



API Reference

Amazon Bedrock



Amazon Bedrock: API Reference

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

Amazon Bedrock API Reference	1
Amazon Bedrock endpoints	1
AWS Command Line Interface references	2
AWS SDK references	2
Actions	4
Amazon Bedrock	7
Agents for Amazon Bedrock	142
Agents for Amazon Bedrock Runtime	327
Amazon Bedrock Runtime	356
Data Types	372
Amazon Bedrock	378
Agents for Amazon Bedrock	460
Agents for Amazon Bedrock Runtime	567
Amazon Bedrock Runtime	673
Common Parameters	676
Common Errors	678

Amazon Bedrock API Reference

This document provides detailed information about the Bedrock API actions and their parameters. For more information about setting up the Amazon Bedrock APIs, see [Set up the Amazon Bedrock API](#).

For information about the IAM access control permissions you need to use the APIs, see [Identity-based policy examples for Amazon Bedrock](#).

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](#).

Note

Check that you're using the correct endpoint when making an API request.

AWS Command Line Interface references

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](#)

AWS SDK references

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](#).

SDK documentation	Code examples	Amazon Bedrock prefix	Amazon Bedrock runtime prefix	Agents for Amazon Bedrock prefix	Agents for Amazon Bedrock runtime prefix
AWS SDK for C++	AWS SDK for C++ code examples	bedrock	bedrock-runtime	bedrock-agent	bedrock-agent-runtime

SDK documentation	Code examples	Amazon Bedrock prefix	Amazon Bedrock runtime prefix	Agents for Amazon Bedrock prefix	Agents for Amazon Bedrock runtime prefix
AWS SDK for Go	AWS SDK for Go code examples	bedrock	bedrockruntime	bedrockagent	bedrockagentruntime
AWS SDK for Java	AWS SDK for Java code examples	bedrock	bedrockruntime	bedrockagent	bedrockagentruntime
AWS SDK for JavaScript	AWS SDK for JavaScript code examples	bedrock	bedrock-runtime	bedrock-agent	bedrock-agent-runtime
AWS SDK for Kotlin	AWS SDK for Kotlin code examples	bedrock	bedrockruntime	bedrockagent	bedrockagentruntime
AWS SDK for .NET	AWS SDK for .NET code examples	Bedrock	BedrockRuntime	BedrockAgent	BedrockAgentRuntime
AWS SDK for PHP	AWS SDK for PHP code examples	Bedrock	BedrockRuntime	BedrockAgent	BedrockAgentRuntime
AWS SDK for Python (Boto3)	AWS SDK for Python (Boto3) code examples	bedrock	bedrock-runtime	bedrock-agent	bedrock-agent-runtime

SDK documentation	Code examples	Amazon Bedrock prefix	Amazon Bedrock runtime prefix	Agents for Amazon Bedrock prefix	Agents for Amazon Bedrock runtime prefix
AWS SDK for Ruby	AWS SDK for Ruby code examples	Bedrock	BedrockRuntime	BedrockAgent	BedrockAgentRuntime
AWS SDK for Rust	AWS SDK for Rust code examples	aws-sdk-bedrock	aws-sdk-bedrockruntime	aws-sdk-bedrockagent	aws-sdk-bedrockagentruntime
AWS SDK for SAP ABAP	AWS SDK for SAP ABAP code examples	BDK	BDR	BDA	BDZ
AWS SDK for Swift	AWS SDK for Swift code examples	AWSBedrock	AWSBedrockRuntime	AWSBedrockAgent	AWSBedrockAgentRuntime

Topics

- [Actions](#)
- [Data Types](#)
- [Common Parameters](#)
- [Common Errors](#)

Actions

The following actions are supported by Amazon Bedrock:

- [CreateEvaluationJob](#)
- [CreateGuardrail](#)

