an excerpt from the cited source for that part of the response. Choose the link to navigate to the S3 object containing the file.

   • To see details about the chunks cited for each footnote, select **Show source details**. You can carry out the following actions in the **Source details** pane:

     • To see the configurations that you set for query, expand **Query configurations**.

     • To view details about a source chunk, expand it by choosing the right arrow ( '>' ) next to it. You can see the following information:

       • The raw text from the source chunk. To copy this text, choose the copy icon ( 'COPY' ). To navigate to the S3 object containing the file, choose the external link icon ( 'FILE' ).

       • The metadata associated with the source chunk. The attribute keys and values are defined in the `.metadata.json` file that's associated with the source document. For more information, see [Metadata file requirements](#).
Chat options

1. If you are generating responses, you can select **Change model** to use a different model for response generation. If you change the model, the text in the chat window will be completely cleared.

2. Switch between generating responses for your query and returning direct quotations by selecting or clearing **Generate responses**. If you change the setting, the text in the chat window will be completely cleared.

3. To clear the chat window, select the broom icon
   (🧹).

4. To copy all the output in the chat window, select the copy icon
   (📋).

API

**Retrieve**

To query a knowledge base and only return relevant text from data sources, send a **Retrieve** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock runtime endpoint**.

The following list briefly describes the parameters and request body. For detailed information and the request structure, see see [Retrieve request syntax](#):

<table>
<thead>
<tr>
<th>Variable</th>
<th>Required?</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>knowledgeBaseld</td>
<td>Yes</td>
<td>To specify the knowledge base to query</td>
</tr>
<tr>
<td>retrievalQuery</td>
<td>Yes</td>
<td>Contains a text field to specify the query</td>
</tr>
<tr>
<td>nextToken</td>
<td>No</td>
<td>To return the next batch of responses</td>
</tr>
<tr>
<td>Variable</td>
<td>Required?</td>
<td>Use case</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>retrievalConfiguration</td>
<td>No</td>
<td>To include query configurations for customizing the vector search.</td>
</tr>
</tbody>
</table>

The following list briefly describes the response body. For detailed information and the response structure, see [Retrieve response syntax](#):  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrievalResults</td>
<td>Contains the source chunks, Amazon S3 location of the source, and a relevancy score for the chunk.</td>
</tr>
<tr>
<td>nextToken</td>
<td>To use in another request to return the next batch of results.</td>
</tr>
</tbody>
</table>

**RetrieveAndGenerate**

To query a knowledge base and use a foundation model to generate responses based off the results from the data sources, send a RetrieveAndGenerate request with a Agents for Amazon Bedrock runtime endpoint.

The following list briefly describes the parameters and request body. For detailed information and the request structure, see [RetrieveAndGenerate request syntax](#):  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Required?</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td>Yes</td>
<td>Contains a text field to specify the query</td>
</tr>
<tr>
<td>retrieveAndGenerateConfiguration</td>
<td>Yes</td>
<td>For specifying the knowledge base to query, the model to use for response generation, and</td>
</tr>
<tr>
<td>Variable</td>
<td>Required?</td>
<td>Use case</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>sessionld</td>
<td>No</td>
<td>Use the same value to continue the same session and maintain information</td>
</tr>
<tr>
<td>sessionConfiguration</td>
<td>No</td>
<td>To include a KMS key for encryption of the session</td>
</tr>
</tbody>
</table>

The following list briefly describes the response body. For detailed information and the response structure, see [Retrieve response syntax](#).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>citations</td>
<td>Contains parts of the generated response in each object within the generated ResponsePart, and the source chunk in the content object and the Amazon S3 location of the source in the location object of the retrievedReferences object.</td>
</tr>
<tr>
<td>output</td>
<td>Contains the whole generated response.</td>
</tr>
<tr>
<td>sessionld</td>
<td>Contains the ID of the session, which you can reuse in another request to maintain the same conversation</td>
</tr>
</tbody>
</table>

**Note**

If you receive an error that the prompt exceeds the character limit while generating responses, you can shorten the prompt in the following ways:
• Reduce the maximum number of source chunks (this shortens what is filled in for the $search_results$ placeholder in the Knowledge base prompt template).

• Recreate the data source with a chunking strategy that uses smaller chunks (this shortens what is filled in for the $search_results$ placeholder in the Knowledge base prompt template).

• Shorten the prompt template.

• Shorten the user query (this shortens what is filled in for the $query$ placeholder in the Knowledge base prompt template).

Query configurations

You can modify configurations when you query the knowledge base to customize retrieval and response generation. To learn more about a configuration and how to modify it in the console or the API, select from the following topics.

Search type

The search type defines how data sources in the knowledge base are queried. The following search types are possible:

• **Default** – Amazon Bedrock decides the search strategy for you.

• **Hybrid** – Combines searching vector embeddings (semantic search) with searching through the raw text. Hybrid search is currently only supported for Amazon OpenSearch Serverless vector stores that contain a filterable text field. If you use a different vector store or your Amazon OpenSearch Serverless vector store doesn't contain a filterable text field, the query uses semantic search.

• **Semantic** – Only searches vector embeddings.

To learn how to define the search type, select the tab corresponding to your method of choice and follow the steps.

Console

Follow the console steps at Query the knowledge base and return results or generate responses. When you open the Configurations pane, you'll see the following options for Search type:
• **Default** – Amazon Bedrock decides which search strategy is best-suited for your vector store configuration.

• **Hybrid** – Amazon Bedrock queries the knowledge base using both the vector embeddings and the raw text. This option is only available if you're using an Amazon OpenSearch Serverless vector store configured with a filterable text field.

• **Semantic** – Amazon Bedrock queries the knowledge base using its vector embeddings.

**API**

When you make a [Retrieve](#) or [RetrieveAndGenerate](#) request, include a `retrievalConfiguration` field, mapped to a `KnowledgeBaseRetrievalConfiguration` object. To see the location of this field, refer to the [Retrieve](#) and [RetrieveAndGenerate](#) request bodies in the API reference.

The following JSON object shows the minimal fields required in the `KnowledgeBaseRetrievalConfiguration` object to set search type configurations:

```json
"retrievalConfiguration": {
   "vectorSearchConfiguration": {
      "overrideSearchType": "HYBRID | SEMANTIC"
   }
}
```

Specify the search type in the `overrideSearchType` field. You have the following options:

• If you don't specify a value, Amazon Bedrock decides which search strategy is best-suited for your vector store configuration.

• **HYBRID** – Amazon Bedrock queries the knowledge base using both the vector embeddings and the raw text. This option is only available if you're using an Amazon OpenSearch Serverless vector store configured with a filterable text field.

• **SEMANTIC** – Amazon Bedrock queries the knowledge base using its vector embeddings.

**Maximum number of source chunks**

When you query a knowledge base, Amazon Bedrock returns up to five source chunks in the response. To modify the maximum number of source chunks to return, select the tab corresponding to your method of choice and follow the steps.
Console

Follow the console steps at [Query the knowledge base and return results or generate responses](#). In the Configurations pane, expand the Maximum number of source chunks.

API

When you make a Retrieve or RetrieveAndGenerate request, include a retrievalConfiguration field, mapped to a KnowledgeBaseRetrievalConfiguration object. To see the location of this field, refer to the Retrieve and RetrieveAndGenerate request bodies in the API reference.

The following JSON object shows the minimal fields required in the KnowledgeBaseRetrievalConfiguration object to set the maximum number of source chunks to return:

```json
"retrievalConfiguration": {
  "vectorSearchConfiguration": {
    "numberOfResults": number
  }
}
```

Specify the maximum number of source chunks (see the numberOfResults field in KnowledgeBaseRetrievalConfiguration for the range of accepted values) to return in the numberOfResults field.

Metadata and filtering

Your data sources can includes metadata files associated with the source documents. A metadata file contains attributes as key-value pairs that you define for a source document. For more information about creating metadata for your data source files, see [Add metadata to your files to allow for filtering](#). To use filters during knowledge base query, check that your knowledge base fulfills the following requirements:

- The Amazon S3 bucket containing your data source includes at least one .metadata.json file with the same name as the source document it's associated with.
- If your knowledge base's vector index is in an Amazon OpenSearch Serverless vector store, check that the vector index is configured with the faiss engine. If the vector index is configured with the nmslib engine, you'll have to do one of the following:
• **Create a new knowledge base** in the console and let Amazon Bedrock automatically create a vector index in Amazon OpenSearch Serverless for you.

• **Create another vector index** in the vector store and select **faiss** as the **Engine**. Then **Create a new knowledge base** and specify the new vector index.

You can use the following filtering operators when modifying query configurations for filtering:

### Filtering operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Console</th>
<th>API filter name</th>
<th>Supported attribute data types</th>
<th>Filtered results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equals</td>
<td>=</td>
<td>equals</td>
<td>string, number, boolean</td>
<td>Attribute matches the value you provide</td>
</tr>
<tr>
<td>Not equals</td>
<td>!=</td>
<td>notEquals</td>
<td>string, number, boolean</td>
<td>Attribute doesn't match the value you provide</td>
</tr>
<tr>
<td>Greater than</td>
<td>&gt;</td>
<td>greaterThan</td>
<td>number</td>
<td>Attribute is greater than the value you provide</td>
</tr>
<tr>
<td>Greater than or equals</td>
<td>&gt;=</td>
<td>greaterThanOrEquals</td>
<td>number</td>
<td>Attribute is greater than or equal to the value you provide</td>
</tr>
<tr>
<td>Less than</td>
<td>&lt;</td>
<td>lessThan</td>
<td>number</td>
<td>Attribute is less than the value you provide</td>
</tr>
<tr>
<td>Operator</td>
<td>Console</td>
<td>API filter name</td>
<td>Supported attribute data types</td>
<td>Filtered results</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------</td>
<td>-----------------</td>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Less than or equals</td>
<td>&lt;=</td>
<td>lessThanOrEquals</td>
<td>number</td>
<td>Attribute is less than or equal to the value you provide</td>
</tr>
<tr>
<td>In</td>
<td>:</td>
<td>in</td>
<td>string list</td>
<td>Attribute is in the list you provide</td>
</tr>
<tr>
<td>Not in</td>
<td>!:</td>
<td>notIn</td>
<td>string list</td>
<td>Attribute isn't in the list you provide</td>
</tr>
<tr>
<td>Starts with</td>
<td>^</td>
<td>startsWith</td>
<td>string</td>
<td>Attribute starts with the string you provide (only supported for Amazon OpenSearch Serverless vector stores)</td>
</tr>
</tbody>
</table>

To combine filtering operators, you can use the following logical operators:

**Logical operators**

<table>
<thead>
<tr>
<th>Operator</th>
<th>Console</th>
<th>API filter field name</th>
<th>Filtered results</th>
</tr>
</thead>
<tbody>
<tr>
<td>And</td>
<td>and</td>
<td>andAll</td>
<td>Results fulfill all of the filtering expressions in the group</td>
</tr>
<tr>
<td>Operator</td>
<td>Console</td>
<td>API filter field name</td>
<td>Filtered results</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>----------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Or</td>
<td>or</td>
<td>orAll</td>
<td>Results fulfill at least one of the filtering expressions in the group</td>
</tr>
</tbody>
</table>

To learn how to filter results using metadata, select the tab corresponding to your method of choice and follow the steps.

**Console**

Follow the console steps at [Query the knowledge base and return results or generate responses](#). When you open the **Configurations** pane, you'll see a **Filters section**. The following procedures describe different use cases:

- To add a filter, create a filtering expression by entering a metadata attribute, filtering operator, and value in the box. Separate each part of the expression with a whitespace. Press **Enter** to add the filter.

For a list of accepted filtering operators, see the **Filtering operators** table above. You can also see a list of filtering operators when you add a whitespace after the metadata attribute.

**Note**

You must surround strings with quotation marks.

For example, you can filter for results from source documents that contain a `genre` metadata attribute whose value is "entertainment" by adding the following filter: `genre = "entertainment"`.
To add another filter, enter another filtering expression in the box and press Enter. You can add up to 5 filters in the group.
By default, the query will return results that fulfill all the filtering expressions you provide. To return results that fulfill at least one of the filtering expressions, choose the and dropdown menu between any two filtering operations and select or.

To combine different logical operators, select + Add Group to add a filter group. Enter filtering expressions in the new group. You can add up to 5 filter groups.
To change the logical operator used between all the filtering groups, choose the AND dropdown menu between any two filter groups and select OR.
• To edit a filter, select it, modify the filtering operation, and choose Apply.
To remove a filter group, choose the trash can icon next to the group. To remove a filter, choose the delete icon next to the filter.
The following image shows an example filter configuration that returns all documents written after 2018 whose genre is "entertainment", in addition to documents whose genre is "cooking" or "sports" and whose author starts with "C".
When you make a Retrieve or RetrieveAndGenerate request, include a retrievalConfiguration field, mapped to a KnowledgeBaseRetrievalConfiguration object. To see the location of this field, refer to the Retrieve and RetrieveAndGenerate request bodies in the API reference.

The following JSON objects show the minimal fields required in the KnowledgeBaseRetrievalConfiguration object to set filters for different use cases:

1. Use one filtering operator (see the Filtering operators table above).

```json
"retrievalConfiguration": {
"vectorSearchConfiguration": {
"filter": {
"<filter-type>": {
"key": "string",
"value": "string" | number | boolean | ["string", "string", ...]
}
}
}
```
2. Use a logical operator (see the Logical operators table above) to combine up to 5 filtering operators into a filter group, and a second logical operator to combine that filter group with another filtering operator.

```
"retrievalConfiguration": {
    "vectorSearchConfiguration": {
        "filter": {
            "andAll | orAll": [
                "<filter-type>": {
                    "key": "string",
                    "value": "string" | number | boolean | ["string",
                    "string", ...]
                },
                "<filter-type>": {
                    "key": "string",
                    "value": "string" | number | boolean | ["string",
                    "string", ...]
                },
                ...
            ]
        }
    }
}
```

3. Use a logical operator to combine up to 5 filtering operators into a filter group, and a second logical operator to combine that filter group with another filtering operator.

```
"retrievalConfiguration": {
    "vectorSearchConfiguration": {
        "filter": {
            "andAll | orAll": [
                "andAll | orAll": [
                    "<filter-type>": {
                        "key": "string",
                        "value": "string" | number | boolean | ["string",
                        "string", ...]
                    },
                    "<filter-type>": {
                        "key": "string",
                        "value": "string" | number | boolean | ["string",
                        "string", ...]
                    }
                ]
            }
        }
    }
}
```
4. Combine up to 5 filter groups by embedding them within another logical operator. You can create one level of embedding.

```
"retrievalConfiguration": {
  "vectorSearchConfiguration": {
    "filter": {
      "andAll | orAll": [
        "andAll | orAll": {
          "<filter-type>": {
            "key": "string",
            "value": "string" | number | boolean | ["string",
            "string", ...]
          },
          "<filter-type>": {
            "key": "string",
            "value": "string" | number | boolean | ["string",
            "string", ...]
          },
          ...
        },
        "andAll | orAll": {
          "<filter-type>": {
            "key": "string",
            "value": "string" | number | boolean | ["string",
            "string", ...]
          },
          "<filter-type>": {
            "key": "string",
            "value": "string" | number | boolean | ["string",
            "string", ...]
          },
          ...
        }
      }
    }
  }
}
```
The following table describes the filter types that you can use:

<table>
<thead>
<tr>
<th>Field</th>
<th>Supported value data types</th>
<th>Filtered results</th>
</tr>
</thead>
<tbody>
<tr>
<td>equals</td>
<td>string, number, boolean</td>
<td>Attribute matches the value you provide</td>
</tr>
<tr>
<td>notEquals</td>
<td>string, number, boolean</td>
<td>Attribute doesn't match the value you provide</td>
</tr>
<tr>
<td>greaterThan</td>
<td>number</td>
<td>Attribute is greater than the value you provide</td>
</tr>
<tr>
<td>greaterThanOrEquals</td>
<td>number</td>
<td>Attribute is greater than or equal to the value you provide</td>
</tr>
<tr>
<td>lessThan</td>
<td>number</td>
<td>Attribute is less than the value you provide</td>
</tr>
<tr>
<td>lessThanOrEquals</td>
<td>number</td>
<td>Attribute is less than or equal to the value you provide</td>
</tr>
<tr>
<td>in</td>
<td>list of strings</td>
<td>Attribute is in the list you provide</td>
</tr>
<tr>
<td>notIn</td>
<td>list of strings</td>
<td>Attribute isn't in the list you provide</td>
</tr>
</tbody>
</table>
### Field and Supported value data types

<table>
<thead>
<tr>
<th>Field</th>
<th>Supported value data types</th>
<th>Filtered results</th>
</tr>
</thead>
<tbody>
<tr>
<td>startsWith</td>
<td>string</td>
<td>Attribute starts with the string you provide (only supported for Amazon OpenSearch Serverless vector stores)</td>
</tr>
</tbody>
</table>

To combine filter types, you can use one of the following logical operators:

<table>
<thead>
<tr>
<th>Field</th>
<th>Maps to</th>
<th>Filtered results</th>
</tr>
</thead>
<tbody>
<tr>
<td>andAll</td>
<td>List of up to 5 filter types</td>
<td>Results fulfill all of the filtering expressions in the group</td>
</tr>
<tr>
<td>orAll</td>
<td>List of up to 5 filter types</td>
<td>Results fulfill at least one of the filtering expressions in the group</td>
</tr>
</tbody>
</table>

For examples, see [Send a query and include filters (Retrieve)](Send_a_query_and_include_filters_Retrieve) and [Send a query and include filters (RetrieveAndGenerate)](Send_a_query_and_include_filters_RetrieveAndGenerate).

### Knowledge base prompt template

When you query a knowledge base and request response generation, Amazon Bedrock uses a prompt template that combines instructions and context with the user query to construct the prompt that's sent to the model for response generation. You can engineer the prompt template with the following tools:

- **Prompt placeholders** – Pre-defined variables in Knowledge bases for Amazon Bedrock that are dynamically filled in at runtime during knowledge base query. In the system prompt, you'll see these placeholders surrounded by the $ symbol. The following list describes the placeholders you can use:
<table>
<thead>
<tr>
<th>Variable</th>
<th>Replaced by</th>
<th>Model</th>
<th>Required?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$query$</td>
<td>The user query sent to the knowledge base.</td>
<td>Anthropic Claude Instant, Anthropic Claude v2.x</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anthropic Claude 3 Sonnet</td>
<td>No (automatically included in model input)</td>
</tr>
<tr>
<td>$search_results$</td>
<td>The retrieved results for the user query.</td>
<td>All</td>
<td>Yes</td>
</tr>
<tr>
<td>$output_format_instructions$</td>
<td>Underlying instructions for formatting the response generation and citations. Differs by model. If you define your own formatting instructions, we suggest that you remove this placeholder. Without this placeholder, the response won't contain citations.</td>
<td>All</td>
<td>No</td>
</tr>
<tr>
<td>$current_time$</td>
<td>The current time.</td>
<td>All</td>
<td>No</td>
</tr>
</tbody>
</table>

**XML tags** – Anthropic models support the use of XML tags to structure and delineate your prompts. Use descriptive tag names for optimal results. For example, in the default system prompt, you'll see the `<database>` tag used to delineate a database of previously asked questions. For more information, see [Use XML tags](#) in the [Anthropic user guide](#).

For general prompt engineering guidelines, see [Prompt engineering guidelines](#).
Select the tab corresponding to your method of choice and follow the steps.

Console

Follow the console steps at Query the knowledge base and return results or generate responses. In the test window, turn on Generate responses. Then, in the Configurations pane, expand the Knowledge base prompt template section.

1. Choose Edit.
2. Edit the system prompt in the text editor, including prompt placeholders and XML tags as necessary. To revert to the default prompt template, choose Reset to default.
3. When you're finished editing, choose Save changes. To exit without saving the system prompt, choose Discard changes.

API

When you make a RetrieveAndGenerate request, include a generationConfiguration field, mapped to a GenerationConfiguration object. To see the location of this field, refer to the RetrieveAndGenerate request body in the API reference.

The following JSON object shows the minimal fields required in the GenerationConfiguration object to set the maximum number of source chunks to return:

```json
"generationConfiguration": {
  "promptTemplate": {
    "textPromptTemplate": "string"
  }
}
```

Enter your custom prompt template in the textPromptTemplate field, including prompt placeholders and XML tags as necessary. For the maximum number of characters allowed in the system prompt, see the textPromptTemplate field in GenerationConfiguration.

Manage a data source

After you create a data source, you can view details about it, update it, or delete it.
View information about a data source

You can view information about your data source and its sync history. Select the tab corresponding to your method of choice and follow the steps.

Console

To view information about a data source

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select Knowledge base.
3. In the Data source section, select the data source for which you want to view details.
4. The Data source overview contains details about the data source.
5. The Sync history contains details about when the data source was synced. To see reasons for why a sync event failed, select a sync event and choose View warnings.

API

To get information about a data source, send a GetDataSource request with a Agents for Amazon Bedrock build-time endpoint and specify the dataSourceId and the knowledgeBaseId of the knowledge base that it belongs to.

To list information about a knowledge base's data sources, send a ListDataSources request with a Agents for Amazon Bedrock build-time endpoint and specify the ID of the knowledge base.

- To set the maximum number of results to return in a response, use the maxResults field.
- If there are more results than the number you set, the response returns a nextToken. You can use this value in another ListDataSources request to see the next batch of results.

To get information a sync event for a data source, send a GetIngestionJob request with a Agents for Amazon Bedrock build-time endpoint. Specify the dataSourceId, knowledgeBaseId, and ingestionJobId.

To list the sync history for a data source in a knowledge base, send a ListIngestionJobs request with a Agents for Amazon Bedrock build-time endpoint. Specify the ID of the knowledge base and data source. You can set the following specifications.
• Filter for results by specifying a status to search for in the filters object.

• Sort by the time that the job was started or the status of a job by specifying the sortBy object. You can sort in ascending or descending order.

• Set the maximum number of results to return in a response in the maxResults field. If there are more results than the number you set, the response returns a nextToken that you can send in another ListIngestionJobs request to see the next batch of jobs.

Update a data source

You can update a data source in the following ways:

• Add, change, or remove files from the S3 bucket that contains the files for the data source.

• Change the name or S3 bucket for the data source, or the KMS key to use for encrypting transient data during data ingestion.

Each time you add, modify, or remove files from the S3 bucket for a data source, you must sync the data source so that it is re-indexed to the knowledge base. Syncing is incremental, so Amazon Bedrock only processes the objects in your S3 bucket that have been added, modified, or deleted since the last sync. Before you begin ingestion, check that your data source fulfills the following conditions:

• The files are in supported formats. For more information, see Set up a vector index for your knowledge base in a supported vector store.

• The files don't exceed the maximum file size of 50 MB. For more information, see Knowledge base quotas.

• If your data source contains metadata files, check the following conditions to ensure that the metadata files aren't ignored:
  • Each .metadata.json file shares the same name as the source file that it's associated with.
  • If the vector index for your knowledge base is in an Amazon OpenSearch Serverless vector store, check that the vector index is configured with the faiss engine. If the vector index is configured with the nmslib engine, you'll have to do one of the following:
    • Create a new knowledge base in the console and let Amazon Bedrock automatically create a vector index in Amazon OpenSearch Serverless for you.
    • Create another vector index in the vector store and select faiss as the Engine. Then create a new knowledge base and specify the new vector index.
If the vector index for your knowledge base is in an Amazon Aurora database cluster, check that the table for your index contains a column for each metadata property in your metadata files before starting ingestion.

To learn how to update a data source, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To update a data source**

1. (Optional) Make the necessary changes to the files in the S3 bucket that contains the files for the data source.
2. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
3. From the left navigation pane, select **Knowledge base**.
4. In the **Data source** section, select the radio button next to the data source that you want to sync.
5. (Optional) Choose **Edit**, change any configurations necessary, and select **Submit**.
6. Choose **Sync**.
7. A green banner appears when the sync is complete and the **Status** becomes **Ready**.

**API**

**To update a data source**

1. (Optional) Make the necessary changes to the files in the S3 bucket that contains the files for the data source.
2. (Optional) Send an **UpdateDataSource** request with a **Agents for Amazon Bedrock build-time endpoint**, changing the necessary configurations and specifying the same configurations you don't want to change.

**Note**

You can't change the chunkingConfiguration. Send the request with the existing chunkingConfiguration.
3. Send a StartIngestionJob request with a Agents for Amazon Bedrock build-time endpoint, specifying the dataSourceId and the knowledgeBaseId.

Delete a data source

If you no longer need a data source, you can delete it. Select the tab corresponding to your method of choice and follow the steps.

Console

To delete a data source

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select Knowledge base.
3. In the Data source section, select the radio button next to the data source that you want to delete.
4. Choose Delete.
5. A green banner appears when the data source is successfully deleted.

API

To delete a data source from a knowledge base, send a DeleteDataSource request, specifying the dataSourceId and knowledgeBaseId.

Manage a knowledge base

After you set up a knowledge base, you can view information about it, modify it, or delete it. Select the tab corresponding to your method of choice and follow the steps.

View information about a knowledge base

You can view information about a knowledge base. Select the tab corresponding to your method of choice and follow the steps.
Console

To view information about a knowledge base

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select Knowledge base.
3. To view details for a knowledge base, either select the Name of the source or choose the radio button next to the source and select Edit.
4. On the details page, you can carry out the following actions:
   - To change the details of the knowledge base, select Edit in the Knowledge base overview section.
   - To update the tags attached to the knowledge base, select Manage tags in the Tags section.
   - If you update the data source from which the knowledge base was created and need to sync the changes, select Sync in the Data source section.
   - To view the details of a data source, select a Data source name. Within the details, you can choose the radio button next to a sync event in the Sync history section and select View warnings to see why files in the data ingestion job failed to sync.
   - To manage the embeddings model used for the knowledge base, select Edit Provisioned Throughput.
   - Select Save changes when you are finished editing.

API

To get information about a knowledge base, send a GetKnowledgeBase request with a Agents for Amazon Bedrock build-time endpoint, specifying the knowledgeBaseId.

To list information about your knowledge bases, send a ListKnowledgeBases request with a Agents for Amazon Bedrock build-time endpoint. You can set the maximum number of results to return in a response. If there are more results than the number you set, the response returns a nextToken. You can use this value in the nextToken field of another ListKnowledgeBases request to see the next batch of results.
Update a knowledge base

Console

To update a knowledge base

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select Knowledge base.
3. Select a knowledge base to view details about it, or choose the radio button next to the knowledge base and select Edit.
4. You can modify the knowledge base in the following ways.
   - Change configurations for the knowledge base by choosing Edit in the Knowledge base overview section.
   - Change the tags attached to the knowledge base by choosing Manage tags in the Tags section
   - Manage the data source in the Data source section. For more information, see Manage a data source.
5. Select Save changes when you are finished editing.

API

To update a knowledge base, send an UpdateKnowledgeBase request with a Agents for Amazon Bedrock build-time endpoint. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same.

Delete a knowledge base

If you no longer need a knowledge base, you can delete it. When you delete a knowledge base, you should also carry out the following actions to fully delete all resources associated with the knowledge base.

- Dissociate the knowledge base from any agents it is associated with.
• The underlying data that was indexed from your knowledge base remains in the vector store you set up and can still be retrieved. To delete the data, you also need to delete the vector index containing the data embeddings.

Select the tab corresponding to your method of choice and follow the steps.

**Console**

**To delete a knowledge base**

1. Before the following steps, make sure to delete the knowledge base from any agents that it's associated with. To do this, carry out the following steps:
   a. From the left navigation pane, select **Agents**.
   b. Choose the **Name** of the agent that you want to delete the knowledge base from.
   c. A red banner appears to warn you to delete the reference to the knowledge base, which no longer exists, from the agent.
   d. Select the radio button next to the knowledge base that you want to remove. Select **More** and then choose **Delete**.

2. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).

3. From the left navigation pane, select **Knowledge base**.

4. Choose a knowledge base or select the radio button next to a knowledge base. Then choose **Delete**.

5. Review the warnings for deleting a knowledge base. If you accept these conditions, enter **delete** in the input box and select **Delete** to confirm.

6. To fully delete the vector embeddings for your knowledge base, you need to delete the vector index containing the data embeddings.

**API**

Before deleting a knowledge base, disassociate the knowledge base from any agents that it is associated with by making a [DisassociateAgentKnowledgeBase](https://console.aws.amazon.com/bedrock/) request with a [Agents for Amazon Bedrock build-time endpoint](https://console.aws.amazon.com/bedrock/).
To delete the knowledge base, send a `DeleteKnowledgeBase` request with a [Agents for Amazon Bedrock build-time endpoint](https://console.aws.amazon.com/bedrock/).

To fully delete the vector embeddings for your knowledge base, you need to delete the vector index containing the data embeddings.

## Deploy a knowledge base

To deploy a knowledge base in your application, set it up to make `Retrieve` or `RetrieveAndGenerate` requests to the knowledge base. To see how to use these API operations, select the API tab in [Test a knowledge base in Amazon Bedrock](https://console.aws.amazon.com/bedrock/).

You can also associate the knowledge base with an agent and the agent will invoke it when necessary during orchestration. For more information, see [Agents for Amazon Bedrock](https://console.aws.amazon.com/bedrock/). Select the tab corresponding to your method of choice and follow the steps.

### Console

**To associate a knowledge base with an agent**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select **Agents**.
3. Choose the agent to which you want to add a knowledge base.
4. In the Working draft section, choose **Working draft**.
5. In the Knowledge bases section, select **Add**.
6. Choose a knowledge base from the dropdown list under **Select knowledge base** and specify the instructions for the agent regarding how it should interact with the knowledge base and return results.

**To dissociate a knowledge base with an agent**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, select **Agents**.
3. Choose the agent to which you want to add a knowledge base.
4. In the **Working draft** section, choose **Working draft**.
5. In the **Knowledge bases** section, choose a knowledge base.
6. Select **Delete**.

**API**

To associate a knowledge base with an agent, send an [AssociateAgentKnowledgeBase](#) request.

- Include a detailed description to provide instructions for how the agent should interact with the knowledge base and return results.
- Set the `knowledgeBaseState` to `ENABLED` to allow the agent to query the knowledge base.

You can update an knowledge base that is associated with an agent by sending an [UpdateAgentKnowledgeBase](#) request. For example, you might want to set the `knowledgeBaseState` to `ENABLED` to troubleshoot an issue. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same.

To dissociate a knowledge base with an agent, send a [DisassociateAgentKnowledgeBase](#) request.
Agents for Amazon Bedrock

Agents for Amazon Bedrock offers you the ability to build and configure autonomous agents in your application. An agent helps your end-users complete actions based on organization data and user input. Agents orchestrate interactions between foundation models (FMs), data sources, software applications, and user conversations. In addition, agents automatically call APIs to take actions and invoke knowledge bases to supplement information for these actions. Developers can save weeks of development effort by integrating agents to accelerate the delivery of generative artificial intelligence (generative AI) applications.

With agents, you can automate tasks for your customers and answer questions for them. For example, you can create an agent that helps customers process insurance claims or an agent that helps customers make travel reservations. You don't have to provision capacity, manage infrastructure, or write custom code. Amazon Bedrock manages prompt engineering, memory, monitoring, encryption, user permissions, and API invocation.

Agents perform the following tasks:

- Extend foundation models to understand user requests and break down the tasks that the agent must perform into smaller steps.
- Collect additional information from a user through natural conversation.
- Take actions to fulfill a customer's request by making API calls to your company systems.
- Augment performance and accuracy by querying data sources.

To use an agent, you perform the following steps:

1. (Optional) Create a knowledge base to store your private data in that database. For more information, see Knowledge bases for Amazon Bedrock.
2. Configure an agent for your use case and add actions the agent can perform. To define how the agent handles the actions, write Lambda functions in a programming language of your choice.
3. Associate a knowledge base with the agent to augment the agent's performance. For more information, see Create an agent in Amazon Bedrock.
4. (Optional) To customize the agent's behavior to your specific use-case, modify prompt templates for the pre-processing, orchestration, knowledge base response generation, and
post-processing steps that the agent performs. For more information, see Advanced prompts in Amazon Bedrock.

5. Test your agent in the Amazon Bedrock console or through API calls to the TSTALIASID. Modify the configurations as necessary. Use traces to examine your agent's reasoning process at each step of its orchestration. For more information, see Test an Amazon Bedrock agent and Trace events in Amazon Bedrock.

6. When you have sufficiently modified your agent and it's ready to be deployed to your application, create an alias to point to a version of your agent. For more information, see Deploy an Amazon Bedrock agent.

7. Set up your application to make API calls to your agent alias.

8. Iterate on your agent and create more versions and aliases as necessary.

Topics

- How Agents for Amazon Bedrock works
- Supported regions and models for Agents for Amazon Bedrock
- Prerequisites for Agents for Amazon Bedrock
- Create an agent in Amazon Bedrock
- Create an action group for an Amazon Bedrock agent
- Associate a knowledge base with an Amazon Bedrock agent
- Test an Amazon Bedrock agent
- Manage an Amazon Bedrock agent
- Customize an Amazon Bedrock agent
- Deploy an Amazon Bedrock agent

How Agents for Amazon Bedrock works

Agents for Amazon Bedrock consists of the following two main sets of API operations to help you set up and run an agent:

- **Build-time API operations** to create, configure, and manage your agents and their related resources
- **Runtime API operations** to invoke your agent with user input and to initiate orchestration to carry out a task.
Build-time configuration

An agent consists of the following components:

- **Foundation model** – You choose a foundation model (FM) that the agent invokes to interpret user input and subsequent prompts in its orchestration process. The agent also invokes the FM to generate responses and follow-up steps in its process.

- **Instructions** – You write instructions that describe what the agent is designed to do. With advanced prompts, you can further customize instructions for the agent at every step of orchestration and include Lambda functions to parse each step's output.

- **Action groups** (Optional) – You define the actions that the agent should perform through providing the following resources:
  - An OpenAPI schema to define the API operations that the agent can invoke to perform its tasks.
  - A Lambda function with the following input and output:
    - Input – The API operation and parameters identified during orchestration.
    - Output – The result of the API invocation.

- **Knowledge bases** (Optional) – Associate knowledge bases with an agent. The agent queries the knowledge base for extra context to augment response generation and input into steps of the orchestration process.

- **Prompt templates** – Prompt templates are the basis for creating prompts to be provided to the FM. Agents for Amazon Bedrock exposes the default four base prompt templates that are used during the pre-processing, orchestration, knowledge base response generation, and post-processing. You can optionally edit these base prompt templates to customize your agent's behavior at each step of its sequence. You can also turn off steps for troubleshooting purposes or if you decide that a step is unnecessary. For more information, see [Advanced prompts in Amazon Bedrock](#).

At build-time, all these components are gathered to construct base prompts for the agent to perform orchestration until the user request is completed. With advanced prompts, you can modify these base prompts with additional logic and few-shot examples to improve accuracy for each step of agent invocation. The base prompt templates contain instructions, action descriptions, knowledge base descriptions, and conversation history, all of which you can customize to modify the agent to meet your needs. You then **prepare** your agent, which packages all the components of the agents, including security configurations. Preparing the agent brings it into a state where it
can be tested in runtime. The following image shows how build-time API operations construct your agent.

**Runtime process**

Runtime is managed by the `InvokeAgent` API operation. This operation starts the agent sequence, which consists of the following three main steps.

1. **Pre-processing** – Manages how the agent contextualizes and categorizes user input and can be used to validate input.

2. **Orchestration** – Interprets the user input, invokes action groups and queries knowledge bases, and returns output to the user or as input to continued orchestration. Orchestration consists of the following steps:
a. The agent interprets the input with a foundation model and generates a *rationale* that lays out the logic for the next step it should take.

b. The agent invokes action groups and queries knowledge bases (*Knowledge base response generation*) to retrieve additional context and summarize the data to augment its generation.

c. The agent generates an output, known as an *observation*, from invoking action groups and summarizing results from knowledge bases. The agent uses the observation to augment the base prompt, which is then interpreted with a foundation model. The agent then determines if it needs to reiterate the orchestration process.

d. This loop continues until the agent returns a response to the user or until it needs to prompt the user for extra information.

During orchestration, the base prompt template is augmented with the agent instructions, action groups, and knowledge bases that you added to the agent. Then, the augmented base prompt is used to invoke the FM. The FM predicts the best possible steps and trajectory to fulfill the user input. At each iteration of orchestration, the FM predicts the API operation to invoke or the knowledge base to query.

3. **Post-processing** – The agent formats the final response to return to the user. This step is turned off by default.

When you invoke your agent, you can turn on a *trace* at runtime. With the trace, you can track the agent's rationale, actions, queries, and observations at each step of the agent sequence. The trace includes the full prompt sent to the foundation model at each step and the outputs from the foundation model, API responses, and knowledge base queries. You can use the trace to understand the agent's reasoning at each step. For more information, see [Trace events in Amazon Bedrock](#)

As the user session with the agent continues through more `InvokeAgent` requests, the conversation history is preserved. The conversation history continually augments the orchestration base prompt template with context, helping improve the agent's accuracy and performance. The following diagram shows the agent's process during runtime:
Supported regions and models for Agents for Amazon Bedrock

Agents for Amazon Bedrock is supported in the following regions:
You can use Agents for Amazon Bedrock with the following models:

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropic Claude Instant v1</td>
<td>anthropic.claude-instant-v1</td>
</tr>
<tr>
<td>Anthropic Claude v2.0</td>
<td>anthropic.claude-v2</td>
</tr>
<tr>
<td>Anthropic Claude v2.1</td>
<td>anthropic.claude-v2:1</td>
</tr>
</tbody>
</table>

Prerequisites for Agents for Amazon Bedrock

Ensure that your IAM role has the necessary permissions to perform actions related to Agents for Amazon Bedrock.

An agent uses action groups and knowledge bases to help your customer perform tasks. Following is a short description of each type of resource:

- **Action group** – Defines an action that the agent can help the user perform. Includes the APIs that can be called, how to handle the action, and how to return the response.
- **Knowledge base** – Provides a repository of information that the agent can query to answer customer queries and improve its generated responses.

Before creating an agent, review the following prerequisites and determine which ones you need to fulfill:

1. **Set up an action group.** For your agent to orchestrate API calls to your systems, you must add at least one action group. You can skip this prerequisite if you want to add an action group later or if you plan to have no action groups for your agent.
2. **Set up a knowledge base.** To improve responses to customer queries by using private data sources, you can associate at least one knowledge base. You can skip this prerequisite if you plan for your agent to have no knowledge bases associated with it.

3. **Create a custom AWS Identity and Access Management (IAM) service role for your agent with the proper permissions.** You can skip this prerequisite if you plan to use the AWS Management Console to automatically create a service role for you.

## Create an agent in Amazon Bedrock

To create an agent with Amazon Bedrock, you set up the following components:

- The configuration of the agent, which defines the purpose of the agent and indicates the foundation model (FM) that it uses to generate prompts and responses.
- (Optional) Action groups that define what actions the agent is designed to perform.
- (Optional) A knowledge base of data sources to augment the generative capabilities of the agent.

You can create an agent in either the console or the API. Select the tab corresponding to your method of choice and follow the steps.

**Console**

Depending on how

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane.
3. In the **Agents** section, choose **Create Agent**.
4. Choose **Create**.
5. You will be taken to the **Agent Builder** for your newly created agent, where you can configure your agent.
6. In the **Agent details** section, you can set up the following configurations:
   a. You can edit the **Name** or **Description**.
   b. For **Select model**, choose **Change** to select an FM for your agent to invoke during orchestration.
c. In **Instructions for the Agent**, enter details to tell the agent what it should do and how it should interact with users. The instructions replace the $instructions$ placeholder in the **orchestration prompt template**. Following is an example of instructions:

```
You are an office assistant in an insurance agency. You are friendly and polite. You help with managing insurance claims and coordinating pending paperwork.
```

d. If you expand **Additional settings**, you can modify the following configurations:

**User input** – Choose whether to allow the agent to request more information from the user if it doesn't have enough information.

- If you choose **Yes**, the agent returns an **Observation** reprompting the user for more information if it needs to invoke an API in an action group, but doesn't have enough information to complete the API request.

- If you choose **No**, the agent doesn't request the user for additional details and instead informs the user that it doesn't have enough information to complete the task.

- **Idle session timeout** – By default, if a user hasn't responded for 30 minutes in a session with a Amazon Bedrock agent, the agent no longer maintains the conversation history. Conversation history is used to both resume an interaction and to augment responses with context from the conversation. To change this default length of time, enter a number in the **Session timeout** field and choose a unit of time.

e. When you are done setting up the agent configuration, select **Next**.

7. In the **Action groups** section, you can add action groups to your agent. For more information on setting up action groups, see the section called “Create an action group”. To add an action group, choose **Create** or **Create action group**.

a. In the **Action group invocation** section, you set up what the agent does after it predicts the API or function that it should invoke and the parameters that it must pass. You can do this by defining a Lambda function, which provides the business logic that's performed upon invoking the action. Choose one of the following options for the Lambda function:
b. To set up another action group for your agent, select **Add another action group**. When you're done with adding action groups, choose **Next**.

**API**

To create an agent, send a [CreateAgent](#) request (see link for request and response formats and field details) with an [Agents for Amazon Bedrock build-time endpoint](#).

**See code examples**

The following list describes the fields in the request:

- Minimally, provide the following required fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>agentName</td>
<td>Name for the agent</td>
</tr>
<tr>
<td>agentResourceRoleArn</td>
<td>ARN of the service role for the agent</td>
</tr>
<tr>
<td>foundationModel</td>
<td>Foundation model for the agent to orchestrate with</td>
</tr>
</tbody>
</table>

- The following fields are optional but recommended:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>Describes what the agent does</td>
</tr>
<tr>
<td>instruction</td>
<td>Instructions to tell the agent what to do. Used in the orchestration prompt template.</td>
</tr>
</tbody>
</table>

- The following fields can be specified for security purposes:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>clientToken</td>
<td>Identifier to ensure the API request completes only once.</td>
</tr>
<tr>
<td>Field</td>
<td>Short description</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------------------------------------------------</td>
</tr>
<tr>
<td>customerEncryptionKeyArn</td>
<td>ARN of a KMS key to encrypt agent resources</td>
</tr>
<tr>
<td>idleSessionTTLInSeconds</td>
<td>Duration after which the agent ends the session and deletes any stored information.</td>
</tr>
</tbody>
</table>

- To customize your agent's behavior by overriding the default prompt templates, include a promptOverrideConfiguration object. For more information, see Advanced prompts in Amazon Bedrock.
- To attach tags to your agent, use the Tags field. For more information, see Tag resources.

If your agent fails to be created, the CreateAgent object in the response returns a list of failureReasons and a list of recommendedActions for you to troubleshoot.

Create an action group for an Amazon Bedrock agent

An action group defines an action that the agent can help the user perform. You define an action group by specifying APIs in your systems that can be called and by writing a Lambda function to handle the action and how to return the response. To create an action group, prepare the following components:

- **Set up an OpenAPI schema** with the API description, structure, and parameters for the action group. You can add the API schema to the action group in one of the following ways:
  - Upload the schema that you create to an Amazon Simple Storage Service (Amazon S3) bucket.
  - Write the schema in the inline OpenAPI schema editor in the AWS Management Console when you add the action group. This option is only available after the agent that the action group belongs to has already been created.
- **Create a Lambda function** that defines the business logic for the action group.

To learn more about the components of an action group and how to create the action group after you set it up, select from the following topics:

**Topics**
- Define OpenAPI schemas for your agent's action groups in Amazon Bedrock
Define Lambda functions for your agent's action groups in Amazon Bedrock

Add an action group to your agent in Amazon Bedrock

**Define OpenAPI schemas for your agent's action groups in Amazon Bedrock**

When you create an action group in Amazon Bedrock, you must define the API operations that the agent can invoke. To define the API operations, create an OpenAPI schema in JSON or YAML format. You can create OpenAPI schema files and upload them to Amazon Simple Storage Service (Amazon S3). Alternatively, you can use the OpenAPI text editor in the console, which will validate your schema. After you create an agent, you can use the text editor when you add an action group to the agent or edit an existing action group. For more information, see [Edit an agent](https://docs.aws.amazon.com/bedrock/latest/userguide/edit-agent.html).

For more information about API schemas, see the following resources:

- For more details about OpenAPI schemas, see [OpenAPI specification](https://swagger.io/) on the Swagger website.
- For best practices in writing API schemas, see [Best practices in API design](https://swagger.io/) on the Swagger website.

The following is the general format of an OpenAPI schema for an action group.

```json
{
   "openapi": "3.0.0",
   "paths": {
      "/path": {
         "method": {
            "description": "string",
            "operationId": "string",
            "parameters": [ ... ],
            "requestBody": { ... },
            "responses": { ... }
         }
      }
   }
}
```

The following list describes fields in the OpenAPI schema.
• **openapi** – (Required) The version of OpenAPI that's being used. This value must be "3.0.0" or higher for the action group to work.

• **paths** – (Required) Contains relative paths to individual endpoints. Each path must begin with a forward slash (/).

• **method** – (Required) Defines the method to use.

Minimally, each method requires the following fields:

• **description** – A description of the API operation. Use this field to inform the agent when to call this API operation and what the operation does.

• **responses** – Contains properties that the agent returns in the API response. The agent uses the response properties to construct prompts, accurately process the results of an API call, and determine a correct set of steps for performing a task. The agent can use response values from one operation as inputs for subsequent steps in the orchestration.

The fields within the following two objects provide more information for your agent to effectively take advantage of your action group. For each field, set the value of the `required` field to `true` if required and to `false` if optional.

• **parameters** – Contains information about parameters that can be included in the request.

• **requestBody** – Contains the fields in the request body for the operation. Don't include this field for GET and DELETE methods.

To learn more about a structure, select from the following tabs.

**responses**

```json
"responses": {
  "200": {
    "content": {
      "<media type>": {
        "schema": {
          "properties": {
            "<property>": {
              "type": "string",
              "description": "string"
            }
          }
        }
      }
    }
  }
}
```
Each key in the `responses` object is a response code, which describes the status of the response. The response code maps to an object that contains the following information for the response:

- **content** – (Required for each response) The content of the response.
- **<media type>** – The format of the response body. For more information, see [Media types](#) on the Swagger website.
- **schema** – (Required for each media type) Defines the data type of the response body and its fields.
- **properties** – (Required if there are items in the schema) Your agent uses properties that you define in the schema to determine the information it needs to return to the end user in order to fulfill a task. Each property contains the following fields:
  - **type** – (Required for each property) The data type of the response field.
  - **description** – (Optional) Describes the property. The agent can use this information to determine the information that it needs to return to the end user.

### parameters

```json
"parameters": [ 
    
    
    ]
```
Your agent uses the following fields to determine the information it must get from the end user to perform the action group's requirements.

- **name** – (Required) The name of the parameter.
- **description** – (Required) A description of the parameter. Use this field to help the agent understand how to elicit this parameter from the agent user or determine that it already has that parameter value from prior actions or from the user's request to the agent.
- **required** – (Optional) Whether the parameter is required for the API request. Use this field to indicate to the agent whether this parameter is needed for every invocation or if it's optional.
- **schema** – (Optional) The definition of input and output data types. For more information, see [Data Models (Schemas)](https://swagger.io) on the Swagger website.

**requestBody**

Following is the general structure of a `requestBody` field:

```json
"requestBody": {  
  "required": boolean,  
  "content": {  
    "<media type>": {  
      "schema": {  
        "properties": {  
          "<property>": {  
            "type": "string",  
            "description": "string"  
          },  
          ...  
        }  
      }  
    }  
  }  
}
```

The following list describes each field:

- **required** – (Optional) Whether the request body is required for the API request.
- **content** – (Required) The content of the request body.
• **<media type>** – (Optional) The format of the request body. For more information, see [Media types](#) on the Swagger website.

• **schema** – (Optional) Defines the data type of the request body and its fields.

• **properties** – (Optional) Your agent uses properties that you define in the schema to determine the information it must get from the end user to make the API request. Each property contains the following fields:
  
  • **type** – (Optional) The data type of the request field.

  • **description** – (Optional) Describes the property. The agent can use this information to determine the information it needs to return to the end user.


## Example API schema

The following example API schema defines a group of API operations that help handle insurance claims. Three APIs are defined as follows:

• **getAllOpenClaims** – Your agent can use the description field to determine that it should call this API operation if a list of open claims is needed. The properties in the responses specify to return the ID and the policy holder and the status of the claim. The agent returns this information to the agent user or uses some or all of the response as input to subsequent API calls.

• **identifyMissingDocuments** – Your agent can use the description field to determine that it should call this API operation if missing documents must be identified for an insurance claim. The name, description, and required fields tell the agent that it must elicit the unique identifier of the open claim from the customer. The properties in the responses specify to return the IDs of the open insurance claims. The agent returns this information to the end user or uses some or all of the response as input to subsequent API calls.

• **sendReminders** – Your agent can use the description field to determine that it should call this API operation if there is a need to send reminders to the customer. For example, a reminder about pending documents that they have for open claims. The properties in the requestBody tell the agent that it must find the claim IDs and the pending documents. The properties in the responses specify to return an ID of the reminder and its status. The agent returns this information to the end user or uses some or all of the response as input to subsequent API calls.
"openapi": "3.0.0",
"info": {
    "title": "Insurance Claims Automation API",
    "version": "1.0.0",
    "description": "APIs for managing insurance claims by pulling a list of open claims, identifying outstanding paperwork for each claim, and sending reminders to policy holders."
},
"paths": {
    "/claims": {
        "get": {
            "summary": "Get a list of all open claims",
            "description": "Get the list of all open insurance claims. Return all the open claimIds."
        }
    },
    "operationId": "getAllOpenClaims",
    "responses": {
        "200": {
            "description": "Gets the list of all open insurance claims for policy holders",
            "content": {
                "application/json": {
                    "schema": {
                        "type": "array",
                        "items": {
                            "type": "object",
                            "properties": {
                                "claimId": {
                                    "type": "string",
                                    "description": "Unique ID of the claim."
                                },
                                "policyHolderId": {
                                    "type": "string",
                                    "description": "Unique ID of the policy holder who has filed the claim."
                                },
                                "claimStatus": {
                                    "type": "string",
                                    "description": "The status of the claim. Claim can be in Open or Closed state"
                                }
                            }
                        }
                    }
                }
            }
        }
    }
}


"get": {
   "summary": "Identify missing documents for a specific claim",
   "description": "Get the list of pending documents that need to be uploaded by policy holder before the claim can be processed. The API takes in only one claim id and returns the list of documents that are pending to be uploaded by policy holder for that claim. This API should be called for each claim id",
   "operationId": "identifyMissingDocuments",
   "parameters": [{
      "name": "claimId",
      "in": "path",
      "description": "Unique ID of the open insurance claim",
      "required": true,
      "schema": {
         "type": "string"
      }
   }],
   "responses": {
      "200": {
         "description": "List of documents that are pending to be uploaded by policy holder for insurance claim",
         "content": {
            "application/json": {
               "schema": {
                  "type": "object",
                  "properties": {
                     "pendingDocuments": {
                        "type": "string",
                        "description": "The list of pending documents for the claim."
                     }
                  }
               }
            }
         }
      }
   }
}
"/send-reminders": {
    "post": {
        "summary": "API to send reminder to the customer about pending documents for open claim",
        "description": "Send reminder to the customer about pending documents for open claim. The API takes in only one claim id and its pending documents at a time, sends the reminder and returns the tracking details for the reminder. This API should be called for each claim id you want to send reminders for.",
        "operationId": "sendReminders",
        "requestBody": {
            "required": true,
            "content": {
                "application/json": {
                    "schema": {
                        "type": "object",
                        "properties": {
                            "claimId": {
                                "type": "string",
                                "description": "Unique ID of open claims to send reminders for."
                            },
                            "pendingDocuments": {
                                "type": "string",
                                "description": "The list of pending documents for the claim."
                            }
                        },
                        "required": ["claimId", "pendingDocuments"
                    ]
                }
            }
        }
    },
    "responses": {
        "200": {
            "description": "Reminders sent successfully",
            "content": {
                "application/json": {
                    "schema": {
                        "type": "object",
                        "properties": {
                            "claimId": {
                                "type": "string",
                                "description": "Unique ID of open claims to send reminders for."
                            },
                            "pendingDocuments": {
                                "type": "string",
                                "description": "The list of pending documents for the claim."
                            }
                        },
                        "required": ["claimId", "pendingDocuments"
                    ]
                }
            }
        }
    }
}
Define Lambda functions for your agent's action groups in Amazon Bedrock

You must define a Lambda function to program the business logic for an action group. After a Amazon Bedrock agent determines the API operation that it needs to invoke in an action group, it sends information from the API schema alongside relevant metadata as an input event to the Lambda function. To write your function, you must understand the following components of the Lambda function:

- **Input event** – Contains relevant metadata and populated fields from the request body of the API operation that the agent determines must be called.

• **Response** – Contains relevant metadata and populated fields for the response body returned from the API operation.

You write your Lambda function to define how to handle an action group and to customize how you want the API response to be returned. You use the variables from the input event to define your functions and return a response to the agent.

**Note**

An action group can contain up to 5 API operations, but you can only write one Lambda function. Because the Lambda function can only receive an input event and return a response for one API operation at a time, you should write the function considering the different API operations that may be invoked.

For your agent to use a Lambda function, you must attach a resource-based policy to the function to provide permissions for the agent. For more information, follow the steps at [Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function](#). For more information about resource-based policies in Lambda, see [Using resource-based policies for Lambda](#) in the AWS Lambda Developer Guide.

**Topics**

- [Lambda input event from Amazon Bedrock](#)
- [Lambda response event to Amazon Bedrock](#)
- [Action group Lambda function example](#)

**Lambda input event from Amazon Bedrock**

When an action group using a Lambda function is invoked, Amazon Bedrock sends a Lambda input event of the following general format. You can define your Lambda function to use any of the input event fields to manipulate the business logic within the function to successfully perform the action. For more information about Lambda functions, see [Event-driven invocation](#) in the AWS Lambda Developer Guide.

```json
{
    "messageVersion": "1.0",
    "agent": {
```
The following list describes the input event fields:

- name: string
- id: string
- alias: string
- version: string

inputText: string

sessionId: string

actionGroup: string

apiPath: string

httpMethod: string

parameters: [ ]

requestBody: {
  content: {
    
    "<content_type>": {
      properties: [{
        name: string,
        type: string,
        value: string
      }, ...]
    }
  }
}

sessionAttributes: {
  string: string,
}

promptSessionAttributes: {
  string: string
}
- **messageVersion** – The version of the message that identifies the format of the event data going into the Lambda function and the expected format of the response from a Lambda function. Amazon Bedrock only supports version 1.0.

- **agent** – Contains information about the name, ID, alias, and version of the agent that the action group belongs to.

- **inputText** – The user input for the conversation turn.

- **sessionId** – The unique identifier of the agent session.

- **actionGroup** – The name of the action group.

- **apiPath** – The path to the API operation, as defined in the OpenAPI schema.

- **httpMethod** – The method of the API operation, as defined in the OpenAPI schema.

- **parameters** – Contains a list of objects. Each object contains the name, type, and value of a parameter in the API operation, as defined in the OpenAPI schema.

- **requestBody** – Contains the request body and its properties, as defined in the OpenAPI schema.

- **sessionAttributes** – Contains session attributes and their values. These attributes are stored over a session and provide context for the agent.

- **promptSessionAttributes** – Contains prompt session attributes and their values. These attributes are stored over a turn and provide context for the agent.

### Lambda response event to Amazon Bedrock

Amazon Bedrock expects a response from your Lambda function that matches the following format. The response consists of parameters returned from the API operation. The agent can use the response from the Lambda function for further orchestration or to help it return a response to the customer.

```json
{
    "messageVersion": "1.0",
    "response": {
        "actionGroup": "string",
```
The following list describes the response fields:

- **messageVersion** – The version of the message that identifies the format of the event data going into the Lambda function and the expected format of the response from a Lambda function. Amazon Bedrock only supports version 1.0.
- **response** – Contains the following information about the API response.
  - **actionGroup** – The name of the action group.
  - **apiPath** – The path to the API operation, as defined in the OpenAPI schema.
  - **httpMethod** – The method of the API operation, as defined in the OpenAPI schema.
  - **responseBody** – Contains the response body, as defined in the OpenAPI schema.
- **(Optional) sessionAttributes** – Contains session attributes and their values.
- **(Optional) promptSessionAttributes** – Contains prompt attributes and their values.

### Action group Lambda function example

The following is an minimal example of how the Lambda function can be defined in Python.

```python
def lambda_handler(event, context):
    response_body = {
        'application/json': {
            'body': "sample response"
        }
    }
```
Add an action group to your agent in Amazon Bedrock

After setting up the OpenAPI schema and Lambda function for your action group, you can create the action group. Select the tab corresponding to your method of choice and follow the steps.

Console

When you create an agent, you can add action groups to the working draft.

After an agent is created, you can add action groups to it by doing the following steps:

To add an action group to an already created agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose an agent from the Agents section and then choose the Working draft in the Working Draft section.
4. Select Add in the Action groups section.
5. Fill out the action group details

6. To define the schema for the action group with the in-line OpenAPI schema editor, carry out the following actions. For more information about API schemas for your action group, see Define OpenAPI schemas for your agent's action groups in Amazon Bedrock.

   a. Choose Define with in-line OpenAPI schema editor under Select API schema. A sample schema appears that you can edit.

   b. Select the format for the schema by using the dropdown menu next to Format.

   c. To import an existing schema from S3 to edit, select Import schema, provide the S3 URI, and select Import.

   d. To restore the schema to the original sample schema, select Reset and then confirm the message that appears by selecting Reset again.

7. Select Add. A green success banner appears if there are no issues. If there are issues validating the schema, a red banner appears. The following issues are identified by the validation process.

   - Scroll through the schema to see the lines where an error or warning about formatting exists. An X indicates a formatting error, while an exclamation mark indicates a warning about formatting.
   - Select View details in the red banner to see a list of errors about the content of the API schema.

8. Select Prepare to apply the changes that you have made to the agent before testing it.

API

To create an action group, send a CreateAgentActionGroup request with a Agents for Amazon Bedrock build-time endpoint.

See code examples

The following list describes the fields in the request:

- The following fields are required:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
</table>

Add an action group
<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The following fields are optional:

- Specify the OpenAPI schema defining the APIs to be invoked by the action group in the `apiSchema` object. You can either provide the schema directly as a JSON payload or provide the Amazon S3 bucket containing the OpenAPI schema.

- (Optional) Specify the ARN of the Lambda function to carry out the business logic carried out upon invoking the action.

- Set the `actionGroupState` to `ENABLED` to allow the agent to invoke the action.

- To allow your agent to request the user for additional information when trying to complete a task, add an action group with the `parentActionGroupSignature` field set to `AMAZON.UserInput`. You must leave the `description`, `apiSchema`, and `actionGroupExecutor` fields blank for this action group.

  During orchestration, if your agent determines that it needs to invoke an API in an action group, but doesn't have enough information to complete the API request, it will invoke this action group instead and return an `Observation` reprompting the user for more information.

### Associate a knowledge base with an Amazon Bedrock agent

If you haven’t yet created a knowledge base, see [Knowledge bases for Amazon Bedrock](#) to learn about knowledge bases and create one. You can associate a knowledge base during agent creation or after an agent has been created. To associate a knowledge base to an existing agent, select the tab corresponding to your method of choice and follow the steps.
Console

To add a knowledge base

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. On the agent details page, choose the working draft from the Working draft section.
4. For the Knowledge bases section, choose Add.
5. Choose a knowledge base that you have created and provide instructions for how the agent should interact with it.
6. Choose Add. A success banner appears at the top.
7. To apply the changes that you made to the agent before testing it, choose Prepare before testing it.

API

To associate a knowledge base with an agent, send an AssociateAgentKnowledgeBase request with a Agents for Amazon Bedrock build-time endpoint.

The following list describes the fields in the request:

- The following fields are required:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>agentId</td>
<td>ID of the agent</td>
</tr>
<tr>
<td>agentVersion</td>
<td>Version of the agent</td>
</tr>
<tr>
<td>knowledgeBaseId</td>
<td>ID of the knowledge base</td>
</tr>
</tbody>
</table>

- The following fields are optional:
### Test an Amazon Bedrock agent

After you create an agent, you will have a *working draft*. The working draft is a version of the agent that you can use to iteratively build the agent. Each time you make changes to your agent, the working draft is updated. When you're satisfied with your agent's configurations, you can create a *version*, which is a snapshot of your agent, and an *alias*, which points to the version. You can then deploy your agent to your applications by calling the alias. For more information, see [Deploy an Amazon Bedrock agent](#).

The following list describes how you test your agent:

- In the Amazon Bedrock console, you open up the test window on the side and send input for your agent to respond to. You can select the working draft or a version that you've created.
- In the API, the working draft is the DRAFT version. You send input to your agent by using `InvokeAgent` with the test alias, TSTALIASID, or a different alias pointing to a static version.

To help troubleshoot your agent's behavior, Agents for Amazon Bedrock provides the ability to view the *trace* during a session with your agent. The trace shows the agent's step-by-step reasoning process. For more information about the trace, see [Trace events in Amazon Bedrock](#).

Following are steps for testing your agent. Select the tab corresponding to your method of choice and follow the steps.

**Console**

**To test an agent**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.

3. In the **Agents** section, select the link for the agent that you want to test from the list of agents.

4. The **Test** window appears in a pane on the right.

5. After you create an agent, you must package it with the working draft changes by preparing it in one of the following ways:
   - In the **Test** window, select **Prepare**.
   - In the **Working draft** page, select **Prepare** at the top of the page.

6. To choose an alias and associated version to test, use the dropdown menu at the top of the **Test window**. By default, the **TestAlias: Working draft** combination is selected.

7. To test the agent, enter a message and choose **Run**. While you wait for the response to generate or after it is generated, you have the following options:
   - To view details for each step of the agent's orchestration process, including the prompt, inference configurations, and agent's reasoning process for each step and usage of its action groups and knowledge bases, select **Show trace**. The trace is updated in real-time so you can view it before the response is returned. To expand or collapse the trace for a step, select an arrow next to a step. For more information about the **Trace** window and details that appear, see **Trace events in Amazon Bedrock**.
   - If the agent invokes a knowledge base, the response contains footnotes. To view the link to the S3 object containing the cited information for a specific part of the response, select the relevant footnote.
You can perform the following actions in the **Test** window:

- To start a new conversation with the agent, select the refresh icon.
- To view the **Trace** window, select the expand icon. To close the **Trace** window, select the shrink icon.
- To close the **Test** window, select the right arrow icon.

You can enable or disable action groups and knowledge bases. Use this feature to troubleshoot your agent by isolating which action groups or knowledge bases need to be updated by assessing its behavior with different settings.

**To enable an action group or knowledge base**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. In the **Agents** section, select the link for the agent that you want to test from the list of agents.
4. On the agent's details page, in the **Working draft** section, select the link for the **Working draft**.
5. In the **Action groups** or **Knowledge bases** section, hover over the **State** of the action group or knowledge base whose state you want to change.
6. An edit button appears. Select the edit icon and then choose from the dropdown menu whether the action group or knowledge base is **Enabled** or **Disabled**.
7. If an action group is **Disabled**, the agent doesn't use the action group. If a knowledge base is **Disabled**, the agent doesn't use the knowledge base. Enable or disable action groups or knowledge bases and then use the **Test** window to troubleshoot your agent.
8. Choose **Prepare** to apply the changes that you have made to the agent before testing it.

**API**

Before you test your agent for the first time, you must package it with the working draft changes by sending a **PrepareAgent** request (see link for request and response formats and field
details) with an Agents for Amazon Bedrock build-time endpoint. Include the agentId in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.

See code examples

ℹ️ Note

Every time you update the working draft, you must prepare the agent to package the agent with your latest changes. As a best practice, we recommend that you send a GetAgent request (see link for request and response formats and field details) with a Agents for Amazon Bedrock build-time endpoint and check the preparedAt time for your agent to verify that you’re testing your agent with the latest configurations.

To test your agent, send an InvokeAgent request (see link for request and response formats and field details) with an Agents for Amazon Bedrock runtime endpoint.

See code examples

ℹ️ Note

The AWS CLI doesn’t support InvokeAgent.

See code examples

The following fields exist in the request:

- Minimally, provide the following required fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>agentId</td>
<td>ID of the agent</td>
</tr>
<tr>
<td>agentAliasId</td>
<td>ID of the alias. Use TSTALIASID to invoke the DRAFT version</td>
</tr>
<tr>
<td>sessionId</td>
<td>Alphanumeric ID for the session (2–100 characters)</td>
</tr>
</tbody>
</table>
The following fields are optional:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>enableTrace</td>
<td>Specify TRUE to view the trace.</td>
</tr>
<tr>
<td>endSession</td>
<td>Specify TRUE to end the session with the agent after this request.</td>
</tr>
<tr>
<td>sessionState</td>
<td>Includes sessionAttributes that remain throughout the session and/or promptSessionAttributes that remain throughout the InvokeAgent turn. These attributes give the agent context.</td>
</tr>
</tbody>
</table>

The response is returned in bytes in the chunk object. If the agent queried a knowledge base, the chunk includes citations. If you enabled a trace, a trace object is also returned. If an error occurs, a field is returned with the error message. For more information about how to read the trace, see Trace events in Amazon Bedrock.

**Trace events in Amazon Bedrock**

Each response from an Amazon Bedrock agent is accompanied by a trace that details the steps being orchestrated by the agent. The trace helps you follow the agent's reasoning process that leads it to the response it gives at that point in the conversation.

Use the trace to track the agent's path from the user input to the response it returns. The trace provides information about the inputs to the action groups that the agent invokes and the knowledge bases that it queries to respond to the user. In addition, the trace provides information about the outputs that the action groups and knowledge bases return. You can view the reasoning that the agent uses to determine the action that it takes or the query that it makes to a knowledge base. If a step in the trace fails, the trace returns a reason for the failure. Use the detailed information in the trace to troubleshoot your agent. You can identify steps at which the agent has
trouble or at which it yields unexpected behavior. Then, you can use this information to consider ways in which you can improve the agent’s behavior.

View the trace

The following describes how to view the trace. Select the tab corresponding to your method of choice and follow the steps.

Console

To view the trace during a conversation with an agent

Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.

1. In the Agents section, select the link for the agent that you want to test from the list of agents.
2. The Test window appears in a pane on the right.
3. Enter a message and choose Run. While the response is generating or after it finishes generating, select Show trace.
4. You can view the trace for each Step in real-time as your agent performs orchestration.

API

To view the trace, send an InvokeAgent request with a Agents for Amazon Bedrock runtime endpoint and set the enableTrace field to TRUE. By default, the trace is disabled.

If you enable the trace, in the InvokeAgent response, each chunk in the stream is accompanied by a trace field that maps to a TracePart object. Within the TracePart is a trace field that maps to a Trace object.

Structure of the trace

The trace is shown as a JSON object in both the console and the API. Each Step in the console or Trace in the API can be one of the following traces:

- PreProcessingTrace – Traces the input and output of the pre-processing step, in which the agent contextualizes and categorizes user input and determines if it is valid.
- **Orchestration** – Traces the input and output of the orchestration step, in which the agent interprets the input and invokes API operations and queries knowledge bases. Then the agent returns output to either continue orchestration or to respond to the user.

- **PostProcessingTrace** – Traces the input and output of the post-processing step, in which the agent handles the final output of the orchestration and determines how to return the response to the user.

- **FailureTrace** – Traces the reason that a step failed.

Each of the traces (except FailureTrace) contains a **ModelInvocationInput** object. The **ModelInvocationInput** object contains configurations set in the prompt template for the step, alongside the prompt provided to the agent at this step. For more information about how to modify prompt templates, see Advanced prompts in Amazon Bedrock. The structure of the **ModelInvocationInput** object is as follows:

```json
{
    "traceId": "string",
    "text": "string",
    "type": "PRE_PROCESSING | ORCHESTRATION | KNOWLEDGE_BASE_RESPONSE_GENERATION | POST_PROCESSING",
    "inferenceConfiguration": {
        "maximumLength": number,
        "stopSequences": ["string"],
        "temperature": float,
        "topK": float,
        "topP": float
    },
    "promptCreationMode": "DEFAULT | OVERRIDDEN",
    "parserMode": "DEFAULT | OVERRIDDEN",
    "overrideLambda": "string"
}
```

The following list describes the fields of the **ModelInvocationInput** object:

- **traceId** – The unique identifier of the trace.
- **text** – The text from the prompt provided to the agent at this step.
- **type** – The current step in the agent's process.
- **inferenceConfiguration** – Inference parameters that influence response generation. For more information, see Inference parameters.
- promptCreationMode – Whether the agent's default base prompt template was overridden for this step. For more information, see Advanced prompts in Amazon Bedrock.

- parserMode – Whether the agent's default response parser was overridden for this step. For more information, see Advanced prompts in Amazon Bedrock.

- overrideLambda – The Amazon Resource Name (ARN) of the parser Lambda function used to parse the response, if the default parser was overridden. For more information, see Advanced prompts in Amazon Bedrock.

For more information about each trace type, see the following sections:

PreProcessingTrace

```json
{
    "modelInvocationInput": { // see above for details }
    "modelInvocationOutput": {
        "parsedResponse": {
            "isValid": boolean,
            "rationale": "string"
        },
        "traceId": "string"
    }
}
```

The PreProcessingTrace consists of a ModelInvocationInput object and a PreProcessingModelInvocationOutput object. The PreProcessingModelInvocationOutput contains the following fields.

- parsedResponse – Contains the following details about the parsed user prompt.
  - isValid – Specifies whether the user prompt is valid.
  - rationale – Specifies the agent's reasoning for the next steps to take.
  - traceId – The unique identifier of the trace.

OrchestrationTrace

The Orchestration consists of the ModelInvocationInput object and any combination of the Rationale, InvocationInput, and Observation objects. For more information about each object, select from the following tabs:
Rationale

The Rationale object contains the reasoning of the agent given the user input. Following is the structure:

```
{
    "traceId": "string",
    "text": "string"
}
```

The following list describes the fields of the Rationale object:

- **traceId** – The unique identifier of the trace step.
- **text** – The reasoning process of the agent, based on the input prompt.

InvocationInput

The InvocationInput object contains information that will be input to the action group or knowledge base that is to be invoked or queried. Following is the structure:

```
{
    "traceId": "string",
    "invocationType": "ACTION_GROUP | KNOWLEDGE_BASE | FINISH",
    "actionGroupInvocationInput": {
        "actionGroupName": "string",
        "apiPath": "string",
        "verb": "string",
        "parameters": [
            {
                "name": "string",
                "type": "string",
                "value": "string"
            }
        ]
    }
}
```
The following list describes the fields of the `InvocationInput` object:

- **traceId** – The unique identifier of the trace.
- **invocationType** – Specifies whether the agent is invoking an action group or a knowledge base, or ending the session.
- **actionGroupInvocationInput** – Appears if the type is `ACTION_GROUP`. For more information, see [Define OpenAPI schemas for your agent's action groups in Amazon Bedrock](#). Contains the following input for the action group that is invoked:
  - **actionGroupName** – The name of the action group that the agent will invoke.
  - **apiPath** – The path to the API operation to call, according to the API schema.
  - **verb** – The API method being used, according to the API schema.
  - **parameters** – Contains a list of objects. Each object contains the name, type, and value of a parameter in the API operation, as defined in the API schema.
  - **requestBody** – Contains the request body and its properties, as defined in the API schema.
- **knowledgeBaseLookupInput** – Appears if the type is `KNOWLEDGE_BASE`. For more information, see [Knowledge bases for Amazon Bedrock](#). Contains the following information about the knowledge base and the search query for the knowledge base:
• knowledgeBaseId – The unique identifier of the knowledge base that the agent will look up.

• text – The query to be made to the knowledge base.

Observation

The Observation object contains the result or output of an action group or knowledge base, or the response to the user. Following is the structure:

```
{
    "traceId": "string",
    "type": "ACTION_GROUP | KNOWLEDGE_BASE | REPROMPT | ASK_USER | FINISH"
    "actionGroupInvocation": {
        "text": "JSON-formatted string"
    },
    "knowledgeBaseLookupOutput": {
        "retrievedReferences": [
            {
                "content": {
                    "text": "string"
                },
                "location": {
                    "type": "S3",
                    "s3Location": {
                        "uri": "string"
                    }
                }
            },
            ...
        ]
    },
    "repromptResponse": {
        "source": "ACTION_GROUP | KNOWLEDGE_BASE | PARSER",
        "text": "string"
    },
    "finalResponse": {
        "text"
    }
}
```

The following list describes the fields of the Observation object:
• **traceId** – The unique identifier of the trace.

• **type** – Specifies whether the agent's observation is returned from the result of an action group or a knowledge base, if the agent is reprompting the user, requesting more information, or ending the conversation.

• **actionGroupInvocationOutput** – Contains the JSON-formatted string returned by the API operation that was invoked by the action group. Appears if the type is ACTION_GROUP. For more information, see [Define OpenAPI schemas for your agent's action groups in Amazon Bedrock](#).

• **knowledgeBaseLookupOutput** – Contains text retrieved from the knowledge base that is relevant to responding to the prompt, alongside the Amazon S3 location of the data source. Appears if the type is KNOWLEDGE_BASE. For more information, see [Knowledge bases for Amazon Bedrock](#). Each object in the list of retrievedReferences contains the following fields:
  
  • **content** – Contains text from the knowledge base that is returned from the knowledge base query.
  
  • **location** – Contains the Amazon S3 URI of the data source from which the returned text was found.

• **repromptResponse** – Appears if the type is REPROMPT. Contains the text that asks for a prompt again alongside the source of why the agent needs to reprompt.

• **finalResponse** – Appears if the type is ASK_USER or FINISH. Contains the text that asks the user for more information or is a response to the user.

**PostProcessingTrace**

```json
{
  "modelInvocationInput": { // see above for details }
  "modelInvocationOutput": {
    "parsedResponse": {
      "text": "string"
    },
    "traceId": "string"
  }
}
```

---

**Trace events**

385
The `PostProcessingTrace` consists of a `ModelInvocationInput` object and a `PostProcessingModelInvocationOutput` object. The `PostProcessingModelInvocationOutput` contains the following fields:

- `parsedResponse` – Contains the text to return to the user after the text is processed by the parser function.
- `traceId` – The unique identifier of the trace.

`FailureTrace`

```json
{
  "failureReason": "string",
  "traceId": "string"
}
```

The following list describes the fields of the `FailureTrace` object:

- `failureReason` – The reason that the step failed.
- `traceId` – The unique identifier of the trace.

**Manage an Amazon Bedrock agent**

After you create an agent, you can view or update its configuration as required. The configuration applies to the working draft. If you no longer need an agent, you can delete it.

**Topics**

- [View information about an agent](#)
- [Edit an agent](#)
- [Delete an agent](#)
- [Manage the action groups of an agent](#)
- [Manage agent-knowledge bases associations](#)

**View information about an agent**

To learn how to view information about an agent, select the tab corresponding to your method of choice and follow the steps.
To view information about an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. On the agent details, you can view the following information:

   - The Agent overview section contains the agent configuration.
   - The Tags section contains tags that are associated with the agent. For more information, see Tag resources.
   - The Working draft section contains the working draft. If you select the working draft, you can view the following information:
     - The Model details section contains model and instructions used by the agent's working draft.
     - The Action groups section contains the action groups that the agent uses. For more information, see Create an action group for an Amazon Bedrock agent and Manage the action groups of an agent.
     - The Knowledge bases section contains the knowledge bases associated with the agent. For more information, see Associate a knowledge base with an Amazon Bedrock agent and Manage agent-knowledge bases associations.
     - The Advanced prompts section contains the prompt templates for each step of the agent's orchestration. For more information, see Advanced prompts in Amazon Bedrock.
   - The Versions and Aliases sections contain versions and aliases of the agent that you can use for deployment to your applications. For more information, see Deploy an Amazon Bedrock agent.

API

To get information about an agent, send a GetAgent request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and specify the agentId. See code examples.
To list information about your agents, send a `ListAgents` request (see link for request and response formats and field details) with an `Agents for Amazon Bedrock build-time endpoint`. See code examples. You can specify the following optional parameters:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxResults</td>
<td>The maximum number of results to return in a response.</td>
</tr>
<tr>
<td>nextToken</td>
<td>If there are more results than the number you specified in the maxResults field,</td>
</tr>
<tr>
<td></td>
<td>the response returns a nextToken value.</td>
</tr>
<tr>
<td></td>
<td>To see the next batch of results, send the nextToken value in another request.</td>
</tr>
</tbody>
</table>

To list all the tags for an agent, send a `ListTagsForResource` request (see link for request and response formats and field details) with an `Agents for Amazon Bedrock build-time endpoint` and include the Amazon Resource Name (ARN) of the agent.

**Edit an agent**

To learn how to edit an agent, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To edit the agent configuration**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. In the Agent overview section, choose Edit.
4. Edit the existing information in the fields as necessary.
5. When you're done editing the information, choose Save to remain in the same window or Save and exit to return to the agent details page. A success banner appears at the top. To apply the new configurations to your agent, select Prepare in the banner.
You might want to try different foundation models for your agent or change the instructions for the agent. These changes apply only to the working draft.

To change the foundation model that your agent uses or the instructions to the agent.

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose an agent in the Agents section.
4. On the agent details page, for the Working draft section, choose the working draft.
5. In the Model details section, choose Edit
6. Select a different model or edit the instructions to the agent as necessary.

Note

If you change the foundation model, any prompt templates that you modified will be set to default for that model.

7. When you're done editing the information, choose Save to remain in the same window or Save and exit to return to the agent details page. A success banner appears at the top.
8. To apply the changes that you made to the agent before testing it, choose Prepare in the Test window or at the top of the Working draft page.

To edit the tags associated with an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose an agent in the Agents section.
4. In the Tags section, choose Manage tags.
5. To add a tag, choose Add new tag. Then enter a Key and optionally enter a Value. To remove a tag, choose Remove. For more information, see Tag resources.
6. When you're done editing tags, choose Submit.
To edit an agent, send an UpdateAgent request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same. For more information about required and optional fields, see Create an agent in Amazon Bedrock.

To apply the changes to the working draft, send a PrepareAgent request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Include the agentId in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.

To add tags to an agent, send a TagResource request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and include the Amazon Resource Name (ARN) of the agent. The request body contains a tags field, which is an object containing a key-value pair that you specify for each tag.

To remove tags from an agent, send an UntagResource request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and include the Amazon Resource Name (ARN) of the agent. The tagKeys request parameter is a list containing the keys for the tags that you want to remove.

Delete an agent

To learn how to delete an agent, select the tab corresponding to your method of choice and follow the steps.

Console

To delete an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane.
3. To delete an agent, choose the option button that’s next to the agent you want to delete.
4. A dialog box appears warning you about the consequences of deletion. To confirm that you want to delete the agent, enter delete in the input field and then select Delete.
5. When deletion is complete, a success banner appears.

API

To delete an agent, send a `DeleteAgent` request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and specify the agentId.

By default, the `skipResourceInUseCheck` parameter is `false` and deletion is stopped if the resource is in use. If you set `skipResourceInUseCheck` to `true`, the resource will be deleted even if the resource is in use.

See code examples

Select a topic to learn about how to manage the action groups or knowledge bases for an agent.

Topics

- Manage the action groups of an agent
- Manage agent-knowledge bases associations

Manage the action groups of an agent

After creating an action group, you can view, edit, or delete it. The changes apply to the working draft version of the agent.

Topics

- View information about an action group
- Edit an action group
- Delete an action group

View information about an action group

To learn how to view information about an action group, select the tab corresponding to your method of choice and follow the steps.
To view information about an action group

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose an agent in the Agents section.
4. On the agent details page, for the Working draft section, choose the working draft.
5. In the Action groups section, choose an action group for which to view information.

API

To get information about an action group, send a GetAgentActionGroup request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and specify the actionGroupId, agentId, and agentVersion.

To list information about an agent's action groups, send a ListAgentActionGroups request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Specify the agentId and agentVersion for which you want to see action groups. You can include the following optional parameters:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxResults</td>
<td>The maximum number of results to return in a response.</td>
</tr>
<tr>
<td>nextToken</td>
<td>If there are more results than the number you specified in the maxResults field, the response returns a nextToken value. To see the next batch of results, send the nextToken value in another request.</td>
</tr>
</tbody>
</table>

See code examples
Edit an action group

To learn how to edit an action group, select the tab corresponding to your method of choice and follow the steps.

Console

**To edit an action group**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. Choose an agent in the **Agents** section.
4. On the agent details page, for the **Working draft** section, choose the working draft.
5. In the **Action groups** section, select an action group to edit. Then choose **Edit**.
6. Edit the existing fields as necessary. For more information, see [Create an action group for an Amazon Bedrock agent](#).
7. To define the schema for the action group with the in-line OpenAPI schema editor, for **Select API schema**, choose **Define with in-line OpenAPI schema editor**. A sample schema appears that you can edit. You can configure the following options:
   - To import an existing schema from Amazon S3 to edit, choose **Import schema**, provide the Amazon S3 URI, and select **Import**.
   - To restore the schema to the original sample schema, choose **Reset** and then confirm the message that appears by choosing **Confirm**.
   - To select a different format for the schema, use the dropdown menu labeled **JSON**.
   - To change the visual appearance of the schema, choose the gear icon below the schema.
8. To control whether the agent can use the action group, select **Enable** or **Disable**. Use this function to help troubleshoot your agent's behavior.
9. To remain in the same window so that you can test your change, choose **Save**. To return to the action group details page, choose **Save and exit**.
10. A success banner appears if there are no issues. If there are issues validating the schema, an error banner appears. To see a list of errors, choose **Show details** in the banner.
11. To apply the changes that you made to the agent before testing it, choose **Prepare** in the **Test** window or at the top of the **Working draft** page.
API

To edit an action group, send an `UpdateAgentActionGroup` request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same. You must specify the `agentVersion` as DRAFT. For more information about required and optional fields, see [Create an action group for an Amazon Bedrock agent](#).

To apply the changes to the working draft, send a `PrepareAgent` request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. Include the `agentId` in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.

Delete an action group

To learn how to delete an action group, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To delete an action group**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. Choose an agent in the **Agents** section.
4. On the agent details page, for the **Working draft** section, choose the working draft.
5. In the **Action groups** section, choose the option button that's next to the action group you want to delete.
6. A dialog box appears warning you about the consequences of deletion. To confirm that you want to delete the action group, enter `delete` in the input field and then select **Delete**.
7. When deletion is complete, a success banner appears.
8. To apply the changes that you made to the agent before testing it, choose **Prepare** in the **Test** window or at the top of the **Working draft** page.
API

To delete an action group, send a `DeleteAgentActionGroup` request. Specify the actionGroupId and the agentId and agentVersion from which to delete it. By default, the skipResourceInUseCheck parameter is `false` and deletion is stopped if the resource is in use. If you set skipResourceInUseCheck to `true`, the resource will be deleted even if the resource is in use.

To apply the changes to the working draft, send a `PrepareAgent` request (see link for request and response formats and field details) with an `Agents for Amazon Bedrock build-time endpoint`. Include the agentId in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.

Manage agent-knowledge bases associations

After creating an agent, you can add more knowledge bases or edit them. Adding and editing take place within the working draft. To carry out these operations, choose an agent from the `Agents` section and then choose the `Working draft` in the `Working Draft` section.

Topics

- View information about an agent-knowledge base association
- Edit an agent-knowledge base association
- Disassociate a knowledge base from an agent

View information about an agent-knowledge base association

To learn how to view information about a knowledge base, select the tab corresponding to your method of choice and follow the steps.

Console

To view information about a knowledge base that's associated with an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. Choose an agent in the **Agents** section.
4. On the agent details page, for the **Working draft** section, choose the working draft.

5. In the **Knowledge bases** section, select the knowledge base for which you want to view information.

**API**

To get information about a knowledge base associated with an agent, send a
**GetAgentKnowledgeBase** request (see link for request and response formats and field details)
with an **Agents for Amazon Bedrock build-time endpoint**. Specify the following fields:

To list information about the knowledge bases associated with an agent, send a
**ListAgentKnowledgeBases** request (see link for request and response formats and field
details) with an **Agents for Amazon Bedrock build-time endpoint**. Specify the agentId and
agentVersion for which you want to see associated knowledge bases.

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxResults</td>
<td>The maximum number of results to return in a response.</td>
</tr>
<tr>
<td>nextToken</td>
<td>If there are more results than the number you specified in the maxResults field,</td>
</tr>
<tr>
<td></td>
<td>the response returns a nextToken value.</td>
</tr>
<tr>
<td></td>
<td>To see the next batch of results, send the nextToken value in another request.</td>
</tr>
</tbody>
</table>

[See code examples](#)

**Edit an agent-knowledge base association**

To learn how to edit an agent-knowledge base association, select the tab corresponding to your
method of choice and follow the steps.
To edit an agent-knowledge base association

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. Choose an agent in the **Agents** section.
4. On the agent details page, for the **Working draft** section, choose the working draft.
5. In the **Action groups** section, select an action group to edit. Then choose **Edit**.
6. Edit the existing fields as necessary. For more information, see [Associate a knowledge base with an Amazon Bedrock agent](#).
7. To control whether the agent can use the knowledge base, select **Enabled** or **Disabled**. Use this function to help troubleshoot your agent's behavior.
8. To remain in the same window so that you can test your change, choose **Save**. To return to the **Working draft** page, choose **Save and exit**.
9. To apply the changes that you made to the agent before testing it, choose **Prepare** in the **Test** window or at the top of the **Working draft** page.

API

To edit the configuration of a knowledge base associated with an agent, send an **UpdateAgentKnowledgeBase** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same. You must specify the **agentVersion** as DRAFT. For more information about required and optional fields, see [Associate a knowledge base with an Amazon Bedrock agent](#).

To apply the changes to the working draft, send a **PrepareAgent** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. Include the **agentId** in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.
Disassociate a knowledge base from an agent

To learn how to disassociate a knowledge base from an agent, select the tab corresponding to your method of choice and follow the steps.

Console

To dissociate a knowledge base from an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose an agent in the Agents section.
4. On the agent details page, for the Working draft section, choose the working draft.
5. In the Knowledge bases section, choose the option button that's next to the knowledge base that you want to delete. Then choose Delete.
6. Confirm the message that appears and then choose Delete.
7. To apply the changes that you made to the agent before testing it, choose Prepare in the Test window or at the top of the Working draft page.

API

To disassociate a knowledge base from an agent, send a DisassociateAgentKnowledgeBase request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Specify the knowledgeBaseId and the agentId and agentVersion of the agent from which to disassociate it.

To apply the changes to the working draft, send a PrepareAgent request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Include the agentId in the request. The changes apply to the DRAFT version, which the TSTALIASID alias points to.

Customize an Amazon Bedrock agent

After you have set up your agent, you can further customize its behavior with the following features:
• **Advanced prompts** let you modify prompt templates to determine the prompt that is sent to the agent at each step of runtime.

• **Session state** is a field that contains attributes that you can define in build-time when sending a `CreateAgent` request or send at runtime with an `InvokeAgent` request. You can use these attributes to provide and manage context in a conversation between users and the agent.

Select a topic to learn more about that feature.

**Topics**

• **Advanced prompts in Amazon Bedrock**

• **Control session context**

---

**Advanced prompts in Amazon Bedrock**

After creation, an agent is configured with the following four default **base prompt templates**, which outline how the agent constructs prompts to send to the foundation model at each step of the agent sequence. For details about what each step encompasses, see **Runtime process**.

• Pre-processing

• Orchestration

• Knowledge base response generation

• Post-processing (disabled by default)

Prompt templates define how the agent does the following:

• Processes user input text and output prompts from foundation models (FMs)

• Orchestrates between the FM, action groups, and knowledge bases

• Formats and returns responses to the user

By using advanced prompts, you can enhance your agent's accuracy through modifying these prompt templates to provide detailed configurations. You can also provide hand-curated examples for **few-shot prompting**, in which you improve model performance by providing labeled examples for a specific task.

Select a topic to learn more about advanced prompts.
Topics

- **Advanced prompts terminology**
- **Configure the prompt templates**
- **Placeholder variables in Amazon Bedrock agent prompt templates**
- **Parser Lambda function in Agents for Amazon Bedrock**

Advanced prompts terminology

The following terminology is helpful in understanding how advanced prompts work.

- **Session** – A group of `InvokeAgent` requests made to the same agent with the same session ID. When you make an `InvokeAgent` request, you can reuse a `sessionId` that was returned from the response of a previous call in order to continue the same session with an agent. As long as the `idleSessionTTLInSeconds` time in the `Agent` configuration hasn't expired, you maintain the same session with the agent.

- **Turn** – A single `InvokeAgent` call. A session consists of one or more turns.

- **Iteration** – A sequence of the following actions:
  1. (Required) A call to the foundation model
  2. (Optional) An action group invocation
  3. (Optional) A knowledge base invocation
  4. (Optional) A response to the user asking for more information

  An action might be skipped, depending on the configuration of the agent or the agent's requirement at that moment. A turn consists of one or more iterations.

- **Prompt** – A prompt consists of the instructions to the agent, context, and text input. The text input can come from a user or from the output of another step in the agent sequence. The prompt is provided to the foundation model to determine the next step that the agent takes in responding to user input.

- **Base prompt template** – The structural elements that make up a prompt. The template consists of placeholders that are filled in with user input, the agent configuration, and context at runtime to create a prompt for the foundation model to process when the agent reaches that step. For more information about these placeholders, see [Placeholder variables in Amazon Bedrock agent prompt templates](#). With advanced prompts, you can edit these templates.
Configure the prompt templates

With advanced prompts, you can do the following:

- Turn on or turn off invocation for different steps in the agent sequence.
- Configure their inference parameters.
- Edit the default base prompt templates that the agent uses. By overriding the logic with your own configurations, you can customize your agent's behavior.

For each step of the agent sequence, you can edit the following parts:

- **Prompt template** – Describes how the agent should evaluate and use the prompt that it receives at the step for which you're editing the template. When editing a template, you can engineer the prompt with the following tools:
  - **Prompt template placeholders** – Pre-defined variables in Agents for Amazon Bedrock that are dynamically filled in at runtime during agent invocation. In the prompt templates, you'll see these placeholders surrounded by $ (for example, $instructions$). For information about the placeholder variables that you can use in a template, see Placeholder variables in Amazon Bedrock agent prompt templates.
  - **XML tags** – Anthropic models support the use of XML tags to structure and delineate your prompts. Use descriptive tag names for optimal results. For example, in the default orchestration prompt template, you'll see the <examples> tag used to delineate few-shot examples). For more information, see Use XML tags in the Anthropic user guide.

You can enable or disable any step in the agent sequence. The following table shows the default states for each step.

<table>
<thead>
<tr>
<th>Prompt template</th>
<th>Default setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-processing</td>
<td>Enabled</td>
</tr>
<tr>
<td>Orchestration</td>
<td>Enabled</td>
</tr>
<tr>
<td>Knowledge base response generation</td>
<td>Enabled</td>
</tr>
<tr>
<td>Post-processing</td>
<td>Disabled</td>
</tr>
</tbody>
</table>
Note
If you disable the orchestration step, the agent sends the raw user input to the foundation model and doesn't use the base prompt template for orchestration. If you disable any of the other steps, the agent skips that step entirely.

- **Inference configurations** – Influences the response generated by the model that you use. For definitions of the inference parameters and more details about the parameters that different models support, see [Inference parameters for foundation models](#).

- **(Optional) Parser Lambda function** – Defines how to parse the raw foundation model output and how to use it in the runtime flow. This function acts on the output from the steps in which you enable it and returns the parsed response as you define it in the function.

Depending on how you customized the base prompt template, the raw foundation model output might be specific to the template. As a result, the agent's default parser might have difficulty parsing the output correctly. By writing a custom parser Lambda function, you can help the agent parse the raw foundation model output based on your use-case. For more information about the parser Lambda function and how to write it, see [Parser Lambda function in Agents for Amazon Bedrock](#).

Note
You can define one parser Lambda function for all of the base templates, but you can configure whether to invoke the function in each step. Be sure to configure a resource-based policy for your Lambda function so that your agent can invoke it. For more information, see [Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function](#).

After you edit the prompt templates, you can test your agent. To analyze the step-by-step process of the agent and determine if it is working as you intend, turn on the trace and examine it. For more information, see [Trace events in Amazon Bedrock](#).

You can configure advanced prompts in either the AWS Management Console or through the API.
In the console, you can configure advanced prompts after you have created the agent. You configure them while editing the agent.

**To view or edit advanced prompts for your agent**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).

2. In the left navigation pane, choose **Agents**. Then choose an agent in the **Agents** section.

3. On the agent details page, in the **Working draft** section, select **Working draft**.

4. On the **Working draft** page, in the **Advanced prompts** section, choose **Edit**.

5. On the **Edit advanced prompts** page, choose the tab corresponding to the step of the agent sequence that you want to edit.

6. To enable editing of the template, turn on **Override template defaults**. In the **Override template defaults** dialog box, choose **Confirm**.

7. **Warning**
   
   If you turn off **Override template defaults** or change the model, the default Amazon Bedrock template is used and your template will be immediately deleted. To confirm, enter **confirm** in the text box to confirm the message that appears.

8. To allow the agent to use the template when generating responses, turn on **Activate template**. If this configuration is turned off, the agent doesn't use the template.

9. To modify the example prompt template, use the **Prompt template editor**.

10. In **Configurations**, you can modify inference parameters for the prompt. For definitions of parameters and more information about parameters for different models, see [Inference parameters for foundation models](#).

11. (Optional) To use a Lambda function that you have defined to parse the raw foundation model output, perform the following actions:

    **Note**

    One Lambda function is used for all the prompt templates.
a. In the **Configurations** section, select **Use Lambda function for parsing**. If you clear this setting, your agent will use the default parser for the prompt.

b. For the **Parser Lambda function**, select a Lambda function from the dropdown menu.

### Note

You must attach permissions for your agent so that it can access the Lambda function. For more information, see [Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function](#).

11. To save your settings, choose one of the following options:

a. To remain in the same window so that you can dynamically update the prompt settings while testing your updated agent, choose **Save**.

b. To save your settings and return to the **Working draft** page, choose **Save and exit**.

12. To test the updated settings, choose **Prepare** in the **Test** window.
To configure advanced prompts by using the API operations, you send a `CreateAgent` or `UpdateAgent` call and modify the following `promptOverrideConfiguration` object.

```
"promptOverrideConfiguration": {  
  "overrideLambda": "string",  
  "promptConfigurations": [  
    {  
      "basePromptTemplate": "string",  
      "inferenceConfiguration": {  
        "maximumLength": int,  
        "stopSequences": [ "string" ],  
        "temperature": float,  
        "topK": float,  
        "topP": float  
      },  
      "parserMode": "DEFAULT | OVERRIDDEN",  
      "promptCreationMode": "DEFAULT | OVERRIDDEN",  
      "promptState": "ENABLED | DISABLED",  
      "promptType": "PRE_PROCESSING | ORCHESTRATION | KNOWLEDGE_BASE_RESPONSE_GENERATION | POST_PROCESSING"  
    }  
  ]  
}
```

1. In the `promptConfigurations` list, include a `promptConfiguration` object for each prompt template that you want to edit.
2. Specify the prompt to modify in the `promptType` field.
3. Modify the prompt template through the following steps:
   a. Specify the `basePromptTemplate` fields with your prompt template.
   b. Include inference parameters in the `inferenceConfiguration` objects. For more information about inference configurations, see [Inference parameters for foundation models](#).
4. To enable the prompt template, set the `promptCreationMode` to `OVERRIDDEN`.
5. To allow or prevent the agent from performing the step in the `promptType` field, modify the `promptState` value. This setting can be useful for troubleshooting the agent's behavior.
• If you set promptState to DISABLED for the PRE_PROCESSING, KNOWLEDGE_BASE_RESPONSE_GENERATION, or POST_PROCESSING steps, the agent skips that step.

• If you set promptState to DISABLED for the ORCHESTRATION step, the agent sends only the user input to the foundation model in orchestration. In addition, the agent returns the response as is without orchestrating calls between API operations and knowledge bases.

• By default, the POST_PROCESSING step is DISABLED. By default, the PRE_PROCESSING, ORCHESTRATION, and KNOWLEDGE_BASE_RESPONSE_GENERATION steps are ENABLED.

6. To use a Lambda function that you have defined to parse the raw foundation model output, perform the following steps:

a. For each prompt template that you want to enable the Lambda function for, set parserMode to OVERRIDDEN.

b. Specify the Amazon Resource Name (ARN) of the Lambda function in the overrideLambda field in the promptOverrideConfiguration object.

## Placeholder variables in Amazon Bedrock agent prompt templates

You can use placeholder variables in agent prompt templates. The variables will be populated by pre-existing configurations when the prompt template is called. Select a tab to see variables that you can use for each prompt template.

### Pre-processing

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$functions$</td>
<td>Anthropic Claude Instant, Claude v2.0</td>
<td>Action group API operations and knowledge bases configured for the agent.</td>
</tr>
<tr>
<td>$tools$</td>
<td>Anthropic Claude v2.1</td>
<td></td>
</tr>
<tr>
<td>$conversation_history$</td>
<td>All</td>
<td>Conversation history for the current session</td>
</tr>
</tbody>
</table>
### Variable Models supported Replaced by

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$question$</td>
<td>All</td>
<td>User input for the current InvokeAgent call in the session.</td>
</tr>
</tbody>
</table>

### Orchestration

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$functions$</td>
<td>Anthropic Claude Instant, Claude v2.0</td>
<td>Action group API operation s and knowledge bases configured for the agent.</td>
</tr>
<tr>
<td>$tools$</td>
<td>Anthropic Claude v2.1</td>
<td></td>
</tr>
<tr>
<td>$agent_scratchpad$</td>
<td>All</td>
<td>Designates an area for the model to write down its thoughts and actions it has taken. Replaced by predictions and output of the previous iterations in the current turn. Provides the model with context of what has been achieved for the given user input and what the next step should be.</td>
</tr>
<tr>
<td>$any_function_name$</td>
<td>Anthropic Claude Instant, Claude v2.0</td>
<td>A randomly chosen API name from the API names that exist in the agent's action groups.</td>
</tr>
<tr>
<td>$conversation_history$</td>
<td>All</td>
<td>Conversation history for the current session.</td>
</tr>
<tr>
<td>$instruction$</td>
<td>All</td>
<td>Model instructions configured for the agent.</td>
</tr>
</tbody>
</table>
### Variable Models supported Replaced by

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$prompt_session_attributes$</td>
<td>All</td>
<td>Session attributes preserved across a prompt</td>
</tr>
<tr>
<td>$question$</td>
<td>All</td>
<td>User input for the current InvokeAgent call in the session.</td>
</tr>
</tbody>
</table>

You can use the following placeholder variables if you allow the agent to ask the user for more information by doing one of the following actions:

- In the console, set in the **User input** in the agent details.
- Set the `parentActionGroupSignature` to `AMAZON.UserInput` with a [CreateAgentActionGroup](#) or [UpdateAgentActionGroup](#) request.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ask_user_missing_parameters$</td>
<td>Anthropic Claude Instant,</td>
<td>Instructions for the model to ask the user to provide required missing</td>
</tr>
<tr>
<td></td>
<td>Claude v2.0</td>
<td>information</td>
</tr>
<tr>
<td>$ask_user_missing_information$</td>
<td>Anthropic Claude v2.1</td>
<td></td>
</tr>
<tr>
<td>$ask_user_confirm_parameters$</td>
<td>All</td>
<td>Instructions for the model to ask the user to confirm parameters that the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>agent hasn't yet received or is unsure of.</td>
</tr>
<tr>
<td>$ask_user_function$</td>
<td>All</td>
<td>A function to ask the user a question.</td>
</tr>
<tr>
<td>$ask_user_function_format$</td>
<td>All</td>
<td>The format of the function to ask the user a question.</td>
</tr>
</tbody>
</table>
### Variable Models supported Replaced by

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models supported</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ask_user_input_examples$</td>
<td>All</td>
<td>Few-shot examples to inform the model how to predict when it should ask the user a question.</td>
</tr>
</tbody>
</table>

### Knowledge base response generation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$query$</td>
<td>All</td>
<td>The query generated by the orchestration prompt model response when it predicts the next step to be knowledge base querying.</td>
</tr>
<tr>
<td>$search_results$</td>
<td>All</td>
<td>The retrieved results for the user query</td>
</tr>
</tbody>
</table>

### Post-processing

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
<th>Replaced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>$latest_response$</td>
<td>All</td>
<td>The last orchestration prompt model response</td>
</tr>
<tr>
<td>$question$</td>
<td>All</td>
<td>User input for the current InvokeAgent call in the session.</td>
</tr>
<tr>
<td>$responses$</td>
<td>All</td>
<td>The action group and knowledge base outputs from the current turn.</td>
</tr>
</tbody>
</table>
Parser Lambda function in Agents for Amazon Bedrock

Each prompt template includes a parser Lambda function that you can modify. To write a custom parser Lambda function, you must understand the input event that your agent sends and the response that the agent expects as the output from the Lambda function. You write a handler function to manipulate variables from the input event and to return the response. For more information about how AWS Lambda works, see Event-driven invocation in the AWS Lambda Developer Guide.

Topics
- Parser Lambda input event
- Parser Lambda response
- Parser Lambda examples

Parser Lambda input event

The following is the general structure of the input event from the agent. Use the fields to write your Lambda handler function.

```json
{
    "messageVersion": "1.0",
    "agent": {
        "name": "string",
        "id": "string",
        "alias": "string",
        "version": "string"
    },
    "invokeModelRawResponse": "string",
    "promptType": "ORCHESTRATION | POST_PROCESSING | PRE_PROCESSING | KNOWLEDGE_BASE_RESPONSE_GENERATION ",
    "overrideType": "OUTPUT_PARSER"
}
```

The following list describes the input event fields:

- **messageVersion** – The version of the message that identifies the format of the event data going into the Lambda function and the expected format of the response from the Lambda function. Agents for Amazon Bedrock only supports version 1.0.
• agent – Contains information about the name, ID, alias, and version of the agent that the prompts belongs to.

• invokeModelRawResponse – The raw foundation model output of the prompt whose output is to be parsed.

• promptType – The prompt type whose output is to be parsed.

• overrideType – The artifacts that this Lambda function overrides. Currently, only OUTPUT_PARSER is supported, which indicates that the default parser is to be overridden.

**Parser Lambda response**

Your agent expects a response from your Lambda function that matches the following format. The agent uses the response for further orchestration or to help it return a response to the user. Use the Lambda function response fields to configure how the output is returned.

```json
{
    "messageVersion": "1.0",
    "promptType": "ORCHESTRATION | PRE_PROCESSING | POST_PROCESSING | KNOWLEDGE_BASE_RESPONSE_GENERATION",
    "preProcessingParsedResponse": {
        "isValidInput": "boolean",
        "rationale": "string"
    },
    "orchestrationParsedResponse": {
        "rationale": "string",
        "parsingErrorDetails": {
            "repromptResponse": "string"
        },
        "responseDetails": {
            "invocationType": "ACTION_GROUP | KNOWLEDGE_BASE | FINISH | ASK_USER",
            "agentAskUser": {
                "responseText": "string"
            },
            "actionGroupInvocation": {
                "actionGroupName": "string",
                "apiName": "string",
                "verb": "string",
                "actionGroupInput": {
                    "<parameter>": {
                        "value": "string"
                    }
                }
            }
        }
    }
}
```
Advanced prompts
The following list describes the Lambda response fields:

- **messageVersion** – The version of the message that identifies the format of the event data going into the Lambda function and the expected format of the response from a Lambda function. Agents for Amazon Bedrock only supports version 1.0.

- **promptType** – The prompt type of the current turn.

- **preProcessingParsedResponse** – The parsed response for the PRE_PROCESSING prompt type.

- **orchestrationParsedResponse** – The parsed response for the ORCHESTRATION prompt type. See below for more details.

- **knowledgeBaseResponseGenerationParsedResponse** – The parsed response for the KNOWLEDGE_BASE_RESPONSE_GENERATION prompt type.

- **postProcessingParsedResponse** – The parsed response for the POST_PROCESSING prompt type.

For more details about the parsed responses for the four prompt templates, see the following tabs.

**preProcessingParsedResponse**

```json
{
   "isValidInput": "boolean",
   "rationale": "string"
}
```

The preProcessingParsedResponse contains the following fields.

- **isValidInput** – Specifies whether the user input is valid or not. You can define the function to determine how to characterize the validity of user input.

- **rationale** – The reasoning for the user input categorization. This rationale is provided by the model in the raw response, the Lambda function parses it, and the agent presents it in the trace for pre-processing.

**orchestrationResponse**

```json
{
}
```
"rationale": "string",
"parsingErrorDetails": {
  "repromptResponse": "string"
},
"responseDetails": {
  "invocationType": "ACTION_GROUP | KNOWLEDGE_BASE | FINISH | ASK_USER",
  "agentAskUser": {
    "responseText": "string"
  },
  "actionGroupInvocation": {
    "actionGroupName": "string",
    "apiName": "string",
    "verb": "string",
    "actionGroupInput": {
      "<parameter>": {
        "value": "string"
      },
      ...
    },
  },
  "agentKnowledgeBase": {
    "knowledgeBaseId": "string",
    "searchQuery": {
      "value": "string"
    }
  },
  "agentFinalResponse": {
    "responseText": "string",
    "citations": {
      "generatedResponseParts": [
        {
          "text": "string",
          "references": [
            {"sourceId": "string"},
            ...
          ]
        },
        ...
      ]
    }
  }
}
The orchestrationParsedResponse contains the following fields:

- **rationale** – The reasoning for what to do next, based on the foundation model output. You can define the function to parse from the model output.
- **parsingErrorDetails** – Contains the repromptResponse, which is the message to reprompt the model to update its raw response when the model response can’t be parsed. You can define the function to manipulate how to reprompt the model.
- **responseDetails** – Contains the details for how to handle the output of the foundation model. Contains an **invocationType**, which is the next step for the agent to take, and a second field that should match the invocationType. The following objects are possible.
  - **agentAskUser** – Compatible with the ASK_USER invocation type. This invocation type ends the orchestration step. Contains the responseText to ask the user for more information. You can define your function to manipulate this field.
  - **actionGroupInvocation** – Compatible with the ACTION_GROUP invocation type. You can define your function to determine action groups to invoke and parameters to pass. Contains the following fields:
    - **actionGroupName** – The action group to invoke.
    - **apiName** – The name of the API operation to invoke in the action group.
    - **verb** – The method of the API operation to use.
    - **actionGroupInput** – Contains parameters to specify in the API operation request.
  - **agentKnowledgeBase** – Compatible with the KNOWLEDGE_BASE invocation type. You can define your function to determine how to query knowledge bases. Contains the following fields:
    - **knowledgeBaseId** – The unique identifier of the knowledge base.
    - **searchQuery** – Contains the query to send to the knowledge base in the value field.
  - **agentFinalResponse** – Compatible with the FINISH invocation type. This invocation type ends the orchestration step. Contains the response to the user in the responseText field and citations for the response in the citations object.

knowledgeBaseResponseGenerationParsedResponse

```json
{
    "generatedResponse": {
        "generatedResponseParts": [
        {
```
The `knowledgeBaseResponseGenerationParsedResponse` contains the generatedResponse from querying the knowledge base and references for the data sources.

`postProcessingParsedResponse`

```json
{
  "responseText": "string",
  "citations": {
    "generatedResponseParts": [
      {
        "text": "string",
        "references": [
          {
            "sourceId": "string"
          },
          ...
        ],
        ...
      },
      ...
    ]
  }
}
```

The `postProcessingParsedResponse` contains the following fields:

- `responseText` – The response to return to the end user. You can define the function to format the response.

- `citations` – Contains a list of citations for the response. Each citation shows the cited text and its references.
Parser Lambda examples

To see an example parser Lambda function for a specific prompt template, select from the following tabs. Also shown are example input events sent to the function and responses from it. The `lambda_handler` function returns the parsed response to the agent.

Pre-processing

Example function

```python
import json
import re
import logging

PRE_PROCESSING_RATIONALE_REGEX = "<thinking>(.*?)</thinking>"
PREPROCESSING_CATEGORY_REGEX = "<category>(.*?)</category>"
PREPROCESSING_PROMPT_TYPE = "PRE_PROCESSING"
PRE_PROCESSING_RATIONALE_PATTERN = re.compile(PRE_PROCESSING_RATIONALE_REGEX, re.DOTALL)
PREPROCESSING_CATEGORY_PATTERN = re.compile(PREPROCESSING_CATEGORY_REGEX, re.DOTALL)

logger = logging.getLogger()

# This parser lambda is an example of how to parse the LLM output for the default PreProcessing prompt
def lambda_handler(event, context):

    print("Lambda input: " + str(event))
    logger.info("Lambda input: " + str(event))

    prompt_type = event["promptType"]

    # Sanitize LLM response
    model_response = sanitize_response(event['invokeModelRawResponse'])

    if event["promptType"] == PREPROCESSING_PROMPT_TYPE:
        return parse_pre_processing(model_response)

def parse_pre_processing(model_response):

    category_matches = re.finditer(PREPROCESSING_CATEGORY_PATTERN, model_response)
    rationale_matches = re.finditer(PRE_PROCESSING_RATIONALE_PATTERN, model_response)
```

Advanced prompts
category = next((match.group(1) for match in category_matches), None)
rationale = next((match.group(1) for match in rationale_matches), None)

return {
    "promptType": "PRE_PROCESSING",
    "preProcessingParsedResponse": {
        "rationale": rationale,
        "isValidInput": get_is_valid_input(category)
    }
}

def sanitize_response(text):
    pattern = r'(\n*)'
    text = re.sub(pattern, r'
', text)
    return text

def get_is_valid_input(category):
    if category is not None and category.strip().upper() == "D" or
    category.strip().upper() == "E":
        return True
    return False


"preProcessingParsedResponse": {
   "rationale": "The user is asking about the instructions provided to the function calling agent. This input is trying to gather information about what functions/API's or instructions our function calling agent has access to. Based on the categories provided, this input belongs in Category B."
   "isValidInput": false
}
}

Orchestration

The following are example functions for Anthropic Claude 2 and Anthropic Claude 2.1.

Anthropic Claude 2

```python
import json
import re
import logging

RATIONALE_REGEX_LIST = [
   "(.?)(<function_call>)",
   "(.?)(<answer>)"
]
RATIONALE_PATTERNS = [re.compile(regex, re.DOTALL) for regex in RATIONALE_REGEX_LIST]

RATIONALE_VALUE_REGEX_LIST = [
   "<scratchpad>(.*?)(</scratchpad>)",
   "(.*?)(</scratchpad>)",
   "(<scratchpad>)(.*?)(</scratchpad>)"
]
RATIONALE_VALUE_PATTERNS = [re.compile(regex, re.DOTALL) for regex in RATIONALE_VALUE_REGEX_LIST]

ANSWER_REGEX = r"(?<=<answer>)(.*)"
ANSWER_PATTERN = re.compile(ANSWER_REGEX, re.DOTALL)

ANSWER_TAG = "<answer>"
FUNCTION_CALL_TAG = "<function_call>"

ASK_USER_FUNCTION_CALL_REGEX = r"(<function_call>user::askuser)(.*?)(</function_call>)"
ASK_USER_FUNCTION_CALL_PATTERN = re.compile(ASK_USER_FUNCTION_CALL_REGEX, re.DOTALL)
```
ASK_USER_FUNCTION_PARAMETER_REGEX = r"(?<=askuser=")(.*?)""
ASK_USER_FUNCTION_PARAMETER_PATTERN = re.compile(ASK_USER_FUNCTION_PARAMETER_REGEX, re.DOTALL)

KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX = "x_amz_knowledgebase_"

FUNCTION_CALL_REGEX = r"<function_call>(\w+):(\w+):(.+)(.+)\""

ANSWER_PART_REGEX = "<answer_part\s?s?>(.+)</answer_part\s?s?>"
ANSWER_TEXT_PART_REGEX = "<text\s?s?>(.+)</text\s?s?>"
ANSWER_REFERENCE_PART_REGEX = "<source\s?s?>(.+)</source\s?s?>"

ANSWER_PART_PATTERN = re.compile(ANSWER_PART_REGEX, re.DOTALL)
ANSWER_TEXT_PART_PATTERN = re.compile(ANSWER_TEXT_PART_REGEX, re.DOTALL)
ANSWER_REFERENCE_PART_PATTERN = re.compile(ANSWER_REFERENCE_PART_REGEX, re.DOTALL)

# You can provide messages to reprompt the LLM in case the LLM output is not in the expected format
MISSING_API_INPUT_FOR_USER_REPROMPT_MESSAGE = "Missing the argument askuser for user::askuser function call. Please try again with the correct argument added"
ASK_USER_FUNCTION_CALL_STRUCTURE_REMPROMPT_MESSAGE = "The function call format is incorrect. The format for function calls to the askuser function must be: <function_call>user::askuser(askuser="$ASK_USER_INPUT")</function_call>.
FUNCTION_CALL_STRUCTURE_REPROMPT_MESSAGE = 'The function call format is incorrect. The format for function calls must be: <function_call>$FUNCTION_NAME($FUNCTION_ARGUMENT_NAME="$FUNCTION_ARGUMENT_NAME")</function_call>.'

logger = logging.getLogger()

# This parser lambda is an example of how to parse the LLM output for the default orchestration prompt
def lambda_handler(event, context):
    logger.info("Lambda input: " + str(event))

    # Sanitize LLM response
    sanitized_response = sanitize_response(event['invokeModelRawResponse'])

    # Parse LLM response for any rationale
    rationale = parse_rationale(sanitized_response)

    # Construct response fields common to all invocation types
    parsed_response = {
        'promptType': 'ORCHESTRATION',
        'orchestrationParsedResponse': {
            # Additional response fields
        }
    }
'rationale': rationale

# Check if there is a final answer
try:
    final_answer, generated_response_parts = parse_answer(sanitized_response)
except ValueError as e:
    addRepromptResponse(parsed_response, e)
    return parsed_response

if final_answer:
    parsed_response['orchestrationParsedResponse']['responseDetails'] = {
        'invocationType': 'FINISH',
        'agentFinalResponse': {
            'responseText': final_answer
        }
    }

if generated_response_parts:
    parsed_response['orchestrationParsedResponse']['responseDetails']['agentFinalResponse']['citations'] = {
        'generatedResponseParts': generated_response_parts
    }

    logger.info("Final answer parsed response: " + str(parsed_response))
    return parsed_response

# Check if there is an ask user
try:
    ask_user = parse_ask_user(sanitized_response)
    if ask_user:
        parsed_response['orchestrationParsedResponse']['responseDetails'] = {
            'invocationType': 'ASK_USER',
            'agentAskUser': {
                'responseText': ask_user
            }
        }

        logger.info("Ask user parsed response: " + str(parsed_response))
        return parsed_response
except ValueError as e:
    addRepromptResponse(parsed_response, e)
    return parsed_response

Advanced prompts
# Check if there is an agent action
try:
    parsed_response = parse_function_call(sanitized_response, parsed_response)
    logger.info("Function call parsed response: " + str(parsed_response))
    return parsed_response
except ValueError as e:
    addRepromptResponse(parsed_response, e)
    return parsed_response

addRepromptResponse(parsed_response, 'Failed to parse the LLM output')
logger.info(parsed_response)
return parsed_response

raise Exception("unrecognized prompt type")

def sanitize_response(text):
    pattern = r"(\n*)"
    text = re.sub(pattern, r"\n", text)
    return text

def parse_rationale(sanitized_response):
    # Checks for strings that are not required for orchestration
    rationale_matcher = next((pattern.search(sanitized_response) for pattern in RATIONALE_PATTERNS if pattern.search(sanitized_response)), None)
    if rationale_matcher:
        rationale = rationale_matcher.group(1).strip()

        # Check if there is a formatted rationale that we can parse from the string
        rationale_value_matcher = next((pattern.search(rationale) for pattern in RATIONALE_VALUE_PATTERNS if pattern.search(rationale)), None)
        if rationale_value_matcher:
            return rationale_value_matcher.group(1).strip()

    return rationale

    return None

def parse_answer(sanitized_llm_response):
    if has_generated_response(sanitized_llm_response):
        return parse_generated_response(sanitized_llm_response)

    answer_match = ANSWER_PATTERN.search(sanitized_llm_response)
if answer_match and is_answer(sanitized_llm_response):
    return answer_match.group(0).strip(), None

return None, None

def is_answer(llm_response):
    return llm_response.rfind(ANSWER_TAG) > llm_response.rfind(FUNCTION_CALL_TAG)

def parse_generated_response(sanitized_llm_response):
    results = []
    for match in ANSWER_PART_PATTERN.finditer(sanitized_llm_response):
        part = match.group(1).strip()
        text_match = ANSWER_TEXT_PART_PATTERN.search(part)
        if not text_match:
            raise ValueError("Could not parse generated response")
        text = text_match.group(1).strip()
        references = parse_references(sanitized_llm_response, part)
        results.append((text, references))
    final_response = " ".join([r[0] for r in results])
    generated_response_parts = []
    for text, references in results:
        generatedResponsePart = {
            'text': text,
            'references': references
        }
        generated_response_parts.append(generatedResponsePart)
    return final_response, generated_response_parts

def has_generated_response(raw_response):
    return ANSWER_PART_PATTERN.search(raw_response) is not None

def parse_references(raw_response, answer_part):
    references = []
    for match in ANSWER_REFERENCE_PART_PATTERN.finditer(answer_part):
        reference = match.group(1).strip()
        references.append({'sourceId': reference})
    return references
def parse_ask_user(sanitized_llm_response):
    ask_user_matcher = ASK_USER_FUNCTION_CALL_PATTERN.search(sanitized_llm_response)
    if ask_user_matcher:
        try:
            ask_user = ask_user_matcher.group(2).strip()
            ask_user_question_matcher = ASK_USER_FUNCTION_PARAMETER_PATTERN.search(ask_user)
            if ask_user_question_matcher:
                return ask_user_question_matcher.group(1).strip()
            raise ValueError(MISSING_API_INPUT_FOR_USER_REPROMPT_MESSAGE)
        except ValueError as ex:
            raise ex
        except Exception as ex:
            raise Exception(ASK_USER_FUNCTION_CALL_STRUCTURE_REMPROMPT_MESSAGE)
    return None

def parse_function_call(sanitized_response, parsed_response):
    match = re.search(FUNCTION_CALL_REGEX, sanitized_response)
    if not match:
        raise ValueError(FUNCTION_CALL_STRUCTURE_REPROMPT_MESSAGE)
    verb, resource_name, function = match.group(1), match.group(2), match.group(3)
    parameters = {}
    for arg in match.group(4).split(','):
        key, value = arg.split('=', 1)
        parameters[key.strip()] = {'value': value.strip('" ')}
    parsed_response['orchestrationParsedResponse']['responseDetails'] = {}
    # Function calls can either invoke an action group or a knowledge base.
    # Mapping to the correct variable names accordingly
    if resource_name.lower().startswith(KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX):
        parsed_response['orchestrationParsedResponse']['responseDetails']['invocationType'] = 'KNOWLEDGE_BASE'
        parsed_response['orchestrationParsedResponse']['responseDetails']['agentKnowledgeBase'] = {
            'searchQuery': parameters['searchQuery'],
            'knowledgeBaseId': resource_name.replace(KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX, '')
        }
import logging
import re
import xml.etree.ElementTree as ET

RATIONALE_REGEX_LIST = [
    "(.+?)(<function_calls>)", 
    "(.+?)(<answer>)"
]
RATIONALE_PATTERNS = [re.compile(regex, re.DOTALL) for regex in RATIONALE_REGEX_LIST]

RATIONALE_VALUE_REGEX_LIST = [
    "<scratchpad>(.*?)(</scratchpad>)", 
    "((.+?)</scratchpad>)", 
    "(<scratchpad>)(.*?)"
]
RATIONALE_VALUE_PATTERNS = [re.compile(regex, re.DOTALL) for regex in RATIONALE_VALUE_REGEX_LIST]
Advanced prompts

**FUNCTION_CALL_REGEX = r'(?<=<function_calls>)(.*)'**

ANSWER_TAG = "<answer>

**FUNCTION_CALL_TAG = "<function_calls>"**

**ASK_USER_FUNCTION_CALL_REGEX = r"<tool_name>user::askuser</tool_name>"**

**ASK_USER_FUNCTION_CALL_PATTERN = re.compile(ASK_USER_FUNCTION_CALL_REGEX, re.DOTALL)**

**ASK_USER_TOOL_NAME_REGEX = r"<tool_name>((.|
)*?)</tool_name>"**

**ASK_USER_TOOL_NAME_PATTERN = re.compile(ASK_USER_TOOL_NAME_REGEX, re.DOTALL)**

**TOOL_PARAMETERS_REGEX = r"<parameters>((.|
)*?)</parameters>"**

**TOOL_PARAMETERS_PATTERN = re.compile(TOOL_PARAMETERS_REGEX, re.DOTALL)**

**ASK_USER_TOOL_PARAMETER_REGEX = r"<question>((.|
)*?)</question>"**

**ASK_USER_TOOL_PARAMETER_PATTERN = re.compile(ASK_USER_TOOL_PARAMETER_REGEX, re.DOTALL)**

**KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX = "x_amz_knowledgebase_"**

**FUNCTION_CALL_REGEX = r"(?<=<function_calls>)(.*)"**

**ANSWER_PART_REGEX = "<answer_part>(.+?)</answer_part>"**

**ANSWER_TEXT_PART_REGEX = "<text>(.+?)</text>"**

**ANSWER_REFERENCE_PART_REGEX = "<source>(.+?)</source>"**

**ANSWER_PART_PATTERN = re.compile(ANSWER_PART_REGEX, re.DOTALL)**

**ANSWER_TEXT_PART_PATTERN = re.compile(ANSWER_TEXT_PART_REGEX, re.DOTALL)**

**ANSWER_REFERENCE_PART_PATTERN = re.compile(ANSWER_REFERENCE_PART_REGEX, re.DOTALL)**

# You can provide messages to reprompt the LLM in case the LLM output is not in the expected format

**MISSING_API_INPUT_FOR_USER_REPROMPT_MESSAGE = "Missing the parameter 'question' for user::askuser function call. Please try again with the correct argument added."**

**ASK_USER_FUNCTION_CALL_STRUCTURE_REMPROMPT_MESSAGE = "The function call format is incorrect. The format for function calls to the askuser function must be: <invoke> <tool_name>user::askuser</tool_name> <parameters> <question>$QUESTION</question> </parameters> </invoke>."**

**FUNCTION_CALL_STRUCTURE_REPROMPT_MESSAGE = "The function call format is incorrect. The format for function calls must be: <invoke> <tool_name>$TOOL_NAME</tool_name> <parameters> <$PARAMETER_NAME>$PARAMETER_VALUE</$PARAMETER_NAME>... </parameters> </invoke>."**
logger = logging.getLogger()

# This parser lambda is an example of how to parse the LLM output for the default orchestration prompt
def lambda_handler(event, context):
    logger.info("Lambda input: " + str(event))

    # Sanitize LLM response
    sanitized_response = sanitize_response(event["invokeModelRawResponse"])

    # Parse LLM response for any rationale
    rationale = parse_rationale(sanitized_response)

    # Construct response fields common to all invocation types
    parsed_response = {
        'promptType': "ORCHESTRATION",
        'orchestrationParsedResponse': {
            'rationale': rationale
        }
    }

    # Check if there is a final answer
    try:
        final_answer, generated_response_parts = parse_answer(sanitized_response)
    except ValueError as e:
        addRepromptResponse(parsed_response, e)
        return parsed_response

    if final_answer:
        parsed_response['orchestrationParsedResponse']['responseDetails'] = {
            'invitationType': 'FINISH',
            'agentFinalResponse': {
                'responseText': final_answer
            }
        }

        if generated_response_parts:
            parsed_response['orchestrationParsedResponse']['responseDetails']
            ['agentFinalResponse']['citations'] = {
                'generatedResponseParts': generated_response_parts
            }

    logger.info("Final answer parsed response: " + str(parsed_response))
def sanitize_response(text):
    pattern = r"((\n*)"
    text = re.sub(pattern, r"\n", text)
    return text

def parse_rationale(sanitized_response):
    # Checks for strings that are not required for orchestration
    rationale_matcher = next("unrecognized prompt type")
(pattern.search(sanitized_response) for pattern in RATIONALE_PATTERNS if pattern.search(sanitized_response)),
None)

if rationale_matcher:
    rationale = rationale_matcher.group(1).strip()

# Check if there is a formatted rationale that we can parse from the string
rationale_value_matcher = next((pattern.search(rationale) for pattern in RATIONALE_VALUE_PATTERNS if pattern.search(rationale)), None)
if rationale_value_matcher:
    return rationale_value_matcher.group(1).strip()

return rationale

return None

def parse_answer(sanitized_llm_response):
    if has_generated_response(sanitized_llm_response):
        return parse_generated_response(sanitized_llm_response)

    answer_match = ANSWER_PATTERN.search(sanitized_llm_response)
    if answer_match and is_answer(sanitized_llm_response):
        return answer_match.group(0).strip(), None

    return None, None

def is_answer(llm_response):
    return llm_response.rfind(ANSWER_TAG) > llm_response.rfind(FUNCTION_CALL_TAG)

def parse_generated_response(sanitized_llm_response):
    results = []

    for match in ANSWER_PART_PATTERN.finditer(sanitized_llm_response):
        part = match.group(1).strip()

        text_match = ANSWER_TEXT_PART_PATTERN.search(part)
        if not text_match:
            raise ValueError("Could not parse generated response")
text = text_match.group(1).strip()
references = parse_references(sanitized_llm_response, part)
results.append((text, references))

final_response = " ".join([r[0] for r in results])

generated_response_parts = []
for text, references in results:
    generatedResponsePart = {
        'text': text,
        'references': references
    }
    generated_response_parts.append(generatedResponsePart)

return final_response, generated_response_parts

def has_generated_response(raw_response):
    return ANSWER_PART_PATTERN.search(raw_response) is not None

def parse_references(raw_response, answer_part):
    references = []
    for match in ANSWER_REFERENCE_PART_PATTERN.finditer(answer_part):
        reference = match.group(1).strip()
        references.append({"sourceId": reference})
    return references

def parse_ask_user(sanitized_llm_response):
    ask_user_matcher = ASK_USER_FUNCTION_CALL_PATTERN.search(sanitized_llm_response)
    if ask_user_matcher:
        try:
            parameters_matches = TOOL_PARAMETERS_PATTERN.search(sanitized_llm_response)
            params = parameters_matches.group(1).strip()
            ask_user_question_matcher = ASK_USER_TOOL_PARAMETER_PATTERN.search(params)
            if ask_user_question_matcher:
                ask_user_question = ask_user_question_matcher.group(1)
                return ask_user_question
        except ValueError as ex:
            raise ex
except Exception as ex:
    raise Exception(ASK_USER_FUNCTION_CALL_STRUCTURE_REMPROMPT_MESSAGE)

return None

def parse_function_call(sanitized_response, parsed_response):
    match = re.search(FUNCTION_CALL_REGEX, sanitized_response)
    if not match:
        raise ValueError(FUNCTION_CALL_STRUCTURE_REPROMPT_MESSAGE)

    tool_name_matches = ASK_USER_TOOL_NAME_PATTERN.search(sanitized_response)
    tool_name = tool_name_matches.group(1)
    parameters_matches = TOOL_PARAMETERS_PATTERN.search(sanitized_response)
    params = parameters_matches.group(1).strip()

    action_split = tool_name.split('::')
    verb = action_split[0].strip()
    resource_name = action_split[1].strip()
    function = action_split[2].strip()

    xml_tree = ET.ElementTree(ET.fromstring("<parameters>{}
parameters">
parameters>".format(params)))
    parameters = {}
    for elem in xml_tree.iter():
        if elem.text:
            parameters[elem.tag] = {'value': elem.text.strip('" ')}

    parsed_response['orchestrationParsedResponse']['responseDetails'] = {}

    # Function calls can either invoke an action group or a knowledge base.
    # Mapping to the correct variable names accordingly
    if resource_name.lower().startswith(KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX):
        parsed_response['orchestrationParsedResponse']['responseDetails']
        ['invocationType'] = 'KNOWLEDGE_BASE'
        parsed_response['orchestrationParsedResponse']['responseDetails']
        ['agentKnowledgeBase'] = {
            'searchQuery': parameters['searchQuery'],
            'knowledgeBaseId':
            resource_name.replace(KNOWLEDGE_STORE_SEARCH_ACTION_PREFIX, '')
        }

    return parsed_response
Example request

```json
{
    'agent': {
        'alias': 'TSTALIASID',
        'id': 'AGENTID123',
        'name': 'InsuranceAgent',
        'version': 'DRAFT'
    },
    'invokeModelRawResponse': 'To answer this question, I will:
1. Call the GET::x_amz_knowledgebase_KBID123456::Search function to search for a phone number to call.
2. I have checked that I have access to the GET::x_amz_knowledgebase_KBID23456::Search function.

<scratchpad>

<function_call>GET::x_amz_knowledgebase_KBID123456::Search(searchQuery="What is the phone number I can call?")
```

Example response

Advanced prompts
'{
  'promptType': 'ORCHESTRATION',
  'orchestrationParsedResponse': {
    'rationale': 'To answer this question, I will:

1. Call the GET::x_amz_knowledgebase_KBID123456::Search function to search for a phone number to call Farmers.

I have checked that I have access to the GET::x_amz_knowledgebase_KBID123456::Search function.',
    'responseDetails': {
      'invocationType': 'KNOWLEDGE_BASE',
      'agentKnowledgeBase': {
        'searchQuery': {'value': 'What is the phone number I can call?'},
        'knowledgeBaseId': 'KBID123456'
      }
    }
  }
}

Knowledge base response generation

Example function

```python
import json
import re
import logging

ANSWER_PART_REGEX = "<answer_part\s?>(.+?)</answer_part\s?>"
ANSWER_TEXT_PART_REGEX = "<text\s?>(.+?)</text\s?>"
ANSWER_REFERENCE_PART_REGEX = "<source\s?>(.+?)</source\s?>"
ANSWER_PART_PATTERN = re.compile(ANSWER_PART_REGEX, re.DOTALL)
ANSWER_TEXT_PART_PATTERN = re.compile(ANSWER_TEXT_PART_REGEX, re.DOTALL)
ANSWER_REFERENCE_PART_PATTERN = re.compile(ANSWER_REFERENCE_PART_REGEX, re.DOTALL)

logger = logging.getLogger()

# This parser lambda is an example of how to parse the LLM output for the default KB response generation prompt
def lambda_handler(event, context):
    logger.info("Lambda input: " + str(event))
    raw_response = event['invokeModelRawResponse']

    parsed_response = {
      'promptType': 'KNOWLEDGE_BASE_RESPONSE_GENERATION',
      'knowledgeBaseResponseGenerationParsedResponse': {
```
'generatedResponse': parse_generated_response(raw_response)
}
}

logger.info(parsed_response)
return parsed_response

def parse_generated_response(sanitized_llm_response):
    results = []

    for match in ANSWER_PART_PATTERN.finditer(sanitized_llm_response):
        part = match.group(1).strip()

        text_match = ANSWER_TEXT_PART_PATTERN.search(part)
        if not text_match:
            raise ValueError("Could not parse generated response")

        text = text_match.group(1).strip()
        references = parse_references(sanitized_llm_response, part)
        results.append(((text, references))

    generated_response_parts = []
    for text, references in results:
        generatedResponsePart = {
            'text': text,
            'references': references
        }
        generated_response_parts.append(generatedResponsePart)

    return {
            'generatedResponseParts': generated_response_parts
    }

def parse_references(raw_response, answer_part):
    references = []
    for match in ANSWER_REFERENCE_PART_PATTERN.finditer(answer_part):
        reference = match.group(1).strip()
        references.append({'sourceId': reference})

    return references

Example request

{
Example response

```json
{
    'promptType': 'KNOWLEDGE_BASE_RESPONSE_GENERATION',
    'knowledgeBaseResponseGenerationParsedResponse': {
        'generatedResponse': {
            'generatedResponseParts': [
                {
                    'text': 'The search results contain information about different types of insurance benefits, including personal injury protection (PIP), medical payments coverage, and lost wages coverage. PIP typically covers reasonable medical expenses for injuries caused by an accident, as well as income continuation, child care, loss of services, and funerals. Medical payments coverage provides payment for medical treatment resulting from a car accident. Who pays lost wages due to injuries depends on the laws in your state and the coverage purchased.',
                    'references': [
                        {'sourceId': '1234567-1234-1234-1234-123456789abc'},
                        {'sourceId': '2345678-2345-2345-2345-23456789abcd'},
                        {'sourceId': '3456789-3456-3456-3456-3456789abcde'}
                    ]
                }
            ]
        }
    }
}
```
Post-processing

Example function

```python
import json
import re
import logging

FINAL_RESPONSE_REGEX = r"<final_response>\([\s\S]*?\)</final_response>"
FINAL_RESPONSE_PATTERN = re.compile(FINAL_RESPONSE_REGEX, re.DOTALL)

logger = logging.getLogger()

# This parser lambda is an example of how to parse the LLM output for the default PostProcessing prompt
def lambda_handler(event, context):
    logger.info("Lambda input: " + str(event))
    raw_response = event['invokeModelRawResponse']

    parsed_response = {
        'promptType': 'POST_PROCESSING',
        'postProcessingParsedResponse': {}
    }

    matcher = FINAL_RESPONSE_PATTERN.search(raw_response)
    if not matcher:
        raise Exception("Could not parse raw LLM output")
    response_text = matcher.group(1).strip()

    parsed_response['postProcessingParsedResponse']['responseText'] = response_text

    logger.info(parsed_response)
    return parsed_response
```

Example request
Based on your request, I searched our insurance benefit information database for details. The search results indicate that insurance policies may cover different types of benefits, depending on the policy and state laws. Specifically, the results discussed personal injury protection (PIP) coverage, which typically covers medical expenses for insured individuals injured in an accident (cited sources: 1234567-1234-1234-1234-123456789abc, 2345678-2345-2345-2345-23456789abcd). PIP may pay for costs like medical care, lost income replacement, childcare expenses, and funeral costs. Medical payments coverage was also mentioned as another option that similarly covers medical treatment costs for the policyholder and others injured in a vehicle accident involving the insured vehicle. The search results further noted that whether lost wages are covered depends on the state and coverage purchased. Please let me know if you need any clarification or have additional questions.