- [CreateGuardrailVersion](#)
- [CreateModelCustomizationJob](#)
- [CreateProvisionedModelThroughput](#)
- [DeleteCustomModel](#)
- [DeleteGuardrail](#)
- [DeleteModelInvocationLoggingConfiguration](#)
- [DeleteProvisionedModelThroughput](#)
- [GetCustomModel](#)
- [GetEvaluationJob](#)
- [GetFoundationModel](#)
- [GetGuardrail](#)
- [GetModelCustomizationJob](#)
- [GetModelInvocationLoggingConfiguration](#)
- [GetProvisionedModelThroughput](#)
- [ListCustomModels](#)
- [ListEvaluationJobs](#)
- [ListFoundationModels](#)
- [ListGuardrails](#)
- [ListModelCustomizationJobs](#)
- [ListProvisionedModelThroughputs](#)
- [ListTagsForResource](#)
- [PutModelInvocationLoggingConfiguration](#)
- [StopEvaluationJob](#)
- [StopModelCustomizationJob](#)
- [TagResource](#)
- [UntagResource](#)
- [UpdateGuardrail](#)
- [UpdateProvisionedModelThroughput](#)

The following actions are supported by Agents for Amazon Bedrock:

- [AssociateAgentKnowledgeBase](#)
- [CreateAgent](#)
- [CreateAgentActionGroup](#)
- [CreateAgentAlias](#)
- [CreateDataSource](#)
- [CreateKnowledgeBase](#)
- [DeleteAgent](#)
- [DeleteAgentActionGroup](#)
- [DeleteAgentAlias](#)
- [DeleteAgentVersion](#)
- [DeleteDataSource](#)
- [DeleteKnowledgeBase](#)
- [DisassociateAgentKnowledgeBase](#)
- [GetAgent](#)
- [GetAgentActionGroup](#)
- [GetAgentAlias](#)
- [GetAgentKnowledgeBase](#)
- [GetAgentVersion](#)
- [GetDataSource](#)
- [GetIngestionJob](#)
- [GetKnowledgeBase](#)
- [ListAgentActionGroups](#)
- [ListAgentAliases](#)
- [ListAgentKnowledgeBases](#)
- [ListAgents](#)
- [ListAgentVersions](#)
- [ListDataSources](#)
- [ListIngestionJobs](#)
- [ListKnowledgeBases](#)
- [ListTagsForResource](#)

- [PrepareAgent](#)
- [StartIngestionJob](#)
- [TagResource](#)
- [UntagResource](#)
- [UpdateAgent](#)
- [UpdateAgentActionGroup](#)
- [UpdateAgentAlias](#)
- [UpdateAgentKnowledgeBase](#)
- [UpdateDataSource](#)
- [UpdateKnowledgeBase](#)

The following actions are supported by Agents for Amazon Bedrock Runtime:

- [InvokeAgent](#)
- [Retrieve](#)
- [RetrieveAndGenerate](#)

The following actions are supported by Amazon Bedrock Runtime:

- [InvokeModel](#)
- [InvokeModelWithResponseStream](#)

Amazon Bedrock

The following actions are supported by Amazon Bedrock:

- [CreateEvaluationJob](#)
- [CreateGuardrail](#)
- [CreateGuardrailVersion](#)
- [CreateModelCustomizationJob](#)
- [CreateProvisionedModelThroughput](#)
- [DeleteCustomModel](#)
- [DeleteGuardrail](#)

- [DeleteModelInvocationLoggingConfiguration](#)
- [DeleteProvisionedModelThroughput](#)
- [GetCustomModel](#)
- [GetEvaluationJob](#)
- [GetFoundationModel](#)
- [GetGuardrail](#)
- [GetModelCustomizationJob](#)
- [GetModelInvocationLoggingConfiguration](#)
- [GetProvisionedModelThroughput](#)
- [ListCustomModels](#)
- [ListEvaluationJobs](#)
- [ListFoundationModels](#)
- [ListGuardrails](#)
- [ListModelCustomizationJobs](#)
- [ListProvisionedModelThroughputs](#)
- [ListTagsForResource](#)
- [PutModelInvocationLoggingConfiguration](#)
- [StopEvaluationJob](#)
- [StopModelCustomizationJob](#)
- [TagResource](#)
- [UntagResource](#)
- [UpdateGuardrail](#)
- [UpdateProvisionedModelThroughput](#)

CreateEvaluationJob

Service: Amazon Bedrock

API operation for creating and managing Amazon Bedrock automatic model evaluation jobs and model evaluation jobs that use human workers. To learn more about the requirements for creating a model evaluation job see, [Model evaluations](#).