Advanced prompts
Control session context

For greater control of session context, you can modify the SessionState object in your agent. The SessionState object contains two types of attributes that you can use to provide conversational context for the agent during user conversations.

- **sessionAttributes** – Attributes that persist over a session between a user and agent. All InvokeAgent requests made with the same session Id belong to the same session, as long as the session time limit (the idleSessionTTLinSeconds) has not been surpassed.

- **promptSessionAttributes** – Attributes that persist over a single turn (one InvokeAgent call). You can use the $prompt_session_attributes$ placeholder when you edit the orchestration base prompt template. This placeholder will be populated at runtime with the attributes that you specify in the promptSessionAttributes field.

The general format of the SessionState object is as follows.

```json
{
    "sessionAttributes": {
        "<attributeName1>": "<attributeValue1>",
        "<attributeName2>": "<attributeValue2>",
        ...
    },
    "promptSessionAttributes": {
        "<attributeName3>": "<attributeValue3>",
        "<attributeName4>": "<attributeValue4>",
        ...
    }
}
```

You can define the session state attributes at two different steps.

- When you set up an action group and write the Lambda function, include sessionAttributes or promptSessionAttributes in the response event that is returned to Amazon Bedrock.
During runtime, when you send an `InvokeAgent` request, include a `sessionState` object in the request body to dynamically change the session state attributes in the middle of the conversation.

**Session attribute example**

The following example uses a session attribute to personalize a message to your user.

1. Write your application code to ask the user to provide their first name and the request they want to make to the agent and to store the answers as the variables `<first_name>` and `<request>`.
2. Write your application code to send an `InvokeAgent` request with the following body:

   ```json
   {
       "inputText": "<request>",
       "sessionState": {
           "sessionAttributes": {
               "firstName": "<first_name>"
           }
       }
   }
   ```

3. When a user uses your application and provides their first name, your code will send the first name as a session attribute and the agent will store their first name for the duration of the session.

4. Because session attributes are sent in the Lambda input event, you can refer to these session attributes in a Lambda function for an action group. For example, if the action API schema requires a first name in the request body, you can use the `firstName` session attribute when writing the Lambda function for an action group to automatically populate that field when sending the API request.

**Prompt session attribute example**

The following general example uses a prompt session attribute to provide temporal context for the agent.

1. Write your application code to store the user request in a variable called `<request>`.
2. Write your application code to retrieve the time zone at the user's location if the user uses a word indicating relative time (such as "tomorrow") in the `<request>`, and store in a variable called `<timezone>`.

3. Write your application to send an InvokeAgent request with the following body:

```json
{
    "inputText": "<request>",
    "sessionState": {
        "promptSessionAttributes": {
            "timeZone": "<timezone>
        }
    }
}
```

4. If a user uses a word indicating relative time, your code will send the timeZone prompt session attribute and the agent will store it for the duration of the turn.

5. For example, if a user asks I need to book a hotel for tomorrow, your code sends the user's time zone to the agent and the agent can determine the exact date that "tomorrow" refers to.

6. The prompt session attribute can be used at the following steps.

   - If you include the $prompt_session_attributes$ placeholder in the orchestration prompt template, the orchestration prompt to the FM includes the prompt session attributes.
   - Prompt session attributes are sent in the Lambda input event and can be used to help populate API requests or returned in the response.

**Deploy an Amazon Bedrock agent**

When you first create an Amazon Bedrock agent, you have a working draft version (DRAFT) and a test alias (TSTALIASID) that points to the working draft version. When you make changes to your agent, the changes apply to the working draft. You iterate on your working draft until you're satisfied with the behavior of your agent. Then, you can set up your agent for deployment and integration into your application by creating aliases of your agent.

To deploy your agent, you must create an alias. During alias creation, Amazon Bedrock creates a version of your agent automatically. The alias points to this newly created version. Alternatively, you can point the alias to a previously created version of your agent. Then, you configure your application to make API calls to that alias.
A version is a snapshot that preserves the resource as it exists at the time it was created. You can continue to modify the working draft and create new aliases (and consequently, versions) of your agent as necessary. In Amazon Bedrock, you create a new version of your agent by creating an alias that points to the new version by default. Amazon Bedrock creates versions in numerical order, starting from 1.

Versions are immutable because they act as a snapshot of your agent at the time you created it. To make updates to an agent in production, you must create a new version and set up your application to make calls to the alias that points to that version.

With aliases, you can switch efficiently between different versions of your agent without requiring the application to keep track of the version. For example, you can change an alias to point to a previous version of your agent if there are changes that you need to revert quickly.

To deploy your agent

1. Create an alias and version of your agent. Select the tab corresponding to your method of choice and follow the steps.

   **Console**

   To create an alias (and optionally a new version)

   1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).

   2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.

   3. In the **Aliases** section, choose **Create**.

   4. Enter a unique name for the alias and provide an optional description.

   5. Choose one of the following options:

      - To create a new version, choose **Create a new version and to associate it to this alias**.
      - To use an existing version, choose **Use an existing version to associate this alias**. From the dropdown menu, choose the version that you want to associate the alias to.

   6. Select **Create alias**. A success banner appears at the top.
To create an alias for an agent, send a **CreateAgentAlias** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. To create a new version and associate this alias with it, leave the **routingConfiguration** object unspecified.

**See code examples**

2. Deploy your agent by setting up your application to make an **InvokeAgent** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock runtime endpoint**. In the **agentAliasId** field, specify the ID of the alias pointing to the version of the agent that you want to use.

To learn how to manage versions and aliases of agents, select from the following topics.

**Topics**

- **Manage versions of agents in Amazon Bedrock**
- **Manage aliases of agents in Amazon Bedrock**

**Manage versions of agents in Amazon Bedrock**

After you create a version of your agent, you can view information about it or delete it. You can only create a new version of an agent by creating a new alias.

**Topics**

- **View information about versions of agents in Amazon Bedrock**
- **Delete a version of an agent in Amazon Bedrock**

**View information about versions of agents in Amazon Bedrock**

To learn how to view information about the versions of an agent, select the tab corresponding to your method of choice and follow the steps.
To view information about a version of an agent

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose the version to view from the Versions section.
4. To view details about the model, action groups, or knowledge bases attached to version of the agent, choose the name of the information that you want to view. You can't modify any part of a version. To make modifications to the agent, use the working draft and create a new version.

API

To get information about an agent version, send a GetAgentVersion request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint. Specify the agentId and agentVersion.

To list information about an agent's versions, send a ListAgentVersions request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and specify the agentId. You can specify the following optional parameters:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxResults</td>
<td>The maximum number of results to return in a response.</td>
</tr>
<tr>
<td>nextToken</td>
<td>If there are more results than the number you specified in the maxResults field, the response returns a nextToken value. To see the next batch of results, send the nextToken value in another request.</td>
</tr>
</tbody>
</table>
Delete a version of an agent in Amazon Bedrock

To learn how to delete a version of an agent, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To delete a version of an agent**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. To choose the version for deletion, in the **Versions** section, choose the option button next to the version that you want to delete.
4. Choose **Delete**.
5. A dialog box appears warning you about the consequences of deletion. To confirm that you want to delete the version, enter **delete** in the input field and choose **Delete**.
6. A banner appears to inform you that the version is being deleted. When deletion is complete, a success banner appears.

**API**

To delete a version of an agent, send a **DeleteAgentVersion** request (see link for request and response formats and field details) with an **Agents for Amazon Bedrock build-time endpoint**. By default, the **skipResourceInUseCheck** parameter is **false** and deletion is stopped if the resource is in use. If you set **skipResourceInUseCheck** to **true**, the resource will be deleted even if the resource is in use.

Manage aliases of agents in Amazon Bedrock

After you create an alias of your agent, you can view information about it, edit it, or delete it.

**Topics**

- View information about aliases of agents in Amazon Bedrock
- Edit an alias of an agent in Amazon Bedrock
- Delete an alias of an agent in Amazon Bedrock
View information about aliases of agents in Amazon Bedrock

To learn how to view information about the aliases of an agent, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To view the details of an alias**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at [https://console.aws.amazon.com/bedrock/](https://console.aws.amazon.com/bedrock/).
2. Select **Agents** from the left navigation pane. Then, choose an agent in the **Agents** section.
3. Choose the alias to view from the **Aliases** section.
4. You can view the name and description of the alias and tags that are associated with the alias.

**API**

To get information about an agent alias, send a [GetAgentAlias](https://docs.aws.amazon.com/bedrock/latest/APIReference/API_GetAgentAlias.html) request (see link for request and response formats and field details) with an [Agents for Amazon Bedrock build-time endpoint](https://docs.aws.amazon.com/bedrock/latest/APIReference/API_GetAgentAlias.html). Specify the `agentId` and `agentAliasId`.

To list information about an agent's aliases, send a [ListAgentVersions](https://docs.aws.amazon.com/bedrock/latest/APIReference/API_ListAgentVersions.html) request (see link for request and response formats and field details) with an [Agents for Amazon Bedrock build-time endpoint](https://docs.aws.amazon.com/bedrock/latest/APIReference/API_ListAgentVersions.html) and specify the `agentId`. You can specify the following optional parameters:

<table>
<thead>
<tr>
<th>Field</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxResults</td>
<td>The maximum number of results to return in a response.</td>
</tr>
<tr>
<td>nextToken</td>
<td>If there are more results than the number you specified in the <code>maxResults</code> field,</td>
</tr>
<tr>
<td></td>
<td>the response returns a <code>nextToken</code> value. To see the next batch of results, send</td>
</tr>
<tr>
<td></td>
<td>the <code>nextToken</code> value in another request.</td>
</tr>
</tbody>
</table>
To view all the tags for an alias, send a ListTagsForResource request (see link for request and response formats and field details) with an Agents for Amazon Bedrock build-time endpoint and include the Amazon Resource Name (ARN) of the alias.

**Edit an alias of an agent in Amazon Bedrock**

To learn how to edit an alias of an agent, select the tab corresponding to your method of choice and follow the steps.

**Console**

**To edit an alias**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. In the Aliases section, choose the option button next to the alias that you want to edit.
4. You can edit the name and description of the alias. Additionally, you can perform one of the following actions:
   - To create a new version and associate this alias with that version, choose Create a new version and associate it to this alias.
   - To associate this alias with a different existing version, choose Use an existing version and associate this alias.

**To add or remove tags associated with an alias**

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. Choose the alias for which you want to manage tags from the Aliases section.
4. In the Tags section, choose Manage tags.
5. To add a tag, choose Add new tag. Then enter a Key and optionally enter a Value. To remove a tag, choose Remove. For more information, see Tag resources.
6. When you're done editing tags, choose Submit.
API

To edit an agent alias, send an `UpdateAgentAlias` request. Because all fields will be overwritten, include both fields that you want to update as well as fields that you want to keep the same.

To add tags to an alias, send a `TagResource` request (see link for request and response formats and field details) with an `Agents for Amazon Bedrock build-time endpoint` and include the Amazon Resource Name (ARN) of the alias. The request body contains a `tags` field, which is an object containing a key-value pair that you specify for each tag.

To remove tags from an alias, send an `UntagResource` request (see link for request and response formats and field details) with an `Agents for Amazon Bedrock build-time endpoint` and include the Amazon Resource Name (ARN) of the alias. The `tagKeys` request parameter is a list containing the keys for the tags that you want to remove.

Delete an alias of an agent in Amazon Bedrock

To learn how to delete an alias of an agent, select the tab corresponding to your method of choice and follow the steps.

Console

To delete an alias

1. Sign in to the AWS Management Console, and open the Amazon Bedrock console at `https://console.aws.amazon.com/bedrock/`.
2. Select Agents from the left navigation pane. Then, choose an agent in the Agents section.
3. To choose the alias for deletion, in the Aliases section, choose the option button next to the alias that you want to delete.
4. Choose Delete.
5. A dialog box appears warning you about the consequences of deletion. To confirm that you want to delete the alias, enter delete in the input field and choose Delete.
6. A banner appears to inform you that the alias is being deleted. When deletion is complete, a success banner appears.
API

To delete an alias of an agent, send a DeleteAgentAlias request (see link for request and response formats and field details) with an [Agents for Amazon Bedrock build-time endpoint](#). By default, the skipResourceInUseCheck parameter is false and deletion is stopped if the resource is in use. If you set skipResourceInUseCheck to true, the resource will be deleted even if the resource is in use.

[See code examples](#)
Custom models

Model customization is the process of providing training data to a model in order to improve its performance for specific use-cases. You can customize Amazon Bedrock foundation models in order to improve their performance and create a better customer experience. Amazon Bedrock currently provides the following customization methods.

- **Continued Pre-training**

  Provide *unlabeled* data to pre-train a foundation model by familiarizing it with certain types of inputs. You can provide data from specific topics in order to expose a model to those areas. The Continued Pre-training process will tweak the model parameters to accommodate the input data and improve its domain knowledge.

  For example, you can train a model with private data, such as business documents, that are not publically available for training large language models. Additionally, you can continue to improve the model by retraining the model with more unlabeled data as it becomes available.

- **Fine-tuning**

  Provide *labeled* data in order to train a model to improve performance on specific tasks. By providing a training dataset of labeled examples, the model learns to associate what types of outputs should be generated for certain types of inputs. The model parameters are adjusted in the process and the model's performance is improved for the tasks represented by the training dataset.

For information about model customization quotas, see [Model customization quotas](#).

### Note

You are charged for model training based on the number of tokens processed by the model (number of tokens in training data corpus × number of epochs) and model storage charged per month per model. For more information, see [Amazon Bedrock pricing](#).

You carry out the following steps in model customization.

1. **Create a training and, if applicable, a validation dataset** for your customization task.
2. If you plan to use a new custom IAM role, set up IAM permissions to access the S3 buckets for your data. You can also use an existing role or let the console automatically create a role with the proper permissions.

3. (Optional) Configure KMS keys and/or VPC for extra security.

4. Create a Fine-tuning or Continued Pre-training job, controlling the training process by adjusting the hyperparameter values.

5. Analyze the results by looking at the training or validation metrics or by using model evaluation.

6. Purchase Provisioned Throughput for your newly created custom model.

7. Use your custom model as you would a base model in Amazon Bedrock tasks, such as model inference.

**Topics**

- Supported regions and models for model customization
- Prerequisites for model customization
- Submit a model customization job
- Manage a model customization job
- Analyze the results of a model customization job
- Use a custom model
- Code samples for model customization
- Guidelines for model customization
- Troubleshooting

**Supported regions and models for model customization**

The following table shows regional support for each customization method:

<table>
<thead>
<tr>
<th>Region</th>
<th>Fine-tuning</th>
<th>Continued pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>US East (N. Virginia)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>US West (Oregon)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>AWS GovCloud (US-West)</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
The following table shows model support for each customization method:

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID</th>
<th>Fine-tuning</th>
<th>Continued pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>amazon.titan-text-express-v1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Lite</td>
<td>amazon.titan-text-lite-v1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon Titan Image Generator G1</td>
<td>amazon.titan-image-generator-v1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Multimodal Embeddings G1 G1</td>
<td>amazon.titan-embed-image-v1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>cohere.command-text-v14</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>cohere.command-light-text-v14</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>meta.llama2-13b-chat-v1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>meta.llama2-70b-chat-v1</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Prerequisites for model customization**

Before you can start a model customization job, you need to fulfill the following prerequisites:

1. Determine whether you plan to carry out a Fine-tuning or Continued Pre-training job and which model you plan to use. The choice you make determines the format of the datasets that you feed into the customization job.
2. Prepare the training dataset file. If the customization method and model that you choose supports a validation dataset, you can also prepare a validation dataset file. Follow the steps below in Prepare the datasets and then upload the files to an Amazon S3 bucket.

3. (Optional) Create a custom AWS Identity and Access Management (IAM) service role with the proper permissions by following the instructions at Create a service role for model customization to set up the role. You can skip this prerequisite if you plan to use the AWS Management Console to automatically create a service role for you.

4. (Optional) Set up extra security configurations.
   - You can encrypt input and output data, customization jobs, or inference requests made to custom models. For more information, see Encryption of model customization jobs and artifacts.
   - You can create a virtual private cloud (VPC) to protect your customization jobs. For more information, see Protect model customization jobs using a VPC.

Topics
- Prepare the datasets
- Protect model customization jobs using a VPC

Prepare the datasets

Before you can begin a model customization job, you need to minimally prepare a training dataset. Whether a validation dataset is supported and the format of your training and validation dataset depend on the following factors.

- The type of customization job (fine-tuning or Continued Pre-training).
- The input and output modalities of the data.

To see dataset and file requirements for different models, see Model customization quotas.

Select the tab that is relevant to your use-case.

Fine-tuning: Text-to-text

To fine-tune a text-to-text model, prepare a training and optional validation dataset by creating a JSONL file with multiple JSON lines. Each JSON line is a sample containing both a prompt
and completion field. Use 6 characters per token as an approximation for the number of tokens. The format is as follows.

```json
[{
   "prompt": "<prompt1>",
   "completion": "<expected generated text>",
}
{
   "prompt": "<prompt2>",
   "completion": "<expected generated text>",
}
{
   "prompt": "<prompt3>",
   "completion": "<expected generated text>",
}
]
```

The following is an example item for a question-answer task:

```json
[{
   "prompt": "what is AWS",
   "completion": "it's Amazon Web Services"
}
]
```

**Fine-tuning: Text-to-image & Image-to-embeddings**

To fine-tune a text-to-image or image-to-embedding model, prepare a training dataset by create a JSONL file with multiple JSON lines. Validation datasets are not supported. Each JSON line is a sample containing an image-ref, the Amazon S3 URI for an image, and a caption that could be a prompt for the image.

The images must be in JPEG or PNG format.

```json
[{
   "image-ref": "s3://bucket/path/to/image001.png",
   "caption": "<prompt text>",
}
{
   "image-ref": "s3://bucket/path/to/image002.png",
   "caption": "<prompt text>",
}
{
   "image-ref": "s3://bucket/path/to/image003.png",
   "caption": "<prompt text>",
}
]
```

The following is an example item:

```json
[{
   "image-ref": "s3://my-bucket/my-pets/cat.png",
   "caption": "an orange cat with white spots"
}
]
```

To allow Amazon Bedrock access to the image files, add an IAM policy similar to the one in Permissions to access training and validation files and to write output files in S3 to the Amazon Bedrock model customization service role that you set up or that was automatically set up for you in the console. The Amazon S3 paths you provide in the training dataset must be in folders that you specify in the policy.

**Continued Pre-training: Text-to-text**

To carry out Continued Pre-training on a text-to-text model, prepare a training and optional validation dataset by creating a JSONL file with multiple JSON lines. Because Continued Pre-training involves unlabeled data, each JSON line is a sample containing only an input field. Use 6 characters per token as an approximation for the number of tokens. The format is as follows.
Protect model customization jobs using a VPC

When you run a model customization job, the job accesses your Amazon S3 bucket to download the input data and to upload job metrics. To control access to your data, we recommend that you use a virtual private cloud (VPC) with Amazon VPC. You can further protect your data by configuring your VPC so that your data isn’t available over the internet and instead creating a VPC interface endpoint with AWS PrivateLink to establish a private connection to your data. For more information about how Amazon VPC and AWS PrivateLink integrate with Amazon Bedrock, see Protect your data using Amazon VPC and AWS PrivateLink.

Carry out the following steps to configure and use a VPC for the training, validation, and output data for your model customization jobs.

Topics
- Set up a VPC
- Create an Amazon S3 VPC Endpoint
- (Optional) Use IAM policies to restrict access to your S3 files
- Attach VPC permissions to a model customization role
- Add the VPC configuration when submitting a model customization job

Set up a VPC

You can use a default VPC for your model customization data or create a new VPC by following the guidance at Get started with Amazon VPC and Create a VPC.

When you create your VPC, we recommend that you use the default DNS settings for your endpoint route table, so that standard Amazon S3 URLs (for example, http://s3-aws-region.amazonaws.com/training-bucket) resolve.
Create an Amazon S3 VPC Endpoint

If you configure your VPC with no internet access, you need to create an Amazon S3 VPC endpoint to allow your model customization jobs to access the S3 buckets that store your training and validation data and that will store the model artifacts.

Create the S3 VPC endpoint by following the steps at Create a gateway endpoint for Amazon S3.

**(Optional) Use IAM policies to restrict access to your S3 files**

You can use resource-based policies to more tightly control access to your S3 files. You can use any combination of the following types of resource-based policies.

- **Endpoint policies** – Endpoint policies restrict access through the VPC endpoint. The default endpoint policy allows full access to Amazon S3 for any user or service in your VPC. While creating or after you create the endpoint, you can optionally attach a resource-based policy to the endpoint to add restrictions, such as only allowing the endpoint to access a specific bucket or only allowing a specific IAM role to access the endpoint. For examples, see Edit the VPC endpoint policy.

The following is an example policy you can attach to your VPC endpoint to only allow it to access the bucket containing your training data.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "RestrictAccessToTrainingBucket",
            "Effect": "Allow",
            "Principal": "*",
            "Action": ["s3:GetObject",
```
Bucket policies – Bucket policies restrict access to S3 buckets. You can use a bucket policy to restrict access to traffic that comes from your VPC. To attach a bucket policy, follow the steps at Using bucket policies and use the aws:sourceVpc, aws:sourceVpce, or aws:VpcSourceIp condition keys. For examples, see Control access using bucket policies.

The following is an example policy you can attach to the S3 bucket that will contain your output data to deny all traffic to the bucket unless it comes from your VPC.

```json
{
    "Version": "2012-10-17",
    "Statement": [{
        "Sid": "RestrictAccessToOutputBucket",
        "Effect": "Deny",
        "Principal": "*",
        "Action": [
            "s3:GetObject",
            "s3:PutObject",
            "s3:ListBucket"
        ],
        "Resource": [
            "arn:aws:s3:::output-bucket",
            "arn:aws:s3:::output-bucket/*"
        ],
        "Condition": {
            "StringNotEquals": {
                "aws:sourceVpc": "your-vpc-id"
            }
        }
    }
    ]
}
```
Attach VPC permissions to a model customization role

After you finish setting up your VPC and endpoint, you need to attach the following permissions to your model customization IAM role. Modify this policy to allow access to only the VPC resources that your job needs. Replace the subnet-ids and security-group-id with the values from your VPC.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ec2:DescribeNetworkInterfaces",
                "ec2:DescribeVpcs",
                "ec2:DescribeDhcpOptions",
                "ec2:DescribeSubnets",
                "ec2:DescribeSecurityGroups"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Allow",
            "Action": [
                "ec2:CreateNetworkInterface"
            ],
            "Resource": [
            ],
            "Condition": {
                "StringEquals": {
                    "aws:RequestTag/BreadcrumbManaged": ["true"]
                },
                "ArnEquals": {
                    "aws:RequestTag/BreadcrumbModelCustomizationJobArn": [
                    ]
                }
            }
        }
    ]
}
```
["arn:aws:ec2:region:account-id:subnet/subnet-id",
"arn:aws:ec2:region:account-id:security-group/security-group-id"
]
},
{
"Effect": "Allow",
"Action": [
"ec2:CreateNetworkInterfacePermission",
"ec2:DeleteNetworkInterface",
"ec2:DeleteNetworkInterfacePermission",
],
"Resource": "*",
"Condition": {
"ArnEquals": {
"ec2:Subnet": [
"arn:aws:ec2:region:account-id:subnet/subnet-id2"
],
"ec2:ResourceTag/BedrockModelCustomizationJobArn": [
},
"StringEquals": {
"ec2:ResourceTag/BedrockManaged": "true"
}
}
},
{
"Effect": "Allow",
"Action": [
"ec2:CreateTags"
],
"Condition": {
"StringEquals": {
"ec2:CreateAction": [
"CreateNetworkInterface"
]
},
"ForAllValues:StringEquals": {
"aws:TagKeys": [
"BedrockManaged",
"BedrockManaged"]
)}
Add the VPC configuration when submitting a model customization job

After you configure the VPC and the required roles and permissions as described in the previous sections, you can create a model customization job that uses this VPC.

When you specify the VPC subnets and security groups for a job, Amazon Bedrock creates *elastic network interfaces* (ENIs) that are associated with your security groups in one of the subnets. ENIs allow the Amazon Bedrock job to connect to resources in your VPC. For information about ENIs, see *Elastic Network Interfaces* in the *Amazon VPC User Guide*. Amazon Bedrock tags ENIs that it creates with *BedrockManaged* and *BedrockModelCustomizationJobArn* tags.

We recommend that you provide at least one subnet in each Availability Zone.

You can use security groups to establish rules for controlling Amazon Bedrock access to your VPC resources.

You can configure the VPC to use in either the console or through the API. Select the tab corresponding to your method of choice and follow the steps.

**Console**

For the Amazon Bedrock console, you specify VPC subnets and security groups in the optional *VPC settings* section when you create the model customization job. For more information about configuring jobs, see *Submit a model customization job*.

**Note**

For a job that includes VPC configuration, the console can't automatically create a service role for you. Follow the guidance at *Create a service role for model customization* to create a custom role.
When you submit a `CreateModelCustomizationJob` request, you can include a VpcConfig as a request parameter to specify the VPC subnets and security groups to use, as in the following example.

```
"VpcConfig": {
  "SecurityGroupIds": [
    "sg-0123456789abcdef0"
  ],
  "Subnets": [
    "subnet-0123456789abcdef0",
    "subnet-0123456789abcdef1",
    "subnet-0123456789abcdef2"
  ]
}
```

Submit a model customization job

You can create a custom model by using Fine-tuning or Continued Pre-training in the Amazon Bedrock console or API. The customization job can take several hours. The duration of the job depends on the size of the training data (number of records, input tokens, and output tokens), number of epochs, and batch size. Select the tab corresponding to your method of choice and follow the steps.

See code examples

Console

To submit a model customization job in the console, carry out the following steps.

1. In the Amazon Bedrock console, choose Custom models under Foundation models from the left navigation pane.
2. In the Models tab, choose Customize model and then Create Fine-tuning job or Create Continued Pre-training job, depending on the type of model you want to train.
3. In the Model details section, do the following.
   a. Choose the model that you want to customize with your own data and give your resulting model a name.
b. (Optional) By default, Amazon Bedrock encrypts your model with a key owned and managed by AWS. To use a custom KMS key, select Model encryption and choose a key.

c. (Optional) To associate tags with the custom model, expand the Tags section and select Add new tag.

4. In the Job configuration section, enter a name for the job and optionally add any tags to associate with the job.

5. (Optional) To use a virtual private cloud (VPC) to protect your training data and customization job, select a VPC that contains the input data and output data Amazon S3 locations, its subnets, and security groups in the VPC settings section.

   Note
   If you include a VPC configuration, the console cannot create a new service role for the job. Create a custom service role and add permissions similar to the example described in Attach VPC permissions to a model customization role.

6. In the Input data section, select the S3 location of the training dataset file and, if applicable, the validation dataset file.

7. In the Hyperparameters section, input values for hyperparameters to use in training.

8. In the Output data section, enter the Amazon S3 location where Amazon Bedrock should save the output of the job. Amazon Bedrock stores the training loss metrics and validation loss metrics for each epoch in separate files in the location that you specify.

9. In the Service access section, select one of the following:

   • Use an existing service role – Select a service role from the drop-down list. For more information on setting up a custom role with the appropriate permissions, see Create a service role for model customization.

   • Create and use a new service role – Enter a name for the service role.

10. Choose Fine-tune model or Create Continued Pre-training job to begin the job.

API

Request
Send a CreateModelCustomizationJob (see link for request and response formats and field details) request with an Amazon Bedrock control plane endpoint to submit a model customization job. Minimally, you must provide the following fields.

- **roleArn** – The ARN of the service role with permissions to customize models. Amazon Bedrock can automatically create a role with the appropriate permissions if you use the console, or you can create a custom role by following the steps at Create a service role for model customization.

  **Note** If you include a vpcConfig field, make sure that the role has the proper permissions to access the VPC. For an example, see Attach VPC permissions to a model customization role.

- **baseModelIdentifier** – The model ID or ARN of the foundation model to customize.
- **customModelName** – The name to give the newly customized model.
- **jobName** – The name to give the training job.
- **hyperParameters** – Hyperparameters that affect the model customization process.
- **trainingDataConfig** – An object containing the Amazon S3 URI of the training dataset. Depending on the customization method and model, you can also include a validationDataConfig. For more information about preparing the datasets, see Prepare the datasets.
- **outputDataConfig** – An object containing the Amazon S3 URI to write the output data to.

If you don't specify the customizationType, the model customization method defaults to FINE_TUNING.

To prevent the request from completing more than once, include a clientRequestToken.

You can include the following optional fields for extra configurations.

- **jobTags** and/or **customModelTags** – Associate tags with the customization job or resulting custom model.
- **customModelKmsKeyId** – Include a custom KMS key to encrypt your custom model.
- **vpcConfig** – Include the configuration for a virtual private cloud (VPC) to protect your training data and customization job.
Response

The response returns a jobArn that you can use to monitor or stop the job.

Manage a model customization job

Once you start a model customization job, you can track its progress or stop it. If you do so through the API, you will need the jobArn. You can find it in one of the following ways:

1. In the Amazon Bedrock console
   1. Select Custom models under Foundation models from the left navigation pane.
   2. Choose the job from the Training jobs table to see details, including the ARN of the job.
2. Look in the jobArn field in the response returned from the CreateModelCustomizationJob call that created the job or from a CreateModelCustomizationJob call.

Monitor a model customization job

After you begin a job, you can monitor its progress in the console or API. Select the tab corresponding to your method of choice and follow the steps.

See code examples

Console

To monitor the status of your fine-tuning jobs

1. In the Amazon Bedrock console, choose Custom models under Foundation models from the left navigation pane.
2. Select the Training jobs tab to display the fine-tuning jobs that you have initiated. Look at the Status column to monitor the progress of the job.
3. Select a job to view the details you input for training.

API

To list information about all your model customization jobs, send a CreateModelCustomizationJob request with an Amazon Bedrock control plane endpoint. Refer to CreateModelCustomizationJob for filters that you can use.
To monitor the status of a model customization job, send a `GetModelCustomizationJob` request with an Amazon Bedrock control plane endpoint with the jobArn of the job.

**Stop a model customization job**

You can stop an Amazon Bedrock model customization job while it's in progress. Select the tab corresponding to your method of choice and follow the steps.

**See code examples**

⚠️ **Warning**

You can't resume a stopped job. Amazon Bedrock charges for the tokens that it used to train the model before you stopped the job. Amazon Bedrock doesn't create an intermediate custom model for a stopped job.

**Console**

**To stop a model customization job**

1. In the Amazon Bedrock console, choose **Custom models** under **Foundation models** from the left navigation pane.
2. In the **Training Jobs** tab, choose the radio button next to the job to stop or select the job to stop to navigate to the details page.
3. Select the **Stop job** button. You can only stop a job if its status is **Training**.
4. A modal appears to warn you that you can't resume the training job if you stop it. Select **Stop job** to confirm.

**API**

To stop a model customization job, send a `CreateModelCustomizationJob` (see link for request and response formats and field details) request with a Amazon Bedrock control plane endpoint, using the jobArn of the job.

You can only stop a job if its status is **IN_PROGRESS**. Check the status with a `GetModelCustomizationJob` request. The system marks the job for termination and sets the state to **STOPPING**. Once the job is stopped, the state becomes **STOPPED**.
Analyze the results of a model customization job

After a model customization job completes, you can analyze the results of the training process by looking at the files in the output S3 folder that you specified when you submitted the job or view details about the model. Amazon Bedrock stores your customized models in AWS-managed storage scoped to your account.

You can also evaluate your model by running a model evaluation job. For more information, see Model evaluation.

The S3 output for a model customization job contains the following output files in your S3 folder. The validation artifacts only appear if you included a validation dataset.

- model-customization-job-*training-job-id*/
  - training_artifacts/
    - step_wise_training_metrics.csv
  - validation_artifacts/
    - post_fine_tuning_validation/
      - validation_metrics.csv

Use the step_wise_training_metrics.csv and the validation_metrics.csv files to analyze the model customization job and to help you adjust the model as necessary.

The columns in the step_wise_training_metrics.csv file are as follows.

- step_number – The step in the training process. Starts from 0.
- epoch_number – The epoch in the training process.
- training_loss – Indicates how well the model fits the training data. A lower value indicates a better fit.
- perplexity – Indicates how well the model can predict a sequence of tokens. A lower value indicates better predictive ability.

The columns in the validation_metrics.csv file are the same as the training file, except that validation_loss (how well the model fits the validation data) appears in place of training_loss.
You can find the output files by opening up the [https://console.aws.amazon.com/s3](https://console.aws.amazon.com/s3) directly or by finding the link to the output folder within your model details. Select the tab corresponding to your method of choice and follow the steps.

**See code examples**

**Console**

1. In the Amazon Bedrock console, choose **Custom models** under **Foundation models** from the left navigation pane.
2. In the **Models** tab, select a model to view its details. The **Job name** can be found in the **Model details** section.
3. To view the output S3 files, select the **S3 location** in the **Output data** section.
4. Find the training and validation metrics files in the folder whose name matches the **Job name** for the model.

**API**

To list information about all your custom models, send a [ListCustomModels](https://console.aws.amazon.com/s3) (see link for request and response formats and field details) request with an [Amazon Bedrock control plane endpoint](https://console.aws.amazon.com/s3). Refer to [ListCustomModels](https://console.aws.amazon.com/s3) for filters that you can use.

To monitor the status of a model customization job, send a [GetCustomModel](https://console.aws.amazon.com/s3) (see link for request and response formats and field details) request with an [Amazon Bedrock control plane endpoint](https://console.aws.amazon.com/s3) with the **modelIdentifier**, which is either of the following.

- The name that you gave the model.
- The ARN of the model.

You can see **trainingMetrics** and **validationMetrics** for a model customization job in either the [GetModelCustomizationJob](https://console.aws.amazon.com/s3) or [GetCustomModel](https://console.aws.amazon.com/s3) response.

To download the training and validation metrics files, follow the steps at [Downloading objects](https://console.aws.amazon.com/s3). Use the S3 URI you provided in the outputDataConfig.
Use a custom model

Before you can use a customized model, you need to purchase Provisioned Throughput for it. For more information about Provisioned Throughput, see Provisioned Throughput. You can then use the resulting provisioned model for inference. Select the tab corresponding to your method of choice and follow the steps.

See code examples

Console

To purchase Provisioned Throughput for a custom model.

1. In the Amazon Bedrock console, choose Custom models under Foundation models from the left navigation pane.
2. In the Models tab, choose the radio button next to the model for which you want to buy Provisioned Throughput or select the model name to navigate to the details page.
3. Select Purchase Provisioned Throughput.

When you carry out any operation that supports usage of custom models, you will see your custom model as an option in the model selection menu.

API

To purchase Provisioned Throughput for a custom model, send a CreateProvisionedModelThroughput (see link for request and response formats and field details) request with an Amazon Bedrock control plane endpoint, using the name or ARN of your custom model as the modelId. The response returns a provisionedModelArn that you can use as the modelId when making an InvokeModel or InvokeModelWithResponseStream request.

Code samples for model customization

The following code samples show how to prepare a basic dataset, set up permissions, create a custom model, view the output files, purchase throughput for the model, and run inference on the model. You can modify these code snippets to your specific use-case.

1. Prepare the training dataset.
a. Create a training dataset file containing the following one line and name it \textit{train.jsonl}.

\begin{verbatim}
{"prompt": "what is AWS", "completion": "it's Amazon Web Services"}
\end{verbatim}

b. Create an S3 bucket for your training data and another one for your output data (the names must be unique).

c. Upload \textit{train.jsonl} into the training data bucket.

2. Create a policy to access your training and attach it to an IAM role with a Amazon Bedrock trust relationship. Select the tab corresponding to your method of choice and follow the steps.

\textbf{Console}

1. Create the S3 policy.

a. Navigate to the IAM console at \url{https://console.aws.amazon.com/iam} and choose \textbf{Policies} from the left navigation pane.

b. Select \textbf{Create policy} and then choose \textbf{JSON} to open the \textbf{Policy editor}.

c. Paste the following policy, replacing \textit{training-bucket} and \textit{output-bucket} with your bucket names, and then select \textbf{Next}.

\begin{verbatim}
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::training-bucket",
                "arn:aws:s3:::training-bucket/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:PutObject",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::training-bucket",
                "arn:aws:s3:::training-bucket/*"
            ]
        }
    ]
}
\end{verbatim}
d. Name the policy `MyFineTuningDataAccess` and select Create policy.

2. Create an IAM role and attach the policy.

   a. From the left navigation pane, choose Roles and then select Create role.

   b. Select Custom trust policy, paste the following policy, and select Next.

   ```json
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Effect": "Allow",
         "Principal": {
           "Service": "bedrock.amazonaws.com"
         },
         "Action": "sts:AssumeRole"
       }
     ]
   }
   ```

   c. Search for the `MyFineTuningDataAccess` policy you created, select the checkbox, and choose Next.

   d. Name the role `MyCustomizationRole` and select Create role.

**CLI**

1. Create a file called `BedrockTrust.json` and paste the following policy into it.

   ```json
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Effect": "Allow",
         "Principal": {
           "Service": "bedrock.amazonaws.com"
         },
         "Action": "sts:AssumeRole"
       }
     ]
   }
   ```
2. Create another file called `MyFineTuningDataAccess.json` and paste the following policy into it, replacing `training-bucket` and `output-bucket` with your bucket names.

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3:::training-bucket",
        "arn:aws:s3:::training-bucket/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3:::training-bucket",
        "arn:aws:s3:::training-bucket/*"
      ]
    }
  ]
}
```

3. In a terminal, navigate to the folder containing the policies you created.
4. Make a `CreateRole` request to create an IAM role called `MyCustomizationRole` and attach the `BedrockTrust.json` trust policy that you created.

```
aws iam create-role \
    --role-name MyCustomizationRole \
    --assume-role-policy-document file://BedrockTrust.json
```

5. Make a `CreatePolicy` request to create the S3 data access policy with the `MyFineTuningDataAccess.json` file you created. The response returns an Arn for the policy.

```
aws iam create-policy \
    --policy-name MyFineTuningDataAccess \
    --policy-document file://myFineTuningDataAccess.json
```

6. Make an `AttachRolePolicy` request to attach the S3 data access policy to your role, replacing the `policy-arn` with the ARN in the response from the previous step:

```
aws iam attach-role-policy \
    --role-name MyCustomizationRole \
    --policy-arn policy-arn
```

**Python**

1. Run the following code to make a `CreateRole` request to create an IAM role called `MyCustomizationRole` and to make a `CreatePolicy` request to create an S3 data access policy called `MyFineTuningDataAccess`. For the S3 data access policy, replace `training-bucket` and `output-bucket` with your S3 bucket names.

```python
import boto3
import json

iam = boto3.client("iam")

iam.create_role(
    RoleName="MyCustomizationRole",
    AssumeRolePolicyDocument=json.dumps({
        "Version": "2012-10-17",
        "Statement": [

```
"Effect": "Allow",
"Principal": {
    "Service": "bedrock.amazonaws.com"
},
"Action": "sts:AssumeRole"
}
)}
)
)

iam.create_policy(
    PolicyName="MyFineTuningDataAccess",
    PolicyDocument=json.dumps({
        "Version": "2012-10-17",
        "Statement": [
            {
                "Effect": "Allow",
                "Action": [
                    "s3:GetObject",
                    "s3:ListBucket"
                ],
                "Resource": [
                    "arn:aws:s3:::training-bucket",
                    "arn:aws:s3:::training-bucket/*"
                ]
            },
            {
                "Effect": "Allow",
                "Action": [
                    "s3:GetObject",
                    "s3:PutObject",
                    "s3:ListBucket"
                ],
                "Resource": [
                    "arn:aws:s3:::output-bucket",
                    "arn:aws:s3:::output-bucket/*"
                ]
            }
        ]
    })
)
)

2. An Arn is returned in the response. Run the following code snippet to make an AttachRolePolicy request, replacing policy-arn with the returned Arn.
iam.attach_role_policy(
    RoleName="MyCustomizationRole",
    PolicyArn="policy-arn"
)

3. Select a language to see code samples to call the model customization API operations.

**CLI**

First, create a text file named `FineTuningData.json`. Copy the JSON code from below into the text file, replacing `training-bucket` and `output-bucket` with your S3 bucket names.

```json
{
    "trainingDataConfig": {
        "s3Uri": "s3://training-bucket/train.jsonl"
    },
    "outputDataConfig": {
        "s3Uri": "s3://output-bucket"
    }
}
```

To submit a model customization job, navigate to the folder containing `FineTuningData.json` in a terminal and run the following command in the command line, replacing `your-customization-role-arn` with the model customization role that you set up.

```bash
aws bedrock create-model-customization-job \
    --customization-type FINE_TUNING \
    --base-model-identifier arn:aws:bedrock:us-east-1::foundation-model/\n    amazon.titan-text-express-v1 \
    --role-arn your-customization-role-arn \
    --job-name myFineTuningJob \
    --custom-model-name myCustomModel \
    --hyper-parameters
    epochCount=1,batchSize=1,learningRate=.0005,learningRateWarmupSteps=0 \
    --cli-input-json file://FineTuningData.json
```

The response returns a `jobArn`. Allow the job some time to complete. You can check its status with the following command.
When the status is COMPLETE, you can see the trainingMetrics in the response. You can download the artifacts to the current folder by running the following command, replacing `aet.et-bucket` with your output bucket name and `jobId` with the ID of the customization job (the sequence following the last slash in the jobArn).

```
aws s3 cp s3://output-bucket/model-customization-job-jobId . --recursive
```

Purchase a no-commitment Provisioned Throughput for your custom model with the following command.

```
aws bedrock create-provisioned-model-throughput
   --model-id myCustomModel
   --provisioned-model-name myProvisionedCustomModel
   --model-units 1
```

The response returns a `provisionedModelArn` that you can use as a `modelId` in inference. Allow the Provisioned Throughput some time to be created. You can check its status with the following command.

```
aws bedrock get-provisioned-model-throughput
   --provisioned-model-id provisionedModelArn
```

When the status is InService, you can run inference with your custom model with the following command. Replace the `provisionedModelArn` with the one returned to you. The output is written to a file named `output.txt` in your current folder.

```
aws bedrock-runtime invoke-model
   --model-id provisionedModelArn
```
```python
import boto3
import json

bedrock = boto3.client(service_name='bedrock')

# Set parameters
customizationType = "FINE_TUNING"
baseModelIdentifier = "arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-text-express-v1"
roleArn = "your-customization-role-arn"
jobName = "myFineTuningJob"
customModelName = "myCustomModel"
hyperParameters = {
    "epochCount": "1",
    "batchSize": "1",
    "learningRate": ".0005",
    "learningRateWarmupSteps": "0"
}
trainingDataConfig = {
    "s3Uri": "s3://training-bucket/myInputData/train.jsonl"
}
outputDataConfig = {
    "s3Uri": "s3://output-bucket/myOutputData"
}

# Create job
response_ft = bedrock.create_model_customization_job(
    jobName=jobName,
    customModelName=customModelName,
    roleArn=roleArn,
    baseModelIdentifier=baseModelIdentifier,
    hyperParameters=hyperParameters,
    trainingDataConfig=trainingDataConfig,
    outputDataConfig=outputDataConfig
)
```

Run the following code snippet to submit a fine-tuning job. Replace `your-customization-role-arn` with the ARN of the `myCustomizationRole` that you set up and replace `training-bucket` and `output-bucket` with your S3 bucket names.
jobArn = response_ft.get('jobArn')

The response returns a **jobArn**. Allow the job some time to complete. You can check its status with the following command.

```
bedrock.get_model_customization_job(jobIdentifier=jobArn).get('status')
```

When the status is **COMPLETE**, you can see the **trainingMetrics** in the **GetModelCustomizationJob** response. You can also follow the steps at **Downloading objects** to download the metrics.

Purchase a no-commitment Provisioned Throughput for your custom model with the following command.

```
response_pt = bedrock.create_provisioned_model_throughput(
    modelId="myCustomModel",
    provisionedModelName="myProvisionedCustomModel"
    modelUnits="1"
)
provisionedModelArn = response_pt.get('provisionedModelArn')
```

The response returns a **provisionedModelArn** that you can use as a **modelId** in inference. Allow the Provisioned Throughput some time to be created. You can check its status with the following command.

```
bedrock.get_provisioned_model_throughput(provisionedModelId=provisionedModelArn)
```

When the status is **InService**, you can run inference with your custom model with the following command.

```
import json
import logging
import boto3

from botocore.exceptions import ClientError

class ImageError(Exception):
    "Custom exception for errors returned by the model"
```
def __init__(self, message):
    self.message = message

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO)

def generate_text(model_id, body):
    """
    Generate text using your provisioned custom model.
    Args:
        model_id (str): The model ID to use.
        body (str): The request body to use.
    Returns:
        response (json): The response from the model.
    """

    logger.info(
        "Generating text with your provisioned custom model %s", model_id)

    brt = boto3.client(service_name='bedrock-runtime')

    accept = "application/json"
    content_type = "application/json"

    response = brt.invoke_model(
        body=body, modelId=model_id, accept=accept, contentType=content_type)

    response_body = json.loads(response.get("body").read())

    finish_reason = response_body.get("error")

    if finish_reason is not None:
        raise ImageError(f"Text generation error. Error is {finish_reason}")

    logger.info(
        "Successfully generated text with provisioned custom model %s", model_id)

    return response_body

def main():
    """
Entrypoint for example.

```python
try:
    logging.basicConfig(level=logging.INFO,
                        format="%s: %(levelname)s: %(message)s")

model_id = provisionedModelArn

prompt = "What is A.W.S.?"

body = json.dumps({
    "inputText": "what is AWS?"
})

response_body = generate_text(model_id, body)
print(f"Input token count: {response_body['inputTextTokenCount']}")

for result in response_body['results']:
    print(f"Token count: {result['tokenCount']}")
    print(f"Output text: {result['outputText']}")
    print(f"Completion reason: {result['completionReason']}")

except ClientError as err:
    message = err.response["Error"]['Message']
    logger.error("A client error occurred: %s", message)
    print("A client error occurred: " +
          format(message))
except ImageError as err:
    logger.error(err.message)
    print(err.message)

else:
    print( f"Finished generating text with your provisioned custom model {model_id}" )

if __name__ == "__main__":
    main()
```
Guidelines for model customization

The ideal parameters for customizing a model depend on the dataset and the task for which the model is intended. You should experiment with values to determine which parameters work best for your specific case. To help, evaluate your model by running a model evaluation job. For more information, see Model evaluation.

This topic provides guidelines and recommended values as a baseline for customization of the Amazon Titan Text G1 - Express model. For other models, check the provider's documentation.

Use the training and validation metrics from the output files generated when you submit a fine-tuning job to help you adjust your parameters. Find these files in the Amazon S3 bucket to which you wrote the output, or use the GetCustomModel operation.

Amazon Titan Text G1 - Express

The following guidelines are for the Titan Text Express text-to-text model model. For information about the hyperparameters that you can set, see Amazon Titan text model customization hyperparameters.

Impact on other tasks types

In general, the larger the training dataset, the better the performance for a specific task. However, training for a specific task might make the model perform worse on different tasks, especially if you use a lot of examples. For example, if the training dataset for a summarization task contains 100,000 samples, the model might perform worse on a classification task.

Model size

In general, the larger the model, the better the task performs given limited training data.

If you are using the model for a classification task, you might see relatively small gains for few-shot fine-tuning (less than 100 samples), especially if the number of classes is relatively small (less than 100).

Epochs

We recommend using the following metrics to determine the number of epochs to set:

1. Validation output accuracy – Set the number of epochs to one that yields a high accuracy.
2. **Training and validation loss** – Determine the number of epochs after which the training and validation loss becomes stable. This corresponds to when the model converges. Find the training loss values in the `step_wise_training_metrics.csv` and `validation_metrics.csv` files.

**Batch size**

When you change the batch size, we recommend that you change the learning rate using the following formula:

\[
\text{newLearningRate} = \frac{\text{oldLearningRate} \times \text{newBatchSize}}{\text{oldBatchSize}}
\]

**Learning rate**

In general, use smaller learning rates for larger models. We recommend using a learning rate in the range of 1.00E-06 to 1.00E-05.

The following table shows recommended learning rate values for fine-tuning:

<table>
<thead>
<tr>
<th>Task</th>
<th>Minimum learning rate</th>
<th>Default learning rate</th>
<th>Max learning rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarization</td>
<td>1.00E-06</td>
<td>3.00E-06</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>Classification</td>
<td>5.00E-06</td>
<td>5.00E-05</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>Question-answer</td>
<td>5.00E-06</td>
<td>5.00E-06</td>
<td>5.00E-05</td>
</tr>
</tbody>
</table>

**Learning warmup steps**

We recommend the default value of 5.

**Troubleshooting**

This section summarizes errors that you might encounter and what to check if you do.

**Permissions issues**

If you encounter an issue with permissions to access an Amazon S3 bucket, check that the following are true:
1. If the Amazon S3 bucket uses a CM-KMS key for Server Side encryption, ensure that the IAM role passed to Amazon Bedrock has `kms:Decrypt` permissions for the AWS KMS key. For example, see [Allow a user to encrypt and decrypt with any AWS KMS key in a specific AWS account](#).

2. The Amazon S3 bucket is in the same region as the Amazon Bedrock model customization job.

3. The IAM role trust policy includes the service SP (`bedrock.amazonaws.com`).

The following messages indicate issues with permissions to access training or validation data in an Amazon S3 bucket:

```
Could not validate GetObject permissions to access Amazon S3 bucket: training-data-bucket at key train.jsonl
Could not validate GetObject permissions to access Amazon S3 bucket: validation-data-bucket at key validation.jsonl
```

If you encounter one of the above errors, check that the IAM role passed to the service has `s3:GetObject` and `s3:ListBucket` permissions for the training and validation dataset Amazon S3 URIs.

The following message indicates issues with permissions to write the output data in an Amazon S3 bucket:

```
Amazon S3 perms missing (PutObject): Could not validate PutObject permissions to access S3 bucket: bedrock-output-bucket at key output/write_access_check_file.tmp
```

If you encounter the above error, check that the IAM role passed to the service has `s3:PutObject` permissions for the output data Amazon S3 URI.

**Data issues**

The following errors are related to issues with the training, validation, or output data files:

**Invalid file format**

```
Unable to parse Amazon S3 file: fileName.jsonl. Data files must conform to JSONL format.
```

If you encounter the above error, check that the following are true:
1. Each line is in JSON.

2. Each JSON has two keys, an input and an output, and each key is a string. For example:

   ```json
   {
     "input": "this is my input",
     "output": "this is my output"
   }
   ```

3. There are no additional new lines or empty lines.

Character quota exceeded

Input size exceeded in file `fileName.jsonl` for record starting with...

If you encounter an error beginning with the text above, ensure that the number of characters conforms to the character quota in Model customization quotas.

Token count exceeded

Maximum input token count 4097 exceeds limit of 4096
Maximum output token count 4097 exceeds limit of 4096
Max sum of input and output token length 4097 exceeds total limit of 4096

If you encounter an error similar to the preceding example, make sure that the number of tokens conforms to the token quota in Model customization quotas.

Internal error

Encountered an unexpected error when processing the request, please try again

If you encounter the above error, there might be an issue with the service. Try the job again. If the issue persists, contact AWS Support.
Provisioned Throughput

Note
Provisioned Throughput is only available in the US East (N. Virginia), US West (Oregon), and AWS GovCloud (US-West) Regions.

When you configure Provisioned Throughput for a model, you receive a level of throughput at a fixed cost.

You can use Provisioned Throughput with Amazon and third-party base models, and with customized models.

Provisioned Throughput pricing varies depending on the model that you use and the level of commitment you choose. You receive a discounted rate when you commit to a longer period of time. For details about pricing for each model, see the Model providers page in the Amazon Bedrock console.

Your options for throughput for a model differ depending on whether you run inference on a base model or a custom model.

Note
In the AWS GovCloud (US) region, you can only purchase Provisioned Throughput for custom models with no commitment.

<table>
<thead>
<tr>
<th>Pricing option</th>
<th>Base model</th>
<th>Custom model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provisioned Throughput, no commitment (hourly pricing)</td>
<td>Not available</td>
<td>Available (maximum 2 Provisioned Throughputs per account)</td>
</tr>
<tr>
<td>Provisioned Throughput, 1 month commitment</td>
<td>Available</td>
<td>Available</td>
</tr>
</tbody>
</table>
You specify Provisioned Throughput in Model Units (MU). A model unit delivers a specific throughput level for the specified model. The throughput level of a MU for a given Text model specifies the following:

- **The total number of input tokens per minute** – The number of input tokens that an MU can process across all requests within a span of one minute.

- **The total number of output tokens per minute** – The number of output tokens that an MU can generate across all requests within a span of one minute.

### Note
For more information about what an MU specifies, contact your AWS account manager.

Model unit quotas depend on the level of commitment you specify for the Provisioned Throughput.

- For custom models with no commitment, a quota of one model unit is available for each Provisioned Throughput. You can create up to two Provisioned Throughputs per account.

- For base or custom models with commitment, there is a default quota of 0 model units. To request an increase, use the limit increase form.

### Topics
- **Supported regions and models for Provisioned Throughput**
- **Procedures**
- **Permissions**
- **Provisioned Throughput console procedures**
- **Using the Provisioned Throughput API**
Supported regions and models for Provisioned Throughput

Provisioned Throughput is supported in the following regions:

<table>
<thead>
<tr>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>US East (N. Virginia)</td>
</tr>
<tr>
<td>US West (Oregon)</td>
</tr>
<tr>
<td>AWS GovCloud (US-West)</td>
</tr>
</tbody>
</table>

If you purchase Provisioned Throughput through the Amazon Bedrock API, you must specify contextual variants of Amazon Bedrock FMs for the model ID. The following table shows the models for which you can purchase Provisioned Throughput and the model ID to use when purchasing Provisioned Throughput:

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID for Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>amazon.titan-text-express-v1:0:8k</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Lite</td>
<td>amazon.titan-text-lite-v1:0:4k</td>
</tr>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>amazon.titan-embed-text-v1:2:8k</td>
</tr>
<tr>
<td>Amazon Titan Multimodal Embeddings G1</td>
<td>amazon.titan-embed-image-v1:0</td>
</tr>
<tr>
<td>Amazon Titan Image Generator G1</td>
<td>amazon.titan-image-generator-v1:0</td>
</tr>
<tr>
<td>Model name</td>
<td>Model ID for Provisioned Throughput</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Anthropic Claude V2 18K</td>
<td>anthropic.claude-v 2:0:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2 100K</td>
<td>anthropic.claude-v 2:0:100k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 18K</td>
<td>anthropic.claude-v 2:1:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 200K</td>
<td>anthropic.claude-v 2:1:200k</td>
</tr>
<tr>
<td>Anthropic Claude Instant V1 100K</td>
<td>anthropic.claude-instant-v1:2:100K</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>cohere.command-text-v14:7:4k</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>cohere.command-light-text-v14:7:4k</td>
</tr>
<tr>
<td>Stable Diffusion XL 1.0</td>
<td>stability.stable-diffusion-xl-v1:0</td>
</tr>
<tr>
<td>Meta Llama 2 Chat 13B</td>
<td>meta.llama2-13b-chat-v1:0:4k</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>(see note below)</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>(see note below)</td>
</tr>
</tbody>
</table>
Procedures

Provisioned Throughput supports the following procedures.

Creating

When you create a Provisioned Throughput, the Provisioned Throughput starts in an interim state (Creating) while Bedrock creates the Provisioned Throughput resources. If the creation is successful, the Provisioned Throughput transitions to InService state, and you can start using it for inference.

If the creation fails, the Provisioned Throughput transitions to Failed state.

Updating

You can perform the following types of updates to a Provisioned Throughput:

- Change the name of the Provisioned Throughput.
- Specify a new custom model that uses the same base model as the current custom model.
- Specify a base model. It must be the base model of the current custom model.

When you update a Provisioned Throughput, it transitions to the Updating state while Bedrock performs the update. If the update is successful, the Provisioned Throughput transitions to InService state. During the update, you can run inference using the Provisioned Throughput without disrupting the on-going traffic from your end customers. If the update includes a new model, you may receive output from the old model until the update is fully deployed.

If the update fails, the Provisioned Throughput transitions to Failed state.

Deleting

Deleting a Provisioned Throughput is a synchronous operation. The Provisioned Throughput deletion takes effect immediately.
Running inference

You can run inference using a Provisioned Throughput that is in InService state.

If you request more throughput than is configured for the Provisioned Throughput, the request is throttled (you receive the throughput defined by the Provisioned Throughput).

Permissions

To add Provisioned Throughput to a base model or a model that was customized from a base model, you must have first requested access to the base model by following the steps at Model access.

To add Provisioned Throughput to a custom model that is encrypted, your role must include permissions to carry out the kms:Decrypt action on the model. For an example, see Create a key policy and attach it to the customer managed key.

Provisioned Throughput console procedures

This section describes the console procedures for Provisioned Throughput.

Topics

- View Provisioned Throughput summary
- Purchase Provisioned Throughput
- View details of a Provisioned Throughput
- Edit a Provisioned Throughput
- Delete a Provisioned Throughput

View Provisioned Throughput summary

Use the summary page to review the status of each Provisioned Throughput. The Overview panel displays the number of Provisioned Throughput resources in each state. You can update or delete an active Provisioned Throughput, and you can create a new Provisioned Throughput.

1. Open the Amazon Bedrock console at https://console.aws.amazon.com/bedrock/.
2. From the left navigation pane, choose Provisioned throughput under Deployments.
3. From the **Provisioned Throughput** table, you can review summary information about each Provisioned Throughput.

**Purchase Provisioned Throughput**

You can purchase Provisioned Throughput for a foundation or custom model.


2. From the left menu, choose **Provisioned Throughput** under **Deployments**. Then select **Purchase Provisioned Throughput**. The console opens the **Purchase Provisioned Throughput** page.

3. Under **Provisioned Throughput details**:
   a. Enter a name for the Provisioned Throughput.
   b. Select the model category and model for the Provisioned Throughput.
   c. (Optional) Under **Tags**, you can associate one or more tags with this Provisioned Throughput.

4. Under **Model units & commitment term**:
   a. Enter the desired number of model units.
   b. Choose the amount of time for which you want to commit to using the Provisioned Throughput. To opt out of commitment for a custom model, you must set the number of model units to 1 and then select **No commitment**. With this option, you pay an hourly amount for the allocated throughput until you delete it.

   **Note**
   
   To see limitations on model units and commitment term, see [Provisioned Throughput](#).

5. Under **Estimated purchase summary**, review the estimated cost.

6. Choose **Purchase Provisioned Throughput**.

7. Review the note that appears and acknowledge the commitment duration and price by selecting the checkbox. Then choose **Confirm purchase**.
The console displays the overview page. For the Provisioned Throughput that you just purchased, the console displays it in the table, with the status set to **Creating**.

**View details of a Provisioned Throughput**

From the summary page, you can view the details for any of your Provisioned Throughput resources.

2. From the left menu, choose **Provisioned Throughput**.
3. From the **Provisioned Throughput** table, choose a Provisioned Throughput. The console opens the details page.
4. Under **Provisioned Throughput overview**:
5. Under **Tags**, the console displays the tags that are associated with this Provisioned Throughput. Choose **Manage tags** to add or remove tags for this Provisioned Throughput.

**Edit a Provisioned Throughput**

You can edit only a few fields of a Provisioned Throughput.

2. From the left menu, choose **Provisioned Throughput**.
3. From the **Provisioned Throughput** table, select the Provisioned Throughput to edit.
4. The console displays the Provisioned Throughput fields that you can edit.
5. Choose **Save** to start the update.

**Delete a Provisioned Throughput**

⚠️ **Note**

You can’t delete a Provisioned Throughput with commitment before the commitment term is complete.

2. From the left menu, choose **Provisioned Throughput**.
3. From the **Provisioned Throughput** table, select the Provisioned Throughput to delete.

4. The console displays a modal form to warn you that delete is permanent. Choose **Confirm** to proceed.

**Using the Provisioned Throughput API**

Amazon Bedrock provides API operations to create and manage your Provisioned Throughput resources.

**Topics**

- Create Provisioned Throughput
- Run inference using Provisioned Throughput
- Update Provisioned Throughput
- Get Provisioned Throughput
- Delete Provisioned Throughput
- List Provisioned Throughput resources

**Create Provisioned Throughput**

Use the **CreateProvisionedModelThroughput** operation to create a Provisioned Throughput for a base or custom model.

When you send a CreateProvisionedModelThroughput operation for a base model, Bedrock validates your access permissions for the requested model and checks for capacity availability. If the validations are successful, Bedrock creates the provisioned throughput and returns the ARN of the Provisioned Throughput.

When you send a CreateProvisionedModelThroughput operation for a custom model, Bedrock validates your access permissions for the requested custom model and checks for capacity availability. If the custom model is encrypted using a customer-managed KMS key, your permissions must include `kms:Decrypt` permission for this model.

⚠️ **Note**

To see quotas for your options for `commitmentDuration` and `modelUnits`, see **Provisioned Throughput**.
To request on-demand Provisioned Throughput for a custom model, omit the `commitmentDuration` field in the request.

To create a Provisioned Throughput for an Titan foundation model, use the following IDs instead of the default ones.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model ID for Provisioned Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>amazon.titan-text-express-v1:0:8k</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Lite</td>
<td>amazon.titan-text-lite-v1:0:4k</td>
</tr>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>amazon.titan-embed-text-v1:2:8k</td>
</tr>
<tr>
<td>Amazon Titan Multimodal Embeddings G1</td>
<td>amazon.titan-embed-image-v1:0</td>
</tr>
<tr>
<td>Amazon Titan Image Generator G1</td>
<td>amazon.titan-image-generator-v1:0</td>
</tr>
<tr>
<td>Anthropic Claude V2 18K</td>
<td>anthropic.claude-v2:0:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2 100K</td>
<td>anthropic.claude-v2:0:100k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 18K</td>
<td>anthropic.claude-v2:1:18k</td>
</tr>
<tr>
<td>Anthropic Claude V2.1 200K</td>
<td>anthropic.claude-v2:1:200k</td>
</tr>
<tr>
<td>Model name</td>
<td>Model ID for Provisioned Throughput</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Anthropic Claude Instant V1 100K</td>
<td>anthropic.claude-instant-v1:2:100K</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>cohere.command-text-v14:7:4k</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>cohere.command-light-text-v14:7:4k</td>
</tr>
<tr>
<td>Stable Diffusion XL 1.0</td>
<td>stability.stable-diffusion-xl-v1:0</td>
</tr>
<tr>
<td>Meta Llama 2 Chat 13B</td>
<td>meta.llama2-13b-chat-v1:0:4k</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>(see note below)</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>(see note below)</td>
</tr>
</tbody>
</table>

**Note**

The Meta Llama 2 (non-chat) models can only be used after being customized and after purchasing Provisioned Throughput for them.

The following code examples demonstrate how to create a Provisioned Throughput using the AWS CLI and the Python SDK.

**AWS CLI**

Create the Provisioned Throughput using the following command in the AWS CLI:

```
aws bedrock create-provisioned-model-throughput
   --model-units 1
   --commitment-duration SixMonths
```
Python (Boto)

The following example demonstrates how to create the Provisioned Throughput using Python:

```python
import boto3
bedrock = boto3.client(service_name='bedrock')
bedrock.create_provisioned_model_throughput(
    modelUnits=1,
    commitmentDuration='SixMonths',
    provisionedModelName='test-model',
    modelId='arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-v2'
)
```

Run inference using Provisioned Throughput

Use the `InvokeModel` or `InvokeModelWithResponseStream` operation to run inference using Provisioned Throughput. Specify the provisioned model ARN as the `modelId` parameter.

The following code examples show how to run inference for an Anthropic Claude model with a Provisioned Throughput using the AWS CLI and Python. For information about the prompt format and inference parameters, see [Inference parameters for foundation models](#).

AWS CLI

The following example demonstrates how to run inference for an Anthropic Claude model with a Provisioned Throughput in the AWS CLI.