Request Syntax

```
POST /evaluation-jobs HTTP/1.1
Content-type: application/json

{
  "clientRequestToken": "string",
  "customerEncryptionKeyId": "string",
  "evaluationConfig": { ... },
  "inferenceConfig": { ... },
  "jobDescription": "string",
  "jobName": "string",
  "jobTags": [
    {
      "key": "string",
      "value": "string"
    }
  ],
  "outputDataConfig": {
    "s3Uri": "string"
  },
  "roleArn": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

customerEncryptionKeyId

Specify your customer managed key ARN that will be used to encrypt your model evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[[:^:]]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

Required: No

evaluationConfig

Specifies whether the model evaluation job is automatic or uses human worker.

Type: [EvaluationConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

inferenceConfig

Specify the models you want to use in your model evaluation job. Automatic model evaluation jobs support a single model, and model evaluation job that use human workers support two models.

Type: [EvaluationInferenceConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

jobDescription

A description of the model evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `^ . +$`

Required: No

jobName

The name of the model evaluation job. Model evaluation job names must be unique with your AWS account, and your account's AWS region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9](-*[a-z0-9]){0,62}$`

Required: Yes

jobTags

Tags to attach to the model evaluation job.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

outputDataConfig

An object that defines where the results of model evaluation job will be saved in Amazon S3.

Type: [EvaluationOutputDataConfig](#) object

Required: Yes

roleArn

The Amazon Resource Name (ARN) of an IAM service role that Amazon Bedrock can assume to perform tasks on your behalf. The service role must have Amazon Bedrock as the service principal, and provide access to any Amazon S3 buckets specified in the `EvaluationConfig` object. To pass this role to Amazon Bedrock, the caller of this API must have the `iam:PassRole` permission. To learn more about the required permissions, see [Required permissions](#).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/\.+\$`

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

jobArn

The ARN of the model evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}):evaluation-job/[a-z0-9]{12}\$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateGuardrail

Service: Amazon Bedrock

Creates a guardrail to block topics and to implement safeguards for your generative AI applications. You can configure denied topics to disallow undesirable topics and content filters to block harmful content in model inputs and responses. For more information, see [Guardrails for Amazon Bedrock](#) in the *Amazon Bedrock User Guide*

Request Syntax

```
POST /guardrails HTTP/1.1
Content-type: application/json

{
  "blockedInputMessaging": "string",
  "blockedOutputsMessaging": "string",
  "clientRequestToken": "string",
  "contentPolicyConfig": {
    "filtersConfig": [
      {
        "inputStrength": "string",
        "outputStrength": "string",
        "type": "string"
      }
    ]
  },
  "description": "string",
  "kmsKeyId": "string",
  "name": "string",
  "sensitiveInformationPolicyConfig": {
    "piiEntitiesConfig": [
      {
        "action": "string",
        "type": "string"
      }
    ],
    "regexesConfig": [
      {
        "action": "string",
        "description": "string",
        "name": "string",
        "pattern": "string"
      }
    ]
  }
}
```

```
    ]
  },
  "tags": [
    {
      "key": "string",
      "value": "string"
    }
  ],
  "topicPolicyConfig": {
    "topicsConfig": [
      {
        "definition": "string",
        "examples": [ "string" ],
        "name": "string",
        "type": "string"
      }
    ]
  },
  "wordPolicyConfig": {
    "managedWordListsConfig": [
      {
        "type": "string"
      }
    ],
    "wordsConfig": [
      {
        "text": "string"
      }
    ]
  }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

blockedInputMessaging

The message to return when the guardrail blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

blockedOutputsMessaging

The message to return when the guardrail blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#) in the *Amazon S3 User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

contentPolicyConfig

The content filter policies to configure for the guardrail.

Type: [GuardrailContentPolicyConfig](#) object

Required: No

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

kmsKeyId

The ARN of the AWS KMS key that you use to encrypt the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

Required: No

name

The name to give the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `^[0-9a-zA-Z-_$]+`

Required: Yes

sensitiveInformationPolicyConfig

The sensitive information policy to configure for the guardrail.

Type: [GuardrailSensitiveInformationPolicyConfig](#) object

Required: No

tags

The tags that you want to attach to the guardrail.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

topicPolicyConfig

The topic policies to configure for the guardrail.

Type: [GuardrailTopicPolicyConfig](#) object

Required: No

[wordPolicyConfig](#)

The word policy you configure for the guardrail.

Type: [GuardrailWordPolicyConfig](#) object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "createdAt": "string",
  "guardrailArn": "string",
  "guardrailId": "string",
  "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

[createdAt](#)

The time at which the guardrail was created.

Type: Timestamp

[guardrailArn](#)

The ARN of the guardrail that was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+$`

guardrailId

The unique identifier of the guardrail that was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

version

The version of the guardrail that was created. This value should be 1.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateGuardrailVersion

Service: Amazon Bedrock

Creates a version of the guardrail. Use this API to create a snapshot of the guardrail when you are satisfied with a configuration, or to compare the configuration with another version.

Request Syntax

```
POST /guardrails/guardrailIdentifier HTTP/1.1
Content-type: application/json
```

```
{
  "clientRequestToken": "string",
  "description": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#) in the *Amazon S3 User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

A description of the guardrail version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "guardrailId": "string",
  "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

guardrailId

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

version

The number of the version of the guardrail.

Type: String

Pattern: `^[1-9][0-9]{0,7}$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateModelCustomizationJob

Service: Amazon Bedrock

Creates a fine-tuning job to customize a base model.

You specify the base foundation model and the location of the training data. After the model-customization job completes successfully, your custom model resource will be ready to use. Amazon Bedrock returns validation loss metrics and output generations after the job completes.

For information on the format of training and validation data, see [Prepare the datasets](#).

Model-customization jobs are asynchronous and the completion time depends on the base model and the training/validation data size. To monitor a job, use the `GetModelCustomizationJob` operation to retrieve the job status.

For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "baseModelIdentifier": "string",
  "clientRequestToken": "string",
  "customizationType": "string",
  "customModelKmsKeyId": "string",
  "customModelName": "string",
  "customModelTags": [
    {
      "key": "string",
      "value": "string"
    }
  ],
  "hyperParameters": {
    "string" : "string"
  },
  "jobName": "string",
  "jobTags": [
    {
      "key": "string",
      "value": "string"
    }
  ]
}
```

```

],
"outputDataConfig": {
  "s3Uri": "string"
},
"roleArn": "string",
"trainingDataConfig": {
  "s3Uri": "string"
},
"validationDataConfig": {
  "validators": [
    {
      "s3Uri": "string"
    }
  ]
},
"vpcConfig": {
  "securityGroupIds": [ "string" ],
  "subnetIds": [ "string" ]
}
}

```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

baseModelIdentifier

Name of the base model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^\:]+)?):bedrock:[a-z0-9-]{1,20}:((([0-9]{12}:custom-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|((([0-9a-zA-Z][_]?)+))$`

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

customizationType

The customization type.

Type: String

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

Required: No

customModelKmsKeyId

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

Required: No

customModelName

A name for the resulting custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([\0-9a-zA-Z][_-]?)+$`

Required: Yes

customModelTags

Tags to attach to the resulting custom model.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

hyperParameters

Parameters related to tuning the model. For details on the format for different models, see [Custom model hyperparameters](#).