```bash
aws bedrock-runtime invoke-model
    --model-id provisioned-model-arn \\
    --body "{"prompt": "\n\nHuman: story of two dogs\n\nAssistant:"}"
    invoke-model-output.txt
```

Python (Boto)

The following example demonstrates how to run inference for an Anthropic Claude model with a Provisioned Throughput using Python:

```python
import boto3
```
import json

bedrock = boto3.client(service_name='bedrock-runtime')

body = json.dumps({'prompt': 'explain black holes to 8th graders

Assistant:

modelId = 'provisioned-model-arn'
accept = 'application/json'
contentType = 'application/json'

response = bedrock.invoke_model(body=body, modelId=modelId, accept=accept, contentType=contentType)
response_body = json.loads(response.get('body').read())

# text
print(response_body.get('results')[0].get('outputText'))

# embedding
print(response_body.get('embedding'))

---

**Update Provisioned Throughput**

Use the [UpdateProvisionedModelThroughput](#) operation to update the specified Provisioned Throughput.

The following code examples demonstrate how to update a Provisioned Throughput using the AWS CLI and Python.

**AWS CLI**

Update the Provisioned Throughput using the following command in the AWS CLI:

```
aws bedrock update-provisioned-model-throughput
  --provisioned-model-id provisioned-model-arn | provisioned-model-name
  --model-arn custom-model-arn | foundation-model-arn
```

**Python (Boto)**

The following example demonstrates how to update the Provisioned Throughput using Python:

```python
import boto3
```
Get Provisioned Throughput

Use the **GetProvisionedModelThroughput** operation to retrieve information about the specified Provisioned Throughput.

The following code examples demonstrates how to retrieve information using the AWS CLI and Python.

**AWS CLI**

Retrieve information about the Provisioned Throughput using the following command in the AWS CLI:

```bash
aws bedrock get-provisioned-model-throughput
--provisioned-model-id provisioned-model-arn | provisioned-model-name
```

**Python (Boto)**

The following example demonstrates how to retrieve information about the Provisioned Throughput using Python:

```python
import boto3
bedrock = boto3.client(service_name='bedrock')

bedrock.get_provisioned_model_throughput(
    provisionedModelName='my-provisioned-model-arn | provisioned_model_name'
)
```

Delete Provisioned Throughput

Use the **DeleteProvisionedModelThroughput** operation to delete the specified provisioned throughput.
The following code examples demonstrate how to delete a Provisioned Throughput using the AWS CLI and Python.

AWS CLI

Delete the Provisioned Throughput using the following command in the AWS CLI:

```bash
aws bedrock delete-provisioned-model-throughput
  --provisioned-model-id provisioned-model-arn | provisioned-model-name
```

Python (Boto)

The following example demonstrates how to delete the Provisioned Throughput using Python:

```python
import boto3
bedrock = boto3.client(service_name='bedrock')

bedrock.delete_provisioned_model_throughput(
    provisionedModelName='my-provisioned-model-arn | provisioned model name'
)
```

List Provisioned Throughput resources

Use the ListProvisionedModelThroughputs operation to list the Provisioned Throughput resources that you have defined.

The following code examples demonstrate how to list the Provisioned Throughput resources using the AWS CLI and Python.

AWS CLI

List the Provisioned Throughput resources using the following command in the AWS CLI:

```bash
aws bedrock list-provisioned-model-throughputs
```

Python (Boto)

The following example demonstrates how to list the Provisioned Throughput resources using Python:

```python
import boto3
```
```python
bedrock = boto3.client(service_name='bedrock')

bedrock.list_provisioned_model_throughputs()
```
Tag resources

To help you manage your Amazon Bedrock resources, you can assign metadata to each resource as tags. A tag is a label that you assign to an AWS resource. Each tag consists of a key and a value.

Tags enable you to categorize your AWS resources in different ways, for example, by purpose, owner, or application. Tags help you to do the following:

- Identify and organize your AWS resources. Many AWS resources support tagging, so you can assign the same tag to resources in different services to indicate that the resources are the same.
- Allocate costs. You activate tags on the AWS Billing and Cost Management dashboard. AWS uses the tags to categorize your costs and deliver a monthly cost allocation report to you. For more information, see Use cost allocation tags in the AWS Billing and Cost Management User Guide.
- Control access to your resources. You can use tags with Amazon Bedrock to create policies to control access to Amazon Bedrock resources. These policies can be attached to an IAM role or user to enable tag-based access control.

The Amazon Bedrock resources that you can tag are:

- Custom models
- Model customization jobs
- Provisioned models
- Batch inference jobs (API only)
- Agents
- Agent aliases
- Knowledge bases
- Model evaluations (console only)

Topics

- Use the console
- Use the API
- Best practices and restrictions
Use the console

You can add, modify, and remove tags at any time while creating or editing a supported resource.

Use the API

To carry out tagging operations, you need the Amazon Resource Name (ARN) of the resource on which you want to carry out a tagging operation. There are two sets of tagging operations, depending on the resource for which you are adding or managing tags.

1. The following resources use the Amazon Bedrock `TagResource`, `UntagResource`, and `ListTagsForResource` operations.
   - Custom models
   - Model customization jobs
   - Provisioned models
   - Batch inference jobs

2. The following resources use the Agents for Amazon Bedrock `TagResource`, `UntagResource`, and `ListTagsForResource` operations.
   - Agents
   - Agent aliases
   - Knowledge bases

To add tags to a resource, send a Amazon Bedrock `TagResource` or Agents for Amazon Bedrock `TagResource` request.

To untag a resource, send an `UntagResource` or `UntagResource` request.

To list the tags for a resource, send a `ListTagsForResource` or `ListTagsForResource` request.

Select a tab to see code examples in an interface or language.

AWS CLI

Add two tags to an agent. Separate key/value pairs with a space.

```bash
aws bedrock-agent tag-resource \
  --resource-arn "arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345" \
```
Remove the tags from the agent. Separate keys with a space.

```
aws bedrock-agent untag-resource \
   --resource-arn "arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345" \
   --tag-keys key=department facing
```

List the tags for the agent.

```
aws bedrock-agent list-tags-for-resource \
   --resource-arn "arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345"
```

Python (Boto)

Add two tags to an agent.

```python
import boto3

bedrock = boto3.client(service_name='bedrock-agent')

tags = [
   {
      'key': 'department',
      'value': 'billing'
   },
   {
      'key': 'facing',
      'value': 'internal'
   }
]

bedrock.tag_resource(resourceArn='arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345', tags=tags)
```

Remove the tags from the agent.

```python
bedrock.untag_resource(
   resourceArn='arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345',
   tagKeys=['department', 'facing'])
```
List the tags for the agent.

```python
correct bedrock.list_tags_for_resource(resourceArn='arn:aws:bedrock:us-east-1:123456789012:agent/AGENT12345')
```

**Best practices and restrictions**

For best practices and restrictions on tagging, see [Tagging your AWS resources](#).
Amazon Titan Models

Amazon Titan foundation models (FMs) are a family of FMs pretrained by AWS on large datasets, making them powerful, general-purpose models built to support a variety of use cases. Use them as-is or privately customize them with your own data.

Amazon Titan supports the following models for Amazon Bedrock.

- Amazon Titan Text
- Amazon Titan Embeddings G1 - Text
- Amazon Titan Multimodal Embeddings G1
- Amazon Titan Image Generator G1 (preview)

Note
Amazon Titan Image Generator G1 is in public preview release.

Topics
- Amazon Titan Text models
- Amazon Titan Embeddings G1 - Text models
- Amazon Titan Multimodal Embeddings G1 model
- Amazon Titan Image Generator G1 model

Amazon Titan Text models

Amazon Titan text models include Amazon Titan Text G1 - Express and Amazon Titan Text G1 - Lite.

Amazon Titan Text G1 - Express

Amazon Titan Text G1 - Express is a large language model for text generation. It is useful for a wide range of advanced, general language tasks such as open-ended text generation and conversational chat, as well as support within Retrieval Augmented Generation (RAG). At launch, the model is
optimized for English, with multilingual support for more than 100 additional languages available in preview.

- **Model ID** – amazon.titan-text-express-v1
- **Max tokens** – 8K
- **Languages** – English (GA), 100 additional languages (Preview)
- **Supported use cases** – Retrieval augmented generation, open-ended text generation, brainstorming, summarizations, code generation, table creation, data formatting, paraphrasing, chain of thought, rewrite, extraction, QnA, and chat.

**Amazon Titan Text G1 - Lite**

Amazon Titan Text G1 - Lite is a light weight efficient model, ideal for fine-tuning of English-language tasks, including like summarizations and copy writing, where customers want a smaller, more cost-effective model that is also highly customizable.

- **Model ID** – amazon.titan-text-lite-v1
- **Max tokens** – 4K
- **Languages** – English
- **Supported use cases** – Open-ended text generation, brainstorming, summarizations, code generation, table creation, data formatting, paraphrasing, chain of thought, rewrite, extraction, QnA, and chat.

**Amazon Titan Text Model Customization**

For more information on customizing Amazon Titan text models, see the following pages.

- [Prepare the datasets](#)
- [Amazon Titan text model customization hyperparameters](#)

**Amazon Titan Text Prompt Engineering Guidelines**

Amazon Titan text models can be used in a wide variety of applications for different use cases. Amazon Titan Text models have prompt engineering guidelines for the following applications including:
• Chatbot
• Text2SQL
• Function Calling
• RAG (Retrieval Augmented Generation)

For more information on Amazon Titan Text prompt engineering guidelines, see Amazon Titan Text Prompt Engineering Guidelines.

For general prompt engineering guidelines, see Prompt Engineering Guidelines.

AWS AI Service Card - Amazon Titan Text

AI Service Cards provide transparency and document the intended use cases and fairness considerations for our AWS AI services. AI Service Cards provide a single place to find information on the intended use cases, responsible AI design choices, best practices, and performance for a set of AI service use cases.

Amazon Titan Embeddings G1 - Text models

Amazon Titan Embeddings text models include Amazon Titan Embeddings G1 - Text G1.

Text embeddings represent meaningful vector representations of unstructured text such as documents, paragraphs, and sentences. You input a body of text and the output is a (1 x n) vector. You can use embedding vectors for a wide variety of applications.

The Amazon Titan Embeddings G1 - Text model (amazon.titan-embed-text-v1). The Amazon Titan Embeddings G1 - Text – Text v1.2 can intake up to 8k tokens and outputs a vector of 1,536 dimensions. The model also works in 25+ different languages. The model is optimized for text retrieval tasks, but can also perform additional tasks, such as semantic similarity and clustering. Amazon Titan Embeddings G1 - Text – Text v1.2 also supports long documents, however, for retrieval tasks it is recommended to segment documents into logical segments (such as paragraphs or sections). In line with our recommendation.

Note

Titan Embeddings G1 - Text model doesn’t support inference parameters such as maxTokenCount or topP.
To use the text or image embeddings models, use the Invoke Model API operation with `amazon.titan-embed-text-v1` or `amazon.titan-embed-image-v1` as the model Id and retrieve the embedding object in the response.

To see Jupyter notebook examples:

2. From the left-side menu, choose Base models.
3. Scroll down and select the Amazon Titan Embeddings G1 - Text model.
4. In the Amazon Titan Embeddings G1 - Text tab (depending on which model you chose), select View example notebook to see example notebooks for embeddings.

For more information on preparing your dataset for multimodal training, see Preparing your dataset.

**Amazon Titan Multimodal Embeddings G1 model**

Amazon Titan Multimodal Embeddings G1 Generation 1 (G1) is a multimodal embeddings model for use cases like searching images by text, image, or a combination of text and image. Designed for high accuracy and fast responses, this model is an ideal choice for search and recommendations use cases.

- **Model ID** – `amazon.titan-embed-image-v1`
- **Max input text tokens** – 128
- **Languages** – English
- **Max input image size** – 5 MB
- **Output vector size** – 1,024 (default), 384, 256
- **Inference types** – On-Demand, Provisioned Throughput
- **Supported use cases** – image search, recommendations, and personalization

**Embedding length**

Setting a custom embedding length is optional. The embedding default length is 1024 characters which will work for most use cases. The embedding length can be set to 256, 384, or 1024 characters. Larger embedding sizes create more detailed responses, but will also increase the
computational time. Shorter embedding lengths are less detailed but will improve the response time.

```python
# EmbeddingConfig Shape
{
    'outputEmbeddingLength': int // Optional, One of: [256, 384, 1024], default: 1024
}

# Updated API Payload Example
body = json.dumps(
    {
        "inputText": "hi",
        "inputImage": image_string,
        "embeddingConfig": {
            "outputEmbeddingLength": 256
        }
    }
)
```

**Finetuning**

- Input to the Amazon Titan Multimodal Embeddings G1 finetuning is image-text pairs.
- Image formats: PNG, JPEG
- Input image size limit: 5 MB
- Image dimensions: min: 128 px, max: 4,096 px
- Max number of tokens in caption: 128
- Training dataset size range: 1000 - 500,000
- Validation dataset size range: 8 - 50,000
- Caption length in characters: 0 - 2,560
- Maximum total pixels per image: 2048*2048*3
- Aspect ratio (w/h): min: 0.25, max: 4

**Preparing datasets**

For the training dataset, create a .jsonl file with multiple JSON lines. Each JSON line contains both an image-ref and caption attributes similar to Sagemaker Augmented Manifest format. A validation dataset is required. Auto-captioning is not currently supported.
For both the training and validation datasets, you will create .jsonl files with multiple JSON lines.

The Amazon S3 paths need to be in the same folders where you have provided permissions for Amazon Bedrock to access the data by attaching an IAM policy to your Amazon Bedrock service role. For more information on granting an IAM policies for training data, see [Grant custom jobs access to your training data](#).

**Hyperparameters**

These values can be adjusted for the Multimodal Embeddings model hyperparameters. The default values will work well for most use cases.

- **Learning rate** - (min/max learning rate) – default: 5.00E-05, min: 5.00E-08, max: 1
- **Batch size** - Effective batch size – default: 576, min: 256, max: 9,216
- **Max epochs** – default: "auto", min: 1, max: 100

**Amazon Titan Image Generator G1 model**

Amazon Titan Image Generator G1 is an image generation model. It generates images from text, and allows users to upload and edit an existing image. Users can edit an image with a text prompt (without a mask) or parts of an image with an image mask. You can extend the boundaries of an image with outpainting, and fill in an image with inpainting. It can also generate variations of an image based on an optional text prompt. Amazon Titan Image Generator G1 Generator includes watermarking on the output files.

**Note**

Amazon Titan Image Generator G1 is currently in preview release. It is not recommended for production work. Some features may not perform as expected. Images and output data may produce less than accurate results.
For more information on Amazon Titan Image Generator G1 prompt engineering guidelines, see Amazon Titan Image Generator G1 Prompt Engineering Best Practices.

- **Model ID** – amazon.titan-image-generator-v1
- **Max input characters** – 1,024 char
- **Max input image size** – 50 MB (only some specific resolutions are supported)
- **Max image size using in/outpainting** – 1,024 x 1,024 px
- **Max image size using image variation** – 4,096 x 4,096 px
- **Languages** – English
- **Output type** – image
- **Supported image types** – JPEG, JPG, PNG
- **Inference types** – On-Demand, Provisioned Throughput
- **Supported use cases** – image generation, image editing, image variations

**Features**

- **Text-to-image (T2I) generation** – Input a text prompt and generate a new image as output. The generated image captures the concepts described by the text prompt.
- **Finetuning of a T2I model** – Import several images to capture your own style and personalization and then fine tune the core T2I model. The fine-tuned model generates images that follow the style and personalization of a specific user.
- **Image editing options** – includes inpainting, outpainting, generating variations, and automatic editing without an image mask.
- **Inpainting** – Uses an image and a segmentation mask as input (either from the user or estimated by the model) and reconstructs the region within the mask. Use inpainting to remove masked elements and replace them with background pixels.
- **Outpainting** – Uses an image and a segmentation mask as input (either from the user or estimated by the model) and generates new pixels that seamlessly extend the region. Use precise outpainting to preserve the pixels of the masked image when extending the image to the boundaries. Use default outpainting to extend the pixels of the masked image to the image boundaries based on segmentation settings.
- **Image variation** – Uses an image and an optional prompt as input. It generates a new image that preserves the content of the input image, but variates its style and background.
Note

if you are using a fine-tuned model, you cannot use inpainting or outpainting features of the API or the model.

Parameters

For information on Amazon Titan Image Generator G1 inference parameters, see Amazon Titan Image Generator G1 inference parameters.

Fine-tuning

For more information on fine-tuning the Amazon Titan Image Generator G1 model, see the following pages.

- Prepare the datasets
- Amazon Titan Image Generator G1 model customization hyperparameters

Titan Image Generator G1 fine-tuning and pricing

The model uses the following example formula to calculate the total price per job:

Total Price = Steps * Batch size * Price per image seen

Minimum values (auto):

- Minimum steps (auto) - 500
- Minimum batch size - 8
- Default learning rate - 0.00001
- Price per image seen - 0.005

Fine-tuning hyperparameter settings

Steps – The number of times the model is exposed to each batch. There is no default step count set. You must select a number between 10 - 40,000, or a String value of "Auto."

Step settings - Auto – Amazon Bedrock determines a reasonable value based on training information. Select this option to prioritize model performance over training cost. The number of
steps is determined automatically. This number will typically be between 1,000 and 8,000 based on your dataset. Job costs are impacted by the number of steps used to expose the model to the data. Refer to the pricing examples section of pricing details to understand how job cost is calculated. (See example table above to see how step count is related to number of images when Auto is selected.)

**Step settings - Custom** – You can enter the number of steps you want Bedrock to expose your custom model to the training data. This value can be between 10 and 40,000. You can reduce the cost per image produced by the model by using a lower step count value.

**Batch size** – The number of sample processed before model parameters are updated. This value is between 8 and 192 and is a multiple of 8.

**Learning rate** – The rate at which model parameters are updated after each batch of training data. This is a float value between 0 and 1. The learning rate is set to 0.00001 by default.

For more information on the fine-tuning procedure, see [Submit a model customization job](#).

### Output

Titan Image Generator G1 uses the output image size and quality to determine how an image is priced. Titan Image Generator G1 has two pricing segments based on size: one for 512*512 images and another for 1024*1024 images. Pricing is based on image size height*width, less than or equal to 512*512 or greater than 512*512.

For more information on Amazon Bedrock pricing, see [Amazon Bedrock Pricing](#).

### Watermark detection

**Note**

Watermark detection is in public preview release and will only detect a watermark generated from Titan Image Generator G1. This feature is currently only available in the PDX and IAD regions. Watermark detection is a highly accurate detection of the watermark generated by Titan Image Generator G1. Images that are modified from the original image may produce less accurate detection results.

This model adds an invisible watermark to all generated images to reduce the spread of misinformation, assist with copyright protection, and track content usage. In the future, a
watermark detection API will be available to help you confirm whether an image was generated by the Titan Image Generator G1 model, which checks for the existence of this watermark.

You can upload your image to detect if a watermark from Titan Image Generator G1 is present on the image. Use the console to detect a watermark from this model by following the below steps.

**To detect a watermark with Titan Image Generator G1:**

1. Open the Amazon Bedrock console at Amazon Bedrock console
2. Select Overview from the navigation pane in Amazon Bedrock. Choose the Build and Test tab.
3. In the Safeguards section, go to Watermark detection and choose View watermark detection.
4. Select Upload image and locate a file that is in JPG or PNG format. The maximum file size allowed is 5 MB.
5. Once uploaded, a thumbnail of image is shown with the name, file size, and the last date modified. Select X to delete or replace image from the Upload section.
6. Select Analyze to begin watermark detection analysis.
7. The image is previewed under Results, and indicates if a watermark is detected with Watermark detected below the image and a banner across the image. If no watermark is detected, the text below the image will say Watermark NOT detected.
8. To load the next image, select X in the thumbnail of the image in the Upload section and choose a new image to analyze.

**Prompt Engineering Guidelines**

**Mask prompt** – This algorithm classifies pixels into concepts. The user can give a text prompt that will be used to classify the areas of the image to mask, based on the interpretation of the mask prompt. The prompt option can interpret more complex prompts, and encode the mask into the segmentation algorithm.

**Image mask** – You can also use an image mask to set the mask values. The image mask can be combined with prompt input for the mask to improve accuracy. The image mask file must conform to the following parameters:

- Mask image values must be 0 (black) or 255 (white) for the mask image. The image mask area with the value of 0 will be regenerated with the image from the user prompt and/or input image.
• The maskImage field must be a base64 encoded image string.
• Mask image must have the same dimensions as the input image (same height and width).
• Only PNG or JPG files can be used for the input image and the mask image.
• Mask image must only use black and white pixels values.
• Mask image can only use the RGB channels (alpha channel not supported).

For more information on Amazon Titan Image Generator G1 prompt engineering, see Amazon Titan Image Generator G1 Prompt Engineering Best Practices.

For general prompt engineering guidelines, see Prompt Engineering Guidelines.
Security in Amazon Bedrock

Cloud security at AWS is the highest priority. As an AWS customer, you benefit from data centers and network architectures that are built to meet the requirements of the most security-sensitive organizations.

Security is a shared responsibility between AWS and you. The shared responsibility model describes this as security of the cloud and security in the cloud:

- **Security of the cloud** – AWS is responsible for protecting the infrastructure that runs AWS services in the AWS Cloud. AWS also provides you with services that you can use securely. Third-party auditors regularly test and verify the effectiveness of our security as part of the AWS Compliance Programs. To learn about the compliance programs that apply to Amazon Bedrock, see AWS Services in Scope by Compliance Program.

- **Security in the cloud** – Your responsibility is determined by the AWS service that you use. You are also responsible for other factors including the sensitivity of your data, your company’s requirements, and applicable laws and regulations.

This documentation helps you understand how to apply the shared responsibility model when using Amazon Bedrock. The following topics show you how to configure Amazon Bedrock to meet your security and compliance objectives. You also learn how to use other AWS services that help you to monitor and secure your Amazon Bedrock resources.

**Topics**

- Data protection
- Identity and access management for Amazon Bedrock
- Compliance validation for Amazon Bedrock
- Incident response in Amazon Bedrock
- Resilience in Amazon Bedrock
- Infrastructure security in Amazon Bedrock
- Cross-service confused deputy prevention
- Configuration and vulnerability analysis in Amazon Bedrock
- Use interface VPC endpoints (AWS PrivateLink)
Data protection

The AWS shared responsibility model applies to data protection in Amazon Bedrock. As described in this model, AWS is responsible for protecting the global infrastructure that runs all of the AWS Cloud. You are responsible for maintaining control over your content that is hosted on this infrastructure. You are also responsible for the security configuration and management tasks for the AWS services that you use. For more information about data privacy, see the Data Privacy FAQ. For information about data protection in Europe, see the AWS Shared Responsibility Model and GDPR blog post on the AWS Security Blog.

For data protection purposes, we recommend that you protect AWS account credentials and set up individual users with AWS IAM Identity Center or AWS Identity and Access Management (IAM). That way, each user is given only the permissions necessary to fulfill their job duties. We also recommend that you secure your data in the following ways:

- Use multi-factor authentication (MFA) with each account.
- Use SSL/TLS to communicate with AWS resources. We require TLS 1.2 and recommend TLS 1.3.
- Set up API and user activity logging with AWS CloudTrail.
- Use AWS encryption solutions, along with all default security controls within AWS services.
- Use advanced managed security services such as Amazon Macie, which assists in discovering and securing sensitive data that is stored in Amazon S3.
- If you require FIPS 140-2 validated cryptographic modules when accessing AWS through a command line interface or an API, use a FIPS endpoint. For more information about the available FIPS endpoints, see Federal Information Processing Standard (FIPS) 140-2.

We strongly recommend that you never put confidential or sensitive information, such as your customers' email addresses, into tags or free-form text fields such as a Name field. This includes when you work with Amazon Bedrock or other AWS services using the console, API, AWS CLI, or AWS SDKs. Any data that you enter into tags or free-form text fields used for names may be used for billing or diagnostic logs. If you provide a URL to an external server, we strongly recommend that you do not include credentials information in the URL to validate your request to that server.

Data protection in Amazon Bedrock

Amazon Bedrock doesn't use your prompts and continuations to train any AWS models or distribute them to third parties.
Each model provider has an escrow account that they upload their models to. The Amazon Bedrock inference account has permissions to call these models, but the escrow accounts themselves don’t have outbound permissions to Amazon Bedrock accounts. Additionally, model providers don’t have access to Amazon Bedrock logs or access to customer prompts and continuations.

Amazon Bedrock doesn’t store or log your data in its service logs.

**Data protection in Amazon Bedrock model customization**

Your training data isn’t used to train the base Titan models or distributed to third parties. Other usage data, such as usage timestamps, logged account IDs, and other information logged by the service, is also not used to train the models.

Amazon Bedrock uses the fine tuning data you provide only for fine tuning an Amazon Bedrock foundation model. Amazon Bedrock doesn’t use fine tuning data for any other purpose, such as training base foundation models.

Bedrock uses your training data with the `CreateModelCustomizationJob` action, or with the console, to create a custom model which is a fine tuned version of an Amazon Bedrock foundational model. Your custom models are managed and stored by AWS. By default, custom models are encrypted with AWS Key Management Service keys that AWS owns, but you can use your own AWS KMS keys to encrypt your custom models. You encrypt a custom model when you submit a fine tuning job with the console or programmatically with the `CreateModelCustomizationJob` action.

None of the training or validation data you provide for fine tuning is stored in Amazon Bedrock accounts, once the fine tuning job completes. During training, your data exists in AWS Service Management Connector instance memory, but is encrypted on these machines using an XTS-AES-256 cipher that is implemented on a hardware module, on the instance itself.

We don’t recommend using confidential data to train a custom model as the model might generate inference responses based on that confidential data. If you use confidential data to train a custom model, the only way to prevent responses based on that data is to delete the custom model, remove the confidential data from your training dataset, and retrain the custom model.

Custom model metadata (name and Amazon Resource Name) and a provisioned model's metadata is stored in an Amazon DynamoDB table that is encrypted with a key that the Amazon Bedrock service owns.

**Topics**

- [Data encryption](#)
Protect your data using Amazon VPC and AWS PrivateLink

Data encryption

Amazon Bedrock uses encryption to protect data at rest and data in transit.

Topics

- Encryption in transit
- Encryption at rest
- Key management
- Encryption of model customization jobs and artifacts
- Encryption of agent resources
- Encryption of knowledge base resources

Encryption in transit

Within AWS, all inter-network data in transit supports TLS 1.2 encryption.

Requests to the Amazon Bedrock API and console are made over a secure (SSL) connection. You pass AWS Identity and Access Management (IAM) roles to Amazon Bedrock to provide permissions to access resources on your behalf for training and deployment.

Encryption at rest

Amazon Bedrock provides Encryption of model customization jobs and artifacts at rest.

Key management

Use the AWS Key Management Service to manage the keys that you use to encrypt your resources. For more information, see AWS Key Management Service concepts. You can encrypt the following resources with a KMS key.

- Through Amazon Bedrock
  - Model customization jobs and their output custom models – During job creation in the console or by specifying the customModelKmsKeyId field in the CreateModelCustomizationJob API call.
• Agents – During agent creation in the console or by specifying the `field in the `CreateAgent` API call.

• Data source ingestion jobs for knowledge bases – During knowledge base creation in the console or by specifying the `kmsKeyArn` field in the `CreateDataSource` or `UpdateDataSource` API call.

• Vector stores in Amazon OpenSearch Service – During vector store creation. For more information, see Creating, listing, and deleting Amazon OpenSearch Service collections and Encryption of data at rest for Amazon OpenSearch Service.

• Through Amazon S3 – For more information, see Using server-side encryption with AWS KMS keys (SSE-KMS).

• Training, validation, and output data for model customization

• Data sources for knowledge bases

• Through AWS Secrets Manager – For more information, see Secret encryption and decryption in AWS Secrets Manager

• Vector stores for third-party models

After you encrypt a resource, you can find the ARN of the KMS key by selecting a resource and viewing its Details in the console or by using the following Get API calls.

• GetModelCustomizationJob

• GetAgent

• GetIngestionJob

Encryption of model customization jobs and artifacts

By default, Amazon Bedrock encrypts the following model artifacts from your model customization jobs with an AWS managed key.

• The model customization job

• The output files (training and validation metrics) from the model customization job

• The resulting custom model
Optionally, you can encrypt the model artifacts by creating a customer managed key. For more information about AWS KMS keys, see Customer managed keys in the AWS Key Management Service Developer Guide. To use a customer managed key, carry out the following steps.

1. Create a customer managed key with the AWS Key Management Service.
2. Attach a resource-based policy with permissions for the specified-roles to create or use custom models.

**Topics**

- Create a customer managed key
- Create a key policy and attach it to the customer managed key
- Encryption of training, validation, and output data

**Create a customer managed key**

First ensure that you have CreateKey permissions. Then follow the steps at Creating keys to create a customer managed key either in the AWS KMS console or the CreateKey API operation. Make sure to create a symmetric encryption key.

Creation of the key returns an Arn for the key that you can use as the customModelKmsKeyId when submitting a model customization job.

**Create a key policy and attach it to the customer managed key**

Attach the following resource-based policy to the KMS key by following the steps at Creating a key policy. The policy contains two statements.

1. Permissions for a role to encrypt model customization artifacts. Add ARNs of custom model builder roles to the Principal field.
2. Permissions for a role to use a custom model in inference. Add ARNs of custom model user roles to the Principal field.

```
{
    "Version": "2012-10-17",
    "Id": "KMS Key Policy",
    "Statement": [
        {
```

Data encryption
Encryption of training, validation, and output data

When you use Amazon Bedrock to run a model customization job, you store the input (training/validation data) files in your Amazon S3 bucket. When the job completes, Amazon Bedrock stores the output metrics files in the S3 bucket that you specified when creating the job and the resulting custom model artifacts in an Amazon S3 bucket controlled by AWS.

The input and output files are encrypted with Amazon S3 SSE-S3 server-side encryption by default, using an AWS managed key. This type of key is created, managed, and used on your behalf by AWS.

You can instead choose to encrypt these files with a customer managed key that you create, own, and manage yourself. Refer to the preceding sections and the following links to learn how to create customer managed keys and key policies.

- To learn more about Amazon S3 SSE-S3 server-side encryption, see Using server-side encryption with Amazon S3 managed keys (SSE-S3)
- To learn more about customer managed keys for encrypting S3 objects, see Using server-side encryption with AWS KMS keys (SSE-KMS)
Encryption of agent resources

Amazon Bedrock encrypts your agent's session information. By default, Amazon Bedrock encrypts this data using an AWS managed key. Optionally, you can encrypt the agent artifacts using a customer managed key.

For more information about AWS KMS keys, see [Customer managed keys](#) in the [AWS Key Management Service Developer Guide](#).

If you encrypt sessions with your agent with a custom KMS key, you must set up the following identity-based policy and resource-based policy to allow Amazon Bedrock to encrypt and decrypt agent resources on your behalf.

1. Attach the following identity-based policy to an IAM role or user with permissions to make `InvokeAgent` calls. This policy validates the user making an `InvokeAgent` call has KMS permissions. Replace the `region`, `account-id`, `agent-id`, and `key-id` with the appropriate values.

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Allow Amazon Bedrock to encrypt and decrypt Agent resources on behalf of authorized users",
      "Effect": "Allow",
      "Action": [
        "kms:GenerateDataKey",
        "kms:Decrypt"
      ],
      "Resource": "arn:aws:kms:region:account-id:key/key-id",
      "Condition": {
        "StringEquals": {
        }
      }
    }
  ]
}
```
2. Attach the following resource-based policy to your KMS key. Change the scope of the permissions as necessary. Replace the `region`, `account-id`, `agent-id`, and `key-id` with the appropriate values.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "Allows account root to modify the KMS key, not used by Amazon Bedrock. Change it as per your permission requirements.",
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::account-id:root"
            },
            "Action": "kms:*",
            "Resource": "arn:aws:kms:region:account-id:key/key-id"
        },
        {
            "Sid": "Allow Amazon Bedrock to encrypt and decrypt Agent resources on behalf of authorized users",
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": [
                "kms:GenerateDataKey",
                "kms:Decrypt"
            ],
            "Resource": "arn:aws:kms:region:account-id:key/key-id",
            "Condition": {
                "StringEquals": {
                }
            }
        }
    ]
}
```
Encryption of knowledge base resources

Amazon Bedrock encrypts resources related to your knowledge bases. By default, Amazon Bedrock encrypts this data using an AWS managed key. Optionally, you can encrypt the model artifacts using a customer managed key.

Encryption with a KMS key can occur with the following processes:

- Transient data storage while ingesting your data sources
- Passing information to OpenSearch Service if you let Amazon Bedrock set up your vector database
- Querying a knowledge base

The following resources used by your knowledge bases can be encrypted with a KMS key. If you encrypt them, you need to add permissions to decrypt the KMS key.

- Data sources stored in an Amazon S3 bucket
- Third-party vector stores

For more information about AWS KMS keys, see Customer managed keys in the AWS Key Management Service Developer Guide.

Topics

- Encryption of transient data storage during data ingestion
- Encryption of information passed to Amazon OpenSearch Service
- Encryption of knowledge base retrieval
- Permissions to decrypt your AWS KMS key for your data sources in Amazon S3
- Permissions to decrypt an AWS Secrets Manager secret for the vector store containing your knowledge base

Encryption of transient data storage during data ingestion

When you set up a data ingestion job for your knowledge base, you can encrypt the job with a custom KMS key.
To allow the creation of a AWS KMS key for transient data storage in the process of ingesting your data source, attach the following policy to your Amazon Bedrock service role. Replace the `region`, `account-id`, and `key-id` with the appropriate values.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "kms:GenerateDataKey",
                "kms:Decrypt"
            ],
            "Resource": [
                "arn:aws:kms:region:account-id:key/key-id"
            ]
        }
    ]
}
```

**Encryption of information passed to Amazon OpenSearch Service**

If you opt to let Amazon Bedrock create a vector store in Amazon OpenSearch Service for your knowledge base, Amazon Bedrock can pass a KMS key that you choose to Amazon OpenSearch Service for encryption. To learn more about encryption in Amazon OpenSearch Service, see Encryption in Amazon OpenSearch Service.

**Encryption of knowledge base retrieval**

You can encrypt sessions in which you generate responses from querying a knowledge base with a KMS key. To do so, include the ARN of a KMS key in the `kmsKeyArn` field when making a `RetrieveAndGenerate` request. Attach the following policy, replacing the `values` appropriately to allow Amazon Bedrock to encrypt the session context.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": [
                "kms:Encrypt",
                "kms:Decrypt"
            ],
            "Resource": [
                "arn:aws:kms:region:account-id:key/key-id"
            ]
        }
    ]
}
```
Permissions to decrypt your AWS KMS key for your data sources in Amazon S3

You store the data sources for your knowledge base in your Amazon S3 bucket. To encrypt these documents at rest, you can use the Amazon S3 SSE-S3 server-side encryption option. With this option, objects are encrypted with service keys managed by the Amazon S3 service.

For more information, see Protecting data using server-side encryption with Amazon S3-managed encryption keys (SSE-S3) in the Amazon Simple Storage Service User Guide.

If you encrypted your data sources in Amazon S3 with a custom AWS KMS key, attach the following policy to your Amazon Bedrock service role to allow Amazon Bedrock to decrypt your key. Replace `region` and `account-id` with the region and account ID to which the key belongs. Replace `key-id` with the ID of your AWS KMS key.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "KMS:Decrypt"
            ],
            "Resource": [
                "arn:aws:kms:region:account-id:key/key-id"
            ],
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": [
                        "s3.region.amazonaws.com"
                    ]
                }
            }
        }
    ]
}
```
Permissions to decrypt an AWS Secrets Manager secret for the vector store containing your knowledge base

If the vector store containing your knowledge base is configured with an AWS Secrets Manager secret, you can encrypt the secret with a custom AWS KMS key by following the steps at Secret encryption and decryption in AWS Secrets Manager.

If you do so, you attach the following policy to your Amazon Bedrock service role to allow it to decrypt your key. Replace region and account-id with the region and account ID to which the key belongs. Replace key-id with the ID of your AWS KMS key.

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "kms:Decrypt"
         ],
         "Resource": [
            "arn:aws:kms:region:account-id:key/key-id"
         ]
      }
   ]
}
```

Protect your data using Amazon VPC and AWS PrivateLink

To control access to your data, we recommend that you use a virtual private cloud (VPC) with Amazon VPC. Using a VPC protects your data and lets you monitor all network traffic in and out of the AWS job containers by using VPC Flow Logs. You can further protect your data by configuring your VPC so that your data isn’t available over the internet and instead creating a VPC interface endpoint with AWS PrivateLink to establish a private connection to your data.

For an example of using VPC to protect data that you integrate with Amazon Bedrock see Protect model customization jobs using a VPC.

Use interface VPC endpoints (AWS PrivateLink)

You can use AWS PrivateLink to create a private connection between your VPC and Amazon Bedrock. You can access Amazon Bedrock as if it were in your VPC, without the use of an internet
gateway, NAT device, VPN connection, or AWS Direct Connect connection. Instances in your VPC don't need public IP addresses to access Amazon Bedrock.

You establish this private connection by creating an *interface endpoint*, powered by AWS PrivateLink. We create an endpoint network interface in each subnet that you enable for the interface endpoint. These are requester-managed network interfaces that serve as the entry point for traffic destined for Amazon Bedrock.

For more information, see [Access AWS services through AWS PrivateLink](https://docs.aws.amazon.com/privatelink/latest/userguide/) in the *AWS PrivateLink Guide*.

**Considerations for Amazon Bedrock VPC endpoints**

Before you set up an interface endpoint for Amazon Bedrock, review [Considerations](https://docs.aws.amazon.com/privatelink/latest/userguide/) in the *AWS PrivateLink Guide*.

Amazon Bedrock supports making the following API calls through VPC endpoints.

<table>
<thead>
<tr>
<th>Category</th>
<th>Endpoint prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Bedrock Control Plane API actions</td>
<td>bedrock</td>
</tr>
<tr>
<td>Amazon Bedrock Runtime API actions</td>
<td>bedrock-runtime</td>
</tr>
<tr>
<td>Agents for Amazon Bedrock Build-time API actions</td>
<td>bedrock-agent</td>
</tr>
<tr>
<td>Agents for Amazon Bedrock Runtime API actions</td>
<td>bedrock-agent-runtime</td>
</tr>
</tbody>
</table>

**Availability Zones**

Amazon Bedrock and Agents for Amazon Bedrock endpoints are available in multiple Availability Zones.

**Create an interface endpoint for Amazon Bedrock**

You can create an interface endpoint for Amazon Bedrock using either the Amazon VPC console or the AWS Command Line Interface (AWS CLI). For more information, see [Create an interface endpoint](https://docs.aws.amazon.com/privatelink/latest/userguide/) in the *AWS PrivateLink Guide*.
Create an interface endpoint for Amazon Bedrock using any of the following service names:

- com.amazonaws.region.bedrock
- com.amazonaws.region.bedrock-runtime
- com.amazonaws.region.bedrock-agent
- com.amazonaws.region.bedrock-agent-runtime

After you create the endpoint, you have the option to enable a private DNS hostname. Enable this setting by selecting Enable Private DNS Name in the VPC console when you create the VPC endpoint.

If you enable private DNS for the interface endpoint, you can make API requests to Amazon Bedrock using its default Regional DNS name. The following examples show the format of the default Regional DNS names.

- bedrock.region.amazonaws.com
- bedrock-runtime.region.amazonaws.com
- bedrock-agent.region.amazonaws.com
- bedrock-agent-runtime.region.amazonaws.com

Create an endpoint policy for your interface endpoint

An endpoint policy is an IAM resource that you can attach to an interface endpoint. The default endpoint policy allows full access to Amazon Bedrock through the interface endpoint. To control the access allowed to Amazon Bedrock from your VPC, attach a custom endpoint policy to the interface endpoint.

An endpoint policy specifies the following information:

- The principals that can perform actions (AWS accounts, IAM users, and IAM roles).
- The actions that can be performed.
- The resources on which the actions can be performed.

For more information, see Control access to services using endpoint policies in the AWS PrivateLink Guide.
Example: VPC endpoint policy for Amazon Bedrock actions

The following is an example of a custom endpoint policy. When you attach this resource-based policy to your interface endpoint, it grants access to the listed Amazon Bedrock actions for all principals on all resources.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Principal": "*",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel",
                "bedrock:InvokeModelWithResponseStream"
            ],
            "Resource": "*"
        }
    ]
}
```

Identity and access management for Amazon Bedrock

AWS Identity and Access Management (IAM) is an AWS service that helps an administrator securely control access to AWS resources. IAM administrators control who can be authenticated (signed in) and authorized (have permissions) to use Amazon Bedrock resources. IAM is an AWS service that you can use with no additional charge.

Topics

- Audience
- Authenticating with identities
- Managing access using policies
- How Amazon Bedrock works with IAM
- Identity-based policy examples for Amazon Bedrock
- AWS managed policies for Amazon Bedrock
- Service roles
- Troubleshooting Amazon Bedrock identity and access
Audience

How you use AWS Identity and Access Management (IAM) differs, depending on the work that you do in Amazon Bedrock.

**Service user** – If you use the Amazon Bedrock service to do your job, then your administrator provides you with the credentials and permissions that you need. As you use more Amazon Bedrock features to do your work, you might need additional permissions. Understanding how access is managed can help you request the right permissions from your administrator. If you cannot access a feature in Amazon Bedrock, see [Troubleshooting Amazon Bedrock identity and access](#).

**Service administrator** – If you're in charge of Amazon Bedrock resources at your company, you probably have full access to Amazon Bedrock. It's your job to determine which Amazon Bedrock features and resources your service users should access. You must then submit requests to your IAM administrator to change the permissions of your service users. Review the information on this page to understand the basic concepts of IAM. To learn more about how your company can use IAM with Amazon Bedrock, see [How Amazon Bedrock works with IAM](#).

**IAM administrator** – If you're an IAM administrator, you might want to learn details about how you can write policies to manage access to Amazon Bedrock. To view example Amazon Bedrock identity-based policies that you can use in IAM, see [Identity-based policy examples for Amazon Bedrock](#).

Authenticating with identities

Authentication is how you sign in to AWS using your identity credentials. You must be *authenticated* (signed in to AWS) as the AWS account root user, as an IAM user, or by assuming an IAM role.

You can sign in to AWS as a federated identity by using credentials provided through an identity source. AWS IAM Identity Center (IAM Identity Center) users, your company's single sign-on authentication, and your Google or Facebook credentials are examples of federated identities. When you sign in as a federated identity, your administrator previously set up identity federation using IAM roles. When you access AWS by using federation, you are indirectly assuming a role.

Depending on the type of user you are, you can sign in to the AWS Management Console or the AWS access portal. For more information about signing in to AWS, see [How to sign in to your AWS account](#) in the *AWS Sign-In User Guide*.
If you access AWS programmatically, AWS provides a software development kit (SDK) and a command line interface (CLI) to cryptographically sign your requests by using your credentials. If you don't use AWS tools, you must sign requests yourself. For more information about using the recommended method to sign requests yourself, see Signing AWS API requests in the IAM User Guide.

Regardless of the authentication method that you use, you might be required to provide additional security information. For example, AWS recommends that you use multi-factor authentication (MFA) to increase the security of your account. To learn more, see Multi-factor authentication in the AWS IAM Identity Center User Guide and Using multi-factor authentication (MFA) in AWS in the IAM User Guide.

**AWS account root user**

When you create an AWS account, you begin with one sign-in identity that has complete access to all AWS services and resources in the account. This identity is called the AWS account root user and is accessed by signing in with the email address and password that you used to create the account. We strongly recommend that you don't use the root user for your everyday tasks. Safeguard your root user credentials and use them to perform the tasks that only the root user can perform. For the complete list of tasks that require you to sign in as the root user, see Tasks that require root user credentials in the IAM User Guide.

**Federated identity**

As a best practice, require human users, including users that require administrator access, to use federation with an identity provider to access AWS services by using temporary credentials.

A federated identity is a user from your enterprise user directory, a web identity provider, the AWS Directory Service, the Identity Center directory, or any user that accesses AWS services by using credentials provided through an identity source. When federated identities access AWS accounts, they assume roles, and the roles provide temporary credentials.

For centralized access management, we recommend that you use AWS IAM Identity Center. You can create users and groups in IAM Identity Center, or you can connect and synchronize to a set of users and groups in your own identity source for use across all your AWS accounts and applications. For information about IAM Identity Center, see What is IAM Identity Center? in the AWS IAM Identity Center User Guide.
IAM users and groups

An **IAM user** is an identity within your AWS account that has specific permissions for a single person or application. Where possible, we recommend relying on temporary credentials instead of creating IAM users who have long-term credentials such as passwords and access keys. However, if you have specific use cases that require long-term credentials with IAM users, we recommend that you rotate access keys. For more information, see [Rotate access keys regularly for use cases that require long-term credentials](#) in the *IAM User Guide*.

An **IAM group** is an identity that specifies a collection of IAM users. You can't sign in as a group. You can use groups to specify permissions for multiple users at a time. Groups make permissions easier to manage for large sets of users. For example, you could have a group named `IAMAdmins` and give that group permissions to administer IAM resources.

Users are different from roles. A user is uniquely associated with one person or application, but a role is intended to be assumable by anyone who needs it. Users have permanent long-term credentials, but roles provide temporary credentials. To learn more, see [When to create an IAM user (instead of a role)](#) in the *IAM User Guide*.

IAM roles

An **IAM role** is an identity within your AWS account that has specific permissions. It is similar to an IAM user, but is not associated with a specific person. You can temporarily assume an IAM role in the AWS Management Console by **switching roles**. You can assume a role by calling an AWS CLI or AWS API operation or by using a custom URL. For more information about methods for using roles, see [Using IAM roles](#) in the *IAM User Guide*.

IAM roles with temporary credentials are useful in the following situations:

- **Federated user access** – To assign permissions to a federated identity, you create a role and define permissions for the role. When a federated identity authenticates, the identity is associated with the role and is granted the permissions that are defined by the role. For information about roles for federation, see [Creating a role for a third-party Identity Provider](#) in the *IAM User Guide*. If you use IAM Identity Center, you configure a permission set. To control what your identities can access after they authenticate, IAM Identity Center correlates the permission set to a role in IAM. For information about permissions sets, see [Permission sets](#) in the *AWS IAM Identity Center User Guide*.

- **Temporary IAM user permissions** – An IAM user or role can assume an IAM role to temporarily take on different permissions for a specific task.
• **Cross-account access** – You can use an IAM role to allow someone (a trusted principal) in a different account to access resources in your account. Roles are the primary way to grant cross-account access. However, with some AWS services, you can attach a policy directly to a resource (instead of using a role as a proxy). To learn the difference between roles and resource-based policies for cross-account access, see [How IAM roles differ from resource-based policies](#) in the IAM User Guide.

• **Cross-service access** – Some AWS services use features in other AWS services. For example, when you make a call in a service, it's common for that service to run applications in Amazon EC2 or store objects in Amazon S3. A service might do this using the calling principal's permissions, using a service role, or using a service-linked role.

• **Forward access sessions (FAS)** – When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see [Forward access sessions](#).

• **Service role** – A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see [Creating a role to delegate permissions to an AWS service](#) in the IAM User Guide.

• **Service-linked role** – A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.

• **Applications running on Amazon EC2** – You can use an IAM role to manage temporary credentials for applications that are running on an EC2 instance and making AWS CLI or AWS API requests. This is preferable to storing access keys within the EC2 instance. To assign an AWS role to an EC2 instance and make it available to all of its applications, you create an instance profile that is attached to the instance. An instance profile contains the role and enables programs that are running on the EC2 instance to get temporary credentials. For more information, see [Using an IAM role to grant permissions to applications running on Amazon EC2 instances](#) in the IAM User Guide.
To learn whether to use IAM roles or IAM users, see When to create an IAM role (instead of a user) in the IAM User Guide.

Managing access using policies

You control access in AWS by creating policies and attaching them to AWS identities or resources. A policy is an object in AWS that, when associated with an identity or resource, defines their permissions. AWS evaluates these policies when a principal (user, root user, or role session) makes a request. Permissions in the policies determine whether the request is allowed or denied. Most policies are stored in AWS as JSON documents. For more information about the structure and contents of JSON policy documents, see Overview of JSON policies in the IAM User Guide.

Administrators can use AWS JSON policies to specify who has access to what. That is, which principal can perform actions on what resources, and under what conditions.

By default, users and roles have no permissions. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

IAM policies define permissions for an action regardless of the method that you use to perform the operation. For example, suppose that you have a policy that allows the iam:GetRole action. A user with that policy can get role information from the AWS Management Console, the AWS CLI, or the AWS API.

Identity-based policies

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see Creating IAM policies in the IAM User Guide.

Identity-based policies can be further categorized as inline policies or managed policies. Inline policies are embedded directly into a single user, group, or role. Managed policies are standalone policies that you can attach to multiple users, groups, and roles in your AWS account. Managed policies include AWS managed policies and customer managed policies. To learn how to choose between a managed policy or an inline policy, see Choosing between managed policies and inline policies in the IAM User Guide.
Resource-based policies

Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM role trust policies and Amazon S3 bucket policies. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

Resource-based policies are inline policies that are located in that service. You can’t use AWS managed policies from IAM in a resource-based policy.

Access control lists (ACLs)

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

Amazon S3, AWS WAF, and Amazon VPC are examples of services that support ACLs. To learn more about ACLs, see Access control list (ACL) overview in the Amazon Simple Storage Service Developer Guide.

Other policy types

AWS supports additional, less-common policy types. These policy types can set the maximum permissions granted to you by the more common policy types.

- Permissions boundaries – A permissions boundary is an advanced feature in which you set the maximum permissions that an identity-based policy can grant to an IAM entity (IAM user or role). You can set a permissions boundary for an entity. The resulting permissions are the intersection of an entity's identity-based policies and its permissions boundaries. Resource-based policies that specify the user or role in the Principal field are not limited by the permissions boundary. An explicit deny in any of these policies overrides the allow. For more information about permissions boundaries, see Permissions boundaries for IAM entities in the IAM User Guide.

- Service control policies (SCPs) – SCPs are JSON policies that specify the maximum permissions for an organization or organizational unit (OU) in AWS Organizations. AWS Organizations is a service for grouping and centrally managing multiple AWS accounts that your business owns. If
you enable all features in an organization, then you can apply service control policies (SCPs) to any or all of your accounts. The SCP limits permissions for entities in member accounts, including each AWS account root user. For more information about Organizations and SCPs, see How SCPs work in the AWS Organizations User Guide.

- **Session policies** – Session policies are advanced policies that you pass as a parameter when you programmatically create a temporary session for a role or federated user. The resulting session's permissions are the intersection of the user or role's identity-based policies and the session policies. Permissions can also come from a resource-based policy. An explicit deny in any of these policies overrides the allow. For more information, see Session policies in the IAM User Guide.

### Multiple policy types

When multiple types of policies apply to a request, the resulting permissions are more complicated to understand. To learn how AWS determines whether to allow a request when multiple policy types are involved, see Policy evaluation logic in the IAM User Guide.

### How Amazon Bedrock works with IAM

Before you use IAM to manage access to Amazon Bedrock, learn what IAM features are available to use with Amazon Bedrock.

<table>
<thead>
<tr>
<th>IAM feature</th>
<th>Amazon Bedrock support</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identity-based policies</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Resource-based policies</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Policy actions</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Policy resources</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Policy condition keys</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>ACLs</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>ABAC (tags in policies)</strong></td>
<td>Yes</td>
</tr>
<tr>
<td>IAM feature</td>
<td>Amazon Bedrock support</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Temporary credentials</td>
<td>Yes</td>
</tr>
<tr>
<td>Principal permissions</td>
<td>Yes</td>
</tr>
<tr>
<td>Service roles</td>
<td>Yes</td>
</tr>
<tr>
<td>Service-linked roles</td>
<td>No</td>
</tr>
</tbody>
</table>

To get a high-level view of how Amazon Bedrock and other AWS services work with most IAM features, see [AWS services that work with IAM](#) in the IAM User Guide.

### Identity-based policies for Amazon Bedrock

| Supports identity-based policies | Yes |

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see [Creating IAM policies](#) in the IAM User Guide.

With IAM identity-based policies, you can specify allowed or denied actions and resources as well as the conditions under which actions are allowed or denied. You can't specify the principal in an identity-based policy because it applies to the user or role to which it is attached. To learn about all of the elements that you can use in a JSON policy, see [IAM JSON policy elements reference](#) in the IAM User Guide.

### Identity-based policy examples for Amazon Bedrock

To view examples of Amazon Bedrock identity-based policies, see [Identity-based policy examples for Amazon Bedrock](#).

### Resource-based policies within Amazon Bedrock

| Supports resource-based policies | No |
Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM role trust policies and Amazon S3 bucket policies. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

To enable cross-account access, you can specify an entire account or IAM entities in another account as the principal in a resource-based policy. Adding a cross-account principal to a resource-based policy is only half of establishing the trust relationship. When the principal and the resource are in different AWS accounts, an IAM administrator in the trusted account must also grant the principal entity (user or role) permission to access the resource. They grant permission by attaching an identity-based policy to the entity. However, if a resource-based policy grants access to a principal in the same account, no additional identity-based policy is required. For more information, see How IAM roles differ from resource-based policies in the IAM User Guide.

### Policy actions for Amazon Bedrock

| Supports policy actions | Yes |

Administrators can use AWS JSON policies to specify who has access to what. That is, which principal can perform actions on what resources, and under what conditions.

The Action element of a JSON policy describes the actions that you can use to allow or deny access in a policy. Policy actions usually have the same name as the associated AWS API operation. There are some exceptions, such as permission-only actions that don't have a matching API operation. There are also some operations that require multiple actions in a policy. These additional actions are called dependent actions.

Include actions in a policy to grant permissions to perform the associated operation.

To see a list of Amazon Bedrock actions, see Actions defined by Amazon Bedrock in the Service Authorization Reference.

Policy actions in Amazon Bedrock use the following prefix before the action:
To specify multiple actions in a single statement, separate them with commas.

```
"Action": [
  "bedrock:action1",
  "bedrock:action2"
]
```

To view examples of Amazon Bedrock identity-based policies, see Identity-based policy examples for Amazon Bedrock.

### Policy resources for Amazon Bedrock

| Supports policy resources | Yes |

Administrators can use AWS JSON policies to specify who has access to what. That is, which principal can perform actions on what resources, and under what conditions.

The Resource JSON policy element specifies the object or objects to which the action applies. Statements must include either a Resource or a NotResource element. As a best practice, specify a resource using its Amazon Resource Name (ARN). You can do this for actions that support a specific resource type, known as resource-level permissions.

For actions that don't support resource-level permissions, such as listing operations, use a wildcard (*) to indicate that the statement applies to all resources.

```
"Resource": "*"
```

To see a list of Amazon Bedrock resource types and their ARNs, see Resources defined by Amazon Bedrock in the Service Authorization Reference. To learn with which actions you can specify the ARN of each resource, see Actions defined by Amazon Bedrock.

Some Amazon Bedrock API actions support multiple resources. For example, AssociateAgentKnowledgeBase accesses AGENT12345 and KB12345678, so a principal must have
permissions to access both resources. To specify multiple resources in a single statement, separate
the ARNs with commas.

"Resource": [
]

To view examples of Amazon Bedrock identity-based policies, see Identity-based policy examples
for Amazon Bedrock.

**Policy condition keys for Amazon Bedrock**

<table>
<thead>
<tr>
<th>Supports service-specific policy condition keys</th>
<th>Yes</th>
</tr>
</thead>
</table>

Administrators can use AWS JSON policies to specify who has access to what. That is, which
principal can perform **actions** on what **resources**, and under what **conditions**.

The Condition element (or Condition block) lets you specify conditions in which a statement
is in effect. The Condition element is optional. You can create conditional expressions that use
condition operators, such as equals or less than, to match the condition in the policy with values in
the request.

If you specify multiple Condition elements in a statement, or multiple keys in a single
Condition element, AWS evaluates them using a logical AND operation. If you specify multiple
values for a single condition key, AWS evaluates the condition using a logical OR operation. All of
the conditions must be met before the statement's permissions are granted.

You can also use placeholder variables when you specify conditions. For example, you can grant
an IAM user permission to access a resource only if it is tagged with their IAM user name. For more
information, see IAM policy elements: variables and tags in the IAM User Guide.

AWS supports global condition keys and service-specific condition keys. To see all AWS global
condition keys, see AWS global condition context keys in the IAM User Guide.

To see a list of Amazon Bedrock condition keys, see Condition Keys for Amazon Bedrock in the
Service Authorization Reference. To learn with which actions and resources you can use a condition
key, see Actions defined by Amazon Bedrock.
All Amazon Bedrock actions support condition keys using Amazon Bedrock models as the resource.

To view examples of Amazon Bedrock identity-based policies, see Identity-based policy examples for Amazon Bedrock.

**ACLs in Amazon Bedrock**

| Supports ACLs | No |

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

**ABAC with Amazon Bedrock**

| Supports ABAC (tags in policies) | Yes |

Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes. In AWS, these attributes are called tags. You can attach tags to IAM entities (users or roles) and to many AWS resources. Tagging entities and resources is the first step of ABAC. Then you design ABAC policies to allow operations when the principal's tag matches the tag on the resource that they are trying to access.

ABAC is helpful in environments that are growing rapidly and helps with situations where policy management becomes cumbersome.

To control access based on tags, you provide tag information in the condition element of a policy using the `aws:ResourceTag/key-name`, `aws:RequestTag/key-name`, or `aws:TagKeys` condition keys.

If a service supports all three condition keys for every resource type, then the value is Yes for the service. If a service supports all three condition keys for only some resource types, then the value is Partial.

For more information about ABAC, see What is ABAC? in the IAM User Guide. To view a tutorial with steps for setting up ABAC, see Use attribute-based access control (ABAC) in the IAM User Guide.
Using temporary credentials with Amazon Bedrock

<table>
<thead>
<tr>
<th>Supports temporary credentials</th>
<th>Yes</th>
</tr>
</thead>
</table>

Some AWS services don't work when you sign in using temporary credentials. For additional information, including which AWS services work with temporary credentials, see [AWS services that work with IAM](#) in the *IAM User Guide*.

You are using temporary credentials if you sign in to the AWS Management Console using any method except a user name and password. For example, when you access AWS using your company's single sign-on (SSO) link, that process automatically creates temporary credentials. You also automatically create temporary credentials when you sign in to the console as a user and then switch roles. For more information about switching roles, see [Switching to a role (console)](#) in the *IAM User Guide*.

You can manually create temporary credentials using the AWS CLI or AWS API. You can then use those temporary credentials to access AWS. AWS recommends that you dynamically generate temporary credentials instead of using long-term access keys. For more information, see [Temporary security credentials in IAM](#).

Cross-service principal permissions for Amazon Bedrock

<table>
<thead>
<tr>
<th>Supports forward access sessions (FAS)</th>
<th>Yes</th>
</tr>
</thead>
</table>

When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see [Forward access sessions](#).

Service roles for Amazon Bedrock

<table>
<thead>
<tr>
<th>Supports service roles</th>
<th>Yes</th>
</tr>
</thead>
</table>
A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Creating a role to delegate permissions to an AWS service in the IAM User Guide.

⚠️ Warning
Changing the permissions for a service role might break Amazon Bedrock functionality. Edit service roles only when Amazon Bedrock provides guidance to do so.

### Service-linked roles for Amazon Bedrock

| Supports service-linked roles | No |

A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.

### Identity-based policy examples for Amazon Bedrock

By default, users and roles don't have permission to create or modify Amazon Bedrock resources. They also can't perform tasks by using the AWS Management Console, AWS Command Line Interface (AWS CLI), or AWS API. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

To learn how to create an IAM identity-based policy by using these example JSON policy documents, see Creating IAM policies in the IAM User Guide.

For details about actions and resource types defined by Amazon Bedrock, including the format of the ARNs for each of the resource types, see Actions, Resources, and Condition Keys for Amazon Bedrock in the Service Authorization Reference.

### Topics
- Policy best practices
- Use the Amazon Bedrock console
• **Allow users to view their own permissions**

• **Allow access to third-party model subscriptions**

• **Deny access for inference on specific models**

• **Identity-based policy examples for Agents for Amazon Bedrock**

**Policy best practices**

Identity-based policies determine whether someone can create, access, or delete Amazon Bedrock resources in your account. These actions can incur costs for your AWS account. When you create or edit identity-based policies, follow these guidelines and recommendations:

• **Get started with AWS managed policies and move toward least-privilege permissions** – To get started granting permissions to your users and workloads, use the *AWS managed policies* that grant permissions for many common use cases. They are available in your AWS account. We recommend that you reduce permissions further by defining AWS customer managed policies that are specific to your use cases. For more information, see [AWS managed policies](https://docs.aws.amazon.com/IAM/latest/userguide/using-managed-policies.html) or [AWS managed policies for job functions](https://docs.aws.amazon.com/IAM/latest/userguide/using-managed-policies-for-job-functions.html) in the *IAM User Guide*.