Type: String to string map

Required: Yes

jobName

A name for the fine-tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\])*$`

Required: Yes

jobTags

Tags to attach to the job.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

[outputDataConfig](#)

S3 location for the output data.

Type: [OutputDataConfig](#) object

Required: Yes

[roleArn](#)

The Amazon Resource Name (ARN) of an IAM service role that Amazon Bedrock can assume to perform tasks on your behalf. For example, during model training, Amazon Bedrock needs your permission to read input data from an S3 bucket, write model artifacts to an S3 bucket. To pass this role to Amazon Bedrock, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/\.+\$`

Required: Yes

[trainingDataConfig](#)

Information about the training dataset.

Type: [TrainingDataConfig](#) object

Required: Yes

[validationDataConfig](#)

Information about the validation dataset.

Type: [ValidationDataConfig](#) object

Required: No

[vpcConfig](#)

VPC configuration (optional). Configuration parameters for the private Virtual Private Cloud (VPC) that contains the resources you are using for this job.

Type: [VpcConfig](#) object

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
  "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn

Amazon Resource Name (ARN) of the fine tuning job

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateProvisionedModelThroughput

Service: Amazon Bedrock

Creates dedicated throughput for a base or custom model with the model units and for the duration that you specify. For pricing details, see [Amazon Bedrock Pricing](#). For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /provisioned-model-throughput HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "clientRequestToken": "string",
  "commitmentDuration": "string",
  "modelId": "string",
  "modelUnits": number,
  "provisionedModelName": "string",
  "tags": [
    {
      "key": "string",
      "value": "string"
    }
  ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#) in the Amazon S3 User Guide.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

commitmentDuration

The commitment duration requested for the Provisioned Throughput. Billing occurs hourly and is discounted for longer commitment terms. To request a no-commit Provisioned Throughput, omit this field.

Custom models support all levels of commitment. To see which base models support no commitment, see [Supported regions and models for Provisioned Throughput](#) in the Amazon Bedrock User Guide

Type: String

Valid Values: OneMonth | SixMonths

Required: No

modelId

The Amazon Resource Name (ARN) or name of the model to associate with this Provisioned Throughput. For a list of models for which you can purchase Provisioned Throughput, see [Amazon Bedrock model IDs for purchasing Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:((\[[0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((\[[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|((\[[0-9a-zA-Z][_]?)+))$`

Required: Yes

modelUnits

Number of model units to allocate. A model unit delivers a specific throughput level for the specified model. The throughput level of a model unit specifies the total number of input and output tokens that it can process and generate within a span of one minute. By default, your account has no model units for purchasing Provisioned Throughputs with commitment. You must first visit the [AWS support center](#) to request MUs.

For model unit quotas, see [Provisioned Throughput quotas](#) in the Amazon Bedrock User Guide.

For more information about what an MU specifies, contact your AWS account manager.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelName

The name for this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[0-9a-zA-Z][_ -]?+$`

Required: Yes

tags

Tags to associate with this Provisioned Throughput.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
```



```
Content-type: application/json

{
  "provisionedModelArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

provisionedModelArn

The Amazon Resource Name (ARN) for this Provisioned Throughput.

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteCustomModel

Service: Amazon Bedrock

Deletes a custom model that you created earlier. For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name of the model to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|((([0-9a-zA-Z][_]?)+))$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteGuardrail

Service: Amazon Bedrock

Deletes a guardrail.

- To delete a guardrail, only specify the ARN of the guardrail in the `guardrailIdentifier` field. If you delete a guardrail, all of its versions will be deleted.
- To delete a version of a guardrail, specify the ARN of the guardrail in the `guardrailIdentifier` field and the version in the `guardrailVersion` field.