• **Apply least-privilege permissions** – When you set permissions with IAM policies, grant only the permissions required to perform a task. You do this by defining the actions that can be taken on specific resources under specific conditions, also known as *least-privilege permissions*. For more information about using IAM to apply permissions, see [Policies and permissions in IAM](https://docs.aws.amazon.com/IAM/latest/userguide/policies-and-permissions.html) in the *IAM User Guide*.

• **Use conditions in IAM policies to further restrict access** – You can add a condition to your policies to limit access to actions and resources. For example, you can write a policy condition to specify that all requests must be sent using SSL. You can also use conditions to grant access to service actions if they are used through a specific AWS service, such as AWS CloudFormation. For more information, see [IAM JSON policy elements: Condition](https://docs.aws.amazon.com/IAM/latest/userguide/reference-iam-policies-elements.html) in the *IAM User Guide*.

• **Use IAM Access Analyzer to validate your IAM policies to ensure secure and functional permissions** – IAM Access Analyzer validates new and existing policies so that the policies adhere to the IAM policy language (JSON) and IAM best practices. IAM Access Analyzer provides more than 100 policy checks and actionable recommendations to help you author secure and functional policies. For more information, see [IAM Access Analyzer policy validation](https://docs.aws.amazon.com/IAM/latest/userguide/using-access-analyzer.html) in the *IAM User Guide*.

• **Require multi-factor authentication (MFA)** – If you have a scenario that requires IAM users or a root user in your AWS account, turn on MFA for additional security. To require MFA when
API operations are called, add MFA conditions to your policies. For more information, see Configuring MFA-protected API access in the IAM User Guide.

For more information about best practices in IAM, see Security best practices in IAM in the IAM User Guide.

Use the Amazon Bedrock console

To access the Amazon Bedrock console, you must have a minimum set of permissions. These permissions must allow you to list and view details about the Amazon Bedrock resources in your AWS account. If you create an identity-based policy that is more restrictive than the minimum required permissions, the console won't function as intended for entities (users or roles) with that policy.

You don't need to allow minimum console permissions for users that are making calls only to the AWS CLI or the AWS API. Instead, allow access to only the actions that match the API operation that they're trying to perform.

To provide access to the Amazon Bedrock console, attach the following policy to the roles or entities that need access. A few of the actions in the policy are for use by the Amazon Bedrock console only, so IAM might raise Invalid Action errors. You can ignore these errors and choose Next to complete the task.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "BedrockConsole",
            "Effect": "Allow",
            "Action": [
                "bedrock:PutFoundationModelEntitlement",
                "bedrock:GetModelInvocationLoggingConfiguration",
                "bedrock:PutModelInvocationLoggingConfiguration",
                "bedrock:CreateFoundationModelAgreement",
                "bedrock:DeleteFoundationModelAgreement",
                "bedrock:ListFoundationModelAgreementOffers",
                "bedrock:GetUseCaseForModelAccess",
                "bedrock:PutUseCaseForModelAccess",
                "bedrock:ListFoundationModels",
                "bedrock:GetFoundationModel",
                "bedrock:InvokeModel",
            ]
        }
    ]
}
```
"bedrock:InvokeModelWithResponseStream",
"bedrock:CreateModelCustomizationJob",
"bedrock:GetModelCustomizationJob",
"bedrock:GetFoundationModelAvailability",
"bedrock:ListModelCustomizationJobs",
"bedrock:StopModelCustomizationJob",
"bedrock:GetCustomModel",
"bedrock:ListCustomModels",
"bedrock:DeleteCustomModel",
"bedrock:CreateProvisionedModelThroughput",
"bedrock:UpdateProvisionedModelThroughput",
"bedrock:GetProvisionedModelThroughput",
"bedrock:ListProvisionedModelThroughputs",
"bedrock:DeleteProvisionedModelThroughput",
"bedrock:TagResource",
"bedrock:UntagResource",
"bedrock:ListTagsForResource",
"bedrock:CreateAgent",
"bedrock:UpdateAgent",
"bedrock:GetAgent",
"bedrock:ListAgents",
"bedrock:DeleteAgent",
"bedrock:CreateAgentActionGroup",
"bedrock:UpdateAgentActionGroup",
"bedrock:GetAgentActionGroup",
"bedrock:ListAgentActionGroups",
"bedrock:DeleteAgentActionGroup",
"bedrock:GetAgentVersion",
"bedrock:ListAgentVersions",
"bedrock:CreateAgentAlias",
"bedrock:UpdateAgentAlias",
"bedrock:GetAgentAlias",
"bedrock:ListAgentAliases",
"bedrock:DeleteAgentAlias",
"bedrock:AssociateAgentKnowledgeBase",
"bedrock:DisassociateAgentKnowledgeBase",
"bedrock:GetKnowledgeBase",
"bedrock:ListKnowledgeBases",
"bedrock:PrepareAgent",
"bedrock:InvokeAgent",
"bedrock:CreateKnowledgeBase",
"bedrock:UpdateKnowledgeBase",
"bedrock:GetKnowledgeBase",
"bedrock:ListKnowledgeBases"
For more information, see [Adding permissions to a user](https://docs.aws.amazon.com/IAM/latest_USER_Guide/) in the IAM User Guide.

**Allow users to view their own permissions**

This example shows how you might create a policy that allows IAM users to view the inline and managed policies that are attached to their user identity. This policy includes permissions to complete this action on the console or programmatically using the AWS CLI or AWS API.

```json
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ViewOwnUserInfo",
      "Effect": "Allow",
      "Action": [
        "iam:GetUserPolicy",
        "iam:ListGroupsForUser",
        "iam:ListAttachedUserPolicies",
        "iam:ListUserPolicies",
        "iam:GetUser"
      ],
      "Resource": ["arn:aws:iam::*:user/${aws:username}" ]
    },
    {
      "Sid": "NavigateInConsole",
      "Effect": "Allow",
      "Action": [
        "iam:GetUserPolicy",
        "iam:ListGroupsForUser",
        "iam:ListAttachedUserPolicies",
        "iam:ListUserPolicies",
        "iam:GetUser"
      ],
      "Resource": ["arn:aws:iam::*:user/${aws:username}" ]
    }
  ]
}
```
Allow access to third-party model subscriptions

To access the Amazon Bedrock models for the first time, you use the Amazon Bedrock console to subscribe to third-party models. Your IAM user or role that the console user assumes requires permission to access the subscription API operations.

The following example shows an identity-based policy to allow access to the subscription API operations.

Use a condition key, as in the example, to limits the scope of the policy to a subset of the Amazon Bedrock foundation models in the Marketplace. To see a list of product IDs and which foundation models they correspond to, see the table in Control model access permissions.

```json
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "aws-marketplace:Subscribe"
         ],
         "Resource": "*",
         "Condition": {
            "ForAnyValue:StringEquals": {
               "aws-marketplace:ProductId": [
                  "1d288c71-65f9-489a-a3e2-9c7f4f6e6a85",
                  "cc0bde50-279a-40d8-829c-4009b77a1f6c"
               ]
            }
         }
      }
   ]
}
```
Deny access for inference on specific models

The following example shows an identity-based policy that denies access to running inference on a specific model.

```json
{
    "Version": "2012-10-17",
    "Statement": {
        "Sid": "DenyInference",
        "Effect": "Deny",
        "Action": [
            "bedrock:InvokeModel",
            "bedrock:InvokeModelWithResponseStream"
        ],
        "Resource": "arn:aws:bedrock:*::foundation-model/model-id"
    }
}
```
Identity-based policy examples for Agents for Amazon Bedrock

Select a topic to see example IAM policies that you can attach to an IAM role to provision permissions for actions in Agents for Amazon Bedrock.

Topics

- Required permissions for Agents for Amazon Bedrock
- Allow users to view information about and invoke an agent

Required permissions for Agents for Amazon Bedrock

For an IAM identity to use Agents for Amazon Bedrock, you must configure it with the necessary permissions. You can attach the AmazonBedrockFullAccess policy to grant the proper permissions to the role.

To restrict permissions to only actions that are used in Agents for Amazon Bedrock, attach the following identity-based policy to an IAM role:

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AgentsRole",
            "Effect": "Allow",
            "Action": [
                "bedrock:ListFoundationModels",
                "bedrock:GetFoundationModel",
                "bedrock:TagResource",
                "bedrock:UntagResource",
                "bedrock:ListTagsForResource",
                "bedrock:CreateAgent",
                "bedrock:UpdateAgent",
                "bedrock:GetAgent",
                "bedrock:ListAgents",
                "bedrock:DeleteAgent",
                "bedrock:CreateAgentActionGroup",
                "bedrock:UpdateAgentActionGroup",
                "bedrock:GetAgentActionGroup",
                "bedrock:ListAgentActionGroups",
```
Amazon Bedrock
User Guide

"bedrock:DeleteAgentActionGroup",
"bedrock:GetAgentVersion",
"bedrock:ListAgentVersions",
"bedrock:DeleteAgentVersion",
"bedrock:CreateAgentAlias",
"bedrock:UpdateAgentAlias",
"bedrock:GetAgentAlias",
"bedrock:ListAgentAliases",
"bedrock:DeleteAgentAlias",
"bedrock:AssociateAgentKnowledgeBase",
"bedrock:DisassociateAgentKnowledgeBase",
"bedrock:GetKnowledgeBase",
"bedrock:ListKnowledgeBases",
"bedrock:PrepareAgent",
"bedrock:InvokeAgent"
],
"Resource": "*"
}

You can further restrict permissions by omitting actions or specifying resources and condition keys. An IAM identity can call API operations on specific resources. For example, the UpdateAgent operation can only be used on agent resources and the InvokeAgent operation can only be used on alias resources. For API operations that aren't used on a specific resource type (such as CreateAgent), specify * as the Resource. If you specify an API operation that can't be used on the resource specified in the policy, Amazon Bedrock returns an error.

Allow users to view information about and invoke an agent

The following is a sample policy that you can attach to an IAM role to allow it to view information about or edit an agent with the ID AGENT12345 and to interact with its alias with the ID ALIAS12345. For example, you could attach this policy to a role that you want to only have permissions to troubleshoot an agent and update it.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "sid",
            "Effect": "Allow",
            "Action": [
```
AWS managed policies for Amazon Bedrock

To add permissions to users, groups, and roles, it’s easier to use AWS managed policies than to write policies yourself. It takes time and expertise to create IAM customer managed policies that provide your team with only the permissions they need. To get started quickly, you can use our AWS managed policies. These policies cover common use cases and are available in your AWS account. For more information about AWS managed policies, see AWS managed policies in the IAM User Guide.

AWS services maintain and update AWS managed policies. You can't change the permissions in AWS managed policies. Services occasionally add additional permissions to an AWS managed policy to support new features. This type of update affects all identities (users, groups, and roles) where the policy is attached. Services are most likely to update an AWS managed policy when a new feature is launched or when new operations become available. Services do not remove permissions from an AWS managed policy, so policy updates won’t break your existing permissions.

Additionally, AWS supports managed policies for job functions that span multiple services. For example, the ReadOnlyAccess AWS managed policy provides read-only access to all AWS services and resources. When a service launches a new feature, AWS adds read-only permissions for new operations and resources. For a list and descriptions of job function policies, see AWS managed policies for job functions in the IAM User Guide.
AWS managed policy: AmazonBedrockFullAccess

You can attach the AmazonBedrockFullAccess policy to your IAM identities.

This policy grants administrative permissions that allow the user permission to create, read, update, and delete Amazon Bedrock resources.

ℹ️ Note
Fine-tuning and model access require extra permissions. See Allow access to third-party model subscriptions and Permissions to access training and validation files and to write output files in S3 for more information.

Permissions details

This policy includes the following permissions:

- **ec2** (Amazon Elastic Compute Cloud) – Allows permissions to describe VPCs, subnets, and security groups.
- **iam** (AWS Identity and Access Management) – Allows principals to pass roles, but only allows IAM roles with "Amazon Bedrock" in them to be passed to the Amazon Bedrock service. The permissions are restricted to bedrock.amazonaws.com for Amazon Bedrock operations.
- **kms** (AWS Key Management Service) – Allows principals to describe AWS KMS keys and aliases.
- **bedrock** (Amazon Bedrock) – Allows principals read and write access to all actions in the Amazon Bedrock control plane and runtime service.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "BedrockAll",
            "Effect": "Allow",
            "Action": [
                "bedrock:*"
            ],
            "Resource": "*"
        },
        {
```
"Sid": "DescribeKey",
"Effect": "Allow",
"Action": [
    "kms:DescribeKey"
],
"Resource": "arn:*:kms::*::*"
},
{
    "Sid": "APIsWithAllResourceAccess",
    "Effect": "Allow",
    "Action": [
        "iam:ListRoles",
        "ec2:DescribeVpcs",
        "ec2:DescribeSubnets",
        "ec2:DescribeSecurityGroups"
    ],
    "Resource": "*"
},
{
    "Sid": "PassRoleToBedrock",
    "Effect": "Allow",
    "Action": [
        "iam:PassRole"
    ],
    "Resource": "arn:aws:iam::*:role/*AmazonBedrock*",
    "Condition": {
        "StringEquals": {
            "iam:PassedToService": [
                "bedrock.amazonaws.com"
            ]
        }
    }
}

AWS managed policy: AmazonBedrockReadOnly

You can attach the AmazonBedrockReadOnly policy to your IAM identities.

This policy grants read-only permissions that allow users to view all resources in Amazon Bedrock.
Amazon Bedrock updates to AWS managed policies

View details about updates to AWS managed policies for Amazon Bedrock since this service began tracking these changes. For automatic alerts about changes to this page, subscribe to the RSS feed on the Document history for the Amazon Bedrock User Guide.

<table>
<thead>
<tr>
<th>Change</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmazonBedrockFullAccess – New policy</td>
<td>Amazon Bedrock added a new policy to give users permissions to create, read, update, and delete resources.</td>
<td>December 12, 2023</td>
</tr>
<tr>
<td>AmazonBedrockReadOnly – New policy</td>
<td>Amazon Bedrock added a new policy to give users read-only permissions for all actions.</td>
<td>December 12, 2023</td>
</tr>
</tbody>
</table>
Amazon Bedrock started tracking changes for its AWS managed policies.

December 12, 2023

Service roles

Amazon Bedrock uses IAM service roles for the following features to let Amazon Bedrock carry out tasks on your behalf.

The console automatically creates service roles for supported features.

You can also create a custom service role and customize the attached permissions to your specific use-case. If you use the console, you can select this role instead of letting Amazon Bedrock create one for you.

To set up the custom service role, you carry out the following general steps.

1. Create the role by following the steps at Creating a role to delegate permissions to an AWS service.
2. Attach a trust policy.
3. Attach the relevant identity-based permissions.

Refer to the following links for more information about IAM concepts that are relevant to setting service role permissions.

- AWS service role
- Identity-based policies and resource-based policies
- Using resource-based policies for Lambda
- AWS global condition context keys
- Condition keys for Amazon Bedrock

Select a topic to learn more about service roles for a specific feature.

Topics
• Create a service role for model customization
• Create a service role for Agents for Amazon Bedrock
• Create a service role for Knowledge bases for Amazon Bedrock

Create a service role for model customization

To use a custom role for model customization instead of the one Amazon Bedrock automatically creates, create an IAM role and attach the following permissions by following the steps at Creating a role to delegate permissions to an AWS service.

• Trust relationship
• Permissions to access your training and validation data in S3 and to write your output data to S3
• (Optional) If you encrypt any of the following resources with a KMS key, permissions to decrypt the key (see Encryption of model customization jobs and artifacts)
  • A model customization job or the resulting custom model
  • The training, validation, or output data for the model customization job

Topics
• Trust relationship
• Permissions to access training and validation files and to write output files in S3

Trust relationship

The following policy allows Amazon Bedrock to assume this role and carry out the model customization job. The following shows an example policy you can use.

You can optionally restrict the scope of the permission for cross-service confused deputy prevention by using one or more global condition context keys with the Condition field. For more information, see AWS global condition context keys.

• Set the aws:SourceAccount value to your account ID.
• (Optional) Use the ArnEquals or ArnLike condition to restrict the scope to specific model customization jobs in your account ID.
Permissions to access training and validation files and to write output files in S3

Attach the following policy to allow the role to access your training and validation data and the bucket to which to write your output data. Replace the values in the Resource list with your actual bucket names.

To restrict access to a specific folder in a bucket, add an `s3:prefix` condition key with your folder path. You can follow the **User policy** example in [Example 2: Getting a list of objects in a bucket with a specific prefix](#)

```json
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Principal": {
            "Service": "bedrock.amazonaws.com"
         },
         "Action": "sts:AssumeRole",
         "Condition": {
            "StringEquals": {
               "aws:SourceAccount": "account-id"
            },
            "ArnEquals": {
            }
         }
      }
   ]
}
```
Create a service role for Agents for Amazon Bedrock

To use a custom service role for agents instead of the one Amazon Bedrock automatically creates, create an IAM role with the prefix `AmazonBedrockExecutionRoleForAgents_` and attach the following permissions by following the steps at Creating a role to delegate permissions to an AWS service.

- Trust policy
- A policy containing the following identity-based permissions
  - Access to the Amazon Bedrock base models
  - Access to the Amazon S3 objects containing the OpenAPI schemas for the action groups in your agents
  - Permissions for Amazon Bedrock to query knowledge bases that you want to attach to your agents
  - (Optional) If you encrypt your agent with a KMS key, permissions to decrypt the key (see Encryption of agent resources)

Whether you use a custom role or not, you also need to attach a resource-based policy to the Lambda functions for the action groups in your agents to provide permissions for the service role
to access the functions. For more information, see Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function.

Topics

- Trust relationship
- Identity-based permissions for the Agents service role.
- Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function

Trust relationship

The following trust policy allows Amazon Bedrock to assume this role and create and manage agents. Replace the values as necessary. The policy contains optional condition keys (see Condition keys for Amazon Bedrock and AWS global condition context keys) in the Condition field that we recommend you use as a security best practice.

Note

As a best practice for security purposes, replace the * with specific agent IDs after you have created them.

```json
{
    "Version": "2012-10-17",
    "Statement": [[
        "Effect": "Allow",
        "Principal": {
            "Service": "bedrock.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "StringEquals": {
                "aws:SourceAccount": "account-id"
            },
            "ArnLike": {
            }
        }
    ]]
}
```
Identity-based permissions for the Agents service role.

Attach the following policy to provide permissions for the service role, replacing `values` as necessary. The policy contains the following statements. Omit a statement if it isn't applicable to your use-case. The policy contains optional condition keys (see [Condition keys for Amazon Bedrock](#) and [AWS global condition context keys](#)) in the Condition field that we recommend you use as a security best practice.

![Note](https://example.com).

If you encrypt your agent with a customer-managed KMS key, refer to [Encryption of agent resources](#) for further permissions you need to add.

- Permissions to use Amazon Bedrock foundation models to run model inference on prompts used in your agent's orchestration.
- Permissions to access your agent's action group API schemas in Amazon S3. Omit this statement if your agent has no action groups.
- Permissions to access knowledge bases associated with your agent. Omit this statement if your agent has no associated knowledge bases.
- Permissions to access a third-party (Pinecone or Redis Enterprise Cloud) knowledge base associated with your agent. Omit this statement if your knowledge base is first-party (Amazon OpenSearch Serverless or Amazon Aurora) or if your agent has no associated knowledge bases.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "Allow model invocation for orchestration",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel"
            ],
            "Resource": [
                "arn:aws:bedrock:region::foundation-model/anthropic.claude-v2",
                "arn:aws:bedrock:region::foundation-model/anthropic.claude-instant-v1"
            ]
        }
    ]
}
```
{
    "Sid": "Allow access to action group API schemas in S3",
    "Effect": "Allow",
    "Action": [
        "s3:GetObject"
    ],
    "Resource": [
        "arn:aws:s3:::bucket/path/to/schema"
    ],
    "Condition": {
        "StringEquals": {
            "aws:ResourceAccount": "account-id"
        }
    }
},
{
    "Sid": "Query associated knowledge bases",
    "Effect": "Allow",
    "Action": [
        "bedrock:Retrieve",
        "bedrock:RetrieveAndGenerate"
    ],
    "Resource": [
    ]
},
{
    "Sid": "Associate a third-party knowledge base with your agent",
    "Effect": "Allow",
    "Action": [
        "bedrock:AssociateThirdPartyKnowledgeBase"
    ],
    "Condition": {
        "StringEquals": {
            "bedrock:ThirdPartyKnowledgeBaseCredentialsSecretArn": "arn:aws:kms:region:account-id:key/key-id"
        }
    }
}
}
Resource-based policy to allow Amazon Bedrock to invoke an action group Lambda function

Follow the steps at Using resource-based policies for Lambda and attach the following resource-based policy to a Lambda function to allow Amazon Bedrock to access the Lambda function for your agent's action groups, replacing the values as necessary. The policy contains optional condition keys (see Condition keys for Amazon Bedrock and AWS global condition context keys) in the Condition field that we recommend you use as a security best practice.

```json
{
    "Version": "2012-10-17",
    "Statement": [{
        "Sid": "Allow Amazon Bedrock to access action group Lambda function",
        "Effect": "Allow",
        "Principal": {
            "Service": "bedrock.amazonaws.com"
        },
        "Action": "lambda:InvokeFunction",
        "Condition": {
            "StringEquals": {
                "AWS:SourceAccount": "account-id"
            },
            "ArnLike": {
            }
        }
    }]
}
```

Create a service role for Knowledge bases for Amazon Bedrock

To use a custom role for knowledge base instead of the one Amazon Bedrock automatically creates, create an IAM role and attach the following permissions by following the steps at Creating a role to delegate permissions to an AWS service. You can use the same role across your knowledge bases.

- Trust relationship
- Access to the Amazon Bedrock base models
- Access to the Amazon S3 objects containing your data sources
• (If you create a vector database in Amazon OpenSearch Service) Access to your OpenSearch Service collection

• (If you create a vector database in Amazon Aurora)

• (If you create a vector database in Pinecone or Redis Enterprise Cloud) Permissions for AWS Secrets Manager to authenticate your Pinecone or Redis Enterprise Cloud account

• (Optional) If you encrypt any of the following resources with a KMS key, permissions to decrypt the key (see Encryption of knowledge base resources).
  - Your knowledge base
  - Data sources for your knowledge base
  - Your vector database in Amazon OpenSearch Service
  - The secret for your third-party vector database in AWS Secrets Manager
  - A data ingestion job

Topics

• Trust relationship

• Permissions to access Amazon Bedrock models

• Permissions to access your data sources in Amazon S3

• (Optional) Permissions to access your vector database in Amazon OpenSearch Service

• (Optional) Permissions to access your Amazon Aurora database cluster

• (Optional) Permissions to access a vector database configured with an AWS Secrets Manager secret

• (Optional) Permissions for AWS to manage a AWS KMS key for transient data storage during data ingestion

Trust relationship

The following policy allows Amazon Bedrock to assume this role and create and manage knowledge bases. The following shows an example policy you can use. You can restrict the scope of the permission by using one or more global condition context keys. For more information, see AWS global condition context keys. Set the aws:SourceAccount value to your account ID. Use the ArnEquals or ArnLike condition to restrict the scope to specific knowledge bases.
Note

As a best practice for security purposes, replace the * with specific knowledge base IDs after you have created them.

```
{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Principal": {
      "Service": "bedrock.amazonaws.com"
    },
    "Action": "sts:AssumeRole",
    "Condition": {
      "StringEquals": {
        "aws:SourceAccount": "account-id"
      },
      "ArnLike": {
      }
    }
  }]
}
```

Permissions to access Amazon Bedrock models

Attach the following policy to provide permissions for the role to use Amazon Bedrock models to embed your source data.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "bedrock:ListFoundationModels",
        "bedrock:ListCustomModels"
      ],
      "Resource": "*"
    }
  ]
}
```
"Effect": "Allow",
"Action": [
  "bedrock:InvokeModel"
],
"Resource": [
  "arn:aws:bedrock:region::foundation-model/amazon.titan-embed-text-v1",
  "arn:aws:bedrock:region::foundation-model/cohere.embed-multilingual-v3"
}
]

Permissions to access your data sources in Amazon S3

Attach the following policy to provide permissions for the role to access the Amazon S3 URIs containing the data source files for your knowledge base. In the Resource field, provide an Amazon S3 object containing the data sources or add the URI of each data source to the list.

If you encrypted these data sources with a AWS KMS key, attach permissions to decrypt the key to the role by following the steps at Permissions to decrypt your AWS KMS key for your data sources in Amazon S3.

{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Action": [
      "s3:GetObject",
      "s3:ListBucket"
    ],
    "Resource": [
      "arn:aws:s3:::bucket/path/to/folder",
      "arn:aws:s3:::bucket/path/to/folder/*"
    ],
    "Condition": {
      "StringEquals": {
        "aws:PrincipalAccount": "account-id"
      }
    }
  }]
}
(Optional) Permissions to access your vector database in Amazon OpenSearch Service

If you created a vector database in Amazon OpenSearch Service for your knowledge base, attach the following policy to your Knowledge bases for Amazon Bedrock service role to allow access to the collection. Replace `region` and `account-id` with the region and account ID to which the database belongs. Input the ID of your Amazon OpenSearch Service collection in `collection-id`. You can allow access to multiple collections by adding them to the Resource list.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "aoss:APIAccessAll"
            ],
            "Resource": [
            ]
        }
    ]
}
```

(Optional) Permissions to access your Amazon Aurora database cluster

If you created a database (DB) cluster in Amazon Aurora for your knowledge base, attach the following policy to your Knowledge bases for Amazon Bedrock service role to allow access to the DB cluster and to provide read and write permissions on it. Replace `region` and `account-id` with the region and account ID to which the DB cluster belongs. Input the ID of your Amazon Aurora database cluster in `db-cluster-id`. You can allow access to multiple DB clusters by adding them to the Resource list.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "RdsDescribeStatementID",
            "Effect": "Allow",
            "Action": [
                "rds:DescribeDBClusters"
            ],
            "Resource": [
                "arn:aws:rds:region:account-id:cluster:db-cluster-id"
            ]
        }
    ]
}
```
(Optional) Permissions to access a vector database configured with an AWS Secrets Manager secret

If your vector database is configured with an AWS Secrets Manager secret, attach the following policy to your Knowledge bases for Amazon Bedrock service role to allow AWS Secrets Manager to authenticate your account to access the database. Replace `region` and `account-id` with the region and account ID to which the database belongs. Replace `secret-id` with the ID of your secret.

```json
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "secretsmanager:GetSecretValue"
        ],
        "Resource": [
            "arn:aws:secretsmanager:region:account-id:secret:secret-id"
        ]
    }]
}
```

If you encrypted your secret with a AWS KMS key, attach permissions to decrypt the key to the role by following the steps at Permissions to decrypt an AWS Secrets Manager secret for the vector store containing your knowledge base.
(Optional) Permissions for AWS to manage a AWS KMS key for transient data storage during data ingestion

To allow the creation of a AWS KMS key for transient data storage in the process of ingesting your data source, attach the following policy to your Knowledge bases for Amazon Bedrock service role. Replace the region, account-id, and key-id with the appropriate values.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "kms:GenerateDataKey",
                "kms:Decrypt"
            ],
            "Resource": [
                "arn:aws:kms:region:account-id:key/key-id"
            ]
        }
    ]
}
```

Troubleshooting Amazon Bedrock identity and access

Use the following information to help you diagnose and fix common issues that you might encounter when working with Amazon Bedrock and IAM.

Topics

- I am not authorized to perform an action in Amazon Bedrock
- I am not authorized to perform iam:PassRole
- I want to allow people outside of my AWS account to access my Amazon Bedrock resources

I am not authorized to perform an action in Amazon Bedrock

If you receive an error that you're not authorized to perform an action, your policies must be updated to allow you to perform the action.
The following example error occurs when the mateojackson IAM user tries to use the console to view details about a fictional my-example-widget resource but doesn't have the fictional bedrock: GetWidget permissions.

User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform: bedrock:GetWidget on resource: my-example-widget

In this case, the policy for the mateojackson user must be updated to allow access to the my-example-widget resource by using the bedrock: GetWidget action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

**I am not authorized to perform iam:PassRole**

If you receive an error that you're not authorized to perform the iam:PassRole action, your policies must be updated to allow you to pass a role to Amazon Bedrock.

Some AWS services allow you to pass an existing role to that service instead of creating a new service role or service-linked role. To do this, you must have permissions to pass the role to the service.

The following example error occurs when an IAM user named marymajor tries to use the console to perform an action in Amazon Bedrock. However, the action requires the service to have permissions that are granted by a service role. Mary does not have permissions to pass the role to the service.

User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform: iam:PassRole

In this case, Mary's policies must be updated to allow her to perform the iam:PassRole action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

**I want to allow people outside of my AWS account to access my Amazon Bedrock resources**

You can create a role that users in other accounts or people outside of your organization can use to access your resources. You can specify who is trusted to assume the role. For services that support
resource-based policies or access control lists (ACLs), you can use those policies to grant people access to your resources.

To learn more, consult the following:

- To learn whether Amazon Bedrock supports these features, see How Amazon Bedrock works with IAM.
- To learn how to provide access to your resources across AWS accounts that you own, see Providing access to an IAM user in another AWS account that you own in the IAM User Guide.
- To learn how to provide access to your resources to third-party AWS accounts, see Providing access to AWS accounts owned by third parties in the IAM User Guide.
- To learn how to provide access through identity federation, see Providing access to externally authenticated users (identity federation) in the IAM User Guide.
- To learn the difference between using roles and resource-based policies for cross-account access, see How IAM roles differ from resource-based policies in the IAM User Guide.

Compliance validation for Amazon Bedrock

To learn whether an AWS service is within the scope of specific compliance programs, see AWS services in Scope by Compliance Program and choose the compliance program that you are interested in. For general information, see AWS Compliance Programs.

You can download third-party audit reports using AWS Artifact. For more information, see Downloading Reports in AWS Artifact.

Your compliance responsibility when using AWS services is determined by the sensitivity of your data, your company's compliance objectives, and applicable laws and regulations. AWS provides the following resources to help with compliance:

- Security and Compliance Quick Start Guides – These deployment guides discuss architectural considerations and provide steps for deploying baseline environments on AWS that are security and compliance focused.
- Architecting for HIPAA Security and Compliance on Amazon Web Services – This whitepaper describes how companies can use AWS to create HIPAA-eligible applications.
Note

Not all AWS services are HIPAA eligible. For more information, see the HIPAA Eligible Services Reference.

- **AWS Compliance Resources** – This collection of workbooks and guides might apply to your industry and location.

- **AWS Customer Compliance Guides** – Understand the shared responsibility model through the lens of compliance. The guides summarize the best practices for securing AWS services and map the guidance to security controls across multiple frameworks (including National Institute of Standards and Technology (NIST), Payment Card Industry Security Standards Council (PCI), and International Organization for Standardization (ISO)).

- **Evaluating Resources with Rules** in the *AWS Config Developer Guide* – The AWS Config service assesses how well your resource configurations comply with internal practices, industry guidelines, and regulations.

- **AWS Security Hub** – This AWS service provides a comprehensive view of your security state within AWS. Security Hub uses security controls to evaluate your AWS resources and to check your compliance against security industry standards and best practices. For a list of supported services and controls, see Security Hub controls reference.

- **AWS Audit Manager** – This AWS service helps you continuously audit your AWS usage to simplify how you manage risk and compliance with regulations and industry standards.

## Incident response in Amazon Bedrock

Security is the highest priority at AWS. As part of the AWS Cloud shared responsibility model, AWS manages a data center, network, and software architecture that meets the requirements of the most security-sensitive organizations. AWS is responsible for any incident response with respect to the Amazon Bedrock service itself. Also, as an AWS customer, you share a responsibility for maintaining security in the cloud. This means that you control the security you choose to implement from the AWS tools and features you have access to. In addition, you're responsible for incident response on your side of the shared responsibility model.

By establishing a security baseline that meets the objectives for your applications running in the cloud, you're able to detect deviations that you can respond to. To help you understand the impact
that incident response and your choices have on your corporate goals, we encourage you to review the following resources:

- AWS Security Incident Response Guide
- Security Perspective of the AWS Cloud Adoption Framework (CAF) whitepaper

Resilience in Amazon Bedrock

The AWS global infrastructure is built around AWS Regions and Availability Zones. AWS Regions provide multiple physically separated and isolated Availability Zones, which are connected with low-latency, high-throughput, and highly redundant networking. With Availability Zones, you can design and operate applications and databases that automatically fail over between zones without interruption. Availability Zones are more highly available, fault tolerant, and scalable than traditional single or multiple data center infrastructures.

For more information about AWS Regions and Availability Zones, see AWS Global Infrastructure.

Infrastructure security in Amazon Bedrock

As a managed service, Amazon Bedrock is protected by the AWS global network security. For information about AWS security services and how AWS protects infrastructure, see AWS Cloud Security. To design your AWS environment using the best practices for infrastructure security, see Infrastructure Protection in Security Pillar AWS Well-Architected Framework.

You use AWS published API calls to access Amazon Bedrock through the network. Clients must support the following:

- Transport Layer Security (TLS). We require TLS 1.2 and recommend TLS 1.3.
- Cipher suites with perfect forward secrecy (PFS) such as DHE (Ephemeral Diffie-Hellman) or ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). Most modern systems such as Java 7 and later support these modes.

Additionally, requests must be signed by using an access key ID and a secret access key that is associated with an IAM principal. Or you can use the AWS Security Token Service (AWS STS) to generate temporary security credentials to sign requests.
Cross-service confused deputy prevention

The confused deputy problem is a security issue where an entity that doesn't have permission to perform an action can coerce a more-privileged entity to perform the action. In AWS, cross-service impersonation can result in the confused deputy problem. Cross-service impersonation can occur when one service (the calling service) calls another service (the called service). The calling service can be manipulated to use its permissions to act on another customer's resources in a way it should not otherwise have permission to access. To prevent this, AWS provides tools that help you protect your data for all services with service principals that have been given access to resources in your account.

We recommend using the `aws:SourceArn` and `aws:SourceAccount` global condition context keys in resource policies to limit the permissions that Amazon Bedrock gives another service to the resource. Use `aws:SourceArn` if you want only one resource to be associated with the cross-service access. Use `aws:SourceAccount` if you want to allow any resource in that account to be associated with the cross-service use.

The most effective way to protect against the confused deputy problem is to use the `aws:SourceArn` global condition context key with the full ARN of the resource. If you don't know the full ARN of the resource or if you are specifying multiple resources, use the `aws:SourceArn` global context condition key with wildcard characters (*) for the unknown portions of the ARN. For example, `arn:aws:bedrock:*:123456789012:`.

If the `aws:SourceArn` value does not contain the account ID, such as an Amazon S3 bucket ARN, you must use both global condition context keys to limit permissions.

The value of `aws:SourceArn` must be `ResourceDescription`.

The following example shows how you can use the `aws:SourceArn` and `aws:SourceAccount` global condition context keys in Bedrock to prevent the confused deputy problem.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
        }
    ]
}
```
Configuration and vulnerability analysis in Amazon Bedrock

Configuration and IT controls are a shared responsibility between AWS and you, our customer. For more information, see the AWS shared responsibility model.

Use interface VPC endpoints (AWS PrivateLink)

You can use AWS PrivateLink to create a private connection between your VPC and Amazon Bedrock. You can access Amazon Bedrock as if it were in your VPC, without the use of an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Instances in your VPC don't need public IP addresses to access Amazon Bedrock.

You establish this private connection by creating an interface endpoint, powered by AWS PrivateLink. We create an endpoint network interface in each subnet that you enable for the interface endpoint. These are requester-managed network interfaces that serve as the entry point for traffic destined for Amazon Bedrock.

For more information, see Access AWS services through AWS PrivateLink in the AWS PrivateLink Guide.

Considerations for Amazon Bedrock VPC endpoints

Before you set up an interface endpoint for Amazon Bedrock, review Considerations in the AWS PrivateLink Guide.

Amazon Bedrock supports making the following API calls through VPC endpoints.
## Create an interface endpoint for Amazon Bedrock

You can create an interface endpoint for Amazon Bedrock using either the Amazon VPC console or the AWS Command Line Interface (AWS CLI). For more information, see [Create an interface endpoint](#) in the *AWS PrivateLink Guide*.

Create an interface endpoint for Amazon Bedrock using any of the following service names:

- com.amazonaws.
  region .bedrock
- com.amazonaws.
  region .bedrock-runtime
- com.amazonaws.
  region .bedrock-agent
- com.amazonaws.
  region .bedrock-agent-runtime

After you create the endpoint, you have the option to enable a private DNS hostname. Enable this setting by selecting Enable Private DNS Name in the VPC console when you create the VPC endpoint.

If you enable private DNS for the interface endpoint, you can make API requests to Amazon Bedrock using its default Regional DNS name. The following examples show the format of the default Regional DNS names.
Create an endpoint policy for your interface endpoint

An endpoint policy is an IAM resource that you can attach to an interface endpoint. The default endpoint policy allows full access to Amazon Bedrock through the interface endpoint. To control the access allowed to Amazon Bedrock from your VPC, attach a custom endpoint policy to the interface endpoint.

An endpoint policy specifies the following information:

- The principals that can perform actions (AWS accounts, IAM users, and IAM roles).
- The actions that can be performed.
- The resources on which the actions can be performed.

For more information, see Control access to services using endpoint policies in the AWS PrivateLink Guide.

Example: VPC endpoint policy for Amazon Bedrock actions

The following is an example of a custom endpoint policy. When you attach this resource-based policy to your interface endpoint, it grants access to the listed Amazon Bedrock actions for all principals on all resources.

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Principal": "*",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel",
                "bedrock:InvokeModelWithResponseStream"
            ],
            "Resource": "*"
        }
    ]
}
```
Monitor Amazon Bedrock

You can monitor Amazon Bedrock with Amazon CloudWatch and with Amazon EventBridge.

Topics

• Model invocation logging
• Monitor Amazon Bedrock with Amazon CloudWatch
• Monitor Amazon Bedrock events in Amazon EventBridge
• Log Amazon Bedrock API calls using AWS CloudTrail

Model invocation logging

Model invocation logging can be used to collect invocation logs, model input data, and model output data for all invocations in your AWS account used in Amazon Bedrock. By default, logging is disabled.

With invocation logging, you can collect the full request data, response data, and metadata associated with all calls performed in your account. Logging can be configured to provide the destination resources where the log data will be published. Supported destinations include Amazon CloudWatch Logs and Amazon Simple Storage Service (Amazon S3). Only destinations from the same account and region are supported.

Before you can enable invocation logging, you need to set up an Amazon S3 or CloudWatch Logs destination. You can enable invocation logging through either the console or the API.

Topics

• Set up an Amazon S3 destination
• Set up CloudWatch Logs destination
• Using the console
• Using APIs with invocation logging

Set up an Amazon S3 destination

You can set up an S3 destination for logging in Amazon Bedrock with these steps:
1. Create an S3 bucket where the logs will be delivered.

2. Add a bucket policy to it like the one below (Replace values for `accountId`, `region`, `bucketName`, and optionally `prefix`):

   ```json
   {
     "Version": "2012-10-17",
     "Statement": [
       {
         "Sid": "AmazonBedrockLogsWrite",
         "Effect": "Allow",
         "Principal": {
           "Service": "bedrock.amazonaws.com"
         },
         "Action": ["s3:PutObject"],
         "Resource": [
           "arn:aws:s3:::bucketName/prefix/AWSLogs/accountId/BedrockModelInvocationLogs/*"
         ],
         "Condition": {
           "StringEquals": {
             "aws:SourceAccount": "accountId"
           },
           "ArnLike": {
             "aws:SourceArn": "arn:aws:bedrock:region:accountId:*"
           }
         }
       }
     ]
   }
   ```

3. (Optional) If configuring SSE-KMS on the bucket, add the below policy on the KMS key:

   ```json
   ```
For more information on S3 SSE-KMS configurations, see [Specifying KMS Encryption](#).

**Note**
The bucket ACL must be disabled in order for the bucket policy to take effect. For more information, see [Disabling ACLs for all new buckets and enforcing Object Ownership](#).

**Set up CloudWatch Logs destination**

You can set up a Amazon CloudWatch Logs destination for logging in Amazon Bedrock with the following steps:

1. Create a CloudWatch log group where the logs will be published.
2. Create an IAM role with the following permissions for CloudWatch Logs.

**Trusted entity:**

```json
{
    "Version": "2012-10-17",
    "Statement": [
        {
```
Role policy:

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "logs:CreateLogStream",
            "logs:PutLogEvents"
         ],
         "Resource": "arn:aws:logs:"region:accountId":log-group:"logGroupName":log-stream:aws/bedrock/modelinvocations"
      }
   ]
}
```

For more information on setting up SSE for CloudWatch Logs, see Encrypt log data in CloudWatch Logs using AWS Key Management Service.
Using the console

To enable model invocation logging, drag the slider button next to the **Logging** toggle switch in the **Settings** page. Additional configuration settings for logging will appear on the panel.

Choose which data requests and responses you want to publish to the logs. You can choose any combination of the following output options:

- Text
- Image
- Embedding

Choose where to publish the logs:

- Amazon S3 only
- CloudWatch Logs only
- Both Amazon S3 and CloudWatch Logs

Amazon S3 and CloudWatch Logs destinations are supported for invocation logs, and small input and output data. For large input and output data or binary image outputs, only Amazon S3 is supported. The following details summarize how the data will be represented in the target location.

**S3 destination** — Gzipped JSON files, each containing a batch of invocation log records, are delivered to the specified S3 bucket. Similar to a CloudWatch Logs event, each record will contain the invocation metadata, and input and output JSON bodies of up to 100 KB in size. Binary data or JSON bodies larger than 100 KB will be uploaded as individual objects in the specified Amazon S3 bucket under the data prefix. The data can be queried using Amazon S3 Select and Amazon Athena, and can be catalogued for ETL using AWS Glue. The data can be loaded into OpenSearch service, or be processed by any Amazon EventBridge targets.

**CloudWatch Logs destination** — JSON invocation log events are delivered to a specified log group in CloudWatch Logs. The log event contains the invocation metadata, and input and output JSON bodies of up to 100 KB in size. If an Amazon S3 location for large data delivery is provided, binary data or JSON bodies larger than 100 KB will be uploaded to the Amazon S3 bucket under the data prefix instead. Data can be queried using CloudWatch Logs Insights, and can be further streamed to various services in real-time using CloudWatch Logs.
Using APIs with invocation logging

Model invocation logging can be configured using the following APIs:

- PutModelInvocationLoggingConfiguration
- GetModelInvocationLoggingConfiguration
- DeleteModelInvocationLoggingConfiguration

For more information on how to use APIs with invocation logging, see the Bedrock API Guide.

Monitor Amazon Bedrock with Amazon CloudWatch

You can monitor Amazon Bedrock using Amazon CloudWatch, which collects raw data and processes it into readable, near real-time metrics. You can graph the metrics using the CloudWatch console. You can also set alarms that watch for certain thresholds, and send notifications or take actions when values exceed those thresholds.

For more information, see [What is Amazon CloudWatch](https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/what-cloudwatch.html) in the *Amazon CloudWatch User Guide*.

Topics

- [Runtime metrics](#)
- [Logging CloudWatch metrics](#)
- [Use CloudWatch metrics for Amazon Bedrock](#)
- [View Amazon Bedrock metrics](#)

Runtime metrics

The following table describes runtime metrics provided by Amazon Bedrock.

<table>
<thead>
<tr>
<th>Metric name</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invocations</td>
<td>SampleCount</td>
<td>Number of requests to the <code>InvokeModel</code> or <code>InvokeModelWithResponseStream</code> API operations.</td>
</tr>
<tr>
<td>Metric name</td>
<td>Unit</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>InvocationLatency</td>
<td>MilliSeconds</td>
<td>Latency of the invocations.</td>
</tr>
<tr>
<td>InvocationClientErrors</td>
<td>SampleCount</td>
<td>Number of invocations that result in client-side errors.</td>
</tr>
<tr>
<td>InvocationServerErrors</td>
<td>SampleCount</td>
<td>Number of invocations that result in AWS server-side errors.</td>
</tr>
<tr>
<td>InvocationThrottles</td>
<td>SampleCount</td>
<td>Number of invocations that the system throttled.</td>
</tr>
<tr>
<td>InputTokenCount</td>
<td>SampleCount</td>
<td>Number of tokens of text input.</td>
</tr>
<tr>
<td>LegacyModelInvocations</td>
<td>SampleCount</td>
<td>Number of invocations using Legacy models</td>
</tr>
<tr>
<td>OutputTokenCount</td>
<td>SampleCount</td>
<td>Number of tokens of text output.</td>
</tr>
<tr>
<td>OutputImageCount</td>
<td>SampleCount</td>
<td>Number of output images.</td>
</tr>
</tbody>
</table>

Logging CloudWatch metrics

For each delivery success or failure attempt, the following Amazon CloudWatch metrics are emitted under the namespace AWS/Bedrock, and Across all model IDs dimension:

- ModelInvocationLogsCloudWatchDeliverySuccess
- ModelInvocationLogsCloudWatchDeliveryFailure
- ModelInvocationLogsS3DeliverySuccess
- ModelInvocationLogsS3DeliveryFailure
- ModelInvocationLargeDataS3DeliverySuccess
- ModelInvocationLargeDataS3DeliveryFailure
If logs fail to deliver due to permission misconfiguration or transient failures, the delivery is retried periodically for up to 24 hours.

**Use CloudWatch metrics for Amazon Bedrock**

To retrieve metrics for your Amazon Bedrock operations, you specify the following information:

- The metric dimension. A *dimension* is a set of name-value pairs that you use to identify a metric. Amazon Bedrock supports the following dimensions:
  - ModelId – all metrics
  - ModelId + ImageSize + BucketedStepSize – OutputImageCount
  - The metric name, such as InvocationClientErrors.

You can get metrics for Amazon Bedrock with the AWS Management Console, the AWS CLI, or the CloudWatch API. You can use the CloudWatch API through one of the AWS Software Development Kits (SDKs) or the CloudWatch API tools.

You must have the appropriate CloudWatch permissions to monitor Amazon Bedrock with CloudWatch. For more information, see Authentication and Access Control for Amazon CloudWatch in the Amazon CloudWatch User Guide.

**View Amazon Bedrock metrics**

View Amazon Bedrock metrics in the CloudWatch console.

**To view metrics (CloudWatch console)**

2. Choose **Metrics**, choose **All Metrics**, and then search for **ModelId**.

**Monitor Amazon Bedrock events in Amazon EventBridge**

You can use Amazon EventBridge to monitor status change events in Amazon Bedrock. With Amazon EventBridge, you can configure Amazon SageMaker to respond automatically to a model customization job status change in Amazon Bedrock. Events from Amazon Bedrock are delivered to Amazon EventBridge in near real time. You can write simple rules to automate actions when an event matches a rule. If you use Amazon EventBridge with Amazon Bedrock, you can:
• Publish notifications whenever there is a state change event in the model customization you have triggered, whether you add new asynchronous workflows in the future. The event published should give you enough information to react to events in downstream workflows.

• Deliver job status updates without invoking the GetModelCustomizationJob API, which can mean handling API rate limit issues, API updates, and reduction in additional compute resources.

There is no cost to receive AWS events from Amazon EventBridge. For more information about, Amazon EventBridge, see Amazon EventBridge

---

**Note**

• Amazon Bedrock emits events on a best-effort basis. Events are delivered to Amazon EventBridge in near real time. With Amazon EventBridge, you can create rules that trigger programmatic actions in response to an event. For example, you can configure a rule that invokes an SNS topic to send an email notification or invokes a function to take some action. For more information, see the Amazon EventBridge User Guide.

• Amazon Bedrock creates a new event every time there is a state change in a model customization job that you trigger and make best-effort delivery of such event.

---

**Topics**

• [How it works](#)

• [EventBridge schema](#)

• [Rules and targets](#)

• [Create a rule to handle Amazon Bedrock events](#)

---

**How it works**

To receive events from Amazon Bedrock, you need to create rules and targets to match, receive, and handle state change data through Amazon EventBridge. Amazon EventBridge is a serverless event bus that ingests change state events from AWS services, SaaS partners, and customer applications. It processes events based on rules or patterns that you create, and routes these events to one or more “targets” that you choose, such as AWS Lambda, Amazon Simple Queue Service, and Amazon Simple Notification Service.
Amazon Bedrock publishes your events via Amazon EventBridge whenever there is a change in the state of a model customization job. In each case, a new event is created and sent to Amazon EventBridge, which then sends the event to your default event-bus. The event shows which customization job’s state has changed, and the current state of the job. When Amazon EventBridge receives an event that matches a rule that you created, Amazon EventBridge routes it to the target that you specified. When you create a rule, you can configure these targets as well as downstream workflows based on the contents of the event.

**EventBridge schema**

The following event fields in the EventBridge event schema are specific to Amazon Bedrock.

- **jobArn** — The ARN of the model customization job.
- **outputModelArn** — The ARN of the output model. Published when the training job has completed.
- **jobStatus** — The current status of the job.
- **FailureMessage** — A failure message. Published when the training job has failed.

**Event example**

The following is example event JSON for a failed model customization job.

```json
{
    "version": "0",
    "id": "UUID",
    "detail-type": "Model Customization Job State Change",
    "source": "aws.bedrock",
    "account": "123412341234",
    "time": "2023-08-11T12:34:56Z",
    "region": "us-east-1",
    "resources": [ "arn:aws:bedrock:us-east-1:123412341234:model-customization-job/abcdefghwxyz" ],
    "detail": {
        "version": "0.0",
        "jobName": "abcd-wxyz",
        "outputModelName": "dummy-output-model-name",
        "outputModelArn": "arn:aws:bedrock:us-east-1:123412341234:dummy-output-model-name",
    }
}
```
Rules and targets

When an incoming event matches a rule that you created, the event is routed to the target that you specified for that rule, and the target processes these events. Targets support JSON format and can include AWS services such as Amazon EC2 instances, Lambda functions, Kinesis streams, Amazon ECS tasks, Step Functions, Amazon SNS topics, and Amazon SQS. To receive and process events correctly, you need to create rules and targets for matching, receiving, and correctly handling event data. You can create these rules and targets either through the Amazon EventBridge console, or through the AWS CLI.

Example rule

This rule matches an event pattern emitted by: source [“aws.bedrock”]. The rule captures all events sent by Amazon EventBridge that have source “aws.bedrock” to your default event bus.
When creating a rule in Amazon EventBridge, you need to specify a target where EventBridge sends the event that matches your rule pattern. These targets can be a SageMaker pipeline, a Lambda function, an SNS topic, an SQS queue or any of the other targets that EventBridge currently supports. You can refer to the Amazon EventBridge documentation to learn how to set targets for events. For a procedure that shows how to use Amazon Simple Notification Service as a target, see Create a rule to handle Amazon Bedrock events.

Create a rule to handle Amazon Bedrock events

Complete the following procedures in order to receive email notifications about your Amazon Bedrock events.

Create an Amazon Simple Notification Service topic

2. In the navigation pane, choose Topics.
3. Choose Create topic.
4. For Type, choose Standard.
5. For Name, enter a name for your topic.
6. Choose Create topic.
7. Choose Create subscription.
8. For Protocol, choose Email.
9. For Endpoint, enter the email address that receives the notifications.
10. Choose Create subscription.
11. You’ll receive an email message with the following subject line: AWS Notification - Subscription Confirmation. Follow the directions to confirm your subscription.

Use the following procedure to create a rule to handle your Amazon Bedrock events.

To create a rule to handle Amazon Bedrock events

1. Open the Amazon EventBridge console at https://console.aws.amazon.com/events/.
2. Choose **Create rule**.

3. For **Name**, enter a name for your rule.

4. For **Rule type**, choose **Rule with an event pattern**.

5. Choose **Next**.

6. For **Event pattern**, do the following:
   
   a. For **Event source**, choose **AWS services**.
   
   b. For **AWS service**, choose **Amazon Bedrock**.
   
   c. For **Event type**, choose **Model Customization Job State Change**.
   
   d. By default, we send notifications for every event. If you prefer, you can create an event pattern that filters events for a specific job state.
   
   e. Choose **Next**.

7. Specify a target as follows:
   
   a. For **Target types**, choose **AWS service**.
   
   b. For **Select a target**, choose **SNS topic**.
   
   c. For **Topic**, choose the SNS topic that you created for notifications.
   
   d. Choose **Next**.

8. (Optional) Add tags to your rule.

9. Choose **Next**.

10. Choose **Create rule**.

**Log Amazon Bedrock API calls using AWS CloudTrail**

Amazon Bedrock is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in Amazon Bedrock. CloudTrail captures all API calls for Amazon Bedrock as events. The calls captured include calls from the Amazon Bedrock console and code calls to the Amazon Bedrock API operations. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon S3 bucket, including events for Amazon Bedrock. If you don't configure a trail, you can still view the most recent events in the CloudTrail console in **Event history**. Using the information collected by CloudTrail, you can determine the request that was made to Amazon Bedrock, the IP address from which the request was made, who made the request, when it was made, and additional details.
To learn more about CloudTrail, see the AWS CloudTrail User Guide.

Amazon Bedrock information in CloudTrail

CloudTrail is enabled on your AWS account when you create the account. When activity occurs in Amazon Bedrock, that activity is recorded in a CloudTrail event along with other AWS service events in Event history. You can view, search, and download recent events in your AWS account. For more information, see Viewing events with CloudTrail Event history.

For an ongoing record of events in your AWS account, including events for Amazon Bedrock, create a trail. A trail enables CloudTrail to deliver log files to an Amazon S3 bucket. By default, when you create a trail in the console, the trail applies to all AWS Regions. The trail logs events from all Regions in the AWS partition and delivers the log files to the Amazon S3 bucket that you specify. Additionally, you can configure other AWS services to further analyze and act upon the event data collected in CloudTrail logs. For more information, see the following:

- Overview for creating a trail
- CloudTrail supported services and integrations
- Configuring Amazon SNS notifications for CloudTrail
- Receiving CloudTrail log files from multiple regions and Receiving CloudTrail log files from multiple accounts

Every event or log entry contains information about who generated the request. The identity information helps you determine the following:

- Whether the request was made with root or AWS Identity and Access Management (IAM) user credentials.
- Whether the request was made with temporary security credentials for a role or federated user.
- Whether the request was made by another AWS service.

For more information, see the CloudTrail userIdentity element.

Amazon Bedrock data events in CloudTrail

Data events provide information about the resource operations performed on or in a resource (for example, reading or writing to an Amazon S3 object). These are also known as data plane operations. Data events are often high-volume activities that CloudTrail doesn’t log by default.
Amazon Bedrock doesn't log Amazon Bedrock Runtime API operations (InvokeModel and InvokeModelWithResponseStream).

Amazon Bedrock logs all Agents for Amazon Bedrock Runtime API operations actions to CloudTrail as data events.

- To log InvokeAgent calls, configure advanced event selectors to record data events for the AWS::Bedrock::AgentAlias resource type.
- To log Retrieve and RetrieveAndGenerate calls, configure advanced event selectors to record data events for the AWS::Bedrock::KnowledgeBase resource type.

From the CloudTrail console, choose Bedrock agent alias or Bedrock knowledge base for the Data event type. You can additionally filter on the eventName and resources.ARN fields by choosing a custom log selector template. For more information, see Logging data events with the AWS Management Console.

From the AWS CLI, set the resource.type value equal to AWS::Bedrock::AgentAlias or AWS::Bedrock::KnowledgeBase and set the eventCategory equal to Data. For more information, see Logging data events with the AWS CLI.

The following example shows how to configure a trail to log all Amazon Bedrock data events for all Amazon Bedrock resource types in the AWS CLI.

```
aws cloudtrail put-event-selectors --trail-name trailName \
--advanced-event-selectors \
'[
    {
        "Name": "Log all data events on an Agents for Amazon Bedrock agent alias",
        "FieldSelectors": [
            { "Field": "eventCategory", "Equals": ["Data"] },
            { "Field": "resources.type", "Equals": ["AWS::Bedrock::AgentAlias"] }
        ]
    },
    {
        "Name": "Log all data events on an Agents for Amazon Bedrock knowledge base",
        "FieldSelectors": [
            { "Field": "eventCategory", "Equals": ["Data"] },
            { "Field": "resources.type", "Equals": ["AWS::Bedrock::KnowledgeBase"] }
        ]
    }
]
```
You can additionally filter on the `eventName` and `resources.ARN` fields. For more information about these fields, see [AdvancedFieldSelector](#).

Additional charges apply for data events. For more information about CloudTrail pricing, see [AWS CloudTrail Pricing](#).

## Amazon Bedrock management events in CloudTrail

[Management events](#) provide information about management operations that are performed on resources in your AWS account. These are also known as control plane operations. CloudTrail logs management event API operations by default.

Amazon Bedrock logs the remainder of Amazon Bedrock API operations as management events. For a list of the Amazon Bedrock API operations that Amazon Bedrock logs to CloudTrail, see the following pages in the Amazon Bedrock API reference.

All Amazon Bedrock API operations and Agents for Amazon Bedrock API operations are logged by CloudTrail and documented in the [Amazon Bedrock API Reference](#). For example, calls to the `InvokeModel`, `StopModelCustomizationJob`, and `CreateAgent` actions generate entries in the CloudTrail log files.

### Understanding Amazon Bedrock log file entries

A trail is a configuration that enables delivery of events as log files to an Amazon S3 bucket that you specify. CloudTrail log files contain one or more log entries. An event represents a single request from any source and includes information about the requested action, the date and time of the action, request parameters, and so on. CloudTrail log files aren't an ordered stack trace of the public API calls, so they don't appear in any specific order.

The following example shows a CloudTrail log entry that demonstrates the `InvokeModel` action.

```json
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "IAMUser",
        "principalId": "AROAICFHPEXAMPLE",
        "arn": "arn:aws:iam::111122223333:user/userxyz",
        "accountId": "111122223333",
        "aws Account Id": "111122223333",
        "requestID": "34567890abcdef",
        "invocationTime": "2023-03-15T12:00:00Z"
    },
    "eventSource": "com.amazon.bedrock",
    "eventTime": "2023-03-15T12:00:00Z",
    "requestParameters": {
        "query": "query parameter 1",
        "url": "url parameter 1"
    },
    "responseElements": {
        "status": "200",
        "reason": "OK"
    },
    "eventID": "34567890abcdef",
    "readOnly": false
}
```
"accessKeyId": "AKIAIOSFODNN7EXAMPLE",
  "userName": "userxyz"
},
"eventTime": "2023-10-11T21:58:59Z",
"eventSource": "bedrock.amazonaws.com",
"eventName": "InvokeModel",
"awsRegion": "us-west-2",
"sourceIPAddress": "192.0.2.0",
"userAgent": "Boto3/1.28.62 md/Botocore#1.31.62 ua/2.0 os/macos#22.6.0 md/
arch#arm64 lang/python#3.9.6 md/pyimpl#CPython cfg/retry-mode#legacy Botocore/1.31.62",
  "requestParameters": {
    "modelId": "stability.stable-diffusion-xl-v0"
  },
"responseElements": null,
"requestID": "a1b2c3d4-5678-90ab-cdef-EXAMPLE22222",
"eventID": "a1b2c3d4-5678-90ab-cdef-EXAMPLE11111 ",
"readOnly": false,
"eventType": "AwsApiCall",
"managementEvent": true,
"recipientAccountId": "111122223333",
"eventCategory": "Management",
"tlsDetails": {
  "tlsVersion": "TLSv1.2",
  "cipherSuite": "cipher suite",
  "clientProvidedHostHeader": "bedrock-runtime.us-west-2.amazonaws.com"
}
Code examples for Amazon Bedrock using AWS SDKs

The following code examples show how to use Amazon Bedrock with an AWS software development kit (SDK).

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Code examples**

- Code examples for Amazon Bedrock using AWS SDKs
  - Actions for Amazon Bedrock using AWS SDKs
    - Get details about an Amazon Bedrock foundation model using an AWS SDK
    - List the available Amazon Bedrock foundation models using an AWS SDK
  - Scenarios for Amazon Bedrock using AWS SDKs
    - Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions
- Code examples for Amazon Bedrock Runtime using AWS SDKs
  - Actions for Amazon Bedrock Runtime using AWS SDKs
    - Invoke the Amazon Titan Image Generator G1 model on Amazon Bedrock for image generation
    - Invoke the Stability.ai Stable Diffusion XL model on Amazon Bedrock for image generation
    - Invoke Anthropic Claude 3 on Amazon Bedrock with a multimodal prompt
    - Invoke the AI21 Labs Jurassic-2 model on Amazon Bedrock for text generation
    - Invoke the Amazon Titan Text G1 model on Amazon Bedrock for text generation
    - Invoke the Anthropic Claude 2 model on Amazon Bedrock for text generation
    - Invoke Anthropic Claude 3 on Amazon Bedrock to generate text
    - Invoke the Anthropic Claude Instant model on Amazon Bedrock for text generation
    - Invoke Anthropic Claude on Amazon Bedrock for text generation with a response stream
    - Invoke the Meta Llama 2 Chat model on Amazon Bedrock for text generation
    - Invoke the Mistral 7B model on Amazon Bedrock for text generation
    - Invoke the Mixtral 8x7B model on Amazon Bedrock for text generation
  - Scenarios for Amazon Bedrock Runtime using AWS SDKs
• Create a sample application that offers playgrounds to interact with Amazon Bedrock foundation models using an AWS SDK
• An interactive text generation playground for Amazon Bedrock
• Invoke multiple large-language models (LLMs) on Amazon Bedrock
• Invoke multiple foundation models on Amazon Bedrock
• Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

• Code examples for Agents for Amazon Bedrock using AWS SDKs
  • Actions for Agents for Amazon Bedrock using AWS SDKs
    • Create an Amazon Bedrock agent using an AWS SDK
    • Create an Amazon Bedrock agent action group using an AWS SDK
    • Create an Amazon Bedrock agent alias using an AWS SDK
    • Delete an Amazon Bedrock agent using an AWS SDK
    • Delete an Amazon Bedrock agent alias using an AWS SDK
    • Get information about an Amazon Bedrock agent using an AWS SDK
    • List the action groups for a Amazon Bedrock agent using an AWS SDK
    • List the Agents for Amazon Bedrock belonging to an account using an AWS SDK
    • List the knowledge bases associated with a Amazon Bedrock agent using an AWS SDK
    • Prepare an Amazon Bedrock agent using an AWS SDK
  • Scenarios for Agents for Amazon Bedrock using AWS SDKs
    • An end-to-end example showing how to create and invoke Amazon Bedrock agents using an AWS SDK
    • Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions
• Code examples for Agents for Amazon Bedrock Runtime using AWS SDKs
  • Actions for Agents for Amazon Bedrock Runtime using AWS SDKs
    • Invoke an Amazon Bedrock agent using an AWS SDK
  • Scenarios for Agents for Amazon Bedrock Runtime using AWS SDKs
    • Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions
Code examples for Amazon Bedrock using AWS SDKs

The following code examples show how to use Amazon Bedrock with an AWS software development kit (SDK).

*Actions* are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios and cross-service examples.

*Scenarios* are code examples that show you how to accomplish a specific task by calling multiple functions within the same service.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Get started**

**Hello Amazon Bedrock**

The following code examples show how to get started using Amazon Bedrock.