Request Syntax

```
DELETE /guardrails/guardrailIdentifier?guardrailVersion=guardrailVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

Required: Yes

guardrailVersion

The version of the guardrail.

Pattern: `^[1-9][0-9]{0,7}$`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Delete the invocation logging.

Request Syntax

```
DELETE /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteProvisionedModelThroughput

Service: Amazon Bedrock

Deletes a Provisioned Throughput. You can't delete a Provisioned Throughput before the commitment term is over. For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[provisionedModelId](#)

The Amazon Resource Name (ARN) or name of the Provisioned Throughput.

Pattern: `^((([0-9a-zA-Z] [_-]?)+) | (arn:aws(- [^:]+)? : bedrock : [a-z0-9-] { 1, 20 } : [0-9] { 12 } : provisioned-model / [a-z0-9] { 12 })) $`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetCustomModel

Service: Amazon Bedrock

Get the properties associated with a Amazon Bedrock custom model that you have created. For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name or Amazon Resource Name (ARN) of the custom model.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:((\[[0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]?{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/[a-z0-9-]{1,63}[\.]?{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|((\[[a-z0-9-]{1,63}[\.]?{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|((\[[0-9a-zA-Z][_-]?)+))$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "baseModelArn": "string",
```

```

"creationTime": "string",
"customizationType": "string",
"hyperParameters": {
  "string" : "string"
},
"jobArn": "string",
"jobName": "string",
"modelArn": "string",
"modelKmsKeyArn": "string",
"modelName": "string",
"outputDataConfig": {
  "s3Uri": "string"
},
"trainingDataConfig": {
  "s3Uri": "string"
},
"trainingMetrics": {
  "trainingLoss": number
},
"validationDataConfig": {
  "validators": [
    {
      "s3Uri": "string"
    }
  ]
},
"validationMetrics": [
  {
    "validationLoss": number
  }
]
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([[:][a-z0-9-]{1,63}){0,2}))$`

creationTime

Creation time of the model.

Type: Timestamp

customizationType

The type of model customization.

Type: String

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

hyperParameters

Hyperparameter values associated with this model. For details on the format for different models, see [Custom model hyperparameters](#).

Type: String to string map

jobArn

Job Amazon Resource Name (ARN) associated with this model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([[:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})$`

jobName

Job name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\.])*$`

modelArn

Amazon Resource Name (ARN) associated with this model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}:((([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{1}){0,2}[a-z0-9-]{1,63}([[:a-z0-9-]{1,63}){0,2})))$`

modelKmsKeyArn

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

modelName

Model name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

outputDataConfig

Output data configuration associated with this custom model.

Type: [OutputDataConfig](#) object

trainingDataConfig

Contains information about the training dataset.

Type: [TrainingDataConfig](#) object

[trainingMetrics](#)

Contains training metrics from the job creation.

Type: [TrainingMetrics](#) object

[validationDataConfig](#)

Contains information about the validation dataset.

Type: [ValidationDataConfig](#) object

[validationMetrics](#)

The validation metrics from the job creation.

Type: Array of [ValidatorMetric](#) objects

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetEvaluationJob

Service: Amazon Bedrock

Retrieves the properties associated with a model evaluation job, including the status of the job. For more information, see [Model evaluations](#).

Request Syntax

```
GET /evaluation-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The Amazon Resource Name (ARN) of the model evaluation job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^\:]+)?):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "creationTime": "string",
  "customerEncryptionKeyId": "string",
  "evaluationConfig": { ... },
  "failureMessages": [ "string" ],
  "inferenceConfig": { ... },
  "jobArn": "string",
  "jobDescription": "string",
```

```
"jobName": "string",
"jobType": "string",
"lastModifiedTime": "string",
"outputDataConfig": {
  "s3Uri": "string"
},
"roleArn": "string",
"status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

creationTime

When the model evaluation job was created.

Type: Timestamp

customerEncryptionKeyId

The Amazon Resource Name (ARN) of the customer managed key specified when the model evaluation job was created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

evaluationConfig

Contains details about the type of model evaluation job, the metrics used, the task type selected, the datasets used, and any custom metrics you defined.

Type: [EvaluationConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

failureMessages

An array of strings the specify why the model evaluation job has failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

[inferenceConfig](#)

Details about the models you specified in your model evaluation job.

Type: [EvaluationInferenceConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

[jobArn](#)

The Amazon Resource Name (ARN) of the model evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12}$`

[jobDescription](#)

The description of the model evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `^.+ $`

[jobName](#)

The name of the model evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9](-*[a-z0-9]){0,62}$`

[jobType](#)

The type of model evaluation job.

Type: String

Valid Values: Human | Automated

lastModifiedTime

When the model evaluation job was last modified.

Type: Timestamp

outputDataConfig

Amazon S3 location for where output data is saved.

Type: [EvaluationOutputDataConfig](#) object

roleArn

The Amazon Resource Name (ARN) of the IAM service role used in the model evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/\.+\$`

status

The status of the model evaluation job.

Type: String

Valid Values: InProgress | Completed | Failed | Stopping | Stopped

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)


```
"modelArn": "string",
"modelId": "string",
"modelLifecycle": {
  "status": "string"
},
"modelName": "string",
"outputModalities": [ "string" ],
"providerName": "string",
"responseStreamingSupported": boolean
}
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelDetails

Information about the foundation model.

Type: [FoundationModelDetails](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetGuardrail

Service: Amazon Bedrock

Gets details about a guardrail. If you don't specify a version, the response returns details for the DRAFT version.