**.NET**

**AWS SDK for .NET**

```
using Amazon;
using Amazon.Bedrock;
using Amazon.Bedrock.Model;
	namespace ListFoundationModelsExample

{
    /// <summary>
    /// This example shows how to list foundation models.
    /// </summary>
    internal class HelloBedrock
```
/// <summary>
/// Main method to call the ListFoundationModelsAsync method.
/// </summary>
/// <param name="args"> The command line arguments. </param>
static async Task Main(string[] args)
{
    // Specify a region endpoint where Amazon Bedrock is available.
    // For a list of supported region see https://docs.aws.amazon.com/bedrock/latest/
    // userguide/what-is-bedrock.html#bedrock-regions
    AmazonBedrockClient bedrockClient = new(RegionEndpoint.USWest2);

    await ListFoundationModelsAsync(bedrockClient);
}

/// <summary>
/// List foundation models.
/// </summary>
/// <param name="bedrockClient"> The Amazon Bedrock client. </param>
private static async Task ListFoundationModelsAsync(AmazonBedrockClient bedrockClient)
{
    Console.WriteLine("List foundation models with no filter");

    try
    {
        ListFoundationModelsResponse response = await bedrockClient.ListFoundationModelsAsync(new ListFoundationModelsRequest()
        {
        
        
        
        });

        {
            foreach (var fm in response.ModelSummaries)
            {
                WriteToConsole(fm);
            }
        }
        else
        {
            Console.WriteLine("Something wrong happened");
        }
    }
```csharp
    catch (AmazonBedrockException e)
    {
        Console.WriteLine(e.Message);
    }
    }

    /// <summary>
    /// Write the foundation model summary to console.
    /// </summary>
    /// <param name="foundationModel"> The foundation model summary to write to console. </param>
    private static void WriteToConsole(FoundationModelSummary foundationModel)
    {
    }
    }

    // For API details, see ListFoundationModels in AWS SDK for .NET API Reference.

Go

SDK for Go V2

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
const region = "us-east-1"

// main uses the AWS SDK for Go (v2) to create an Amazon Bedrock client and
// list the available foundation models in your account and the chosen region.
// This example uses the default settings specified in your shared credentials
// and config files.
func main() {
    sdkConfig, err := config.LoadDefaultConfig(context.TODO(),
        config.WithRegion(region))
    if err != nil {
        fmt.Println("Couldn't load default configuration. Have you set up your
AWS account?")
        fmt.Println(err)
        return
    }
    bedrockClient := bedrock.NewFromConfig(sdkConfig)
    result, err := bedrockClient.ListFoundationModels(context.TODO(),
        &bedrock.ListFoundationModelsInput{})
    if err != nil {
        fmt.Printf("Couldn't list foundation models. Here's why: %v\n", err)
        return
    }
    if len(result.ModelSummaries) == 0 {
        fmt.Println("There are no foundation models.")
    }
    for _, modelSummary := range result.ModelSummaries {
        fmt.Println(*modelSummary.ModelId)
    }
}

• For API details, see ListFoundationModels in AWS SDK for Go API Reference.
JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import {
    BedrockClient,
    ListFoundationModelsCommand,
} from "@aws-sdk/client-bedrock";

const REGION = "us-east-1";
const client = new BedrockClient({ region: REGION });

export const main = async () => {
    const command = new ListFoundationModelsCommand({});

    const response = await client.send(command);
    const models = response.modelSummaries;

    console.log("Listing the available Bedrock foundation models:");

    for (let model of models) {
        console.log("=".repeat(42));
        console.log(` Model: ${model.modelId}`);
        console.log("-".repeat(42));
        console.log(` Name: ${model.modelName}`);
        console.log(` Provider: ${model.providerName}`);
        console.log(` Model ARN: ${model.modelArn}`);
        console.log(` Input modalities: ${model.inputModalities}`);
        console.log(` Output modalities: ${model.outputModalities}`);
        console.log(` Supported customizations: ${model.customizationsSupported}`);
        console.log(` Supported inference types: ${model.inferenceTypesSupported}`);
    }
```

```javascript
const active = models.filter(
  (m) => m.modelLifecycle.status === "ACTIVE",
).length;
const legacy = models.filter(
  (m) => m.modelLifecycle.status === "LEGACY",
).length;

console.log(`
  `There are ${active} active and ${legacy} legacy foundation models in
 ${REGION}.`,
);

return response;
```

// If this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  await main();
}
```

- For API details, see [ListFoundationModels](https://aws.amazon.com/sdk-for-javascript/api-reference/) in *AWS SDK for JavaScript API Reference*.

### Code examples

- **Actions for Amazon Bedrock using AWS SDKs**
  - Get details about an Amazon Bedrock foundation model using an AWS SDK
  - List the available Amazon Bedrock foundation models using an AWS SDK
- **Scenarios for Amazon Bedrock using AWS SDKs**
  - Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

### Actions for Amazon Bedrock using AWS SDKs

The following code examples demonstrate how to perform individual Amazon Bedrock actions with AWS SDKs. These excerpts call the Amazon Bedrock API and are code excerpts from larger
programs that must be run in context. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the Amazon Bedrock API Reference.

Examples

- Get details about an Amazon Bedrock foundation model using an AWS SDK
- List the available Amazon Bedrock foundation models using an AWS SDK

Get details about an Amazon Bedrock foundation model using an AWS SDK

The following code examples show how to get details about an Amazon Bedrock foundation model.

Java

SDK for Java 2.x

```java
/**
 * Get details about an Amazon Bedrock foundation model.
 *
 * @param bedrockClient   The service client for accessing Amazon Bedrock.
 * @param modelIdentifier The model identifier.
 * @return An object containing the foundation model's details.
 */
public static FoundationModelDetails getFoundationModel(BedrockClient bedrockClient, String modelIdentifier) {
    try {
        GetFoundationModelResponse response = bedrockClient.getFoundationModel(
            r -> r.modelIdentifier(modelIdentifier)
        );
        return response/FoundationModelDetails
    } catch (Exception e) {
        // Handle exception
    }
}
```

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
Get details about a foundation model using the asynchronous Amazon Bedrock client.

```java
/**
 * Get details about an Amazon Bedrock foundation model.
 *
 * @param bedrockClient The async service client for accessing Amazon Bedrock.
 */
```
public static FoundationModelDetails getFoundationModel(BedrockAsyncClient bedrockClient, String modelIdentifier) {
    try {
        CompletableFuture<GetFoundationModelResponse> future =
            bedrockClient.getFoundationModel(
                r -> r.modelIdentifier(modelIdentifier)
            );

        FoundationModelDetails model = future.get().modelDetails();

        System.out.println(" Model ID:", model.modelId());
        System.out.println(" Model ARN:", model.modelArn());
        System.out.println(" Model Name:", model.modelName());
        System.out.println(" Provider Name:", model.providerName());
        System.out.println(" Lifecycle status:", model.modelLifecycle().statusAsString());
        System.out.println(" Input modalities:", model.inputModalities());
        System.out.println(" Output modalities:", model.outputModalities());
        System.out.println(" Supported customizations:", model.customizationsSupported());
        System.out.println(" Supported inference types:", model.inferenceTypesSupported());
        System.out.println(" Response streaming supported:", model.responseStreamingSupported());

        return model;
    } catch (ExecutionException e) {
        if (e.getMessage().contains("ValidationException")) {
            throw new IllegalArgumentException(e.getMessage());
        } else {
            System.err.println(e.getMessage());
            throw new RuntimeException(e);
        }
    }
}
}
Thread.currentThread().interrupt();
System.err.println(e.getMessage());
throw new RuntimeException(e);)
}
}

• For API details, see GetFoundationModel in AWS SDK for Java 2.x API Reference.

JavaScript

SDK for JavaScript (v3)

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Get details about a foundation model.

// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import {
  BedrockClient,
  GetFoundationModelCommand,
} from "@aws-sdk/client-bedrock";

/**
 * Get details about an Amazon Bedrock foundation model.
 *
 * @return {FoundationModelDetails} - The list of available bedrock foundation models.
 */
export const getFoundationModel = async () => {
  const client = new BedrockClient();

  const command = new GetFoundationModelCommand({
    modelIdentifier: "amazon.titan-embed-text-v1",
  }));

  try {
    const response = await client.send(command);
    return response;
• For API details, see `GetFoundationModel` in *AWS SDK for JavaScript API Reference*.

Python

**SDK for Python (Boto3)**

<table>
<thead>
<tr>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>There's more on GitHub. Find the complete example and learn how to set up and run in the <a href="https://github.com/aws-samples/amazon-bedrock-python-sdk">AWS Code Examples Repository</a>.</td>
</tr>
</tbody>
</table>

Get details about a foundation model.

```python
def get_foundation_model(self, model_identifier):
    """
    Get details about an Amazon Bedrock foundation model.
    
    :return: The foundation model's details.
    """

    try:
        return self.bedrock_client.get_foundation_model(
            modelIdentifier=model_identifier
        )['modelDetails']
    except ClientError:
        logger.error(
            f'Couldn't get foundation models details for {model_identifier}'
        )
```
• For API details, see `GetFoundationModel` in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see *Using this service with an AWS SDK*. This topic also includes information about getting started and details about previous SDK versions.

**List the available Amazon Bedrock foundation models using an AWS SDK**

The following code examples show how to list available Amazon Bedrock foundation models.

**.NET**

**AWS SDK for .NET**

---

---

```csharp
/// <summary>
/// List foundation models.
/// </summary>
/// <param name="bedrockClient"> The Amazon Bedrock client. </param>
private static async Task ListFoundationModelsAsync(AmazonBedrockClient bedrockClient)
{
    Console.WriteLine("List foundation models with no filter");

    try
    {
        ListFoundationModelsResponse response = await bedrockClient.ListFoundationModelsAsync(new ListFoundationModelsRequest);
```
```
{
}

{
    foreach (var fm in response.ModelSummaries)
    {
        WriteToConsole(fm);
    }
}
else
{
    Console.WriteLine("Something wrong happened");
}
```

For API details, see ListFoundationModels in AWS SDK for .NET API Reference.

Go

**SDK for Go V2**

- For API details, see ListFoundationModels in AWS SDK for .NET API Reference.

List the available Bedrock foundation models.

```go
// FoundationModelWrapper encapsulates Amazon Bedrock actions used in the examples.
// It contains a Bedrock service client that is used to perform foundation model actions.
```
type FoundationModelWrapper struct {
    BedrockClient *bedrock.Client
}

// ListPolicies lists Bedrock foundation models that you can use.
func (wrapper FoundationModelWrapper) ListFoundationModels() ([]types.FoundationModelSummary, error) {
    var models []types.FoundationModelSummary

    if err != nil {
        log.Printf("Couldn't list foundation models. Here's why: %v\n", err)
    } else {
        models = result.ModelSummaries
    }
    return models, err
}

- For API details, see ListFoundationModels in AWS SDK for Go API Reference.

Java

SDK for Java 2.x

- Note
  There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

List the available Amazon Bedrock foundation models using the synchronous Amazon Bedrock client.
List the available Amazon Bedrock foundation models using the asynchronous Amazon Bedrock client.

```java
public static List<FoundationModelSummary> listFoundationModels(BedrockClient bedrockClient) {
    try {
        ListFoundationModelsResponse response =
            bedrockClient.listFoundationModels(r -> {});

        List<FoundationModelSummary> models = response.modelSummaries();

        if (models.isEmpty()) {
            System.out.println("No available foundation models in " +
                region.toString());
        } else {
            for (FoundationModelSummary model : models) {
                System.out.println("Model ID: " + model.modelId());
                System.out.println("Provider: " + model.providerName());
                System.out.println("Name: " + model.modelName());
            }
        }

        return models;
    } catch (SdkClientException e) {
        System.err.println(e.getMessage());
        throw new RuntimeException(e);
    }
}
```

/*
 * Lists Amazon Bedrock foundation models that you can use.
 * You can filter the results with the request parameters.
 * @param bedrockClient The service client for accessing Amazon Bedrock.
 * @return A list of objects containing the foundation models' details
 */
* @param bedrockClient The async service client for accessing Amazon Bedrock.
* @return A list of objects containing the foundation models' details
* /
    public static List<FoundationModelSummary>
    listFoundationModels(BedrockAsyncClient bedrockClient) {
      try {
        CompletableFuture<ListFoundationModelsResponse> future =
        bedrockClient.listFoundationModels(r -> {});

        List<FoundationModelSummary> models = future.get().modelSummaries();

        if (models.isEmpty()) {
          System.out.println("No available foundation models in " +
          region.toString());
        } else {
          for (FoundationModelSummary model : models) {
            System.out.println("Model ID: " + model.modelId());
            System.out.println("Provider: " + model.providerName());
            System.out.println("Name: " + model.modelName());
            System.out.println();
          }
        }

        return models;
      } catch (InterruptedException e) {
        Thread.currentThread().interrupt();
        System.err.println(e.getMessage());
        throw new RuntimeException(e);
      } catch (ExecutionException e) {
        System.err.println(e.getMessage());
        throw new RuntimeException(e);
      }
    }

  
  • For API details, see ListFoundationModels in AWS SDK for Java 2.x API Reference.
List the available foundation models.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import {
    BedrockClient,
    ListFoundationModelsCommand,
} from "@aws-sdk/client-bedrock";

/**
 * List the available Amazon Bedrock foundation models.
 *
 * @return {FoundationModelSummary[]} - The list of available bedrock foundation models.
 */
export const listFoundationModels = async () => {
    const client = new BedrockClient();

    const input = {
        // byProvider: 'STRING_VALUE',
        // byCustomizationType: 'FINE_TUNING' || 'CONTINUED_PRE_TRAINING',
        // byOutputModality: 'TEXT' || 'IMAGE' || 'EMBEDDING',
        // byInferenceType: 'ON_DEMAND' || 'PROVISIONED',
    };

    const command = new ListFoundationModelsCommand(input);

    const response = await client.send(command);

    return response.modelSummaries;
```
For API details, see [ListFoundationModels](#) in *AWS SDK for JavaScript API Reference*.

**Kotlin**

**SDK for Kotlin**

```kotlin
suspend fun listFoundationModels(): List<FoundationModelSummary>? {
    try {
        val response = BedrockClient { region = "us-east-1" }.use { bedrockClient ->
            bedrockClient.listFoundationModels(ListFoundationModelsRequest {})
        }?.forEach { model ->
            println("Model ID: ${model.modelId}")
            println("Model Name: ${model.modelName}")
            println("Provider: ${model.providerName}")
            println("Input modalities: ${model.inputModalities}")
            println("Output modalities: ${model.outputModalities}")
            println("Supported customizations: ${model.customizationsSupported}")
            println("Supported inference types: ${model.inferenceTypesSupported}")
        }
    } catch (e: Exception) {
        println("Error: ${e.message}")
    }
    return response
}
```

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).
• For API details, see `ListFoundationModels` in AWS SDK for Kotlin API reference.

**PHP**

**SDK for PHP**

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

List the available Amazon Bedrock foundation models.

```php
public function listFoundationModels()
{
    $result = $this->bedrockClient->listFoundationModels();
    return $result;
}
```

• For API details, see `ListFoundationModels` in AWS SDK for PHP API Reference.

**Python**

**SDK for Python (Boto3)**

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

List the available Amazon Bedrock foundation models.

```python
def list_foundation_models(self):
```
List the available Amazon Bedrock foundation models.

:return: The list of available bedrock foundation models.

try:
    response = self.bedrock_client.list_foundation_models()
    models = response["modelSummaries"]
    logger.info("Got %s foundation models.", len(models))
    return models
except ClientError:
    logger.error("Couldn't list foundation models.")
    raise

For API details, see ListFoundationModels in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Scenarios for Amazon Bedrock using AWS SDKs

The following code examples show you how to implement common scenarios in Amazon Bedrock with AWS SDKs. These scenarios show you how to accomplish specific tasks by calling multiple functions within Amazon Bedrock. Each scenario includes a link to GitHub, where you can find instructions on how to set up and run the code.

Examples

- Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

The following code example shows how to build and orchestrate generative AI applications with Amazon Bedrock and Step Functions.
The Amazon Bedrock Serverless Prompt Chaining scenario demonstrates how AWS Step Functions, Amazon Bedrock, and Agents for Amazon Bedrock can be used to build and orchestrate complex, serverless, and highly scalable generative AI applications. It contains the following working examples:

- Write an analysis of a given novel for a literature blog. This example illustrates a simple, sequential chain of prompts.
- Generate a short story about a given topic. This example illustrates how the AI can iteratively process a list of items that it previously generated.
- Create an itinerary for a weekend vacation to a given destination. This example illustrates how to parallelize multiple distinct prompts.
- Pitch movie ideas to a human user acting as a movie producer. This example illustrates how to parallelize the same prompt with different inference parameters, how to backtrack to a previous step in the chain, and how to include human input as part of the workflow.
- Plan a meal based on ingredients the user has at hand. This example illustrates how prompt chains can incorporate two distinct AI conversations, with two AI personas engaging in a debate with each other to improve the final outcome.
- Find and summarize today's highest trending GitHub repository. This example illustrates chaining multiple AI agents that interact with external APIs.

For complete source code and instructions to set up and run, see the full project on GitHub.

**Services used in this example**

- Amazon Bedrock
- Amazon Bedrock Runtime
- Agents for Amazon Bedrock
- Agents for Amazon Bedrock Runtime
- Step Functions

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.
Code examples for Amazon Bedrock Runtime using AWS SDKs

The following code examples show how to use Amazon Bedrock Runtime with an AWS software development kit (SDK).

*Actions* are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios and cross-service examples.

*Scenarios* are code examples that show you how to accomplish a specific task by calling multiple functions within the same service.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Get started**

**Hello Amazon Bedrock**

The following code examples show how to get started using Amazon Bedrock.

**Go**

**SDK for Go V2**

```
package main

import (  
  "context"  
  "encoding/json"  
  "flag"  
  "fmt"  
  "log"  
  "os"

```
// Each model provider defines their own individual request and response formats. // For the format, ranges, and default values for the different models, refer to: // https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters.html

type ClaudeRequest struct {
    Prompt            string `json:"prompt"
    MaxTokensToSample int    `json:"max_tokens_to_sample"
    // Omitting optional request parameters
}

type ClaudeResponse struct {
    Completion string `json:"completion"
}

// main uses the AWS SDK for Go (v2) to create an Amazon Bedrock Runtime client // and invokes Anthropic Claude 2 inside your account and the chosen region. // This example uses the default settings specified in your shared credentials // and config files.
func main() {

    region := flag.String("region", "us-east-1", "The AWS region")
    flag.Parse()

    fmt.Printf("Using AWS region: %s\n", *region)

    sdkConfig, err := config.LoadDefaultConfig(context.Background(),
        config.WithRegion(*region))
    if err != nil {
        fmt.Println("Couldn't load default configuration. Have you set up your AWS account?"
        )
        fmt.Println(err)
        return
    }

    client := bedrockruntime.NewFromConfig(sdkConfig)

    modelId := "anthropic.claude-v2"
prompt := "Hello, how are you today?"

// Anthropic Claude requires you to enclose the prompt as follows:
prefix := "Human: "
postfix := "\n\nAssistant:"
wrappedPrompt := prefix + prompt + postfix

request := ClaudeRequest{
    Prompt:            wrappedPrompt,
    MaxTokensToSample: 200,
}

body, err := json.Marshal(request)
if err != nil {
    log.Panicln("Couldn't marshal the request: ", err)
}

result, err := client.InvokeModel(context.Background(), 
    &bedrockruntime.InvokeModelInput{
        ModelId:     aws.String(modelId),
        ContentType: aws.String("application/json"),
        Body:        body,
    })
if err != nil {
    errMsg := err.Error()
    if strings.Contains(errMsg, "no such host") {
")
    } else if strings.Contains(errMsg, "Could not resolve the foundation model") {
        fmt.Printf("Error: Could not resolve the foundation model from model identifier: "/%v". Please verify that the requested model exists and is accessible within the specified region.\", modelId)
    } else {
    }
    os.Exit(1)
}

var response ClaudeResponse
err = json.Unmarshal(result.Body, &response)
if err != nil {
    log.Fatal("failed to unmarshal", err)
}
fmt.Println("Prompt:\n", prompt)
fmt.Println("Response from Anthropic Claude:\n", response.Completion)

- For API details, see InvokeModel in AWS SDK for Go API Reference.

JavaScript

SDK for JavaScript (v3)

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

/**
 * @typedef {Object} Content
 * @property {string} text
 *
 * @typedef {Object} Usage
 * @property {number} input_tokens
 * @property {number} output_tokens
 *
 * @typedef {Object} ResponseBody
 * @property {Content[]} content
 * @property {Usage} usage
*/

import { fileURLToPath } from "url";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
```
const AWS_REGION = "us-east-1";

const MODEL_ID = "anthropic.claude-3-haiku-20240307-v1:0";
const PROMPT = "Hi. In a short paragraph, explain what you can do.";

const hello = async () => {
    console.log("=".repeat(35));
    console.log("Welcome to the Amazon Bedrock demo!");
    console.log("=".repeat(35));

    console.log("Model: Anthropic Claude 3 Haiku");
    console.log(`Prompt: ${PROMPT}\n`);
    console.log("Invoking model...");

    // Create a new Bedrock Runtime client instance.
    const client = new BedrockRuntimeClient({ region: AWS_REGION });

    // Prepare the payload for the model.
    const payload = {
        anthropic_version: "bedrock-2023-05-31",
        max_tokens: 1000,
        messages: [{ role: "user", content: [{ type: "text", text: PROMPT }] }],
    };

    // Invoke Claude with the payload and wait for the response.
    const apiResponse = await client.send(
        new InvokeModelCommand({
            contentType: "application/json",
            body: JSON.stringify(payload),
            modelId: MODEL_ID,
        }),
    );

    // Decode and return the response(s)
    const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
    /** @type {ResponseBody} */
    const responseBody = JSON.parse(decodedResponseBody);
    const responses = responseBody.content;

    if (responses.length === 1) {
        console.log(`Response: ${responses[0].text}`);
    } else {

For API details, see [InvokeModel](#) in *AWS SDK for JavaScript API Reference*.

### Code examples

- **Actions for Amazon Bedrock Runtime using AWS SDKs**
  - [Invoke the Amazon Titan Image Generator G1 model on Amazon Bedrock for image generation](#)
  - [Invoke the Stability.ai Stable Diffusion XL model on Amazon Bedrock for image generation](#)
  - [Invoke Anthropic Claude 3 on Amazon Bedrock with a multimodal prompt](#)
  - [Invoke the AI21 Labs Jurassic-2 model on Amazon Bedrock for text generation](#)
  - [Invoke the Amazon Titan Text G1 model on Amazon Bedrock for text generation](#)
  - [Invoke the Anthropic Claude 2 model on Amazon Bedrock for text generation](#)
  - [Invoke Anthropic Claude 3 on Amazon Bedrock to generate text](#)
  - [Invoke the Anthropic Claude Instant model on Amazon Bedrock for text generation](#)
  - [Invoke Anthropic Claude on Amazon Bedrock for text generation with a response stream](#)
  - [Invoke the Meta Llama 2 Chat model on Amazon Bedrock for text generation](#)
  - [Invoke the Mistral 7B model on Amazon Bedrock for text generation](#)
  - [Invoke the Mixtral 8x7B model on Amazon Bedrock for text generation](#)

- **Scenarios for Amazon Bedrock Runtime using AWS SDKs**
  - [Create a sample application that offers playgrounds to interact with Amazon Bedrock foundation models using an AWS SDK](#)
  - [An interactive text generation playground for Amazon Bedrock](#)
  - [Invoke multiple large-language models (LLMs) on Amazon Bedrock](#)
  - [Invoke multiple foundation models on Amazon Bedrock](#)
Actions for Amazon Bedrock Runtime using AWS SDKs

The following code examples demonstrate how to perform individual Amazon Bedrock Runtime actions with AWS SDKs. These excerpts call the Amazon Bedrock Runtime API and are code excerpts from larger programs that must be run in context. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the Amazon Bedrock Runtime API Reference.

Examples

• Invoke the Amazon Titan Image Generator G1 model on Amazon Bedrock for image generation
• Invoke the Stability.ai Stable Diffusion XL model on Amazon Bedrock for image generation
• Invoke Anthropic Claude 3 on Amazon Bedrock with a multimodal prompt
• Invoke the AI21 Labs Jurassic-2 model on Amazon Bedrock for text generation
• Invoke the Amazon Titan Text G1 model on Amazon Bedrock for text generation
• Invoke the Anthropic Claude 2 model on Amazon Bedrock for text generation
• Invoke Anthropic Claude 3 on Amazon Bedrock to generate text
• Invoke the Anthropic Claude Instant model on Amazon Bedrock for text generation
• Invoke Anthropic Claude on Amazon Bedrock for text generation with a response stream
• Invoke the Meta Llama 2 Chat model on Amazon Bedrock for text generation
• Invoke the Mistral 7B model on Amazon Bedrock for text generation
• Invoke the Mixtral 8x7B model on Amazon Bedrock for text generation

Invoke the Amazon Titan Image Generator G1 model on Amazon Bedrock for image generation

The following code examples show how to invoke the Amazon Titan Image Generator G1 model on Amazon Bedrock for image generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:
Asynchronously invoke the Amazon Titan Image Generator G1 foundation model to generate images.

```csharp
/// <summary>
/// Asynchronously invokes the Amazon Titan Image Generator G1 model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that describes the image Amazon Titan Image Generator G1 has to generate.</param>
/// <returns>A base-64 encoded image generated by model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
/// For the format, ranges, and default values for Amazon Titan Image Generator G1, refer to:
/// </remarks>

public static async Task<string?> InvokeTitanImageGeneratorG1Async(string prompt, int seed)
{
    string titanImageGeneratorG1ModelId = "amazon.titan-image-generator-v1";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);
```
string payload = new JsonObject()
{
    { "taskType", "TEXT_IMAGE" },
    { "textToImageParams", new JsonObject()
        {
            { "text", prompt }
        }
    },
    { "imageGenerationConfig", new JsonObject()
        {
            { "numberOfImages", 1 },
            { "quality", "standard" },
            { "cfgScale", 8.0f },
            { "height", 512 },
            { "width", 512 },
            { "seed", seed }
        }
    }
}.ToJsonString();

try
{
    InvokeModelResponse response = await client.InvokeModelAsync(new InvokeModelRequest()
    {
        ModelId = titanImageGeneratorG1ModelId,
        Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
        ContentType = "application/json",
        Accept = "application/json"
    });

    {
        var results = JsonNode.ParseAsync(response.Body).Result?["images"].AsArray();
        return results?[0].GetValue<string>();
    }
    else
    {
        Console.WriteLine("InvokeModelAsync failed with status code ".
        + response.HttpStatusCode);
    }
}
catch (AmazonBedrockRuntimeException e)
{
    Console.WriteLine(e.Message);
}
return null;

• For API details, see InvokeModel in AWS SDK for .NET API Reference.

Go

SDK for Go V2

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Amazon Titan Image Generator G1 model to generate images.

type TitanImageRequest struct {
    TaskType            string `json:"taskType"
    TextToImageParams   TextToImageParams `json:"textToImageParams"
    ImageGenerationConfig ImageGenerationConfig `json:"imageGenerationConfig"
}

type TextToImageParams struct {
    Text string `json:"text"
}

type ImageGenerationConfig struct {
    NumberOfImages int `json:"numberOfImages"
    Quality         string `json:"quality"
    CfgScale        float64 `json:"cfgScale"
    Height          int `json:"height"
    Width           int `json:"width"
    Seed            int64 `json:"seed"
}

type TitanImageResponse struct {
Images []string `json:"images"

// Invokes the Titan Image model to create an image using the input provided
// in the request body.
func (wrapper InvokeModelWrapper) InvokeTitanImage(prompt string, seed int64) (string, error) {
    modelId := "amazon.titan-image-generator-v1"

    body, err := json.Marshal(TitanImageRequest {
        TaskType: "TEXT_IMAGE",
        TextToImageParams: TextToImageParams {
            Text: prompt,
        },
        ImageGenerationConfig: ImageGenerationConfig {
            NumberOfImages: 1,
            Quality: "standard",
            CfgScale: 8.0,
            Height: 512,
            Width: 512,
            Seed: seed,
        },
    },

    if err != nil { log.Fatal("failed to marshal", err) }

    output, err := wrapper.BedrockRuntimeClient.InvokeModel(context.TODO(),
                &bedrockruntime.InvokeModelInput{
                    ModelId: aws.String(modelId),
                    ContentType: aws.String("application/json"),
                    Body: body,
                })

    if err != nil { ProcessError(err, modelId) }

    var response TitanImageResponse
    if err := json.Unmarshal(output.Body, &response); err != nil {
        log.Fatal("failed to unmarshal", err)
    }

    base64ImageData := response.Images[0]

    return base64ImageData, nil
Asynchronously invoke the Amazon Titan Image Generator G1 model to generate images.

```java
/**
 * Invokes the Amazon Titan image generation model to create an image using the input provided in the request body.
 * @param prompt The prompt that you want Amazon Titan to use for image generation.
 * @param seed   The random noise seed for image generation (Range: 0 to 2147483647).
 * @return A Base64-encoded string representing the generated image.
 */
public static String invokeTitanImage(String prompt, long seed) {
  String titanImageModelId = "amazon.titan-image-generator-v1";
  BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
      Actions
      InvokeModel in AWS SDK for Go API Reference.

Java

SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
```java
.region(Region.US_EAST_1)
.credentialsProvider(ProfileCredentialsProvider.create())
.build();

var textToImageParams = new JSONObject().put("text", prompt);

var imageGenerationConfig = new JSONObject()
    .put("numberOfImages", 1)
    .put("quality", "standard")
    .put("cfgScale", 8.0)
    .put("height", 512)
    .put("width", 512)
    .put("seed", seed);

JSONObject payload = new JSONObject()
    .put("taskType", "TEXT_IMAGE")
    .put("textToImageParams", textToImageParams)
    .put("imageGenerationConfig", imageGenerationConfig);

InvokeModelRequest request = InvokeModelRequest.builder()
    .body(SdkBytes.fromUtf8String(payload.toString()))
    .modelId(titanImageModelId)
    .contentType("application/json")
    .accept("application/json")
    .build();

CompletableFuture<InvokeModelResponse> completableFuture =
    client.invokeModel(request)
    .whenComplete((response, exception) -> {
        if (exception != null) {
            System.out.println("Model invocation failed: " + exception);
        }
    });

String base64ImageData = "";
try {
    InvokeModelResponse response = completableFuture.get();
    JSONObject responseBody = new JSONObject(response.body().asUtf8String());
    base64ImageData = responseBody
        .getJSONArray("images")
        .getString(0);
}
```
Invoke the Amazon Titan Image Generator G1 model to generate images.

```java
/**
 * Invokes the Amazon Titan image generation model to create an image using the
 * input provided in the request body.
 * @param prompt The prompt that you want Amazon Titan to use for image generation.
 * @param seed The random noise seed for image generation (Range: 0 to 2147483647).
 * @return A Base64-encoded string representing the generated image.
 */
public static String invokeTitanImage(String prompt, long seed) {
    String titanImageModelId = "amazon.titan-image-generator-v1";
    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();
    return base64ImageData;
}
```
```java
var textToImageParams = new JSONObject().put("text", prompt);

var imageGenerationConfig = new JSONObject()
    .put("numberOfImages", 1)
    .put("quality", "standard")
    .put("cfgScale", 8.0)
    .put("height", 512)
    .put("width", 512)
    .put("seed", seed);

JSONObject payload = new JSONObject()
    .put("taskType", "TEXT_IMAGE")
    .put("textToImageParams", textToImageParams)
    .put("imageGenerationConfig", imageGenerationConfig);

InvokeModelRequest request = InvokeModelRequest.builder()
    .body(SdkBytes.fromUtf8String(payload.toString()))
    .modelId(titanImageModelId)
    .contentType("application/json")
    .accept("application/json")
    .build();

InvokeModelResponse response = client.invokeModel(request);

JSONObject responseBody = new JSONObject(response.body().asUtf8String());

String base64ImageData = responseBody
    .getJSONArray("images")
    .getString(0);

return base64ImageData;
```

- For API details, see [InvokeModel](#) in *AWS SDK for Java 2.x API Reference*. 
Invoke the Amazon Titan Image Generator G1 model to generate images.

```php
public function invokeTitanImage(string $prompt, int $seed)
{
    # The different model providers have individual request and response formats.
    # For the format, ranges, and default values for Titan Image models refer to:

    $base64_image_data = "";

    try {
        $modelId = 'amazon.titan-image-generator-v1';

        $request = json_encode([
            'taskType' => 'TEXT_IMAGE',
            'textToImageParams' => [
                'text' => $prompt
            ],
            'imageGenerationConfig' => [
                'numberOfImages' => 1,
                'quality' => 'standard',
                'cfgScale' => 8.0,
                'height' => 512,
                'width' => 512,
                'seed' => $seed
            ]
        ]);

        $result = $this->bedrockRuntimeClient->invokeModel(["contentType" => 'application/json',
            'modelId' => $modelId,
            'request' => $request
        ]);

**Python**

**SDK for Python (Boto3)**

```python
def invoke_titan_image(self, prompt, seed):
    
    """
    Invokes the Titan Image model to create an image using the input provided in the request body.
    
    :param prompt: The prompt that you want Amazon Titan to use for image generation.
    :param seed: Random noise seed (range: 0 to 2147483647)
    :return: Base64-encoded inference response from the model.
    """

    try:
```

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/aws-samples/).
# The different model providers have individual request and response formats.
# For the format, ranges, and default values for Titan Image models refer to:

```python
request = json.dumps(
    {
        "taskType": "TEXT_IMAGE",
        "textToImageParams": {
            "text": prompt,
            "imageGenerationConfig": {
                "numberOfImages": 1,
                "quality": "standard",
                "cfgScale": 8.0,
                "height": 512,
                "width": 512,
                "seed": seed,
            },
        },
    }
)

response = self.bedrock_runtime_client.invoke_model(
    modelId="amazon.titan-image-generator-v1", body=request
)

response_body = json.loads(response["body"].read())
base64_image_data = response_body["images"][0]

return base64_image_data

except ClientError:
    logger.error("Couldn't invoke Titan Image generator")
    raise
```

- For API details, see [InvokeModel](https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-titan-image.html) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-titan-image.html). This topic also includes information about getting started and details about previous SDK versions.
Invoke the Stability.ai Stable Diffusion XL model on Amazon Bedrock for image generation

The following code examples show how to invoke the Stability.ai Stable Diffusion XL model on Amazon Bedrock for image generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- Interactive text playground
- Invoke multiple foundation models on Amazon Bedrock
- Invoke multiple LLMs on Amazon Bedrock

.NET

AWS SDK for .NET

Asynchronously invoke Stability.ai Stable Diffusion XL foundation model to generate images.

```csharp
/// <summary>
/// Asynchronously invokes the Stability.ai Stable Diffusion XL model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that describes the image Stability.ai Stable Diffusion XL has to generate.</param>
/// <returns>A base-64 encoded image generated by model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
/// For the format, ranges, and default values for Stability.ai Stable Diffusion XL, refer to:
```
/// </remarks>
public static async Task<string?> InvokeStableDiffusionXLG1Async(string
prompt, int seed, string? stylePreset = null)
{
    string stableDiffusionXLModelId = "stability.stable-diffusion-xl";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);

    var jsonPayload = new JsonObject()
    {
        { "text_prompts", new JsonArray() 
            {
                new JsonObject()
                {
                    { "text", prompt }
                }
            }
        },
        { "seed", seed }
    };

    if (!string.IsNullOrEmpty(stylePreset))
    {
        jsonPayload.Add("style_preset", stylePreset);
    }

    string payload = jsonPayload.ToString();

    try
    {
        InvokeModelResponse response = await client.InvokeModelAsync(new
InvokeModelRequest()
        {
            ModelId = stableDiffusionXLModelId,
            Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
            ContentType = "application/json",
            Accept = "application/json"
        });

        {
            var results = JsonNode.ParseAsync(response.Body).Result?
["artifacts"]?.AsArray();

            return results?[0]?["base64"]?.GetValue<string>();
        }
    }
}
Asynchronously invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```csharp
} 
else 
{
    Console.WriteLine("InvokeModelAsync failed with status code " + response.HttpStatusCode);
}
}
catch (AmazonBedrockRuntimeException e)
{
    Console.WriteLine(e.Message);
}
return null;
}
```

- For API details, see [InvokeModel](#) in [AWS SDK for .NET API Reference](#).

### Java

**SDK for Java 2.x**

- **Note**
  There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

Asynchronously invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```java
/**
 * Asynchronously invokes the Stability.ai Stable Diffusion XL model to create
 * an image based on the provided input.
 *
 * @param prompt The prompt that guides the Stable Diffusion model.
 * @param seed The random noise seed for image generation (use 0 or omit
 * for a random seed).
 * @param stylePreset The style preset to guide the image model towards a specific style.
 */
```
public static String invokeStableDiffusion(String prompt, long seed, String stylePreset) {
    /*
     * The different model providers have individual request and response formats.
     * For the format, ranges, and available style_presets of Stable Diffusion
     * models refer to:
     */

    String stableDiffusionModelId = "stability.stable-diffusion-xl";

    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
            .region(Region.US_EAST_1)
            .credentialsProvider(ProfileCredentialsProvider.create())
            .build();

    JSONArray wrappedPrompt = new JSONArray().put(new JSONObject().put("text", prompt));
    JSONObject payload = new JSONObject()
            .put("text_prompts", wrappedPrompt)
            .put("seed", seed);

    if (stylePreset != null && !stylePreset.isEmpty()) {
        payload.put("style_preset", stylePreset);
    }

    InvokeModelRequest request = InvokeModelRequest.builder()
            .body(SdkBytes.fromUtf8String(payload.toString()))
            .modelId(stableDiffusionModelId)
            .contentType("application/json")
            .accept("application/json")
            .build();

    CompletableFuture<InvokeModelResponse> completableFuture = client.invokeModel(request)
            .whenComplete((response, exception) -> {
                if (exception != null) {
                    System.out.println("Model invocation failed: " + exception);
                }
            });
Invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```java
/**
 * Invokes the Stability.ai Stable Diffusion XL model to create an image
 * based
 * on the provided input.
 * *
 * @param prompt The prompt that guides the Stable Diffusion model.
 * @param seed The random noise seed for image generation (use 0 or omit
 * for a random seed).
 * @param stylePreset The style preset to guide the image model towards a
 * specific style.
 * @return A Base64-encoded string representing the generated image.
 */
public static String invokeStableDiffusion(String prompt, long seed, String stylePreset) {
    String base64ImageData = "";
    try {
        InvokeModelResponse response = completableFuture.get();
        JSONObject responseBody = new
        JSONObject(response.body().asUtf8String());
        base64ImageData = responseBody
            .getJSONArray("artifacts")
            .getJSONObject(0)
            .getString("base64");
    } catch (InterruptedException e) {
        Thread.currentThread().interrupt();
        System.err.println(e.getMessage());
    } catch (ExecutionException e) {
        System.err.println(e.getMessage());
    }
    return base64ImageData;
}
```
* The different model providers have individual request and response formats.
  * For the format, ranges, and available style_presets of Stable Diffusion
    * models refer to:
  */

String stableDiffusionModelId = "stability.stable-diffusion-xl";

BedrockRuntimeClient client = BedrockRuntimeClient.builder()
  .region(Region.US_EAST_1)
  .credentialsProvider(ProfileCredentialsProvider.create())
  .build();

JSONArray wrappedPrompt = new JSONArray().put(new JSONObject().put("text", prompt));

JSONObject payload = new JSONObject()
  .put("text_prompts", wrappedPrompt)
  .put("seed", seed);

if (!(stylePreset == null || stylePreset.isEmpty())) {
  payload.put("style_preset", stylePreset);
}

InvokeModelRequest request = InvokeModelRequest.builder()
  .body(SdkBytes.fromUtf8String(payload.toString()))
  .modelId(stableDiffusionModelId)
  .contentType("application/json")
  .accept("application/json")
  .build();

InvokeModelResponse response = client.invokeModel(request);

JSONObject responseBody = new JSONObject(response.body().asUtf8String());
Invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```php
public function invokeStableDiffusion(string $prompt, int $seed, string $style_preset)
{
    // The different model providers have individual request and response formats.
    // For the format, ranges, and available style_presets of Stable Diffusion models refer to:

    $base64_image_data = "";

    try {
        $modelId = 'stability.stable-diffusion-xl';

        $body = [
            'text_prompts' => [
                ['text' => $prompt]
            ],
            'seed' => $seed,
            'cfg_scale' => 10,
            'steps' => 30
        ];

        // For API details, see InvokeModel in AWS SDK for Java 2.x API Reference.
        // There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
```
if ($style_preset) {
    $body['style_preset'] = $style_preset;
}

$result = $this->bedrockRuntimeClient->invokeModel(
    'contentType' => 'application/json',
    'body' => json_encode($body),
    'modelId' => $modelId,
);

$response_body = json_decode($result['body']);

$base64_image_data = $response_body->artifacts[0]->base64;
} catch (Exception $e) {
    echo "Error: ({$e->getCode()}) - {$e->getMessage()}
";
}

return $base64_image_data;

- For API details, see [InvokeModel](https://docs.aws.amazon.com/sdk-for-php/v3/developer-guide/api-reference/service-api.html#s3) in *AWS SDK for PHP API Reference*.

**Python**

**SDK for Python (Boto3)**

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/aws-samples/aws-code-examples).

Invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```python
def invoke_stable_diffusion(self, prompt, seed, style_preset=None):
    
    # Invokes the Stability.ai Stable Diffusion XL model to create an image using
    # the input provided in the request body.
```
:param prompt: The prompt that you want Stable Diffusion to use for image generation.
:param seed: Random noise seed (omit this option or use 0 for a random seed)
:param style_preset: Pass in a style preset to guide the image model towards a particular style.
:return: Base64-encoded inference response from the model.

""

try:
    # The different model providers have individual request and response formats.
    # For the format, ranges, and available style_presets of Stable Diffusion models refer to:

    body = {
        "text_prompts": ["text": prompt],
        "seed": seed,
        "cfg_scale": 10,
        "steps": 30,
    }

    if style_preset:
        body["style_preset"] = style_preset

    response = self.bedrock_runtime_client.invoke_model(
        modelId="stability.stable-diffusion-xl", body=json.dumps(body)
    )

    response_body = json.loads(response["body"].read())
    base64_image_data = response_body["artifacts"][0]["base64"]

    return base64_image_data
except ClientError:
    logger.error("Couldn't invoke Stable Diffusion XL")
    raise
- For API details, see `InvokeModel` in *AWS SDK for Python (Boto3) API Reference*.

**SAP ABAP**

**SDK for SAP ABAP**

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/aws-samples/sap-abap-sdk-for-sap-abap).

Invoke the Stability.ai Stable Diffusion XL foundation model to generate images.

```plaintext
"Stable Diffusion Input Parameters should be in a format like this:
*
*   {
*     "text_prompts": [
*       {"text":"Draw a dolphin with a mustache"},
*       {"text":"Make it photorealistic"}
*     ],
*     "cfg_scale":10,
*     "seed":0,
*     "steps":50
*   }
*
TYPES: BEGIN OF prompt_ts,
   text TYPE /aws1/rt_shape_string,
END OF prompt_ts.

DATA: BEGIN OF ls_input,
   text_prompts TYPE STANDARD TABLE OF prompt_ts,
   cfg_scale TYPE /aws1/rt_shape_integer,
   seed TYPE /aws1/rt_shape_integer,
   steps TYPE /aws1/rt_shape_integer,
END OF ls_input.

APPEND VALUE prompt_ts( text = iv_prompt ) TO ls_input-text_prompts.
ls_input-cfg_scale = 10.
ls_input-seed = 0. "or better, choose a random integer."
ls_input-steps = 50.

DATA(lv_json) = /ui2/cl_json=>serialize(
   data = ls_input
)
TRY.
DATA(lo_response) = lo_bdr->invokemodel(
   iv_body = /aws1/cl_rt_util=>string_to_xstring( lv_json )
   iv_modelid = 'stability.stable-diffusion-xl-v0'
   iv_accept = 'application/json'
   iv_contenttype = 'application/json' ).

"Stable Diffusion Result Format:

*       {
*         "result": "success",
*         "artifacts": [
*           {
*             "seed": 0,
*             "base64": "iVBORw0KGgoAAAANSUhEUgAAAgAAA....
*             "finishReason": "SUCCESS"
*           }
*       ]

TYPES: BEGIN OF artifact_ts,
   seed TYPE /aws1/rt_shape_integer,
   base64 TYPE /aws1/rt_shape_string,
   finishreason TYPE /aws1/rt_shape_string,
END OF artifact_ts.

DATA: BEGIN OF ls_response,
   result TYPE /aws1/rt_shape_string,
   artifacts TYPE STANDARD TABLE OF artifact_ts,
END OF ls_response.

/ui2/cl_json=>deserialize(
   EXPORTING jsonx = lo_response->get_body( )
   pretty_name = /ui2/cl_json=>pretty_mode-camel_case
   CHANGING  data  = ls_response ).
IF ls_response-artifacts IS NOT INITIAL.
   DATA(lv_image) =
   cl_http_utility=>if_http_utility~decode_x_base64( ls_response-artifacts[ 1 ]-base64 ).
ENDIF.
CATCH /aws1/cx_bdraccesdeniedex INTO DATA(lo_ex).
   WRITE / lo_ex->get_text( ).
   WRITE / |Don't forget to enable model access at https://
   console.aws.amazon.com/bedrock/home?#/modelaccess|. 
Invoke the Stability.ai Stable Diffusion XL foundation model to generate images using L2 high level client.

TRY.

DATA(lo_bdr_l2_sd) = /aws1/cl_bdr_l2_factory=>create_stable_diffusion_10( lo_bdr ).
" iv_prompt contains a prompt like 'Show me a picture of a unicorn reading an enterprise financial report'.
DATA(lv_image) = lo_bdr_l2_sd->text_to_image( iv_prompt ).
CATCH /aws1/cx_bdraccessdeniedex INTO DATA(lo_ex).
WRITE / lo_ex->get_text( ).
WRITE / |Don't forget to enable model access at https://console.aws.amazon.com/bedrock/home?#/modelaccess|.
ENDTRY.

For API details, see InvokeModel in AWS SDK for SAP ABAP API reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Invoke Anthropic Claude 3 on Amazon Bedrock with a multimodal prompt

The following code example shows how to invoke Anthropic Claude 3 on Amazon Bedrock with a multimodal prompt.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- Interactive text playground
- Invoke multiple foundation models on Amazon Bedrock
- Invoke multiple LLMs on Amazon Bedrock
Invoke Anthropic Claude 3 with a multimodal prompt to analyze an image.

```python
def invoke_claude_3_multimodal(self, prompt, base64_image_data):
    """
    Invokes Anthropic Claude 3 Sonnet to run a multimodal inference using the
    input
    provided in the request body.
    
    :param prompt: The prompt that you want Claude 3 to use.
    :param base64_image_data: The base64-encoded image that you want to add
    to the request.
    :return: Inference response from the model.
    """

    # Initialize the Amazon Bedrock runtime client
    client = self.client or boto3.client(
        service_name="bedrock-runtime", region_name="us-east-1"
    )

    # Invoke the model with the prompt and the encoded image
    model_id = "anthropic.claude-3-sonnet-20240229-v1:0"
    request_body = {
        "anthropic_version": "bedrock-2023-05-31",
        "max_tokens": 2048,
        "messages": [
            {
                "role": "user",
                "content": [
                    {
                        "type": "text",
                        "text": prompt,
                    },
                    {
```

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
try:
    response = client.invoke_model(
        modelId=model_id,
        body=json.dumps(request_body),
    )

    # Process and print the response
    result = json.loads(response.get("body").read())
    input_tokens = result["usage"]["input_tokens"]
    output_tokens = result["usage"]["output_tokens"]
    output_list = result.get("content", [])

    print("Invocation details:")
    print(f"- The input length is {input_tokens} tokens.")
    print(f"- The output length is {output_tokens} tokens.")

    print(f"- The model returned {len(output_list)} response(s):")
    for output in output_list:
        print(output["text"])

    return result
except ClientError as err:
    logger.error(
        "Couldn't invoke Claude 3 Sonnet. Here's why: %s: %s",
        err.response["Error"]["Code"],
        err.response["Error"]["Message"],
    )
    raise

• For API details, see InvokeModel in AWS SDK for Python (Boto3) API Reference.
For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Invoke the AI21 Labs Jurassic-2 model on Amazon Bedrock for text generation

The following code examples show how to invoke the AI21 Labs Jurassic-2 model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- Interactive text playground
- Invoke multiple foundation models on Amazon Bedrock
- Invoke multiple LLMs on Amazon Bedrock

.NET

AWS SDK for .NET

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Asynchronously invoke the AI21 Labs Jurassic-2 foundation model.

```csharp
/// <summary>
/// Asynchronously invokes the AI21 Labs Jurassic-2 model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that you want Claude to complete.</param>
/// <returns>The inference response from the model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
```
public static async Task<string> InvokeJurassic2Async(string prompt)
{
    string jurassic2ModelId = "ai21.j2-mid-v1";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);

    string payload = new JsonObject()
    {
        { "prompt", prompt },
        { "maxTokens", 200 },
        { "temperature", 0.5 }
    }.ToJsonString();

    string generatedText = "";
    try
    {
        InvokeModelResponse response = await client.InvokeModelAsync(new InvokeModelRequest()
        {
            ModelId = jurassic2ModelId,
            Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
            ContentType = "application/json",
            Accept = "application/json"
        });

        {
            return JsonNode.ParseAsync(response.Body)
                .Result?["completions"]?.AsArray()[0]?["data"]?.AsObject()["text"]?.GetValue<string>() ?? "";
        }
        else
        {
            Console.WriteLine("InvokeModelAsync failed with status code "+ response.HttpStatusCode);
        }
    }
    catch (AmazonBedrockRuntimeException e)
For API details, see `InvokeModel` in *AWS SDK for .NET API Reference*.

**Go**

**SDK for Go V2**

>`Note`

There's more on GitHub. Find the complete example and learn how to set up and run in the *AWS Code Examples Repository*.

Invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```go
// Each model provider has their own individual request and response formats. // For the format, ranges, and default values for AI21 Labs Jurassic-2, refer to: // https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-jurassic2.html
type Jurassic2Request struct {
    Prompt      string `json:"prompt"`
    MaxTokens   int     `json:"maxTokens,omitempty"`
    Temperature float64 `json:"temperature,omitempty"`
}

type Jurassic2Response struct {
    Completions []Completion `json:"completions"`
}

type Completion struct {
    Data Data `json:"data"`
}

type Data struct {
    Text string `json:"text"`
}
```
// Invokes AI21 Labs Jurassic-2 on Amazon Bedrock to run an inference using the input provided in the request body.
func (wrapper InvokeModelWrapper) InvokeJurassic2(prompt string) (string, error) {
    modelId := "ai21.j2-mid-v1"

    body, err := json.Marshal(Jurassic2Request {
        Prompt:            prompt,
        MaxTokens:      200,
        Temperature:       0.5,
    })

    if err != nil { log.Fatal("failed to marshal", err) }

    output, err := wrapper.BedrockRuntimeClient.InvokeModel(context.TODO(),
&bedrockruntime.InvokeModelInput{
    ModelId: aws.String(modelId),
    ContentType: aws.String("application/json"),
    Body: body,
})

    if err != nil { ProcessError(err, modelId) }

    var response Jurassic2Response
    if err := json.Unmarshal(output.Body, &response); err != nil { 
        log.Fatal("failed to unmarshal", err)
    }

    return response.Completions[0].Data.Text, nil
}

- For API details, see [InvokeModel](#) in [AWS SDK for Go API Reference](#).
SDK for Java 2.x

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Asynchronously invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```java
/**
 * Asynchronously invokes the AI21 Labs Jurassic-2 model to run an inference
 * based on the provided input.
 * @param prompt The prompt that you want Jurassic to complete.
 * @return The inference response generated by the model.
 */
public static String invokeJurassic2(String prompt) {
    String jurassic2ModelId = "ai21.j2-mid-v1";

    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    String payload = new JSONObject()
        .put("prompt", prompt)
        .put("temperature", 0.5)
        .put("maxTokens", 200)
        .toString();

    InvokeModelRequest request = InvokeModelRequest.builder()
Invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```java
CompletableFuture<InvokeModelResponse> completableFuture =
    client.invokeModel(request)
    .whenComplete((response, exception) -> {
        if (exception != null) {
            System.out.println("Model invocation failed: " +
            exception);
        }
    });

String generatedText = "";
try {
    InvokeModelResponse response = completableFuture.get();
    JSONObject responseBody = new
        JSONObject(response.body().asUtf8String());
    generatedText = responseBody
        .getJSONArray("completions")
        .getJSONObject(0)
        .getJSONObject("data")
        .getString("text");
} catch (InterruptedException e) {
    Thread.currentThread().interrupt();
    System.err.println(e.getMessage());
} catch (ExecutionException e) {
    System.err.println(e.getMessage());
}

return generatedText;
```

/**
 * Invokes the AI21 Labs Jurassic-2 model to run an inference based on
 * provided input.
* @param prompt The prompt for Jurassic to complete.
* @return The generated response.
*/
public static String invokeJurassic2(String prompt) {
    /*
     * The different model providers have individual request and response formats.
     * For the format, ranges, and default values for AI21 Labs Jurassic-2, refer to:
     * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-jurassic2.html
     */

    String jurassic2ModelId = "ai21.j2-mid-v1";

    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    String payload = new JSONObject()
        .put("prompt", prompt)
        .put("temperature", 0.5)
        .put("maxTokens", 200)
        .toString();

    InvokeModelRequest request = InvokeModelRequest.builder()
        .body(SdkBytes.fromUtf8String(payload))
        .modelId(jurassic2ModelId)
        .contentType("application/json")
        .accept("application/json")
        .build();

    InvokeModelResponse response = client.invokeModel(request);

    JSONObject responseBody = new JSONObject(response.body().asUtf8String());

    String generatedText = responseBody
        .getJSONArray("completions")
        .getJSONObject(0)
For API details, see `InvokeModel` in *AWS SDK for Java 2.x API Reference*.

JavaScript

**SDK for JavaScript (v3)**

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "./config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} Data
 * @property {string} text
 *
 * @typedef {Object} Completion
 * @property {Data} data
 *
 * @typedef {Object} ResponseBody
 * @property {Completion[]} completions
 */
```

Invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```javascript
// There's more on GitHub. Find the complete example and learn how to set up and run
```

Note
/**
 * Invokes an AI21 Labs Jurassic-2 model.
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "ai21.j2-mid-v1".
 */
export const invokeModel = async (prompt, modelId = "ai21.j2-mid-v1") => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    prompt,
    maxTokens: 500,
    temperature: 0.5,
  };

  // Invoke the model with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response(s).
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {ResponseBody} */
  const responseBody = JSON.parse(decodedResponseBody);
  return responseBody.completions[0].data.text;
};

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  const prompt =
    'Complete the following in one sentence: "Once upon a time..."';
  const modelId = FoundationModels.JURASSIC2_MID.modelId;
  console.log('Prompt: ${prompt}');
  console.log('Model ID: ${modelId}');

  try {
    console.log("-".repeat(53));
Invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```php
public function invokeJurassic2($prompt)
{
    // The different model providers have individual request and response formats.
    // For the format, ranges, and default values for AI21 Labs Jurassic-2,
    // refer to:

    $completion = "";

    try {
        $modelId = 'ai21.j2-mid-v1';

        $body = [
            'prompt' => $prompt,
            'temperature' => 0.5,
            'maxTokens' => 200,
        ];
```
• For API details, see **InvokeModel** in *AWS SDK for PHP API Reference*.

Python

**SDK for Python (Boto3)**

> Note
> There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

Invoke the AI21 Labs Jurassic-2 foundation model to generate text.

```python
def invoke_jurassic2(self, prompt):
    """
    Invokes the AI21 Labs Jurassic-2 large-language model to run an inference using the input provided in the request body.

    :param prompt: The prompt that you want Jurassic-2 to complete.
    :return: Inference response from the model.
    """

    try:
```
The different model providers have individual request and response formats. For the format, ranges, and default values for AI21 Labs Jurassic-2, refer to:

https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-jurassic2.html

```python
body = {
    "prompt": prompt,
    "temperature": 0.5,
    "maxTokens": 200,
}

response = self.bedrock_runtime_client.invoke_model(
    modelId="ai21.j2-mid-v1", body=json.dumps(body)
)

response_body = json.loads(response["body"].read())
completion = response_body["completions"][0]["data"]["text"]

return completion
```

```python
except ClientError:
    logger.error("Couldn't invoke Jurassic-2")
raise
```

- For API details, see [InvokeModel](https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-jurassic2.html) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](https://docs.aws.amazon.com/bedrock/latest/userguide/using-sdk.html). This topic also includes information about getting started and details about previous SDK versions.

### Invoke the Amazon Titan Text G1 model on Amazon Bedrock for text generation

The following code examples show how to invoke the Amazon Titan Text G1 model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:
Asynchronously invoke the Amazon Titan Text G1 foundation model to generate text.

```csharp
/// <summary>
/// Asynchronously invokes the Amazon Titan Text G1 Express model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that you want Amazon Titan Text G1 Express to complete.</param>
/// <returns>The inference response from the model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
/// For the format, ranges, and default values for Amazon Titan Text G1 Express, refer to:
/// </remarks>
public static async Task<string> InvokeTitanTextG1Async(string prompt)
{
    string titanTextG1ModelId = "amazon.titan-text-express-v1";
    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);
    string payload = new JObject()
    {
        { "inputText", prompt },
    };
```
{ "textGenerationConfig", new JsonObject()
    {
        { "maxTokenCount", 512 },
        { "temperature", 0f },
        { "topP", 1f }
    }
}.ToJsonString();

string generatedText = "";
try
{
    InvokeModelResponse response = await client.InvokeModelAsync(new
    InvokeModelRequest()
    {
        ModelId = titanTextG1ModelId,
        Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
        ContentType = "application/json",
        Accept = "application/json"
    });

    {
        var results = JsonNode.ParseAsync(response.Body).Result?
            ["results"]?.AsArray();

        return results is null ? "" : string.Join(" ",
            results.Select(x => x?["outputText"]?.GetValue<string?>()));
    }
    else
    {
        Console.WriteLine("InvokeModelAsync failed with status code " + response.HttpStatusCode);
    }
}
catch (AmazonBedrockRuntimeException e)
{
    Console.WriteLine(e.Message);
}
return generatedText;
JavaScript

**SDK for JavaScript (v3)**

Invoke the Amazon Titan Text G1 foundation model to generate text.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "./config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} ResponseBody
 * @property {Object[]} results
 */

/**
 * Invokes an Amazon Titan Text generation model.
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "amazon.titan-text-express-v1".
 * @returns
 */
export const invokeModel = async (prompt,
                                  modelId = "amazon.titan-text-express-v1") => {
    // Create a new Bedrock Runtime client instance.
```
const client = new BedrockRuntimeClient({ region: "us-east-1" });

// Prepare the payload for the model.
const payload = {
  inputText: prompt,
  textGenerationConfig: {
    maxTokenCount: 4096,
    stopSequences: [],
    temperature: 0,
    topP: 1,
  },
};

// Invoke the model with the payload and wait for the response.
const command = new InvokeModelCommand({
  contentType: "application/json",
  body: JSON.stringify(payload),
  modelId,
});
const apiResponse = await client.send(command);

// Decode and return the response.
const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
/** @type {ResponseBody} */
const responseBody = JSON.parse(decodedResponseBody);
return responseBody.results[0].outputText;

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  const prompt = 'Complete the following in one sentence: "Once upon a time..."';
  const modelId = FoundationModels.TITAN_TEXT_G1_EXPRESS.modelId;
  console.log(`Prompt: ${prompt}`);
  console.log(`Model ID: ${modelId}`);
  try {
    console.log("-".repeat(53));
    const response = await invokeModel(prompt, modelId);
    console.log(response);
  } catch (err) {
    console.log(err);
  }
}
For API details, see `InvokeModel` in `AWS SDK for JavaScript API Reference`.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

### Invoke the Anthropic Claude 2 model on Amazon Bedrock for text generation

The following code examples show how to invoke the Anthropic Claude 2 model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- Interactive text playground
- Invoke multiple foundation models on Amazon Bedrock
- Invoke multiple LLMs on Amazon Bedrock

**.NET**

**AWS SDK for .NET**

> Asynchronously invoke the Anthropic Claude 2 foundation model to generate text.

```csharp
/// <summary>
/// Asynchronously invokes the Anthropic Claude 2 model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that you want Claude to complete.</param>
/// <returns>The inference response from the model</returns>
```
/// <remarks>
/// The different model providers have individual request and response
/// formats.
/// For the format, ranges, and default values for Anthropic Claude,
/// refer to:
///     https://docs.aws.amazon.com/bedrock/latest/userguide/model-
///     parameters-claude.html
/// </remarks>
public static async Task<string> InvokeClaudeAsync(string prompt)
{
    string claudeModelId = "anthropic.claude-v2";

    // Claude requires you to enclose the prompt as follows:
    string enclosedPrompt = "Human: " + prompt + "\n\nAssistant:";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);

    string payload = new JsonObject()
    {
        { "prompt", enclosedPrompt },
        { "max_tokens_to_sample", 200 },
        { "temperature", 0.5 },
        { "stop_sequences", new JsonArray("\n\nHuman:" )
    }.ToJsonString();

    string generatedText = "";
    try
    {
        InvokeModelResponse response = await client.InvokeModelAsync(new
InvokeModelRequest()
        {
            ModelId = claudeModelId,
            Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
            ContentType = "application/json",
            Accept = "application/json"
        });

        {
            return JsonNode.ParseAsync(response.Body).Result?["completion"]?.GetValue<string>() ?? "";
        }
        else
        {
            Actions
            Actions
        }
Go

SDK for Go V2

Invoke the Anthropic Claude 2 foundation model to generate text.

// Each model provider has their own individual request and response formats. // For the format, ranges, and default values for Anthropic Claude, refer to: // https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-claude.html
type ClaudeRequest struct {
    Prompt     string `json:"prompt"
    MaxTokensToSample int `json:"max_tokens_to_sample"
    Temperature float64 `json:"temperature,omitempty"
    StopSequences []string `json:"stop_sequences,omitempty"
}

type ClaudeResponse struct {
    Completion string `json:"completion"
}
// Invokes Anthropic Claude on Amazon Bedrock to run an inference using the input
// provided in the request body.
func (wrapper InvokeModelWrapper) InvokeClaude(prompt string) (string, error) {
    modelId := "anthropic.claude-v2"

    // Anthropic Claude requires enclosing the prompt as follows:
    enclosedPrompt := "Human: " + prompt + "\n\nAssistant:"

    body, err := json.Marshal(ClaudeRequest {
        Prompt:           enclosedPrompt,
        MaxTokensToSample: 200,
        Temperature:       0.5,
        StopSequences:     []string{"\n\nHuman:"},
    })

    if err != nil { log.Fatal("failed to marshal", err) }

    output, err := wrapper.BedrockRuntimeClient.InvokeModel(context.TODO(),
        &bedrockruntime.InvokeModelInput{
            ModelId: aws.String(modelId),
            ContentType: aws.String("application/json"),
            Body: body,
        })

    if err != nil { ProcessError(err, modelId) }

    var response ClaudeResponse
    if err := json.Unmarshal(output.Body, &response); err != nil {
        log.Fatal("failed to unmarshal", err)
    }

    return response.Completion, nil
}

- For API details, see [InvokeModel](https://aws-sdk-go.amazonaws.com/api/doc/latest/bedrock.html) in [AWS SDK for Go API Reference](https://aws.amazon.com/sdk-for-go/api/bedrock/).
Asynchronously invoke the Anthropic Claude 2 foundation model to generate text.

```java
/**
 * Asynchronously invokes the Anthropic Claude 2 model to run an inference based
 * on the provided input.
 *
 * @param prompt The prompt that you want Claude to complete.
 * @return The inference response from the model.
 */
public static String invokeClaude(String prompt) {
    String claudeModelId = "anthropic.claude-v2";

    // Claude requires you to enclose the prompt as follows:
    String enclosedPrompt = "Human: " + prompt + "\n\nAssistant:";

    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    String payload = new JSONObject()
        .put("prompt", enclosedPrompt)
        .put("max_tokens_to_sample", 200)

    // ...
Invoke the Anthropic Claude 2 foundation model to generate text.

/**
 * Invokes the Anthropic Claude 2 model to run an inference based on the
 * provided input.
 */
public static String invokeClaude(String prompt) {
    String claudeModelId = "anthropic.claude-v2";
    String enclosedPrompt = "Human: " + prompt + "\nAssistant: ";
    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();
    String payload = new JSONObject()
        .put("prompt", enclosedPrompt)
        .put("max_tokens_to_sample", 200)
        .put("temperature", 0.5)
        .put("stop_sequences", List.of("\n\nHuman:"))
        .toString();
    InvokeModelRequest request = InvokeModelRequest.builder()
        .body(SdkBytes.fromUtf8String(payload))
        .modelId(claudeModelId)
        .contentType("application/json")
        .accept("application/json")
        .build();
    InvokeModelResponse response = client.invokeModel(request);
    JSONObject responseBody = new JSONObject(response.body().asUtf8String());
Invoke the Anthropic Claude 2 foundation model to generate text.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "../../config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} ResponseContent
 * @property {string} text
 */

/**
 * @typedef {Object} MessagesResponseBody
 * @property {ResponseContent[]} content
 */

/**
 * @typedef {Object} TextCompletionsResponseBody
 */
```
* @property {completion} text
*/

/**
 * Invokes Anthropic Claude 2.x using the Messages API.
 * To learn more about the Anthropic Messages API, go to:
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "anthropic.claude-v2".
 */
export const invokeModel = async (prompt, modelId = "anthropic.claude-v2") => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the Messages API request.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response(s)
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {MessagesResponseBody} */
  const responseBody = JSON.parse(decodedResponseBody);
  return responseBody.content[0].text;
};
/**
 * Invokes Anthropic Claude 2.x using the Text Completions API.
 * To learn more about the Anthropic Text Completions API, go to:
 * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
anthropic-claude-text-completion.html
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to
 * "anthropic.claude-v2".
 */
export const invokeTextCompletionsApi = async (prompt,
    modelId = "anthropic.claude-v2",
) => {
    // Create a new Bedrock Runtime client instance.
    const client = new BedrockRuntimeClient({ region: "us-east-1" });

    // Prepare the payload for the Text Completions API, using the required prompt
template.
    const enclosedPrompt = `Human: ${prompt}\n\nAssistant:`;
    const payload = {
        prompt: enclosedPrompt,
        max_tokens_to_sample: 500,
        temperature: 0.5,
        stop_sequences: ["\n\nHuman:"],
    };

    // Invoke Claude with the payload and wait for the response.
    const command = new InvokeModelCommand({
        contentType: "application/json",
        body: JSON.stringify(payload),
        modelId,
    });
    const apiResponse = await client.send(command);

    // Decode and return the response.
    const decoded = new TextDecoder().decode(apiResponse.body);
    /** @type {TextCompletionsResponseBody} */
    const responseBody = JSON.parse(decoded);
    return responseBody.completion;
}
// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt =
        'Complete the following in one sentence: "Once upon a time..."';
    const modelId = FoundationModels.CLAUDE_2.modelId;
    console.log('Prompt: ${prompt}');
    console.log('Model ID: ${modelId}');

    try {
        console.log('-'.repeat(53));
        console.log('Using the Messages API:');
        const response = await invokeModel(prompt, modelId);
        console.log(response);
    } catch (err) {
        console.log(err);
    }

    try {
        console.log('-'.repeat(53));
        console.log('Using the Text Completions API:');
        const response = await invokeTextCompletionsApi(prompt, modelId);
        console.log(response);
    } catch (err) {
        console.log(err);
    }
}

- For API details, see InvokeModel in AWS SDK for JavaScript API Reference.

PHP

SDK for PHP

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Anthropic Claude 2 foundation model to generate text.
public function invokeClaude($prompt)
{
    # The different model providers have individual request and response formats.
    # For the format, ranges, and default values for Anthropic Claude, refer to:
    # https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-claude.html

    $completion = "";

    try {
        $modelId = 'anthropic.claude-v2';

        # Claude requires you to enclose the prompt as follows:
        $prompt = "

Human: {$prompt}

Assistant:";

        $body = [
            'prompt' => $prompt,
            'max_tokens_to_sample' => 200,
            'temperature' => 0.5,
            'stop_sequences' => ['

Human:"],
        ];

        $result = $this->bedrockRuntimeClient->invokeModel([%
            'contentType' => 'application/json',
            'body' => json_encode($body),
            'modelId' => $modelId,
        ]);%

        $response_body = json_decode($result['body']);

        $completion = $response_body->completion;
    } catch (Exception $e) {
        echo "Error: ({$e->getCode()}) - {$e->getMessage()}
";
    }

    return $completion;
}

• For API details, see InvokeModel in AWS SDK for PHP API Reference.
Invoke the Anthropic Claude 2 foundation model to generate text.

```python
def invoke_claude(self, prompt):
    
    """
    Invokes the Anthropic Claude 2 model to run an inference using the input
    provided in the request body.

    :param prompt: The prompt that you want Claude to complete.
    :return: Inference response from the model.
    """

    try:
        # The different model providers have individual request and response
        # formats.  
        # For the format, ranges, and default values for Anthropic Claude,
        # refer to: 
        # https://docs.aws.amazon.com/bedrock/latest/userguide/model-
        # parameters-claude.html

        # Claude requires you to enclose the prompt as follows:
        enclosed_prompt = "Human: " + prompt + "\n\nAssistant:"

        body = {
            "prompt": enclosed_prompt,
            "max_tokens_to_sample": 200,
            "temperature": 0.5,
            "stop_sequences": ["\n\nHuman:"],
        }

        response = self.bedrock_runtime_client.invoke_model(
            modelId="anthropic.claude-v2", body=json.dumps(body)
        )
    
```
response_body = json.loads(response["body"].read())
completion = response_body["completion"]

return completion

except ClientError:
    logger.error("Couldn't invoke Anthropic Claude")
raise

• For API details, see InvokeModel in AWS SDK for Python (Boto3) API Reference.

SAP ABAP

SDK for SAP ABAP

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Anthropic Claude 2 foundation model to generate text. This example uses features of /US2/CL_JSON which might not be available on some NetWeaver versions.

"Claude V2 Input Parameters should be in a format like this:
*   {
*     "prompt":"\n\nHuman:Tell me a joke
Assistant:\n",
*     "max_tokens_to_sample":2048,
*     "temperature":0.5,
*     "top_k":250,
*     "top_p":1.0,
*     "stop_sequences":[]
*   }

DATA: BEGIN OF ls_input,
    prompt               TYPE string,
    max_tokens_to_sample TYPE /aws1/rt_shape_integer,
    temperature          TYPE /aws1/rt_shape_float,
    top_k                TYPE /aws1/rt_shape_integer,
    top_p                TYPE /aws1/rt_shape_float,
stop_sequences       TYPE /aws1/rt_stringtab,
END OF ls_input.

"Leave ls_input-stop_sequences empty.
ls_input-prompt = |\n\n\nHuman:\n{ iv_prompt }\n\nAssistant:\n|.
ls_input-max_tokens_to_sample = 2048.
ls_input-temperature = '0.5'.
ls_input-top_k = 250.
ls_input-top_p = 1.

"Serialize into JSON with /ui2/cl_json -- this assumes SAP_UI is installed.
DATA(lv_json) = /ui2/cl_json=>serialize(  
data = ls_input   
pretty_name   = /ui2/cl_json=>pretty_mode-low_case ).

TRY.
DATA(lo_response) = lo_bdr->invokemodel(  
iv_body = /aws1/cl_rt_util=>string_to_xstring( lv_json )  
iv_modelid = 'anthropic.claude-v2'  
iv_accept = 'application/json'  
iv_contenttype = 'application/json' ).

"Claude V2 Response format will be:  
*   {  
*     "completion": "Knock Knock...",  
*     "stop_reason": "stop_sequence"  
*   }
DATA: BEGIN OF ls_response,  
     completion  TYPE string,  
     stop_reason TYPE string,  
END OF ls_response.

/ui2/cl_json=>deserialize(  
  EXPORTING jsonx = lo_response->get_body( )  
  pretty_name = /ui2/cl_json=>pretty_mode-camel_case  
  CHANGING data  = ls_response ).

DATA(lv_answer) = ls_response-completion.
CATCH /aws1/cx_bdraccessdeniedex INTO DATA(lo_ex).
  WRITE / lo_ex->get_text( ).
  WRITE / |Don't forget to enable model access at https://console.aws.amazon.com/bedrock/home#/modelaccess|.
ENDTRY.
Invoke the Anthropic Claude 2 foundation model to generate text using L2 high level client.

```
TRY.
  DATA(lo_bdr_l2_claude) = /aws1/
  cl_bdr_l2_factory=>create_claude_2( lo_bdr ).
  " iv_prompt can contain a prompt like 'tell me a joke about Java
    programmers'.
  DATA(lv_answer) = lo_bdr_l2_claude->prompt_for_text( iv_prompt ).
  CATCH /aws1/cx_bdraccessdeniedex INTO DATA(lo_ex).
  WRITE / lo_ex->get_text( ).
  WRITE / |Don't forget to enable model access at https://
    console.aws.amazon.com/bedrock/home?#/modelaccess|.
ENDTRY.
```

- For API details, see [InvokeModel](http://aws1/) in *AWS SDK for SAP ABAP API reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](http://aws1/). This topic also includes information about getting started and details about previous SDK versions.

**Invoke Anthropic Claude 3 on Amazon Bedrock to generate text**

The following code examples show how to invoke Anthropic Claude 3 on Amazon Bedrock to generate text.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- [Interactive text playground](http://aws1/)
- [Invoke multiple foundation models on Amazon Bedrock](http://aws1/)
- [Invoke multiple LLMs on Amazon Bedrock](http://aws1/)
Invoke Anthropic Claude 3 to generate text.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "../config/foundation_models.js";
import {
  BedrockRuntimeClient,
  InvokeModelCommand,
  InvokeModelWithResponseStreamCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} ResponseContent
 * @property {string} text
 *
 * @typedef {Object} MessagesResponseBody
 * @property {ResponseContent[]} content
 *
 * @typedef {Object} Delta
 * @property {string} text
 *
 * @typedef {Object} Message
 * @property {string} role
 *
 * @typedef {Object} Chunk
 * @property {string} type
 * @property {Delta} delta
 * @property {Message} message
 */
/**
 * Invokes Anthropic Claude 3 using the Messages API.
 *
 * To learn more about the Anthropic Messages API, go to:
 * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
 * anthropic-claude-messages.html
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to
 * "anthropic.claude-3-haiku-20240307-v1:0".
 */

export const invokeModel = async (
  prompt,
  modelId = "anthropic.claude-3-haiku-20240307-v1:0",
) => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response(s)
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {MessagesResponseBody} */
  const responseBody = JSON.parse(decodedResponseBody);
  return responseBody.content[0].text;
};
/**
 * Invokes Anthropic Claude 3 and processes the response stream.
 *
 * To learn more about the Anthropic Messages API, go to:
 * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
 * anthropic-claude-messages.html
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to
 * "anthropic.claude-3-haiku-20240307-v1:0".
 */
export const invokeModelWithResponseStream = async (prompt, modelId = "anthropic.claude-3-haiku-20240307-v1:0") => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the API to respond.
  const command = new InvokeModelWithResponseStreamCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  let completeMessage = "";

  // Decode and process the response stream
  for await (const item of apiResponse.body) {
    /** @type Chunk */
const chunk = JSON.parse(new TextDecoder().decode(item.chunk.bytes));
const chunk_type = chunk.type;

if (chunk_type === "content_block_delta") {
    const text = chunk.delta.text;
    completeMessage = completeMessage + text;
    process.stdout.write(text);
}

// Return the final response
return completeMessage;

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt = 'Write a paragraph starting with: "Once upon a time..."';
    const modelId = FoundationModels.CLAUDE_3_HAIKU.modelId;
    console.log(`Prompt: ${prompt}`);
    console.log(`Model ID: ${modelId}`);

    try {
        console.log("-".repeat(53));
        const response = await invokeModel(prompt, modelId);
        console.log("\n" + "-".repeat(53));
        console.log("Final structured response:");
        console.log(response);
    } catch (err) {
        console.log(`\n${err}`);
    }
}

• For API details, see InvokeModel in AWS SDK for JavaScript API Reference.
Invoke Anthropic Claude 3 to generate text.

```python
def invoke_claude_3_with_text(self, prompt):
    
    """
    Invokes Anthropic Claude 3 Sonnet to run an inference using the input provided in the request body.
    
    :param prompt: The prompt that you want Claude 3 to complete.
    :return: Inference response from the model.
    """
    
    # Initialize the Amazon Bedrock runtime client
    client = self.client or boto3.client(
        service_name="bedrock-runtime", region_name="us-east-1"
    )

    # Invoke Claude 3 with the text prompt
    model_id = "anthropic.claude-3-sonnet-20240229-v1:0"

    try:
        response = client.invoke_model(
            modelId=model_id,
            body=json.dumps(
                {
                    "anthropic_version": "bedrock-2023-05-31",
                    "max_tokens": 1024,
                    "messages": [
                        {
                            "role": "user",
                            "content": [["type": "text", "text": prompt]],
                        }
                    ],
                }
            )
        )
```
For API details, see InvokeModel in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Invoke the Anthropic Claude Instant model on Amazon Bedrock for text generation**

The following code example shows how to invoke the Anthropic Claude Instant model on Amazon Bedrock for text generation.
Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- Interactive text playground
- Invoke multiple foundation models on Amazon Bedrock
- Invoke multiple LLMs on Amazon Bedrock

JavaScript

**SDK for JavaScript (v3)**

```
import { fileURLToPath } from "url";
import { FoundationModels } from "./config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} Content
 * @property {string} text
 *
 * @typedef {Object} MessageApiResponse
 * @property {Content[]} content
 */

/**
 * @typedef {Object} TextCompletionApiResponse
 * @property {string} completion
```
* Invokes Anthropic Claude Instant using the Messages API.
* To learn more about the Anthropic Messages API, go to:
* @param {string} prompt - The input text prompt for the model to complete.
* @param {string} [modelId] - The ID of the model to use. Defaults to "anthropic.claude-instant-v1".
*/

export const invokeModel = async (prompt, modelId = "anthropic.claude-instant-v1") => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the Messages API request.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response(s)
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {MessageApiResponse} */
  const responseBody = JSON.parse(decodedResponseBody);
/**
* Invokes Anthropic Claude Instant using the Text Completions API.
*
* To learn more about the Anthropic Text Completions API, go to:
* https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
* anthropic-claude-text-completion.html
*
* @param {string} prompt - The input text prompt for the model to complete.
* @param {string} [modelId] - The ID of the model to use. Defaults to
* "anthropic.claude-instant-v1".
*/
export const invokeTextCompletionsApi = async (prompt,
modelId = "anthropic.claude-instant-v1",
) => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the Text Completions API, using the required prompt
  template.
  const enclosedPrompt = `Human: ${prompt}\n\nAssistant:`;
  const payload = {
    prompt: enclosedPrompt,
    max_tokens_to_sample: 500,
    temperature: 0.5,
    stop_sequences: ["\n\nHuman:"],
  };

  // Invoke Claude with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response.
  const decoded = new TextDecoder().decode(apiResponse.body);
  /** @type {TextCompletionApiResponse} */
  const responseBody = JSON.parse(decoded);
  return responseBody.completion;
// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt = 
        'Complete the following in one sentence: "Once upon a time..."';
    const modelId = FoundationModels.CLAUDE_INSTANT.modelId;
    console.log('Prompt: ${prompt}');
    console.log('Model ID: ${modelId}');

    try {
        console.log('-'.repeat(53));
        console.log('Using the Messages API:');
        const response = await invokeModel(prompt, modelId);
        console.log(response);
    } catch (err) {
        console.log(err);
    }

    try {
        console.log('-'.repeat(53));
        console.log('Using the Text Completions API:');
        const response = await invokeTextCompletionsApi(prompt, modelId);
        console.log(response);
    } catch (err) {
        console.log(err);
    }
}

- For API details, see InvokeModel in AWS SDK for JavaScript API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Invoke Anthropic Claude on Amazon Bedrock for text generation with a response stream

The following code examples show how to invoke Anthropic Claude model on Amazon Bedrock for text generation with a response stream.
Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- **Invoke multiple foundation models on Amazon Bedrock**

**.NET**

**AWS SDK for .NET**

```
/// <summary>
/// Asynchronously invokes the Anthropic Claude 2 model to run an inference based on the provided input and process the response stream.
/// </summary>
/// <param name="prompt">The prompt that you want Claude to complete.</param>
/// <returns>The inference response from the model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
/// For the format, ranges, and default values for Anthropic Claude, refer to: https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-claude.html
/// </remarks>
public static async IAsyncEnumerable<string> InvokeClaudeWithResponseStreamAsync(string prompt, [EnumeratorCancellation] CancellationToken cancellationToken = default)
{
    string claudeModelId = "anthropic.claude-v2";
    string enclosedPrompt = "Human: " + prompt + "\n\nAssistant:";
    // Claude requires you to enclose the prompt as follows:
    string enclosedPrompt = "Human: " + prompt + "\n\nAssistant:";
```
```csharp
AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);

string payload = new JsonObject()
{
    { "prompt", enclosedPrompt },
    { "max_tokens_to_sample", 200 },
    { "temperature", 0.5 },
    { "stop_sequences", new JsonArray("\n\nHuman:" ) }
}.ToJsonString();

InvokeModelWithResponseStreamResponse? response = null;

try
{
    response = await client.InvokeModelWithResponseStreamAsync(new InvokeModelWithResponseStreamRequest()
    {
        ModelId = claudeModelId,
        Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
        ContentType = "application/json",
        Accept = "application/json"
    });
}
catch (AmazonBedrockRuntimeException e)
{
    Console.WriteLine(e.Message);
}

{
    // create a buffer to write the event in to move from a push mode
to a pull mode
    Channel<string> buffer = Channel.CreateUnbounded<string>();
    bool isStreaming = true;

    response.Body.ChunkReceived += BodyOnChunkReceived;
    response.Body.StartProcessing();

    while (!cancellationToken.IsCancellationRequested && isStreaming) || (!cancellationToken.IsCancellationRequested && buffer.Reader.Count > 0))
    {
```
// pull the completion from the buffer and add it to the
IAsyncEnumerable collection
yield return await
buffer.Reader.ReadAsync(cancellationToken);
}
response.Body.ChunkReceived -= BodyOnChunkReceived;

yield break;

// handle the ChunkReceived events
async void BodyOnChunkReceived(object? sender,
EventStreamEventReceivedArgs<PayloadPart> e)
{
    var streamResponse =
    JsonSerializer.Deserialize<JsonObject>(e.EventStreamEvent.Bytes) ??
    throw new NullReferenceException($"Unable to deserialize
{nameof(e.EventStreamEvent.Bytes)}");

    if (streamResponse["stop_reason"].GetValue<string>!() !=
null)
    {
        isStreaming = false;
    }

    // write the received completion chunk into the buffer
    await
    buffer.Writer.WriteAsync(streamResponse["completion"].GetValue<string>!(),
cancellationToken);
}
else if (response is not null)
{
    Console.WriteLine("InvokeModelAsync failed with status code " +
    response.HttpStatusCode);
}

yield break;
}

• For API details, see InvokeModelWithResponseStream in AWS SDK for .NET API Reference.
Amazon Bedrock

User Guide

Go
SDK for Go V2
Note
There's more on GitHub. Find the complete example and learn how to set up and run
in the AWS Code Examples Repository.

Invoke Anthropic Claude and process the response stream.

// Each model provider defines their own individual request and response formats.
// For the format, ranges, and default values for the different models, refer to:
// https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters.html
type Request struct {
Prompt
string `json:"prompt"`
MaxTokensToSample int
`json:"max_tokens_to_sample"`
Temperature
float64 `json:"temperature,omitempty"`
}
type Response struct {
Completion string `json:"completion"`
}
// Invokes Anthropic Claude on Amazon Bedrock to run an inference and
asynchronously
// process the response stream.
func (wrapper InvokeModelWithResponseStreamWrapper)
InvokeModelWithResponseStream(prompt string) (string, error) {
modelId := "anthropic.claude-v2"
// Anthropic Claude requires you to enclose the prompt as follows:
prefix := "Human: "
postfix := "\n\nAssistant:"
prompt = prefix + prompt + postfix
request := ClaudeRequest{
Prompt:
prompt,
Actions

696


MaxTokensToSample: 200,
Temperature: 0.5,
StopSequences: []string{"\n\nHuman:"},
}

body, err := json.Marshal(request)
if err != nil {
    log.Panicln("Couldn't marshal the request: ", err)
}

output, err :=
wrapper.BedrockRuntimeClient.InvokeModelWithResponseStream(context.Background(),
&bedrockruntime.InvokeModelWithResponseStreamInput{
    Body:        body,
    ModelId:     aws.String(modelId),
    ContentType: aws.String("application/json"),
})
if err != nil {
    errMsg := err.Error()
    if strings.Contains(errMsg, "no such host") {
        log.Printf("The Bedrock service is not available in the selected region.
Please double-check the service availability for your region at https://
aws.amazon.com/about-aws/global-infrastructure/regional-product-services/\n")
    } else if strings.Contains(errMsg, "Could not resolve the foundation model") {
        log.Printf("Could not resolve the foundation model from model identifier: "%v \n". Please verify that the requested model exists and is accessible within the
specified region.\n", modelId)
    } else {
        log.Printf("Couldn't invoke Anthropic Claude. Here's why: %v\n", errMsg)
    }
}

resp, err := processStreamingOutput(output, func(ctx context.Context, part []byte) error {
    fmt.Print(string(part))
    return nil
})
if err != nil {
    log.Fatal("streaming output processing error: ", err)
}

return resp.Completion, nil
type StreamingOutputHandler func(ctx context.Context, part []byte) error

func processStreamingOutput(output *bedrockruntime.InvokeModelWithResponseStreamOutput, handler StreamingOutputHandler) (Response, error) {

    var combinedResult string
    resp := Response{}

    for event := range output.GetStream().Events() {
        switch v := event.(type) {
            case *types.ResponseStreamMemberChunk:

                var resp Response
                if err != nil {
                    return resp, err
                }

                err = handler(context.Background(), []byte(resp.Completion))
                if err != nil {
                    return resp, err
                }

                combinedResult += resp.Completion
            case *types.UnknownUnionMember:
                fmt.Println("unknown tag:", v.Tag)
            default:
                fmt.Println("union is nil or unknown type")
        }
    }

    resp.Completion = combinedResult

    return resp, nil
}
- For API details, see `InvokeModelWithResponseStream` in *AWS SDK for Go API Reference*.

Java

### SDK for Java 2.x

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

Invoke Anthropic Claude and process the response stream.

```java
/**
 * Invokes the Anthropic Claude 2 model and processes the response stream.
 * @param prompt The prompt for Claude to complete.
 * @param silent Suppress console output of the individual response stream chunks.
 * @return The generated response.
 */
public static String invokeClaude(String prompt, boolean silent) {

    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    var finalCompletion = new AtomicReference<"">();
    var payload = new JSONObject()
        .put("prompt", "Human: " + prompt + "Assistant:"),
        .put("temperature", 0.8)
        .put("max_tokens_to_sample", 300);
var request = InvokeModelWithResponseStreamRequest.builder()
    .body(SdkBytes.fromUtf8String(payload))
    .modelId("anthropic.claude-v2")
    .contentType("application/json")
    .accept("application/json")
    .build();

var visitor = 
    InvokeModelWithResponseStreamResponseHandler.Visitor.builder()
    .onChunk(chunk -> {
        var json = new 
            JSONObject(chunk.bytes().asUtf8String());
        var completion = 
            json.getString("completion");
        finalCompletion.set(finalCompletion.get() + completion);
        if (!silent) {
            System.out.print(completion);
        }
    })
    .build();

var handler = 
    InvokeModelWithResponseStreamResponseHandler.builder()
    .onEventStream(stream -> stream.subscribe(event -
        > event.accept(visitor)))
    .onComplete(() -> {
    })
    .onError(e -> System.out.println("\n\nError: " +
        e.getMessage()))
    .build();

client.invokeModelWithResponseStream(request, handler).join();
return finalCompletion.get();

• For API details, see InvokeModelWithResponseStream in AWS SDK for Java 2.x API Reference.
Invoke Anthropic Claude and process the response stream.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "./config/foundation_models.js";
import {
  BedrockRuntimeClient,
  InvokeModelCommand,
  InvokeModelWithResponseStreamCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} ResponseContent
 * @property {string} text
 *
 * @typedef {Object} MessagesResponseBody
 * @property {ResponseContent[]} content
 *
 * @typedef {Object} Delta
 * @property {string} text
 *
 * @typedef {Object} Message
 * @property {string} role
 *
 * @typedef {Object} Chunk
 * @property {string} type
 * @property {Delta} delta
 * @property {Message} message
 */
```
/**
 * Invokes Anthropic Claude 3 using the Messages API.
 *
 * To learn more about the Anthropic Messages API, go to:
 * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
 * anthropic-claude-messages.html
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to
 * "anthropic.claude-3-haiku-20240307-v1:0".
 */
export const invokeModel = async (prompt, modelId = "anthropic.claude-3-haiku-20240307-v1:0",)
=> {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response(s)
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {MessagesResponseBody} */
  const responseBody = JSON.parse(decodedResponseBody);
  return responseBody.content[0].text;
};
```javascript
/**
 * Invokes Anthropic Claude 3 and processes the response stream.
 * To learn more about the Anthropic Messages API, go to:
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to
 * "anthropic.claude-3-haiku-20240307-v1:0".
 */
export const invokeModelWithResponseStream = async (prompt, modelId = "anthropic.claude-3-haiku-20240307-v1:0") => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    anthropic_version: "bedrock-2023-05-31",
    max_tokens: 1000,
    messages: [
      {
        role: "user",
        content: [{ type: "text", text: prompt }],
      },
    ],
  };

  // Invoke Claude with the payload and wait for the API to respond.
  const command = new InvokeModelWithResponseStreamCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);
  let completeMessage = "";

  // Decode and process the response stream
  for await (const item of apiResponse.body) {
    /** @type Chunk */
```
const chunk = JSON.parse(new TextDecoder().decode(item.chunk.bytes));
const chunk_type = chunk.type;

if (chunk_type === "content_block_delta") {
    const text = chunk.delta.text;
    completeMessage = completeMessage + text;
    process.stdout.write(text);
}

// Return the final response
return completeMessage;

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt = 'Write a paragraph starting with: "Once upon a time..."';
    const modelId = FoundationModels.CLAUDE_3_HAIKU.modelId;
    console.log(`Prompt: ${prompt}`);
    console.log(`Model ID: ${modelId}`);

    try {
        console.log("-".repeat(53));
        const response = await invokeModel(prompt, modelId);
        console.log("\n" + "-".repeat(53));
        console.log("Final structured response: ");
        console.log(response);
    } catch (err) {
        console.log(`\n${err}`);
    }
}

- For API details, see InvokeModelWithResponseStream in AWS SDK for JavaScript API Reference.
Invoke Anthropic Claude and process the response stream.

```python
async def invoke_model_with_response_stream(self, prompt):
    """
    Invokes the Anthropic Claude 2 model to run an inference and process the response stream.

    :param prompt: The prompt that you want Claude to complete.
    :return: Inference response from the model.
    """
    try:
        # The different model providers have individual request and response formats.
        # For the format, ranges, and default values for Anthropic Claude, refer to:
        # https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-claude.html

        # Claude requires you to enclose the prompt as follows:
        enclosed_prompt = "Human: " + prompt + "\n\nAssistant:"

        body = {
            "prompt": enclosed_prompt,
            "max_tokens_to_sample": 1024,
            "temperature": 0.5,
            "stop_sequences": ["\n\nHuman:"],
        }

        response =
        self.bedrock_runtime_client.invoke_model_with_response_stream(
            modelId="anthropic.claude-v2", body=json.dumps(body)
        )
```

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
for event in response.get("body"):
    chunk = json.loads(event["chunk"]["bytes"])["completion"]
    yield chunk

except ClientError:
    logger.error("Couldn't invoke Anthropic Claude v2")
    raise

• For API details, see InvokeModelWithResponseStream in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Invoke the Meta Llama 2 Chat model on Amazon Bedrock for text generation

The following code examples show how to invoke the Meta Llama 2 Chat model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

• Interactive text playground
• Invoke multiple foundation models on Amazon Bedrock
• Invoke multiple LLMs on Amazon Bedrock

.AWS SDK for .NET

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
Asynchronously invoke the Meta Llama 2 foundation model to generate text.

```csharp
/// <summary>
/// Asynchronously invokes the Meta Llama 2 Chat model to run an inference based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that you want Llama 2 to complete.</param>
/// <returns>The inference response from the model</returns>
/// <remarks>
/// The different model providers have individual request and response formats.
/// For the format, ranges, and default values for Meta Llama 2 Chat, refer to:
/// </remarks>
public static async Task<string> InvokeLlama2Async(string prompt)
{
    string llama2ModelId = "meta.llama2-13b-chat-v1";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USEast1);

    string payload = new JsonObject()
    {
        { "prompt", prompt },
        { "max_gen_len", 512 },
        { "temperature", 0.5 },
        { "top_p", 0.9 }
    }.ToJsonString();

    string generatedText = "";
    try
    {
        InvokeModelResponse response = await client.InvokeModelAsync(new InvokeModelRequest()
        {
            ModelId = llama2ModelId,
            Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
            ContentType = "application/json",
            Accept = "application/json"
        });
    }
```
{
    return JsonNode.ParseAsync(response.Body)
        .Result?["generation"]?.GetValue<string>() ?? "";
}
else
{
    Console.WriteLine("InvokeModelAsync failed with status code " + response.HttpStatusCode);
}
catch (AmazonBedrockRuntimeException e)
{
    Console.WriteLine(e.Message);
}
return generatedText;

• For API details, see InvokeModel in AWS SDK for .NET API Reference.

Go

SDK for Go V2

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Meta Llama 2 Chat foundation model to generate text.

// Each model provider has their own individual request and response formats.
// For the format, ranges, and default values for Meta Llama 2 Chat, refer to:

type llama2Request struct {
    Prompt string `json:"prompt"`
MaxGenLength    int    `json:"max_gen_len,omitempty"`
Temperature       float64  `json:"temperature,omitempty"

}  


type Llama2Response struct {
    Generation string `json:"generation"
}

// Invokes Meta Llama 2 Chat on Amazon Bedrock to run an inference using the input
// provided in the request body.
func (wrapper InvokeModelWrapper) InvokeLlama2(prompt string) (string, error) {
    modelId := "meta.llama2-13b-chat-v1"

    body, err := json.Marshal(Llama2Request {
        Prompt:            prompt,
        MaxGenLength:    512,
        Temperature:       0.5,
    })

    if err != nil { log.Fatal("failed to marshal", err) }

    output, err := wrapper.BedrockRuntimeClient.InvokeModel(context.TODO(),
        &bedrockruntime.InvokeModelInput{
            ModelId: aws.String(modelId),
            ContentType: aws.String("application/json"),
            Body: body,
        })

    if err != nil { ProcessError(err, modelId) }

    var response Llama2Response
    if err := json.Unmarshal(output.Body, &response); err != nil {
        log.Fatal("failed to unmarshal", err)
    }

    return response.Generation, nil
}

• For API details, see [InvokeModel](https://aws.amazon.com/sdk-for-go/api-reference/) in [AWS SDK for Go API Reference](https://aws.amazon.com/sdk-for-go/api-reference/).
Asynchronously invoke the Meta Llama 2 Chat foundation model to generate text.

```java
/**
 * Asynchronously invokes the Meta Llama 2 Chat model to run an inference
 * based
 * on the provided input.
 * *
 * @param prompt The prompt that you want Llama 2 to complete.
 * @return The inference response generated by the model.
 */
public static String invokeLlama2(String prompt) {
    /*
    * The different model providers have individual request and response
    * formats.
    * For the format, ranges, and default values for Meta Llama 2 Chat,
    * refer to:
    * https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-
    * meta.
    *
    */

    String llama2ModelId = "meta.llama2-13b-chat-v1";

    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    String payload = new JSONObject()
        .put("prompt", prompt)
        .put("max_gen_len", 512)
        .put("temperature", 0.5)
        .put("top_p", 0.9)
```

[Note: There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.]
Invoke the Meta Llama 2 Chat foundation model to generate text.

```java
InvokeModelRequest request = InvokeModelRequest.builder()
    .body(SdkBytes.fromUtf8String(payload))
    .modelId(llama2ModelId)
    .contentType("application/json")
    .accept("application/json")
    .build();

CompletableFuture<InvokeModelResponse> completableFuture =
    client.invokeModel(request)
    .whenComplete((response, exception) -> {
        if (exception != null) {
            System.out.println("Model invocation failed: " +
                exception);
        }
    });

String generatedText = "";
try {
    InvokeModelResponse response = completableFuture.get();
    JSONObject responseBody = new
        JSONObject(response.body().asUtf8String());
    generatedText = responseBody.getString("generation");
} catch (InterruptedException e) {
    Thread.currentThread().interrupt();
    System.err.println(e.getMessage());
} catch (ExecutionException e) {
    System.err.println(e.getMessage());
}

return generatedText;
```

/**
 * Invokes the Meta Llama 2 Chat model to run an inference based on the
 * provided
 * input.
 */
* @param prompt The prompt for Llama 2 to complete.
* @return The generated response.
* /

public static String invokeLlama2(String prompt) {

    String llama2ModelId = "meta.llama2-13b-chat-v1";

    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_EAST_1)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    String payload = new JSONObject()
        .put("prompt", prompt)
        .put("max_gen_len", 512)
        .put("temperature", 0.5)
        .put("top_p", 0.9)
        .toString();

    InvokeModelRequest request = InvokeModelRequest.builder()
        .body(SdkBytes.fromUtf8String(payload))
        .modelId(llama2ModelId)
        .contentType("application/json")
        .accept("application/json")
        .build();

    InvokeModelResponse response = client.invokeModel(request);

    JSONObject responseBody = new JSONObject(response.body().asUtf8String());
    String generatedText = responseBody.getString("generation");

    return generatedText;
• For API details, see [InvokeModel](#) in *AWS SDK for Java 2.x API Reference*.

**JavaScript**

**SDK for JavaScript (v3)**

ℹ️ **Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

Invoke the Meta Llama 2 Chat foundation model to generate text.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "../../config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} ResponseBody
 * @property {generation} text
 */

/**
 * Invokes a Meta Llama 2 Chat model.
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "meta.llama2-13b-chat-v1".
 */
export const invokeModel = async (prompt,
```
modelId = "meta.llama2-13b-chat-v1",
) => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Prepare the payload for the model.
  const payload = {
    prompt,
    temperature: 0.5,
    max_gen_len: 1000,
  };

  // Invoke the model with the payload and wait for the response.
  const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
  });
  const apiResponse = await client.send(command);

  // Decode and return the response.
  const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
  /** @type {ResponseBody} */
  const responseBody = JSON.parse(decodedResponseBody);
  return responseBody.generation;
}

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  const prompt =
    'Complete the following in one sentence: "Once upon a time..."';
  const modelId = FoundationModels.LLAMA2_CHAT_13B.modelId;
  console.log(`Prompt: ${prompt}`);
  console.log(`Model ID: ${modelId}`);

  try {
    console.log("-".repeat(53));
    const response = await invokeModel(prompt, modelId);
    console.log(response);
  } catch (err) {
    console.log(err);
  }
}
For API details, see InvokeModel in AWS SDK for JavaScript API Reference.

PHP

**SDK for PHP**

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Meta Llama 2 Chat foundation model to generate text.

```php
public function invokeLlama2($prompt)
{
    # The different model providers have individual request and response formats.
    # For the format, ranges, and default values for Meta Llama 2 Chat, refer to:
    # https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-meta.html

    $completion = "";

    try {
        $modelId = 'meta.llama2-13b-chat-v1';

        $body = [
            'prompt' => $prompt,
            'temperature' => 0.5,
            'max_gen_len' => 512,
        ];

        $result = $this->bedrockRuntimeClient->invokeModel(["contentType" => 'application/json',
            'body' => json_encode($body),
            'modelId' => $modelId,
        ]);
$response_body = json_decode($result['body']);

$completion = $response_body->generation;
} catch (Exception $e) {
    echo "Error: ({$e->getCode()}) - {$e->getMessage()}\n";
}

return $completion;

• For API details, see InvokeModel in AWS SDK for PHP API Reference.

Python

SDK for Python (Boto3)

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Meta Llama 2 Chat foundation model to generate text.

def invoke_llama2(self, prompt):
    """
    Invokes the Meta Llama 2 large-language model to run an inference using the input provided in the request body.
    
    :param prompt: The prompt that you want Llama 2 to complete.
    :return: Inference response from the model.
    """

    try:
        # The different model providers have individual request and response formats.
        # For the format, ranges, and default values for Meta Llama 2 Chat, refer to:
        # https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-meta.html
body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}

response = self.bedrock_runtime_client.invoke_model(
    modelId="meta.llama2-13b-chat-v1", body=json.dumps(body)
)

response_body = json.loads(response["body"].read())
completion = response_body["generation"]

return completion

except ClientError:
    logger.error("Couldn't invoke Llama 2")
    raise

- For API details, see [InvokeModel](#) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Invoke the Mistral 7B model on Amazon Bedrock for text generation**

The following code examples show how to invoke the Mistral 7B model model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- [Interactive text playground](#)
- [Invoke multiple foundation models on Amazon Bedrock](#)
- [Invoke multiple LLMs on Amazon Bedrock](#)
Asynchronously invoke the Mistral 7B foundation model to generate text.

```csharp
public static async Task<List<string?>> InvokeMistral7BAsync(string prompt)
{
    string mistralModelId = "mistral.mistral-7b-instruct-v0:2";

    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USWest2);

    string payload = new JsonObject()
    {
        { "prompt", prompt },
        { "max_tokens", 200 },
        { "temperature", 0.5 }
    }.ToJsonString();

    List<string?>? generatedText = null;
    try
```
```csharp
{  
InvokeModelResponse response = await client.InvokeModelAsync(new  
InvokeModelRequest()  
{  
ModelId = mistralModelId,  
Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),  
ContentType = "application/json",  
Accept = "application/json"  
});

{  
var results = JsonNode.ParseAsync(response.Body).Result?  
["outputs"]?.AsArray();  

generatedText = results?.Select(x => x?  
["text"]?.GetValue<string?>()?).ToList();  
}  
else  
{  
    Console.WriteLine("InvokeModelAsync failed with status code " + response.HttpStatusCode);
}
}
catch (AmazonBedrockRuntimeException e)  
{  
    Console.WriteLine(e.Message);
}
return generatedText ?? [];
}
```

- For API details, see [InvokeModel](#) in [AWS SDK for .NET API Reference](#).
Amazon Bedrock

User Guide

Java
SDK for Java 2.x
Note
There's more on GitHub. Find the complete example and learn how to set up and run
in the AWS Code Examples Repository.

Asynchronously invoke the Mistral 7B foundation model to generate text.
/**
* Asynchronously invokes the Mistral 7B model to run an inference based on
the provided input.
*
* @param prompt The prompt for Mistral to complete.
* @return The generated response.
*/
public static List<String> invokeMistral7B(String prompt) {
BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
.region(Region.US_WEST_2)
.credentialsProvider(ProfileCredentialsProvider.create())
.build();
// Mistral instruct models provide optimal results when
// embedding the prompt into the following template:
String instruction = "<s>[INST] " + prompt + " [/INST]";
String modelId = "mistral.mistral-7b-instruct-v0:2";
String payload = new JSONObject()
.put("prompt", instruction)
.put("max_tokens", 200)
.put("temperature", 0.5)
.toString();
CompletableFuture<InvokeModelResponse> completableFuture =
client.invokeModel(request -> request
.accept("application/json")
.contentType("application/json")
.body(SdkBytes.fromUtf8String(payload))
.modelId(modelId))
Actions

720


Invoke the Mistral 7B foundation model to generate text.

```java
/**
 * Invokes the Mistral 7B model to run an inference based on the provided input.
 * @param prompt The prompt for Mistral to complete.
 * @return The generated responses.
 */
public static List<String> invokeMistral7B(String prompt) {
    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_WEST_2)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    return List.of();
}
```
String instruction = "<s>[INST] " + prompt + " [/INST]";

String modelId = "mistral.mistral-7b-instruct-v0:2";

String payload = new JSONObject()
    .put("prompt", instruction)
    .put("max_tokens", 200)
    .put("temperature", 0.5)
    .toString();

InvokeModelResponse response = client.invokeModel(request ->
    request
        .accept("application/json")
        .contentType("application/json")
        .body(SdkBytes.fromUtf8String(payload))
        .modelId(modelId));

JSONObject responseBody = new JSONObject(response.body().asUtf8String());
JSONArray outputs = responseBody.getJSONArray("outputs");

return IntStream.range(0, outputs.length())
    .mapToObj(i ->
        outputs.getJSONObject(i).getString("text"))
    .toList();
}

• For API details, see InvokeModel in AWS SDK for Java 2.x API Reference.

JavaScript

SDK for JavaScript (v3)

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Mistral 7B foundation model to generate text.
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved. 
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "../../config/foundation_models.js";
import {
  BedrockRuntimeClient,
  InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} Output
 * @property {string} text
 *
 * @typedef {Object} ResponseBody
 * @property {Output[]} outputs
 */

/**
 * Invokes a Mistral 7B Instruct model.
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "mistral.mistral-7b-instruct-v0:2".
 */

export const invokeModel = async (prompt, modelId = "mistral.mistral-7b-instruct-v0:2", ) => {
  // Create a new Bedrock Runtime client instance.
  const client = new BedrockRuntimeClient({ region: "us-east-1" });

  // Mistral instruct models provide optimal results when embedding
  // the prompt into the following template:
  const instruction = `\<s\>[INST] ${prompt} [/INST]\``;

  // Prepare the payload.
  const payload = {
    prompt: instruction,
    max_tokens: 500,
    temperature: 0.5,
  };
// Invoke the model with the payload and wait for the response.
const command = new InvokeModelCommand({
    contentType: "application/json",
    body: JSON.stringify(payload),
    modelId,
});
const apiResponse = await client.send(command);

// Decode and return the response.
const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
/** @type {ResponseBody} */
const responseBody = JSON.parse(decodedResponseBody);
return responseBody.outputs[0].text;};

// Invoke the function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt =
        'Complete the following in one sentence: "Once upon a time..."';
    const modelId = FoundationModels.MISTRAL_7B.modelId;
    console.log(`Prompt: ${prompt}`);
    console.log(`Model ID: ${modelId}`);

    try {
        console.log("-".repeat(53));
        const response = await invokeModel(prompt, modelId);
        console.log(response);
    } catch (err) {
        console.log(err);
    }
}

- For API details, see InvokeModel in AWS SDK for JavaScript API Reference.
Invoke the Mistral 7B foundation model to generate text.

```python
def invoke_mistral_7b(self, prompt):
    """
    Invokes the Mistral 7B model to run an inference using the input provided in the request body.

    :param prompt: The prompt that you want Mistral to complete.
    :return: List of inference responses from the model.
    """
    try:
        # Mistral instruct models provide optimal results when embedding the prompt into the following template:
        instruction = f"<s>[INST] {prompt} [/INST]"

        model_id = "mistral.mistral-7b-instruct-v0:2"

        body = {
            "prompt": instruction,
            "max_tokens": 200,
            "temperature": 0.5,
        }

        response = self.bedrock_runtime_client.invoke_model(
            modelId=model_id, body=json.dumps(body)
        )

        response_body = json.loads(response["body"]).read()
        outputs = response_body.get("outputs")

        completions = [output["text"] for output in outputs]
```
For API details, see `InvokeModel` in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](https://docs.aws.amazon.com/sdk-for-python/guide/using-with-aws-sdk.html). This topic also includes information about getting started and details about previous SDK versions.

### Invoke the Mixtral 8x7B model on Amazon Bedrock for text generation

The following code examples show how to invoke the Mixtral 8x7B model model on Amazon Bedrock for text generation.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code examples:

- [Interactive text playground](https://github.com/aws/aws-sdk-python/blob/main/examples/mistral.py)
- [Invoke multiple foundation models on Amazon Bedrock](https://github.com/aws/aws-sdk-python/blob/main/examples/mixtral.py)
- [Invoke multiple LLMs on Amazon Bedrock](https://github.com/aws/aws-sdk-python/blob/main/examples/mixtral.py)

### .NET

#### AWS SDK for .NET

```csharp
/// <summary>
    return completions
except ClientError:
    logger.error("Couldn't invoke Mistral 7B")
raise
```
/// Asynchronously invokes the Mixtral 8x7B model to run an inference
/// based on the provided input.
/// </summary>
/// <param name="prompt">The prompt that you want Mixtral 8x7B to
/// complete.</param>
/// <returns>The inference response from the model</returns>
/// <remarks>
/// The different model providers have individual request and response
/// formats.
/// For the format, ranges, and default values for Mixtral 8x7B, refer
/// to:
/// https://docs.aws.amazon.com/bedrock/latest/userguide/model-
/// parameters-mistral.html
/// </remarks>
public static async Task<List<string?>> InvokeMixtral8x7BAsync(string
prompt)
{
    string mixtralModelId = "mistral.mixtral-8x7b-instruct-v0:1";
    AmazonBedrockRuntimeClient client = new(RegionEndpoint.USWest2);

    string payload = new JsonObject()
    {
        { "prompt", prompt },
        { "max_tokens", 200 },
        { "temperature", 0.5 }
    }.ToJsonString();

    List<string?>? generatedText = null;
    try
    {
        InvokeModelResponse response = await client.InvokeModelAsync(new
InvokeModelRequest()
        {
            ModelId = mixtralModelId,
            Body = AWSSDKUtils.GenerateMemoryStreamFromString(payload),
            ContentType = "application/json",
            Accept = "application/json"
        });

        {
            var results = JsonNode.ParseAsync(response.Body).Result?
["outputs"]?.AsArray();
Asynchronously invoke the Mistral 8x7B foundation model to generate text.

```java
/**
 * Asynchronously invokes the Mistral 8x7B model to run an inference based on
 * the provided input.
 * @param prompt The prompt for Mixtral to complete.
 * @return The generated response.
 */
public static List<String> invokeMixtral8x7B(String prompt) {
    BedrockRuntimeAsyncClient client = BedrockRuntimeAsyncClient.builder()
```
```java
.region(Region.US_WEST_2)
.credentialsProvider(ProfileCredentialsProvider.create())
.build();

// Mistral instruct models provide optimal results when
// embedding the prompt into the following template:
String instruction = "<s>[INST] " + prompt + " [/INST]";

String modelId = "mistral.mixtral-8x7b-instruct-v0:1";

String payload = new JSONObject()
    .put("prompt", instruction)
    .put("max_tokens", 200)
    .put("temperature", 0.5)
    .toString();

CompletableFuture<InvokeModelResponse> completableFuture =
client.invokeModel(request -> request
    .accept("application/json")
    .contentType("application/json")
    .body(SdkBytes.fromUtf8String(payload))
    .modelId(modelId))
    .whenComplete((response, exception) -> {
        if (exception != null) {
            System.out.println("Model invocation failed: " +
                exception);
        }
    });

try {
    InvokeModelResponse response = completableFuture.get();
    JSONObject responseBody = new
    JSONObject(response.body().asUtf8String());
    JSONArray outputs = responseBody.getJSONArray("outputs");

    return IntStream.range(0, outputs.length())
        .mapToObj(i -> outputs.getJSONObject(i).getString("text"))
        .toList();
} catch (InterruptedException e) {
    Thread.currentThread().interrupt();
    System.err.println(e.getMessage());
} catch (ExecutionException e) {
    System.err.println(e.getMessage());
}
```
Invoke the Mixtral 8x7B foundation model to generate text.

```java
public static List<String> invokeMixtral8x7B(String prompt) {
    BedrockRuntimeClient client = BedrockRuntimeClient.builder()
        .region(Region.US_WEST_2)
        .credentialsProvider(ProfileCredentialsProvider.create())
        .build();

    // Mistral instruct models provide optimal results when
    // embedding the prompt into the following template:
    String instruction = "<s>[INST] " + prompt + " [/INST]";

    String modelId = "mistral.mistral-8x7b-instruct-v0:1";

    String payload = new JSONObject()
        .put("prompt", instruction)
        .put("max_tokens", 200)
        .put("temperature", 0.5)
        .toString();

    InvokeModelResponse response = client.invokeModel(request ->
        request
            .accept("application/json")
            .contentType("application/json")
            .body(SdkBytes.fromUtf8String(payload))
            .modelId(modelId));

    JSONObject responseBody = new JSONObject(response.body().asUtf8String());
    JSONArray outputs = responseBody.getJSONArray("outputs");

    return IntStream.range(0, outputs.length())
        .mapToObj(i ->
            outputs.getJSONObject(i).getString("text")
                .toString());
}
```
• For API details, see InvokeModel in AWS SDK for Java 2.x API Reference.

JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Invoke the Mixtral 8x7B foundation model to generate text.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved. // SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import { FoundationModels } from "../config/foundation_models.js";
import {
    BedrockRuntimeClient,
    InvokeModelCommand,
} from "@aws-sdk/client-bedrock-runtime";

/**
 * @typedef {Object} Output
 * @property {string} text
 *
 * @typedef {Object} ResponseBody
 * @property {Output[]} outputs
 */

/**
 * Invokes a Mistral 8x7B Instruct model.
 *
 * @param {string} prompt - The input text prompt for the model to complete.
 * @param {string} [modelId] - The ID of the model to use. Defaults to "mistral.mixtral-8x7b-instruct-v0:1".
 */
export const invokeModel = async (prompt,
```
modelId = "mistral.mixtral-8x7b-instruct-v0:1",
  ) => {
    // Create a new Bedrock Runtime client instance.
    const client = new BedrockRuntimeClient({ region: "us-east-1" });

    // Mistral instruct models provide optimal results when embedding
    // the prompt into the following template:
    const instruction = `\<s\>[INST] ${prompt} [/INST]\`;

    // Prepare the payload.
    const payload = {
      prompt: instruction,
      max_tokens: 500,
      temperature: 0.5,
    };

    // Invoke the model with the payload and wait for the response.
    const command = new InvokeModelCommand({
      contentType: "application/json",
      body: JSON.stringify(payload),
      modelId,
    });
    const apiResponse = await client.send(command);

    // Decode and return the response.
    const decodedResponseBody = new TextDecoder().decode(apiResponse.body);
    /** @type {ResponseBody} */
    const responseBody = JSON.parse(decodedResponseBody);
    return responseBody.outputs[0].text;
  }

  // Invoke the function if this file was run directly.
  if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const prompt ='
      Complete the following in one sentence: "Once upon a time..."';
    const modelId = FoundationModels.MIXTRAL_8X7B.modelId;
    console.log('Prompt: ${prompt}');
    console.log('Model ID: ${modelId}');

    try {
      console.log("-".repeat(53));
      const response = await invokeModel(prompt, modelId);
      console.log(response);
    } catch (err) {
console.log(err);
}
}

- For API details, see `InvokeModel` in *AWS SDK for JavaScript API Reference*.

**Python**

**SDK for Python (Boto3)**

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the *AWS Code Examples Repository*.

Invoke the Mixtral 8x7B foundation model to generate text.

```python
def invoke_mixtral_8x7b(self, prompt):
    """
    Invokes the Mixtral 8x7B model to run an inference using the input
    provided in the request body.

    :param prompt: The prompt that you want Mixtral to complete.
    :return: List of inference responses from the model.
    """

    try:
        # Mistral instruct models provide optimal results when
        # embedding the prompt into the following template:
        instruction = f"<s>[INST] {prompt} [/INST]"

        model_id = "mistral.mixtral-8x7b-instruct-v0:1"

        body = {
            "prompt": instruction,
            "max_tokens": 200,
            "temperature": 0.5,
        }

        response = self.bedrock_runtime_client.invoke_model(
```
For API details, see [InvokeModel](https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/boto3-client.html#boto3.client) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](https://docs.aws.amazon.com/sdk-for-python/latest/developerguide/using-aws-sdk-python.html). This topic also includes information about getting started and details about previous SDK versions.

### Scenarios for Amazon Bedrock Runtime using AWS SDKs

The following code examples show you how to implement common scenarios in Amazon Bedrock Runtime with AWS SDKs. These scenarios show you how to accomplish specific tasks by calling multiple functions within Amazon Bedrock Runtime. Each scenario includes a link to GitHub, where you can find instructions on how to set up and run the code.

#### Examples

- [Create a sample application that offers playgrounds to interact with Amazon Bedrock foundation models using an AWS SDK](https://github.com/aws-samples/amazon-bedrock-playground)
- [An interactive text generation playground for Amazon Bedrock](https://github.com/aws-samples/amazon-bedrock-text-generation)
- [Invoke multiple large-language models (LLMs) on Amazon Bedrock](https://github.com/aws-samples/amazon-bedrock-llm)
- [Invoke multiple foundation models on Amazon Bedrock](https://github.com/aws-samples/amazon-bedrock-foundation-models)
- [Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions](https://github.com/aws-samples/amazon-bedrock-orchestration)
Create a sample application that offers playgrounds to interact with Amazon Bedrock foundation models using an AWS SDK

The following code examples show how to create playgrounds to interact with Amazon Bedrock foundation models through different modalities.

.NET

AWS SDK for .NET

.NET Foundation Model (FM) Playground is a .NET MAUI Blazor sample application that showcases how to use Amazon Bedrock from C# code. This example shows how .NET and C# developers can use Amazon Bedrock to build generative AI-enabled applications. You can test and interact with Amazon Bedrock foundation models by using the following four playgrounds:

• A text playground.
• A chat playground.
• A voice chat playground.
• An image playground.

The example also lists and displays the foundation models you have access to and their characteristics. For source code and deployment instructions, see the project in GitHub.

Services used in this example

• Amazon Bedrock Runtime

Java

SDK for Java 2.x

The Java Foundation Model (FM) Playground is a Spring Boot sample application that showcases how to use Amazon Bedrock with Java. This example shows how Java developers can use Amazon Bedrock to build generative AI-enabled applications. You can test and interact with Amazon Bedrock foundation models by using the following three playgrounds:

• A text playground.
• A chat playground.
• An image playground.
The example also lists and displays the foundation models you have access to, along with their characteristics. For source code and deployment instructions, see the project in GitHub.

**Services used in this example**

- Amazon Bedrock Runtime

Python

**SDK for Python (Boto3)**

The Python Foundation Model (FM) Playground is a Python/FastAPI sample application that showcases how to use Amazon Bedrock with Python. This example shows how Python developers can use Amazon Bedrock to build generative AI-enabled applications. You can test and interact with Amazon Bedrock foundation models by using the following three playgrounds:

- A text playground.
- A chat playground.
- An image playground.

The example also lists and displays the foundation models you have access to, along with their characteristics. For source code and deployment instructions, see the project in GitHub.

**Services used in this example**

- Amazon Bedrock Runtime

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**An interactive text generation playground for Amazon Bedrock**

The following code example shows how to prepare and send a prompt to the following large-language models (LLMs) on Amazon Bedrock:

- AI21 Labs: Jurassic2 Mid and Ultra
- Amazon: Titan Text G1 Lite and Express
- Anthropic: Claude Instant
• Anthropic: Claude 2.0 and 2.1
• Anthropic: Claude 3 Haiku and Sonnet
• Meta: Llama2 Chat 13B and 70B
• Mistral AI: Mistral 7B and Mixtral 8x7B

JavaScript

**SDK for JavaScript (v3)**

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import {
  Scenario,
  ScenarioAction,
  ScenarioInput,
  ScenarioOutput,
} from "@aws-doc-sdk-examples/lib/scenario/index.js"
import { FoundationModels } from "./config/foundation_models.js"

/**
 * @typedef {Object} ModelConfig
 * @property {Function} module
 * @property {Function} invoker
 * @property {string} modelId
 * @property {string} modelName
 */

const greeting = new ScenarioOutput(
  "greeting",
  "Welcome to the Amazon Bedrock Runtime client demo!",
  { header: true },
);
```

Note
There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).
const selectModel = new ScenarioInput("model", "First, select a model:", {
  type: "select",
  choices: Object.values(FoundationModels).map((model) => {
    name: model.modelName,
    value: model,
  })),
});

const enterPrompt = new ScenarioInput("prompt", "Now, enter your prompt:", {
  type: "input",
});

const printDetails = new ScenarioOutput("print details",
/**
 * @param {{ model: ModelConfig, prompt: string }} c
 */
(c) => console.log(`Invoking ${c.model.modelName} with '${c.prompt}'...`),
{ slow: false },
);

const invokeModel = new ScenarioAction("invoke model",
/**
 * @param {{ model: ModelConfig, prompt: string, response: string }} c
 */
async (c) => {
  const modelModule = await c.model.module();
  const invoker = c.model.invoker(modelModule);
  c.response = await invoker(c.prompt, c.model.modelId);
},
);

const printResponse = new ScenarioOutput("print response",
/**
 * @param {{ response: string }} c
 */
(c) => c.response,
{ slow: false },
);

const scenario = new Scenario("Amazon Bedrock Runtime Demo", [

greeting,
selectModel,
enterPrompt,
printDetails,
invokeModel,
printResponse,
]);

if (process.argv[1] === fileURLToPath(import.meta.url)) {
  scenario.run();
}

• For API details, see InvokeModel in AWS SDK for JavaScript API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Invoke multiple large-language models (LLMs) on Amazon Bedrock

The following code example shows how to invoke multiple large-language-models (LLMs) on Amazon Bedrock.

• Generate text with Anthropic Claude.
• Generate text with AI21 Labs Jurassic-2.
• Generate text with Meta Llama 2 Chat.

PHP

SDK for PHP

ℹ️ Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
namespace BedrockRuntime;

class GettingStartedWithBedrockRuntime
{
    protected BedrockRuntimeService $bedrockRuntimeService;

    public function runExample()
    {
        echo "\n"
        echo "---------------------------------------------------------------------\n"
        echo "Welcome to the Amazon Bedrock Runtime getting started demo using PHP!\n"
        echo "---------------------------------------------------------------------\n"

        $clientArgs = [
            'region' => 'us-east-1',
            'version' => 'latest',
            'profile' => 'default',
        ];

        $bedrockRuntimeService = new BedrockRuntimeService($clientArgs);

        $prompt = 'In one paragraph, who are you?';
        echo "\nPrompt: " . $prompt;
        echo "\n\nAnthropic Claude:";
        echo $bedrockRuntimeService->invokeClaude($prompt);
        echo "\n\nAI21 Labs Jurassic-2: ";
        echo $bedrockRuntimeService->invokeJurassic2($prompt);
        echo "\n\nMeta Llama 2 Chat: ";
        echo $bedrockRuntimeService->invokeLlama2($prompt);

        echo "\n---------------------------------------------------------------------\n"

        $image_prompt = 'stylized picture of a cute old steampunk robot';
        echo "\nImage prompt: " . $image_prompt;
    }
}
```
echo "\n\nStability.ai Stable Diffusion XL:\n";
$diffusionSeed = rand(0, 4294967295);
$style_preset = 'photographic';
$base64 = $bedrockRuntimeService->invokeStableDiffusion($image_prompt, $diffusionSeed, $style_preset);
$image_path = $this->saveImage($base64, 'stability.stable-diffusion-xl');
echo "The generated images have been saved to $image_path";

echo "\n\nAmazon Titan Image Generation:\n"
$titanSeed = rand(0, 2147483647);
$base64 = $bedrockRuntimeService->invokeTitanImage($image_prompt, $titanSeed);
$image_path = $this->saveImage($base64, 'amazon.titan-image-generator-v1');
echo "The generated images have been saved to $image_path";
}

private function saveImage($base64_image_data, $model_id): string
{
    $output_dir = "output";

    if (!file_exists($output_dir)) {
        mkdir($output_dir);
    }

    $i = 1;
    while (file_exists("$output_dir/$model_id" . '_' . "$i.png")) {
        $i++;
    }

    $image_data = base64_decode($base64_image_data);

    $file_path = "$output_dir/$model_id" . '_' . "$i.png";
    $file = fopen($file_path, 'wb');
    fwrite($file, $image_data);
    fclose($file);

    return $file_path;
}
```
• For API details, see `InvokeModel` in *AWS SDK for PHP API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Invoke multiple foundation models on Amazon Bedrock**

The following code example shows how to invoke multiple foundation models on Amazon Bedrock.

- Generate text with Anthropic Claude.
- Generate text with AI21 Labs Jurassic-2.
- Generate text with Meta Llama 2 Chat.
- Asynchronously process the response stream from Anthropic Claude.
- Generate an image with the Amazon Titan Image Generator.

Go

**SDK for Go V2**

<table>
<thead>
<tr>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>There's more on GitHub. Find the complete example and learn how to set up and run in the <a href="#">AWS Code Examples Repository</a>.</td>
</tr>
</tbody>
</table>

```
// InvokeModelsScenario demonstrates how to use the Amazon Bedrock Runtime client
// to invoke various foundation models for text and image generation

// 1. Generate text with Anthropic Claude 2
// 2. Generate text with AI21 Labs Jurassic-2
// 3. Generate text with Meta Llama 2 Chat
// 4. Generate text and asynchronously process the response stream with Anthropic Claude 2
```
// 5. Generate and image with the Amazon Titan image generation model
type InvokeModelsScenario struct {
    sdkConfig             aws.Config
    invokeModelWrapper    actions.InvokeModelWrapper
    responseStreamWrapper actions.InvokeModelWithResponseStreamWrapper
    questioner            demotools.IQuestioner
}

// NewInvokeModelsScenario constructs an InvokeModelsScenario instance from a
// configuration.
// It uses the specified config to get a Bedrock Runtime client and create
// wrappers for the
// actions used in the scenario.
func NewInvokeModelsScenario(sdkConfig aws.Config, questioner
demotools.IQuestioner) InvokeModelsScenario {
    client := bedrockruntime.NewFromConfig(sdkConfig)
    return InvokeModelsScenario{
        sdkConfig:             sdkConfig,
        invokeModelWrapper:    actions.InvokeModelWrapper{BedrockRuntimeClient: client},
        questioner:            questioner,
    }
}

// Runs the interactive scenario.
func (scenario InvokeModelsScenario) Run() {
    defer func() {
        if r := recover(); r != nil {
            log.Printf("Something went wrong with the demo: %v\n", r)
        }
    }()

    log.Println(strings.Repeat("=", 77))
    log.Println("Welcome to the Amazon Bedrock Runtime model invocation demo.")
    log.Println(strings.Repeat("=", 77))

    log.Printf("First, let's invoke a few large-language models using the
    synchronous client:\n\n")

    text2textPrompt := "In one paragraph, who are you?"

    log.Println(strings.Repeat("-", 77))
func (scenario InvokeModelsScenario) InvokeClaude(prompt string) {
    completion, err := scenario.invokeModelWrapper.InvokeClaude(prompt)
    if err != nil {
        panic(err)
    }
    log.Printf("\nClaude     : %v\n", strings.TrimSpace(completion))
}

func (scenario InvokeModelsScenario) InvokeJurassic2(prompt string) {

    text2ImagePrompt := "stylized picture of a cute old steampunk robot"
    seed := rand.Int63n(2147483648)

    log.Printf("Invoking Amazon Titan with prompt: %v\n", text2ImagePrompt)
    scenario.InvokeTitanImage(text2ImagePrompt, seed)

    log.Println(strings.Repeat("=", 77))
    log.Println(strings.Repeat("=", 77))
}
For API details, see the following topics in AWS SDK for Go API Reference.

- InvokeModel
- InvokeModelWithResponseStream
For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions**

The following code example shows how to build and orchestrate generative AI applications with Amazon Bedrock and Step Functions.

**Python**

**SDK for Python (Boto3)**

The Amazon Bedrock Serverless Prompt Chaining scenario demonstrates how [AWS Step Functions](#), [Amazon Bedrock](#), and [Agents for Amazon Bedrock](#) can be used to build and orchestrate complex, serverless, and highly scalable generative AI applications. It contains the following working examples:

- Write an analysis of a given novel for a literature blog. This example illustrates a simple, sequential chain of prompts.
- Generate a short story about a given topic. This example illustrates how the AI can iteratively process a list of items that it previously generated.
- Create an itinerary for a weekend vacation to a given destination. This example illustrates how to parallelize multiple distinct prompts.
- Pitch movie ideas to a human user acting as a movie producer. This example illustrates how to parallelize the same prompt with different inference parameters, how to backtrack to a previous step in the chain, and how to include human input as part of the workflow.
- Plan a meal based on ingredients the user has at hand. This example illustrates how prompt chains can incorporate two distinct AI conversations, with two AI personas engaging in a debate with each other to improve the final outcome.
- Find and summarize today’s highest trending GitHub repository. This example illustrates chaining multiple AI agents that interact with external APIs.

For complete source code and instructions to set up and run, see the full project on [GitHub](#).

**Services used in this example**

- Amazon Bedrock
For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Code examples for Agents for Amazon Bedrock using AWS SDKs**

The following code examples show how to use Agents for Amazon Bedrock with an AWS software development kit (SDK).

*Actions* are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios and cross-service examples.

*Scenarios* are code examples that show you how to accomplish a specific task by calling multiple functions within the same service.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Get started**

**Hello Agents for Amazon Bedrock**

The following code example shows how to get started using Agents for Amazon Bedrock.

**JavaScript**

**SDK for JavaScript (v3)**

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
import { fileURLToPath } from "url";

import {
    BedrockAgentClient,
    GetAgentCommand,
    paginateListAgents,
} from "@aws-sdk/client-bedrock-agent";

/**
 * @typedef {Object} AgentSummary
 */

/**
 * A simple scenario to demonstrate basic setup and interaction with the Bedrock Agents Client.
 *
 * This function first initializes the Amazon Bedrock Agents client for a specific region.
 * It then retrieves a list of existing agents using the streamlined paginator approach.
 * For each agent found, it retrieves detailed information using a command object.
 *
 * Demonstrates:
 * - Use of the Bedrock Agents client to initialize and communicate with the AWS service.
 * - Listing resources in a paginated response pattern.
 * - Accessing an individual resource using a command object.
 *
 * @returns {Promise<void>} A promise that resolves when the function has completed execution.
 */
export const main = async () => {
    const region = "us-east-1";

    console.log("=".repeat(68));

    console.log(`Initializing Amazon Bedrock Agents client for ${region}...`);
    const client = new BedrockAgentClient({ region });
```javascript
console.log(`Retrieving the list of existing agents...`);
const paginatorConfig = { client);
const pages = paginateListAgents(paginatorConfig, {});

/** @type {AgentSummary[]} */
const agentSummaries = [];
for await (const page of pages) {
    agentSummaries.push(...page.agentSummaries);
}

console.log(`Found ${agentSummaries.length} agents in ${region}.`);

if (agentSummaries.length > 0) {
    for (const agentSummary of agentSummaries) {
        const agentId = agentSummary.agentId;
        console.log("=".repeat(68));
        console.log(`Retrieving agent with ID: ${agentId}:`);
        console.log("-".repeat(68));

        const command = new GetAgentCommand({ agentId });
        const response = await client.send(command);
        const agent = response.agent;

        console.log(` Name: ${agent.agentName}`);
        console.log(` Status: ${agent.agentStatus}`);
        console.log(` ARN: ${agent.agentArn}`);
        console.log(` Foundation model: ${agent.foundationModel}`);
    }
}
console.log("=".repeat(68));

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    await main();
}
```

- For API details, see the following topics in *AWS SDK for JavaScript API Reference*.
  - [GetAgent](#)
  - [ListAgents](#)
Code examples

- **Actions for Agents for Amazon Bedrock using AWS SDKs**
  - Create an Amazon Bedrock agent using an AWS SDK
  - Create an Amazon Bedrock agent action group using an AWS SDK
  - Create an Amazon Bedrock agent alias using an AWS SDK
  - Delete an Amazon Bedrock agent using an AWS SDK
  - Delete an Amazon Bedrock agent alias using an AWS SDK
  - Get information about an Amazon Bedrock agent using an AWS SDK
  - List the action groups for a Amazon Bedrock agent using an AWS SDK
  - List the Agents for Amazon Bedrock belonging to an account using an AWS SDK
  - List the knowledge bases associated with a Amazon Bedrock agent using an AWS SDK
  - Prepare an Amazon Bedrock agent using an AWS SDK

- **Scenarios for Agents for Amazon Bedrock using AWS SDKs**
  - An end-to-end example showing how to create and invoke Amazon Bedrock agents using an AWS SDK
  - Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

**Actions for Agents for Amazon Bedrock using AWS SDKs**

The following code examples demonstrate how to perform individual Agents for Amazon Bedrock actions with AWS SDKs. These excerpts call the Agents for Amazon Bedrock API and are code excerpts from larger programs that must be run in context. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the [Agents for Amazon Bedrock API Reference](#).

**Examples**

- Create an Amazon Bedrock agent using an AWS SDK
- Create an Amazon Bedrock agent action group using an AWS SDK
- Create an Amazon Bedrock agent alias using an AWS SDK
- Delete an Amazon Bedrock agent using an AWS SDK
- Delete an Amazon Bedrock agent alias using an AWS SDK
Create an Amazon Bedrock agent using an AWS SDK

The following code examples show how to create an Amazon Bedrock agent.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

JavaScript

**SDK for JavaScript (v3)**

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import { checkForPlaceholders } from "../lib/utils.js";

import {
   BedrockAgentClient,
   CreateAgentCommand,
} from "@aws-sdk/client-bedrock-agent";

/**

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/aws-samples/aws-code-examples).
* Creates an Amazon Bedrock Agent.
  *
  * @param {string} agentName - A name for the agent that you create.
  * @param {string} foundationModel - The foundation model to be used by the agent you create.
  * @param {string} agentResourceRoleArn - The ARN of the IAM role with permissions required by the agent.
  * @param {string} [region='us-east-1'] - The AWS region in use.
  * @returns {Promise<import("@aws-sdk/client-bedrock-agent").Agent>} An object containing details of the created agent.
  */
export const createAgent = async (agentName, foundationModel, agentResourceRoleArn, region = "us-east-1", ) => {
  const client = new BedrockAgentClient({ region });

  const command = new CreateAgentCommand({
    agentName, foundationModel, agentResourceRoleArn,
    region = "us-east-1",
  });
  const response = await client.send(command);

  return response.agent;
};

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  // Replace the placeholders for agentName and accountId, and roleName with a unique name for the new agent,
  // the id of your AWS account, and the name of an existing execution role that the agent can use inside your account.
  // For foundationModel, specify the desired model. Ensure to remove the brackets '[]' before adding your data.

  // A string (max 100 chars) that can include letters, numbers, dashes '-', and underscores '_'.
  const agentName = "[your-bedrock-agent-name]";

  // Your AWS account id.
  const accountId = "[123456789012]";
// The name of the agent's execution role. It must be prefixed by 'AmazonBedrockExecutionRoleForAgents_'.
const roleName = "[AmazonBedrockExecutionRoleForAgents_your-role-name]";

// The ARN for the agent's execution role.
// Follow the ARN format: 'arn:aws:iam::account-id:role/role-name'
const roleArn = `arn:aws:iam::${accountId}:role/${roleName}`;

// Specify the model for the agent. Change if a different model is preferred.
const foundationModel = "anthropic.claude-v2";

// Check for unresolved placeholders in agentName and roleArn.
checkForPlaceholders([agentName, roleArn]);

custom.sendlog(`Creating a new agent...`);

custom.sendlog(custom.makeAgent(agentName, foundationModel, roleArn));

• For API details, see CreateAgent in AWS SDK for JavaScript API Reference.

Python

SDK for Python (Boto3)

!! Note
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create an agent.

```python
def create_agent(self, agent_name, foundation_model, role_arn, instruction):
    """
    Creates an agent that orchestrates interactions between foundation models,
data sources, software applications, user conversations, and APIs to carry
```
out tasks to help customers.

:param agent_name: A name for the agent.
:param foundation_model: The foundation model to be used for orchestration by the agent.
:param role_arn: The ARN of the IAM role with permissions needed by the agent.
:param instruction: Instructions that tell the agent what it should do and how it should interact with users.
:return: The response from Agents for Bedrock if successful, otherwise raises an exception.

```python
try:
    response = self.client.create_agent(
        agentName=agent_name,
        foundationModel=foundation_model,
        agentResourceRoleArn=role_arn,
        instruction=instruction,
    )
except ClientError as e:
    logger.error(f"Error: Couldn't create agent. Here's why: {e}")
    raise
else:
    return response["agent"]
```

- For API details, see [CreateAgent](aws-sdk-python) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](aws-sdk-python). This topic also includes information about getting started and details about previous SDK versions.

**Create an Amazon Bedrock agent action group using an AWS SDK**

The following code example shows how to create an Amazon Bedrock agent action group.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- [Create and invoke an agent](aws-sdk-python)
Create an agent action group.

```python
def create_agent_action_group(
    self, name, description, agent_id, agent_version, function_arn,
    api_schema
):
    
    Creates an action group for an agent. An action group defines a set of actions that an agent should carry out for the customer.

    :param name: The name to give the action group.
    :param description: The description of the action group.
    :param agent_id: The unique identifier of the agent for which to create the action group.
    :param agent_version: The version of the agent for which to create the action group.
    :param function_arn: The ARN of the Lambda function containing the business logic that is carried out upon invoking the action.
    :param api_schema: Contains the OpenAPI schema for the action group.
    :return: Details about the action group that was created.
    
    try:
        response = self.client.create_agent_action_group(
            actionGroupName=name,
            description=description,
            agentId=agent_id,
            agentVersion=agent_version,
            actionGroupExecutor={"lambda": function_arn},
            apiSchema={"payload": api_schema},
        )
        agent_action_group = response["agentActionGroup"]
```

Note
---
There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
except ClientError as e:
    logger.error(f"Error: Couldn't create agent action group. Here's why: {e}")
    raise
else:
    return agent_action_group

- For API details, see CreateAgentActionGroup in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Create an Amazon Bedrock agent alias using an AWS SDK

The following code example shows how to create an Amazon Bedrock agent alias.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create an agent alias.

```python
def create_agent_alias(self, name, agent_id):
    """
    Creates an alias of an agent that can be used to deploy the agent.
    """
```
For API details, see CreateAgentAlias in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

### Delete an Amazon Bedrock agent using an AWS SDK

The following code examples show how to delete an Amazon Bedrock agent.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

#### JavaScript

**SDK for JavaScript (v3)**

![Note]

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
Delete an agent.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import { checkForPlaceholders } from "../lib/utils.js";

import {
  BedrockAgentClient,
  DeleteAgentCommand,
} from "@aws-sdk/client-bedrock-agent";

/**
 * Deletes an Amazon Bedrock Agent.
 * @param {string} agentId - The unique identifier of the agent to delete.
 * @param {string} [region='us-east-1'] - The AWS region in use.
 * @returns {Promise<import("@aws-sdk/client-bedrock-agent").DeleteAgentCommandOutput>} An object containing the agent id, the status, and some additional metadata.
 */
export const deleteAgent = (agentId, region = "us-east-1") => {
  const client = new BedrockAgentClient({ region });
  const command = new DeleteAgentCommand({ agentId });
  return client.send(command);
};

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  // Replace the placeholders for agentId with an existing agent's id.
  // Ensure to remove the brackets (\[\]) before adding your data.

  // The agentId must be an alphanumeric string with exactly 10 characters.
  const agentId = "[ABC123DE45]";

  // Check for unresolved placeholders in agentId.
  checkForPlaceholders([agentId]);

  console.log(`Deleting agent with ID ${agentId}...`);
  const response = await deleteAgent(agentId);
  console.log(response);
}
```
• For API details, see DeleteAgent in AWS SDK for JavaScript API Reference.

Python

SDK for Python (Boto3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Delete an agent.

```python
def delete_agent(self, agent_id):
    """
    Deletes an Amazon Bedrock agent.

    :param agent_id: The unique identifier of the agent to delete.
    :return: The response from Agents for Bedrock if successful, otherwise raises an exception.
    """

    try:
        response = self.client.delete_agent(
            agentId=agent_id, skipResourceInUseCheck=False
        )
    except ClientError as e:
        logger.error(f"Couldn't delete agent. {e}")
        raise
    else:
        return response
```

• For API details, see DeleteAgent in AWS SDK for Python (Boto3) API Reference.
For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Delete an Amazon Bedrock agent alias using an AWS SDK

The following code example shows how to delete an Amazon Bedrock agent alias.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

Python

SDK for Python (Boto3)

```python
def delete_agent_alias(self, agent_id, agent_alias_id):
    """
    Deletes an alias of an Amazon Bedrock agent.

    :param agent_id: The unique identifier of the agent that the alias belongs to.
    :param agent_alias_id: The unique identifier of the alias to delete.
    :return: The response from Agents for Bedrock if successful, otherwise raises an exception.
    """
    try:
        response = self.client.delete_agent_alias(
            agentId=agent_id, agentAliasId=agent_alias_id
        )
    except ClientError as e:
```

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
logger.error(f"Couldn't delete agent alias. {e}")
raise
else:
    return response

• For API details, see `DeleteAgentAlias` in `AWS SDK for Python (Boto3) API Reference`.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Get information about an Amazon Bedrock agent using an AWS SDK**

The following code examples show how to get information about an Amazon Bedrock agent.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

• [Create and invoke an agent](#)

**JavaScript**

**SDK for JavaScript (v3)**

> Note

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](#).

Get an agent.

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import { checkForPlaceholders } from "../lib/utils.js";
```
import {
    BedrockAgentClient,
    GetAgentCommand,
} from "@aws-sdk/client-bedrock-agent";

/**
 * Retrieves the details of an Amazon Bedrock Agent.
 *
 * @param {string} agentId - The unique identifier of the agent.
 * @param {string} [region='us-east-1'] - The AWS region in use.
 * @returns {Promise<import("@aws-sdk/client-bedrock-agent").Agent>} An object containing the agent details.
 */
export const getAgent = async (agentId, region = "us-east-1") => {
    const client = new BedrockAgentClient({ region });

    const command = new GetAgentCommand({ agentId });
    const response = await client.send(command);
    return response.agent;
};

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    // Replace the placeholders for agentId with an existing agent's id.
    // Ensure to remove the brackets '[]' before adding your data.

    // The agentId must be an alphanumeric string with exactly 10 characters.
    const agentId = "[ABC123DE45]";

    // Check for unresolved placeholders in agentId.
    checkForPlaceholders([agentId]);

    console.log(`Retrieving agent with ID ${agentId}...`);

    const agent = await getAgent(agentId);
    console.log(agent);
}

- For API details, see GetAgent in AWS SDK for JavaScript API Reference.
Python

**SDK for Python (Boto3)**

---

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://aws.amazon.com/code-examples/).

---

Get an agent.

```python
def get_agent(self, agent_id, log_error=True):
    """
    Gets information about an agent.

    :param agent_id: The unique identifier of the agent.
    :param log_error: Whether to log any errors that occur when getting the agent.
        If True, errors will be logged to the logger. If False, errors will still be raised, but not logged.
    :return: The information about the requested agent.
    """

    try:
        response = self.client.get_agent(agentId=agent_id)
        agent = response['agent']
    except ClientError as e:
        if log_error:
            logger.error(f"Couldn't get agent {agent_id}. {e}")
        raise
    else:
        return agent
```

* For API details, see [GetAgent](https://docs.aws.amazon.com/sdk-for-python/api/latest/reference/service/AmazonBedrock.html#AmazonBedrock.Client.get_agent) in *AWS SDK for Python (Boto3) API Reference*. 

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**List the action groups for a Amazon Bedrock agent using an AWS SDK**

The following code examples show how to list the action groups for a Amazon Bedrock agent.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- **Create and invoke an agent**

**JavaScript**

**SDK for JavaScript (v3)**

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";
import { checkForPlaceholders } from "../lib/utils.js";

import {
    BedrockAgentClient,
    ListAgentActionGroupsCommand,
    paginateListAgentActionGroups,
} from "@aws-sdk/client-bedrock-agent";

/**
 * Retrieves a list of Action Groups of an agent utilizing the paginator function.
 */
```

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
This function leverages a paginator, which abstracts the complexity of pagination, providing a straightforward way to handle paginated results inside a `for await...of` loop.

@param {string} agentId - The unique identifier of the agent.
@param {string} agentVersion - The version of the agent.
@param {string} [region='us-east-1'] - The AWS region in use.
@returns {Promise<ActionGroupSummary[]>} An array of action group summaries.

```javascript
export const listAgentActionGroupsWithPaginator = async (agentId, agentVersion, region = "us-east-1") => {
  const client = new BedrockAgentClient({ region });

  // Create a paginator configuration
  const paginatorConfig = {
    client,
    pageSize: 10, // optional, added for demonstration purposes
  };

  const params = { agentId, agentVersion };

  const pages = paginateListAgentActionGroups(paginatorConfig, params);

  // Paginate until there are no more results
  const actionGroupSummaries = [];
  for await (const page of pages) {
    actionGroupSummaries.push(...page.actionGroupSummaries);
  }

  return actionGroupSummaries;
};

/**
* Retrieves a list of Action Groups of an agent utilizing the ListAgentActionGroupsCommand.
*
* This function demonstrates the manual approach, sending a command to the client and processing the response.
* Pagination must manually be managed. For a simplified approach that abstracts away pagination logic, see
```
export const listAgentActionGroupsWithCommandObject = async (agentId, agentVersion, region = "us-east-1") => {
  const client = new BedrockAgentClient({ region });
  let nextToken;
  const actionGroupSummaries = [];
  do {
    const command = new ListAgentActionGroupsCommand({
      agentId,
      agentVersion,
      nextToken,
      maxResults: 10, // optional, added for demonstration purposes
    });

    /** @type {{actionGroupSummaries: ActionGroupSummary[], nextToken?: string}}
    */
    const response = await client.send(command);

    for (const actionGroup of response.actionGroupSummaries || []) {
      actionGroupSummaries.push(actionGroup);
    }

    nextToken = response.nextToken;
  } while (nextToken);

  return actionGroupSummaries;
};

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
  // Replace the placeholders for agentId and agentVersion with an existing agent's id and version.
  // Ensure to remove the brackets '[]' before adding your data.
const agentId = "[ABC123DE45]";

const agentVersion = "[DRAFT]";

checkForPlaceholders([agentId, agentVersion]);

for (const actionGroup of await listAgentActionGroupsWithCommandObject(
    agentId,
    agentVersion,
)) {
    console.log(actionGroup);
}

for (const actionGroup of await listAgentActionGroupsWithPaginator(
    agentId,
    agentVersion,
)) {
    console.log(actionGroup);
}

For API details, see ListAgentActionGroups in AWS SDK for JavaScript API Reference.
List the action groups for an agent.

```python
def list_agent_action_groups(self, agent_id, agent_version):
    
    List the action groups for a version of an Amazon Bedrock Agent.
    
    :param agent_id: The unique identifier of the agent.
    :param agent_version: The version of the agent.
    :return: The list of action group summaries for the version of the agent.
    
    try:
        action_groups = []

        paginator = self.client.get_paginato("list_agent_action_groups")
        for page in paginator.paginate(
            agentId=agent_id,
            agentVersion=agent_version,
            PaginationConfig={"PageSize": 10},
        ):
            action_groups.extend(page["actionGroupSummaries"])

    except ClientError as e:
        logger.error(f"Couldn't list action groups. {e}")
        raise
    else:
        return action_groups
```

- For API details, see [ListAgentActionGroups](#) in [AWS SDK for Python (Boto3) API Reference](#).
For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**List the Agents for Amazon Bedrock belonging to an account using an AWS SDK**

The following code examples show how to list the Agents for Amazon Bedrock belonging to an account.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

**JavaScript**

**SDK for JavaScript (v3)**

```javascript
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import { fileURLToPath } from "url";

import {
    BedrockAgentClient,
    ListAgentsCommand,
    paginateListAgents,
} from "@aws-sdk/client-bedrock-agent";

/**
 * Retrieves a list of available Amazon Bedrock agents utilizing the paginator function.
 */
```

---

**Note**

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
This function leverages a paginator, which abstracts the complexity of pagination, providing a straightforward way to handle paginated results inside a `for await...of` loop.

@param {string} [region='us-east-1'] - The AWS region in use.
@returns {Promise<AgentSummary[]>} An array of agent summaries.

```javascript
export const listAgentsWithPaginator = async (region = "us-east-1") => {
  const client = new BedrockAgentClient({ region });

  const paginatorConfig = {
    client,
    pageSize: 10, // optional, added for demonstration purposes
  };

  const pages = paginateListAgents(paginatorConfig, {});

  // Paginate until there are no more results
  const agentSummaries = [];
  for await (const page of pages) {
    agentSummaries.push(...page.agentSummaries);
  }

  return agentSummaries;
};
```

Retrieves a list of available Amazon Bedrock agents utilizing the ListAgentsCommand.

This function demonstrates the manual approach, sending a command to the client and processing the response. Pagination must manually be managed. For a simplified approach that abstracts away pagination logic, see the `listAgentsWithPaginator()` example below.

```javascript
export const listAgentsWithCommandObject = async (region = "us-east-1") => {
  const client = new BedrockAgentClient({ region });

  let nextToken;
```
const agentSummaries = [];
do {
    const command = new ListAgentsCommand(
        nextToken,
        maxResults: 10, // optional, added for demonstration purposes
    );

    /** @type {{agentSummaries: AgentSummary[], nextToken?: string}} */
    const paginatedResponse = await client.send(command);

    agentSummaries.push(...(paginatedResponse.agentSummaries || []));

    nextToken = paginatedResponse.nextToken;
} while (nextToken);
return agentSummaries;

// Invoke main function if this file was run directly.
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    console.log("=".repeat(68));
    console.log("Listing agents using ListAgentsCommand:");
    for (const agent of await listAgentsWithCommandObject()) {
        console.log(agent);
    }

    console.log("=".repeat(68));
    console.log("Listing agents using the paginateListAgents function:");
    for (const agent of await listAgentsWithPaginator()) {
        console.log(agent);
    }
}

- For API details, see ListAgents in AWS SDK for JavaScript API Reference.
List the agents belonging to an account.

```python
def list_agents(self):
    """
    List the available Amazon Bedrock Agents.

    :return: The list of available bedrock agents.
    """

    try:
        all_agents = []

        paginator = self.client.get_paginator("list_agents")
        for page in paginator.paginate(PaginationConfig={"PageSize": 10}):  
            all_agents.extend(page["agentSummaries"])

        except ClientError as e:
            logger.error(f"Couldn't list agents. {e}")
            raise
        else:
            return all_agents
```

- For API details, see [ListAgents](#) in [AWS SDK for Python (Boto3) API Reference](#).

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.
List the knowledge bases associated with a Amazon Bedrock agent using an AWS SDK

The following code example shows how to list the knowledge bases associated with a Amazon Bedrock agent.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Create and invoke an agent

Python

SDK for Python (Boto3)

```python
def list_agent_knowledge_bases(self, agent_id, agent_version):
    
    List the knowledge bases associated with a version of an Amazon Bedrock Agent.

    :param agent_id: The unique identifier of the agent.
    :param agent_version: The version of the agent.
    :return: The list of knowledge base summaries for the version of the agent.
    
    try:
        knowledge_bases = []

        paginator = self.client.get Paginator("list_agent_knowledge_bases")
        for page in paginator.paginate(
            agentId=agent_id,
            agentVersion=agent_version,
```
• For API details, see [ListAgentKnowledgeBases](https://docs.aws.amazon.com/AmazonBedrock/latest/APIReference/API_ListAgentKnowledgeBases.html) in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](https://docs.aws.amazon.com/sdk-for-python/guide/developer-guide/aws-sdk-developer-guides.html). This topic also includes information about getting started and details about previous SDK versions.

### Prepare an Amazon Bedrock agent using an AWS SDK

The following code example shows how to prepare an Amazon Bedrock agent for internal testing.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- [Create and invoke an agent](https://docs.aws.amazon.com/AmazonBedrock/latest/APIReference/API_CreateAgentInvocation.html)

**Python**

**SDK for Python (Boto3)**

> **Note**
> There's more on GitHub. Find the complete example and learn how to set up and run in the [AWS Code Examples Repository](https://github.com/aws-samples/aws-code-examples).

Prepare an agent for internal testing.
def prepare_agent(self, agent_id):
    """
    Creates a DRAFT version of the agent that can be used for internal
testing.

    :param agent_id: The unique identifier of the agent to prepare.
    :return: The response from Agents for Bedrock if successful, otherwise
raises an exception.
    """
    try:
        prepared_agent_details = self.client.prepare_agent(agentId=agent_id)
    except ClientError as e:
        logger.error(f"Couldn't prepare agent. {e}")
        raise
    else:
        return prepared_agent_details

- For API details, see PrepareAgent in AWS SDK for Python (Boto3) API Reference.

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Scenarios for Agents for Amazon Bedrock using AWS SDKs**

The following code examples show you how to implement common scenarios in Agents for Amazon Bedrock with AWS SDKs. These scenarios show you how to accomplish specific tasks by calling multiple functions within Agents for Amazon Bedrock. Each scenario includes a link to GitHub, where you can find instructions on how to set up and run the code.

**Examples**

- An end-to-end example showing how to create and invoke Amazon Bedrock agents using an AWS SDK
- Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions
An end-to-end example showing how to create and invoke Amazon Bedrock agents using an AWS SDK

The following code example shows how to:

- Create an execution role for the agent.
- Create the agent and deploy a DRAFT version.
- Create a Lambda function that implements the agent's capabilities.
- Create an action group that connects the agent to the Lambda function.
- Deploy the fully configured agent.
- Invoke the agent with user-provided prompts.
- Delete all created resources.

Python

SDK for Python (Boto3)

```
REGION = "us-east-1"
ROLE_POLICY_NAME = "agent_permissions"

class BedrockAgentScenarioWrapper:
    """Runs a scenario that shows how to get started using Agents for Amazon Bedrock."""

    def __init__(
        self, bedrock_agent_client, runtime_client, lambda_client, iam_resource, postfix
    ):
        self.iam_resource = iam_resource
        self.lambda_client = lambda_client
```

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.
self.bedrock_agent_runtime_client = runtime_client
self.postfix = postfix

self.bedrock_wrapper = BedrockAgentWrapper(bedrock_agent_client)

self.agent = None
self.agent_alias = None
self.agent_role = None
self.prepared_agent_details = None
self.lambda_role = None
self.lambda_function = None

def run_scenario(self):
    print("=" * 88)
    print("Welcome to the Amazon Bedrock Agents demo.")
    print("=" * 88)

    # Query input from user
    print("Let's start with creating an agent:")
    print("-" * 40)
    name, foundation_model = self._request_name_and_model_from_user()
    print("-" * 40)

    # Create an execution role for the agent
    self.agent_role = self._create_agent_role(foundation_model)

    # Create the agent
    self.agent = self._create_agent(name, foundation_model)

    # Prepare a DRAFT version of the agent
    self.prepared_agent_details = self._prepare_agent()

    # Create the agent's Lambda function
    self.lambda_function = self._create_lambda_function()

    # Configure permissions for the agent to invoke the Lambda function
    self._allow_agent_to_invoke_function()
    self._let_function_accept_invocations_from_agent()

    # Create an action group to connect the agent with the Lambda function
    self._create_agent_action_group()

    # If the agent has been modified or any components have been added, prepare the agent again
components = [self._get_agent()]
components += self._get_agent_action_groups()
components += self._get_agent_knowledge_bases()

latest_update = max(component["updatedAt"] for component in components)
if latest_update > self.prepared_agent_details["preparedAt"]:
    self.prepared_agent_details = self._prepare_agent()

# Create an agent alias
self.agent_alias = self._create_agent_alias()

# Test the agent
self._chat_with_agent(self.agent_alias)

print("=" * 88)
print("Thanks for running the demo!
")

if q.ask("Do you want to delete the created resources? [y/N] ", q.is_yesno):
    self._delete_resources()
    print("=" * 88)
    print("All demo resources have been deleted. Thanks again for running the demo!")
else:
    self._list_resources()
    print("=" * 88)
    print("Thanks again for running the demo!")

def _request_name_and_model_from_user(self):
    existing_agent_names = [
        agent["agentName"] for agent in self.bedrock_wrapper.list_agents()]

    while True:
        name = q.ask("Enter an agent name: ", self.is_valid_agent_name)
        if name.lower() not in [n.lower() for n in existing_agent_names]:
            break
        print(f"Agent {name} conflicts with an existing agent. Please use a different name.")

models = ["anthropic.claude-instant-v1", "anthropic.claude-v2"]
model_id = models[
    q.choose("Which foundation model would you like to use? ", models)
]

return name, model_id

def _create_agent_role(self, model_id):
    role_name = f"AmazonBedrockExecutionRoleForAgents_{self.postfix}" 
    model_arn = f"arn:aws:bedrock:{REGION}::foundation-model/{model_id}" 

    print("Creating an execution role for the agent...")

    try:
        role = self.iam_resource.create_role(
            RoleName=role_name,
            AssumeRolePolicyDocument=json.dumps(
                {
                    "Version": "2012-10-17",
                    "Statement": [
                        {
                            "Effect": "Allow",
                            "Principal": {"Service": "bedrock.amazonaws.com"},
                            "Action": "sts:AssumeRole",
                        }
                    ],
                }
            ),
        )

        role.Policy(ROLE_POLICY_NAME).put(
            PolicyDocument=json.dumps(
                {
                    "Version": "2012-10-17",
                    "Statement": [
                        {
                            "Effect": "Allow",
                            "Action": "bedrock:InvokeModel",
                            "Resource": model_arn,
                        }
                    ],
                }
            )
        )

    except Exception as e:
        print(f"Error creating execution role: {e}")
except ClientError as e:
    logger.error(f"Couldn't create role {role_name}. Here's why: {e}")
    raise

return role

def _create_agent(self, name, model_id):
    print("Creating the agent...")

    instruction = ""
    You are a friendly chat bot. You have access to a function called that returns
    information about the current date and time. When responding with date or time,
    please make sure to add the timezone UTC.
    ""
    agent = self.bedrock_wrapper.create_agent(
        agent_name=name,
        foundation_model=model_id,
        instruction=instruction,
        role_arn=self.agent_role.arn,
    )
    self._wait_for_agent_status(agent["agentId"], "NOT_PREPARED")

    return agent

def _prepare_agent(self):
    print("Preparing the agent...")

    agent_id = self.agent["agentId"]
    prepared_agent_details = self.bedrock_wrapper.prepare_agent(agent_id)
    self._wait_for_agent_status(agent_id, "PREPARED")

    return prepared_agent_details

def _create_lambda_function(self):
    print("Creating the Lambda function...")

    function_name = f"AmazonBedrockExampleFunction_{self.postfix}"
    self.lambda_role = self._create_lambda_role()

    try:
deployment_package = self._create_deployment_package(function_name)

lambda_function = self.lambda_client.create_function(
    FunctionName=function_name,
    Description="Lambda function for Amazon Bedrock example",
    Runtime="python3.11",
    Role=self.lambda_role.arn,
    Handler=f"{function_name}.lambda_handler",
    Code={"ZipFile": deployment_package},
    Publish=True,
)

waiter = self.lambda_client.get_waiter("function_active_v2")
waiter.wait(FunctionName=function_name)

except ClientError as e:
    logger.error(
        f"Couldn't create Lambda function {function_name}. Here's why: {e}"
    )
    raise

return lambda_function

def _create_lambda_role(self):
    print("Creating an execution role for the Lambda function...")

    role_name = f"AmazonBedrockExecutionRoleForLambda_{self.postfix}"

    try:
        role = self.iam_resource.create_role(
            RoleName=role_name,
            AssumeRolePolicyDocument=json.dumps(
                {
                    "Version": "2012-10-17",
                    "Statement": [
                        {
                            "Effect": "Allow",
                            "Principal": {"Service": "lambda.amazonaws.com"},
                            "Action": "sts:AssumeRole",
                        },
                    ],
                },
            ),
        )
role.attach_policy(
    PolicyArn="arn:aws:iam::aws:policy/service-role/
AWSLambdaBasicExecutionRole"
)

print(f"Created role {role_name}"
except ClientError as e:
    logger.error(f"Couldn't create role {role_name}. Here's why: {e}"
    raise

print("Waiting for the execution role to be fully propagated...")
wait(10)

return role

def _allow_agent_to_invoke_function(self):
    policy = self.iam_resource.RolePolicy(
        self.agent_role.role_name, ROLE_POLICY_NAME
    )
    doc = policy.policy_document
    doc["Statement"][].append(
        {
            "Effect": "Allow",
            "Action": "lambda:InvokeFunction",
            "Resource": self.lambda_function["FunctionArn"],
        }
    )

    self.agent_role.Policy(ROLE_POLICY_NAME).put(PolicyDocument=json.dumps(doc))

def _let_function_accept_invocations_from_agent(self):
    try:
        self.lambda_client.add_permission(
            FunctionName=self.lambda_function["FunctionName"],
            SourceArn=self.agent["agentArn"],
            StatementId="BedrockAccess",
            Action="lambda:InvokeFunction",
            Principal="bedrock.amazonaws.com",
        )
    except ClientError as e:
        logger.error(
            f"Couldn't grant Bedrock permission to invoke the Lambda function. Here's why: {e}"
        )
```python
raise

def _create_agent_action_group(self):
    print("Creating an action group for the agent...")

    try:
        with open("./scenario_resources/api_schema.yaml") as file:
            self.bedrock_wrapper.create_agent_action_group(
                name="current_date_and_time",
                description="Gets the current date and time.",
                agent_id=self.agent["agentId"],
                agent_version=self.prepared_agent_details["agentVersion"],
                function_arn=self.lambda_function["FunctionArn"],
                api_schema=json.dumps(yaml.safe_load(file)),
            )
    except ClientError as e:
        logger.error(f"Couldn't create agent action group. Here's why: {e}")
    raise

def _get_agent(self):
    return self.bedrock_wrapper.get_agent(self.agent["agentId"])

def _get_agent_action_groups(self):
    return self.bedrock_wrapper.list_agent_action_groups(
        self.agent["agentId"], self.prepared_agent_details["agentVersion"]
    )

def _get_agent_knowledge_bases(self):
    return self.bedrock_wrapper.list_agent_knowledge_bases(
        self.agent["agentId"], self.prepared_agent_details["agentVersion"]
    )

def _create_agent_alias(self):
    print("Creating an agent alias...")

    agent_alias_name = "test_agent_alias"
    agent_alias = self.bedrock_wrapper.create_agent_alias(
        agent_alias_name, self.agent["agentId"]
    )

    self._wait_for_agent_status(self.agent["agentId"], "PREPARED")

    return agent_alias
```
def _wait_for_agent_status(self, agent_id, status):
    while self.bedrock_wrapper.get_agent(agent_id)["agentStatus"] != status:
        wait(2)

def _chat_with_agent(self, agent_alias):
    print("-" * 88)
    print("The agent is ready to chat.")
    print("Try asking for the date or time. Type 'exit' to quit.")

    # Create a unique session ID for the conversation
    session_id = uuid.uuid4().hex

    while True:
        prompt = q.ask("Prompt: ", q.non_empty)

        if prompt == "exit":
            break

        response = asyncio.run(self._invoke_agent(agent_alias, prompt, session_id))

        print(f"Agent: {response}"

async def _invoke_agent(self, agent_alias, prompt, session_id):
    response = self.bedrock_agent_runtime_client.invoke_agent(
        agentId=self.agent["agentId"],
        agentAliasId=agent_alias["agentAliasId"],
        sessionId=session_id,
        inputText=prompt,
    )

    completion = ""

    for event in response.get("completion"):
        chunk = event["chunk"]
        completion += chunk["bytes"].decode()

    return completion

def _delete_resources(self):
    if self.agent:
        agent_id = self.agent["agentId"]

    if self.agent_alias:
agent_alias_id = self.agent_alias["agentAliasId"]
print("Deleting agent alias...")
self.bedrock_wrapper.delete_agent_alias(agent_id, agent_alias_id)

print("Deleting agent...")
agent_status = self.bedrock_wrapper.delete_agent(agent_id)["agentStatus"]
while agent_status == "DELETING":
    wait(5)
    try:
        agent_status = self.bedrock_wrapper.get_agent(
            agent_id, log_error=False
        )["agentStatus"]
    except ClientError as err:
        if err.response["Error"]["Code"] == "ResourceNotFoundException":
            agent_status = "DELETED"
else:
    if self.lambda_function:
        name = self.lambda_function["FunctionName"]
        print(f"Deleting function '{name}...'"
        self.lambda_client.delete_function(FunctionName=name)
    elif self.agent_role:
        print(f"Deleting role '{self.agent_role.role_name}...'"
        self.agent_role.Policy(ROLE_POLICY_NAME).delete()
        self.agent_role.delete()
    elif self.lambda_role:
        print(f"Deleting role '{self.lambda_role.role_name}...'"
        for policy in self.lambda_role.attached_policies.all():
            policy.detach_role(RoleName=self.lambda_role.role_name)
        self.lambda_role.delete()

def _list_resources(self):
    print("-" * 40)
    print(f"Here is the list of created resources in '{REGION}'")
    print("Make sure you delete them once you're done to avoid unnecessary costs.
    if self.agent:
        print(f"Bedrock Agent: {self.agent['agentName']}"
    if self.lambda_function:
        print(f"Lambda function: {self.lambda_function['FunctionName']}"
    if self.agent_role:
print(f"IAM role:        {self.agent_role.role_name}")
if self.lambda_role:
    print(f"IAM role:        {self.lambda_role.role_name}")

@staticmethod
def is_valid_agent_name(answer):
    valid_regex = r"^[a-zA-Z0-9-\_]{1,100}$"
    return (answer
        if answer and len(answer) <= 100 and re.match(valid_regex, answer)
        else None,
    )
    "I need a name for the agent, please. Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen).",

@staticmethod
def _create_deployment_package(function_name):
    buffer = io.BytesIO()
    with zipfile.ZipFile(buffer, "w") as zipped:
        zipped.write('./scenario_resources/lambda_function.py', f"{function_name}.py")
    buffer.seek(0)
    return buffer.read()

if __name__ == "__main__":
    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")

    postfix = ".".join(
        random.choice(string.ascii_lowercase + "0123456789") for _ in range(8)
    )
    scenario = BedrockAgentScenarioWrapper(
        bedrock_agent_client=boto3.client(
            service_name="bedrock-agent", region_name=REGION
        ),
        runtime_client=boto3.client(
            service_name="bedrock-agent-runtime", region_name=REGION
        ),
        lambda_client=boto3.client(service_name="lambda", region_name=REGION),
        iam_resource=boto3.resource("iam"),
        postfix=postfix,
    )
    try:
For API details, see the following topics in AWS SDK for Python (Boto3) API Reference.

- CreateAgent
- CreateAgentActionGroup
- CreateAgentAlias
- DeleteAgent
- DeleteAgentAlias
- GetAgent
- ListAgentActionGroups
- ListAgentKnowledgeBases
- ListAgents
- PrepareAgent

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

The following code example shows how to build and orchestrate generative AI applications with Amazon Bedrock and Step Functions.

Python

SDK for Python (Boto3)

The Amazon Bedrock Serverless Prompt Chaining scenario demonstrates how AWS Step Functions, Amazon Bedrock, and Agents for Amazon Bedrock can be used to build and orchestrate complex, serverless, and highly scalable generative AI applications. It contains the following working examples:
• Write an analysis of a given novel for a literature blog. This example illustrates a simple, sequential chain of prompts.

• Generate a short story about a given topic. This example illustrates how the AI can iteratively process a list of items that it previously generated.

• Create an itinerary for a weekend vacation to a given destination. This example illustrates how to parallelize multiple distinct prompts.

• Pitch movie ideas to a human user acting as a movie producer. This example illustrates how to parallelize the same prompt with different inference parameters, how to backtrack to a previous step in the chain, and how to include human input as part of the workflow.

• Plan a meal based on ingredients the user has at hand. This example illustrates how prompt chains can incorporate two distinct AI conversations, with two AI personas engaging in a debate with each other to improve the final outcome.

• Find and summarize today’s highest trending GitHub repository. This example illustrates chaining multiple AI agents that interact with external APIs.

For complete source code and instructions to set up and run, see the full project on GitHub.

**Services used in this example**

• Amazon Bedrock
• Amazon Bedrock Runtime
• Agents for Amazon Bedrock
• Agents for Amazon Bedrock Runtime
• Step Functions

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

**Code examples for Agents for Amazon Bedrock Runtime using AWS SDKs**

The following code examples show how to use Agents for Amazon Bedrock Runtime with an AWS software development kit (SDK).
**Actions** are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios and cross-service examples.

**Scenarios** are code examples that show you how to accomplish a specific task by calling multiple functions within the same service.

For a complete list of AWS SDK developer guides and code examples, see [Using this service with an AWS SDK](#). This topic also includes information about getting started and details about previous SDK versions.

**Code examples**

- [Actions for Agents for Amazon Bedrock Runtime using AWS SDKs](#)
  - Invoke an Amazon Bedrock agent using an AWS SDK
- [Scenarios for Agents for Amazon Bedrock Runtime using AWS SDKs](#)
  - Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions

**Actions for Agents for Amazon Bedrock Runtime using AWS SDKs**

The following code examples demonstrate how to perform individual Agents for Amazon Bedrock Runtime actions with AWS SDKs. These excerpts call the Agents for Amazon Bedrock Runtime API and are code excerpts from larger programs that must be run in context. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the [Agents for Amazon Bedrock Runtime API Reference](#).

**Examples**

- [Invoke an Amazon Bedrock agent using an AWS SDK](#)

**Invoke an Amazon Bedrock agent using an AWS SDK**

The following code examples show how to invoke an Amazon Bedrock agent.
JavaScript

SDK for JavaScript (v3)

Note

There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import {
  BedrockAgentRuntimeClient,
  InvokeAgentCommand,
} from "@aws-sdk/client-bedrock-agent-runtime";

/**
 * @typedef {Object} ResponseBody
 * @property {string} completion
 */

/**
 * Invokes a Bedrock agent to run an inference using the input provided in the request body.
 * @param {string} prompt - The prompt that you want the Agent to complete.
 * @param {string} sessionId - An arbitrary identifier for the session.
 */
export const invokeBedrockAgent = async (prompt, sessionId) => {
  const client = new BedrockAgentRuntimeClient({ region: "us-east-1" });
  // const client = new BedrockAgentRuntimeClient({
  //   region: "us-east-1",
  //   credentials: {
  //     accessKeyId: "accessKeyId", // permission to invoke agent
  //     secretAccessKey: "accessKeySecret",
  //   },
  // });

  const agentId = "AJBHXXILZN";
  const agentAliasId = "AVKP1ITZAA";
const command = new InvokeAgentCommand(
    agentId,
    agentAliasId,
    sessionId,
    inputText: prompt,
  });

try {
    let completion = "";
    const response = await client.send(command);

    if (response.completion === undefined) {
        throw new Error("Completion is undefined");
    }

    for await (let chunkEvent of response.completion) {
        const chunk = chunkEvent.chunk;
        console.log(chunk);
        const decodedResponse = new TextDecoder("utf-8").decode(chunk.bytes);
        completion += decodedResponse;
    }

    return { sessionId: sessionId, completion: completion };}
} catch (err) {
    console.error(err);
}

// Call function if run directly
import { fileURLToPath } from "url";
if (process.argv[1] === fileURLToPath(import.meta.url)) {
    const result = await invokeBedrockAgent("I need help.", "123");
    console.log(result);
}

- For API details, see InvokeAgent in AWS SDK for JavaScript API Reference.
Invoke an agent.

def invoke_agent(self, agent_id, agent_alias_id, session_id, prompt):
    ""
    Sends a prompt for the agent to process and respond to.

    :param agent_id: The unique identifier of the agent to use.
    :param agent_alias_id: The alias of the agent to use.
    :param session_id: The unique identifier of the session. Use the same value across requests
                      to continue the same conversation.
    :param prompt: The prompt that you want Claude to complete.
    :return: Inference response from the model.
    ""

    try:
        response = self.agents_runtime_client.invoke_agent(
            agentId=agent_id,
            agentAliasId=agent_alias_id,
            sessionId=session_id,
            inputText=prompt,
        )

        completion = ""

        for event in response.get("completion"):
            chunk = event["chunk"]
            completion = completion + chunk["bytes"].decode()

    except ClientError as e:
        logger.error(f"Couldn't invoke agent. {e}")
        raise
For API details, see `InvokeAgent` in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see *Using this service with an AWS SDK*. This topic also includes information about getting started and details about previous SDK versions.

**Scenarios for Agents for Amazon Bedrock Runtime using AWS SDKs**

The following code examples show you how to implement common scenarios in Agents for Amazon Bedrock Runtime with AWS SDKs. These scenarios show you how to accomplish specific tasks by calling multiple functions within Agents for Amazon Bedrock Runtime. Each scenario includes a link to GitHub, where you can find instructions on how to set up and run the code.

**Examples**

- **Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions**

**Build and orchestrate generative AI applications with Amazon Bedrock and Step Functions**

The following code example shows how to build and orchestrate generative AI applications with Amazon Bedrock and Step Functions.

**Python**

**SDK for Python (Boto3)**

The Amazon Bedrock Serverless Prompt Chaining scenario demonstrates how *AWS Step Functions*, *Amazon Bedrock*, and *Agents for Amazon Bedrock* can be used to build and orchestrate complex, serverless, and highly scalable generative AI applications. It contains the following working examples:

- Write an analysis of a given novel for a literature blog. This example illustrates a simple, sequential chain of prompts.
- Generate a short story about a given topic. This example illustrates how the AI can iteratively process a list of items that it previously generated.
• Create an itinerary for a weekend vacation to a given destination. This example illustrates how to parallelize multiple distinct prompts.

• Pitch movie ideas to a human user acting as a movie producer. This example illustrates how to parallelize the same prompt with different inference parameters, how to backtrack to a previous step in the chain, and how to include human input as part of the workflow.

• Plan a meal based on ingredients the user has at hand. This example illustrates how prompt chains can incorporate two distinct AI conversations, with two AI personas engaging in a debate with each other to improve the final outcome.

• Find and summarize today’s highest trending GitHub repository. This example illustrates chaining multiple AI agents that interact with external APIs.

For complete source code and instructions to set up and run, see the full project on GitHub.

**Services used in this example**

• Amazon Bedrock
• Amazon Bedrock Runtime
• Agents for Amazon Bedrock
• Agents for Amazon Bedrock Runtime
• Step Functions

For a complete list of AWS SDK developer guides and code examples, see Using this service with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.
Amazon Bedrock abuse detection

AWS is committed to the responsible use of AI. To help prevent potential misuse, Amazon Bedrock implements automated abuse detection mechanisms to identify potential violations of AWS’s Acceptable Use Policy (AUP) and Service Terms, including the Responsible AI Policy or a third-party model provider’s AUP.

Our abuse detection mechanisms are fully automated, so there is no human review of, or access to, user inputs or model outputs.

Automated abuse detection includes:

- **Categorize content** — We use classifiers to detect harmful content (such as content that incites violence) in user inputs and model outputs. A classifier is an algorithm that processes model inputs and outputs, and assigns type of harm and level of confidence. We may run these classifiers on both Titan and third-party model usage. The classification process is automated and does not involve human review of user inputs or model outputs.

- **Identify patterns** — We use classifier metrics to identify potential violations and recurring behavior. We may compile and share anonymized classifier metrics with third-party model providers. Amazon Bedrock does not store user input or model output and does not share these with third-party model providers.

- **Detecting and blocking child sexual abuse material (CSAM)** — You are responsible for the content you (and your end users) upload to Amazon Bedrock and must ensure this content is free from illegal images. To help stop the dissemination of CSAM, Amazon Bedrock may use automated abuse detection mechanisms (such as hash matching technology or classifiers) to detect apparent CSAM. If Amazon Bedrock detects apparent CSAM in your image inputs, Amazon Bedrock will block the request and you will receive an automated error message. Amazon Bedrock may also file a report with the National Center for Missing and Exploited Children (NCMEC) or a relevant authority. We take CSAM seriously and will continue to update our detection, blocking, and reporting mechanisms. You might be required by applicable laws to take additional actions, and you are responsible for those actions.

Once our automated abuse detection mechanisms identify potential violations, we may request information about your use of Amazon Bedrock and compliance with our terms of service or a third-party provider’s AUP. In the event that you are unwilling or unable to comply with these terms or policies, AWS may suspend your access to Amazon Bedrock.
Contact AWS Support if you have additional questions. For more information, see the Amazon Bedrock FAQs.
Quotas for Amazon Bedrock

Your AWS account has default quotas, formerly referred to as limits, for each AWS service. Unless otherwise noted, each quota is Region-specific within your AWS account. Some quotas may be adjustable. The following list explains the meaning of the **Adjustable** column in the following tables:

- If a quota is marked as **Yes**, you can adjust it by following the steps at [Requesting a Quota Increase](#) in the *Service Quotas User Guide*.
- If a quota is marked as **No**, you might be able to request a quota increase in one of the following ways:
  - To request a quota increase for an on-demand runtime quota, contact your AWS account manager. If you don't have an AWS account manager, unfortunately you can't increase your quota at this time.
  - To request other quota increases, submit a request through the [limit increase form](#) to be considered for an increase.

**Note**

Due to overwhelming demand, priority will be given to customers who generate traffic that consumes their existing quota allocation. Your request might be denied if you don't meet this condition.

Some quotas differ by model. Unless specified otherwise, a quota applies to all versions of a model.

Select a topic to learn more about quotas for it.

**Topics**

- [Runtime quotas](#)
- [Batch inference quotas](#)
- [Knowledge base quotas](#)
- [Agent quotas](#)
- [Model customization quotas](#)
- [Provisioned Throughput quotas](#)
Runtime quotas

Latency differs by model and is directly proportional to the following conditions:

- The number of input and output tokens
- The total number of ongoing on-demand requests by all customers at the time.

The following quotas apply when you carry out model inference. These quotas consider the combined sum for `InvokeModel` and `InvokeModelWithResponseStream` requests.

To increase throughput, purchase Provisioned Throughput.

<table>
<thead>
<tr>
<th>Model</th>
<th>Requests processed per minute</th>
<th>Tokens processed per minute</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI21 Labs Jurassic-2 Mid</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>AI21 Labs Jurassic-2 Ultra</td>
<td>100</td>
<td>200,000</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Embeddings G1 - Text</td>
<td>2,000</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Image Generator G1</td>
<td>60</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Multimodal Embeddings G1</td>
<td>2,000</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Express</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Amazon Titan Text G1 - Lite</td>
<td>800</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Model</td>
<td>Requests processed per minute</td>
<td>Tokens processed per minute</td>
<td>Adjustable</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------</td>
<td>-----------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Anthropic Claude Instant</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude 2.x</td>
<td>100</td>
<td>200,000</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude 3 Sonnet</td>
<td>100</td>
<td>200,000</td>
<td>No</td>
</tr>
<tr>
<td>Anthropic Claude 3 Haiku</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Command</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Command Light</td>
<td>800</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Embed (English)</td>
<td>2,000</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Cohere Embed (Multilingual)</td>
<td>2,000</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Meta Llama 2 13B</td>
<td>800</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Meta Llama 2 70B</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Mistral AI Mistral 7B Instruct</td>
<td>800</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Mistral AI Mixtral 8X7B Instruct</td>
<td>400</td>
<td>300,000</td>
<td>No</td>
</tr>
<tr>
<td>Stable Diffusion XL</td>
<td>60</td>
<td>N/A</td>
<td>No</td>
</tr>
</tbody>
</table>

Select a tab to see model-specific inference quotas.
## Amazon Titan Text models

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text prompt length, in characters</td>
<td>42,000</td>
<td>No</td>
</tr>
</tbody>
</table>

## Amazon Titan Image Generator G1

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text prompt length, in characters</td>
<td>1,024</td>
<td>No</td>
</tr>
<tr>
<td>Input image size</td>
<td>5 MB</td>
<td>No</td>
</tr>
<tr>
<td>Input image height in pixels (inpainting/ouptainting)</td>
<td>1,024</td>
<td>No</td>
</tr>
<tr>
<td>Input image width in pixels (inpainting/ouptainting)</td>
<td>1,024</td>
<td>No</td>
</tr>
<tr>
<td>Input image height in pixels (image variation)</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Input image width in pixels (image variation)</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Input image total pixels</td>
<td>12,582,912</td>
<td>No</td>
</tr>
</tbody>
</table>

## Amazon Titan Embeddings G1 - Text

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text input length, in characters</td>
<td>50,000</td>
<td>No</td>
</tr>
</tbody>
</table>
Amazon Titan Multimodal Embeddings G1

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text input length, in characters</td>
<td>100,000</td>
<td>No</td>
</tr>
<tr>
<td>Base64-encoded string of image, in characters</td>
<td>25,000,000</td>
<td>No</td>
</tr>
</tbody>
</table>

**Batch inference quotas**

The following quotas apply when you run batch inference. The quotas depend on the modality of the input and output data.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Minimum file size</th>
<th>Maximum file size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text to embeddings</td>
<td>75 MB</td>
<td>500 MB</td>
</tr>
<tr>
<td>Text to text</td>
<td>20 MB</td>
<td>150 MB</td>
</tr>
<tr>
<td>Text/image to image</td>
<td>1 MB</td>
<td>50 MB</td>
</tr>
</tbody>
</table>

**Knowledge base quotas**

The following quotas apply to Knowledge bases for Amazon Bedrock.

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge bases per account per region</td>
<td>50</td>
<td>No</td>
</tr>
<tr>
<td>Data source file size (source document)</td>
<td>50 MB</td>
<td>No</td>
</tr>
<tr>
<td>Data source file size (metadata file)</td>
<td>10 KB</td>
<td>No</td>
</tr>
</tbody>
</table>
### Description

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source chunk size (Titan Text G1 - Embeddings)</td>
<td>8,192</td>
<td>No</td>
</tr>
<tr>
<td>Data source chunk size (Cohere Embed English)</td>
<td>512</td>
<td>No</td>
</tr>
<tr>
<td>Data source chunk size (Cohere Embed Multilingual)</td>
<td>512</td>
<td>No</td>
</tr>
</tbody>
</table>

### Agent quotas

The following quotas apply to Agents for Amazon Bedrock.

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agents per account</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>Aliases per Agent</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Characters in Agent instructions</td>
<td>1,200</td>
<td>No</td>
</tr>
<tr>
<td>Action groups per Agent</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>Number of APIs across action groups in an Agent</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>Lambda response payload size</td>
<td>25 KB</td>
<td>No</td>
</tr>
<tr>
<td>Associated knowledge bases per Agent</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>

### Model customization quotas

The following quotas apply to model customization.
### Description

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheduled training jobs per account</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Custom models per account</td>
<td>100</td>
<td>Yes</td>
</tr>
</tbody>
</table>

To see hyperparameter quotas, see [Custom model hyperparameters](#).

Select a tab to see model-specific quotas that apply to training and validation datasets used for customizing different foundation models.

#### Amazon Titan Text G1 - Express

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum (Continued Pre-training)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of input and output tokens when batch size is 1</td>
<td>4,096</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Sum of input and output tokens when batch size is 2, 3, or 4</td>
<td>2,048</td>
<td>2,048</td>
<td>No</td>
</tr>
<tr>
<td>Character quota per sample in dataset</td>
<td>Token quota x 6</td>
<td>Token quota x 6</td>
<td>No</td>
</tr>
<tr>
<td>Records in a training dataset</td>
<td>100,000</td>
<td>10,000</td>
<td>Yes</td>
</tr>
<tr>
<td>Records in a validation dataset</td>
<td>1,000</td>
<td>1,000</td>
<td>Yes</td>
</tr>
<tr>
<td>Training dataset file size</td>
<td>10 GB</td>
<td>1 GB</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Amazon Bedrock

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum (Continued Pre-training)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation dataset file size</td>
<td>100 MB</td>
<td>100 MB</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Amazon Titan Text G1 - Lite

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum (Continued Pre-training)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of input and output tokens when batch size is 1 or 2</td>
<td>4,096</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Sum of input and output tokens when batch size is 3, 4, 5, 6 or 6</td>
<td>2,048</td>
<td>2,048</td>
<td>No</td>
</tr>
<tr>
<td>Character quota per sample in dataset</td>
<td>Token quota x 6</td>
<td>Token quota x 6</td>
<td>No</td>
</tr>
<tr>
<td>Records in a training dataset</td>
<td>100,000</td>
<td>10,000</td>
<td>Yes</td>
</tr>
<tr>
<td>Records in a validation dataset</td>
<td>1,000</td>
<td>1,000</td>
<td>Yes</td>
</tr>
<tr>
<td>Training dataset file size</td>
<td>10 GB</td>
<td>1 GB</td>
<td>Yes</td>
</tr>
<tr>
<td>Validation dataset file size</td>
<td>100 MB</td>
<td>100 MB</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Amazon Titan Image Generator G1

<table>
<thead>
<tr>
<th>Description</th>
<th>Minimum (Fine-tuning)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text prompt length in training sample, in characters</td>
<td>3</td>
<td>1,024</td>
<td>No</td>
</tr>
<tr>
<td>Records in a training dataset</td>
<td>5</td>
<td>10,000</td>
<td>No</td>
</tr>
<tr>
<td>Input image size</td>
<td>0</td>
<td>50 MB</td>
<td>No</td>
</tr>
<tr>
<td>Input image height in pixels</td>
<td>512</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Input image width in pixels</td>
<td>512</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Input image total pixels</td>
<td>0</td>
<td>12,582,912</td>
<td>No</td>
</tr>
<tr>
<td>Input image aspect ratio</td>
<td>1:4</td>
<td>4:1</td>
<td>No</td>
</tr>
</tbody>
</table>

### Amazon Titan Multimodal Embeddings G1

<table>
<thead>
<tr>
<th>Description</th>
<th>Minimum (Fine-tuning)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text prompt length in training sample, in characters</td>
<td>0</td>
<td>2,560</td>
<td>No</td>
</tr>
<tr>
<td>Records in a training dataset</td>
<td>1,000</td>
<td>500,000</td>
<td>No</td>
</tr>
</tbody>
</table>
### Amazon Bedrock

<table>
<thead>
<tr>
<th>Description</th>
<th>Minimum (Fine-tuning)</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input image size</td>
<td>0</td>
<td>5 MB</td>
<td>No</td>
</tr>
<tr>
<td>Input image height in pixels</td>
<td>128</td>
<td>4096</td>
<td>No</td>
</tr>
<tr>
<td>Input image width in pixels</td>
<td>128</td>
<td>4096</td>
<td>No</td>
</tr>
<tr>
<td>Input image total pixels</td>
<td>0</td>
<td>12,528,912</td>
<td>No</td>
</tr>
<tr>
<td>Input image aspect ratio</td>
<td>1:4</td>
<td>4:1</td>
<td>No</td>
</tr>
</tbody>
</table>

**Cohere Command & Meta Llama 2**

<table>
<thead>
<tr>
<th>Description</th>
<th>Maximum (Fine-tuning)</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input tokens</td>
<td>4,096</td>
<td>No</td>
</tr>
<tr>
<td>Output tokens</td>
<td>2,048</td>
<td>No</td>
</tr>
<tr>
<td>Character quota per sample in dataset</td>
<td>Token quota x 6</td>
<td>No</td>
</tr>
<tr>
<td>Records in a training dataset</td>
<td>10,000</td>
<td>Yes</td>
</tr>
<tr>
<td>Records in a validation dataset</td>
<td>1,000</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Provisioned Throughput quotas**

The following quotas apply to Provisioned Throughput.
<table>
<thead>
<tr>
<th>Description</th>
<th>Default</th>
<th>Adjustable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model units for a base model</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>Model units for a custom model</td>
<td>2</td>
<td>Yes</td>
</tr>
</tbody>
</table>
API reference

The API Reference can be found here.
## Document history for the Amazon Bedrock User Guide

- **Latest documentation update:** March 27th, 2024

The following table describes important changes in each release of Amazon Bedrock. For notification about updates to this documentation, you can subscribe to an RSS feed.

<table>
<thead>
<tr>
<th>Change</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New feature</strong></td>
<td>You can now create a network access policy in Amazon OpenSearch Serverless to allow your Amazon Bedrock knowledge base to access a private OpenSearch Serverless vector search collection configured with a VPC endpoint.</td>
<td>March 28, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>You can now include metadata for your source documents in Knowledge bases for Amazon Bedrock and filter on the metadata during knowledge base query.</td>
<td>March 27, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>You can now use a prompt template to customize the prompt sent to a model when you query a knowledge base and generate responses.</td>
<td>March 26, 2024</td>
</tr>
<tr>
<td><strong>More model support for querying knowledge bases in Amazon Bedrock</strong></td>
<td>You can now use Anthropic Claude 3 Sonnet for knowledge base response generation.</td>
<td>March 25, 2024</td>
</tr>
<tr>
<td>Category</td>
<td>Description</td>
<td>Date</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>New model</strong></td>
<td>You can now use Anthropic Claude 3 Haiku with Amazon Bedrock.</td>
<td>March 13, 2024</td>
</tr>
<tr>
<td><strong>New model</strong></td>
<td>You can now use Anthropic Claude 3 Sonnet with Amazon Bedrock.</td>
<td>March 4, 2024</td>
</tr>
<tr>
<td><strong>New model</strong></td>
<td>You can now use Mistral AI models with Amazon Bedrock.</td>
<td>March 1, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>You can now customize the search strategy in Knowledge Base for Amazon OpenSearch Serverless vector stores that contain a filterable text field.</td>
<td>February 28, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>You can now detect images with a watermark from Amazon Bedrock Titan Image Generator.</td>
<td>February 14, 2024</td>
</tr>
<tr>
<td><strong>Updated AWS PrivateLink support</strong></td>
<td>You can now use AWS PrivateLink to create interface VPC endpoints for the <strong>Agents for Amazon Bedrock Build-time service.</strong></td>
<td>February 9, 2024</td>
</tr>
<tr>
<td><strong>IAM role update</strong></td>
<td>You can now use the same service role across knowledge bases and use roles without a predefined prefix.</td>
<td>February 9, 2024</td>
</tr>
<tr>
<td><strong>Model in legacy status</strong></td>
<td>Stable Diffusion XL v0.8 is now in legacy status. Migrate to Stable Diffusion XL v1.x before April 30, 2024.</td>
<td>February 2, 2024</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>Date</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td><strong>Code examples chapter added</strong></td>
<td>The Amazon Bedrock guide now includes code examples across a variety of Amazon Bedrock actions and scenarios.</td>
<td>January 25, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>Knowledge bases for Amazon Bedrock now offers you a choice between a production and non-production account when you choose to quick create an Amazon OpenSearch Serverless vector store in the console.</td>
<td>January 24, 2024</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>Agents for Amazon Bedrock now lets you view traces in real-time when you use the test window in the console.</td>
<td>January 18, 2024</td>
</tr>
<tr>
<td><strong>More model support for embedding data sources in Knowledge bases for Amazon Bedrock</strong></td>
<td>Knowledge bases for Amazon Bedrock now supports using the Cohere Embed English and Cohere Embed Multilingual to embed your data sources.</td>
<td>January 17, 2024</td>
</tr>
<tr>
<td><strong>More model support for Agents for Amazon Bedrock and querying knowledge bases in Amazon Bedrock</strong></td>
<td>Agents for Amazon Bedrock and Knowledge bases for Amazon Bedrock response generation now support Anthropic Claude 2.1.</td>
<td>December 27, 2023</td>
</tr>
<tr>
<td>Region expansion</td>
<td>Amazon Bedrock is now available in AWS GovCloud (US-West) (us-gov-west-1). For information on endpoints, see <a href="#">Amazon Bedrock endpoints and quotas</a>.</td>
<td>December 21, 2023</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>New vector store support</td>
<td>You can now create a knowledge base in an Amazon Aurora database cluster. For more information, see <a href="#">Create a vector store in Amazon Aurora</a>.</td>
<td>December 21, 2023</td>
</tr>
<tr>
<td>New managed policies</td>
<td>Amazon Bedrock has added AmazonBedrockFullAccess to give users permission to create, read, update, and delete resources, and AmazonBedrockReadOnly to give users read-only permissions for all actions.</td>
<td>December 12, 2023</td>
</tr>
<tr>
<td>New feature</td>
<td>Amazon Bedrock now supports creating model evaluation jobs using automatic metrics or human workers.</td>
<td>November 29, 2023</td>
</tr>
<tr>
<td>New feature</td>
<td>You can now monitor and customize your <a href="#">model versions</a>.</td>
<td>November 29, 2023</td>
</tr>
<tr>
<td>Feature Type</td>
<td>Description</td>
<td>Date</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td><strong>New Titan models</strong></td>
<td>New models from Titan include Amazon Titan Image Generator G1 and Amazon Titan Multimodal Embeddings G1. For more information, see <a href="#">Titan Models</a>.</td>
<td>November 29, 2023</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>With Continued Pre-training you can teach a model new domain knowledge. For more information, see <a href="#">Custom Models</a>.</td>
<td>November 28, 2023</td>
</tr>
<tr>
<td><strong>New feature</strong></td>
<td>You can now query knowledge bases through the Retrieve and RetrieveAndGenerate APIs. For more information, see <a href="#">Query a knowledge base</a>.</td>
<td>November 28, 2023</td>
</tr>
<tr>
<td><strong>General release</strong></td>
<td>General release of the Knowledge bases for Amazon Bedrock service. For more information, see <a href="#">Knowledge bases for Amazon Bedrock</a>.</td>
<td>November 28, 2023</td>
</tr>
<tr>
<td><strong>General release</strong></td>
<td>General release of the Agents for Amazon Bedrock service. For more information, see <a href="#">Agents for Amazon Bedrock</a>.</td>
<td>November 28, 2023</td>
</tr>
<tr>
<td><strong>Customize more models</strong></td>
<td>You can now customize models from Cohere and Meta. For more information, see <a href="#">Custom Models</a>.</td>
<td>November 28, 2023</td>
</tr>
<tr>
<td><strong>New model releases</strong></td>
<td>Updated documentation to cover new Meta and Cohere models. For more information, see <a href="https://aws.amazon.com/bedrock/">Amazon Bedrock</a>.</td>
<td>November 13, 2023</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>Documentation localization</strong></td>
<td>Amazon Bedrock documentation is now available in <a href="https://aws.amazon.com/bedrock/ja">Japanese</a> and <a href="https://aws.amazon.com/bedrock/de">German</a>.</td>
<td>October 20, 2023</td>
</tr>
<tr>
<td><strong>Region expansion</strong></td>
<td>Amazon Bedrock is now available in Europe (Frankfurt) (eu-central-1). For information on endpoints, see <a href="https://aws.amazon.com/bedrock/quotas/">Amazon Bedrock endpoints and quotas</a>.</td>
<td>October 19, 2023</td>
</tr>
<tr>
<td><strong>Region expansion</strong></td>
<td>Amazon Bedrock is now available in Asia Pacific (Tokyo) (ap-northeast-1). For information on endpoints, see <a href="https://aws.amazon.com/bedrock/quotas/">Amazon Bedrock endpoints and quotas</a>.</td>
<td>October 3, 2023</td>
</tr>
<tr>
<td><strong>Gated general release</strong></td>
<td>Gated general release of the Amazon Bedrock service. For more information, see <a href="https://aws.amazon.com/bedrock/">Amazon Bedrock</a>.</td>
<td>September 28, 2023</td>
</tr>
</tbody>
</table>
AWS Glossary

For the latest AWS terminology, see the AWS glossary in the AWS Glossary Reference.