Request Syntax

```
GET /guardrails/guardrailIdentifier?guardrailVersion=guardrailVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail for which to get details.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

Required: Yes

guardrailVersion

The version of the guardrail for which to get details. If you don't specify a version, the response returns details for the DRAFT version.

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "blockedInputMessaging": "string",
  "blockedOutputsMessaging": "string",
```

```
"contentPolicy": {
  "filters": [
    {
      "inputStrength": "string",
      "outputStrength": "string",
      "type": "string"
    }
  ]
},
"createdAt": "string",
"description": "string",
"failureRecommendations": [ "string" ],
"guardrailArn": "string",
"guardrailId": "string",
"kmsKeyArn": "string",
"name": "string",
"sensitiveInformationPolicy": {
  "piiEntities": [
    {
      "action": "string",
      "type": "string"
    }
  ],
  "regexes": [
    {
      "action": "string",
      "description": "string",
      "name": "string",
      "pattern": "string"
    }
  ]
},
"status": "string",
"statusReasons": [ "string" ],
"topicPolicy": {
  "topics": [
    {
      "definition": "string",
      "examples": [ "string" ],
      "name": "string",
      "type": "string"
    }
  ]
},
```

```
"updatedAt": "string",
"version": "string",
"wordPolicy": {
  "managedWordLists": [
    {
      "type": "string"
    }
  ],
  "words": [
    {
      "text": "string"
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

blockedInputMessaging

The message that the guardrail returns when it blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

blockedOutputsMessaging

The message that the guardrail returns when it blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

contentPolicy

The content policy that was configured for the guardrail.

Type: [GuardrailContentPolicy](#) object

createdAt

The date and time at which the guardrail was created.

Type: Timestamp

description

The description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

failureRecommendations

Appears if the status of the guardrail is FAILED. A list of recommendations to carry out before retrying the request.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 200.

guardrailArn

The ARN of the guardrail that was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+$`

guardrailId

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

kmsKeyArn

The ARN of the AWS KMS key that encrypts the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

name

The name of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `^[0-9a-zA-Z-_$]+`

sensitiveInformationPolicy

The sensitive information policy that was configured for the guardrail.

Type: [GuardrailSensitiveInformationPolicy](#) object

status

The status of the guardrail.

Type: String

Valid Values: CREATING | UPDATING | VERSIONING | READY | FAILED | DELETING

statusReasons

Appears if the status is FAILED. A list of reasons for why the guardrail failed to be created, updated, versioned, or deleted.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 200.

topicPolicy

The topic policy that was configured for the guardrail.

Type: [GuardrailTopicPolicy](#) object

updatedAt

The date and time at which the guardrail was updated.

Type: Timestamp

version

The version of the guardrail.

Type: String

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

wordPolicy

The word policy that was configured for the guardrail.

Type: [GuardrailWordPolicy](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetModelCustomizationJob

Service: Amazon Bedrock

Retrieves the properties associated with a model-customization job, including the status of the job. For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

Identifier for the customization job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^\:]+)??:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[\.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\.])*)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "baseModelArn": "string",
  "clientRequestToken": "string",
  "creationTime": "string",
  "customizationType": "string",
  "endTime": "string",
```

```

"failureMessage": "string",
"hyperParameters": {
  "string" : "string"
},
"jobArn": "string",
"jobName": "string",
"lastModifiedTime": "string",
"outputDataConfig": {
  "s3Uri": "string"
},
"outputModelArn": "string",
"outputModelKmsKeyArn": "string",
"outputModelName": "string",
"roleArn": "string",
"status": "string",
"trainingDataConfig": {
  "s3Uri": "string"
},
"trainingMetrics": {
  "trainingLoss": number
},
"validationDataConfig": {
  "validators": [
    {
      "s3Uri": "string"
    }
  ]
},
"validationMetrics": [
  {
    "validationLoss": number
  }
],
"vpcConfig": {
  "securityGroupIds": [ "string" ],
  "subnetIds": [ "string" ]
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[\.]{1}([a-z0-9-]{1,63}[\.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}$`

clientRequestToken

The token that you specified in the `CreateCustomizationJob` request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

creationTime

Time that the resource was created.

Type: Timestamp

customizationType

The type of model customization.

Type: String

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

endTime

Time that the resource transitioned to terminal state.

Type: Timestamp

failureMessage

Information about why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

hyperParameters

The hyperparameter values for the job. For details on the format for different models, see [Custom model hyperparameters](#).

Type: String to string map

jobArn

The Amazon Resource Name (ARN) of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

jobName

The name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\.])*$`

lastModifiedTime

Time that the resource was last modified.

Type: Timestamp

outputDataConfig

Output data configuration

Type: [OutputDataConfig](#) object

outputModelArn

The Amazon Resource Name (ARN) of the output model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[\.]{1}){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

outputModelKmsKeyArn

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

outputModelName

The name of the output model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_ -]?)+$`

roleArn

The Amazon Resource Name (ARN) of the IAM role.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/.[+]$`

status

The status of the job. A successful job transitions from in-progress to completed when the output model is ready to use. If the job failed, the failure message contains information about why the job failed.

Type: String

Valid Values: `InProgress` | `Completed` | `Failed` | `Stopping` | `Stopped`

[trainingDataConfig](#)

Contains information about the training dataset.

Type: [TrainingDataConfig](#) object

[trainingMetrics](#)

Contains training metrics from the job creation.

Type: [TrainingMetrics](#) object

[validationDataConfig](#)

Contains information about the validation dataset.

Type: [ValidationDataConfig](#) object

[validationMetrics](#)

The loss metric for each validator that you provided in the createjob request.

Type: Array of [ValidatorMetric](#) objects

[vpcConfig](#)

VPC configuration for the custom model job.

Type: [VpcConfig](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Get the current configuration values for model invocation logging.

Request Syntax

```
GET /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "loggingConfig": {
    "cloudWatchConfig": {
      "largeDataDeliveryS3Config": {
        "bucketName": "string",
        "keyPrefix": "string"
      },
      "logGroupName": "string",
      "roleArn": "string"
    },
    "embeddingDataDeliveryEnabled": boolean,
    "imageDataDeliveryEnabled": boolean,
    "s3Config": {
      "bucketName": "string",
      "keyPrefix": "string"
    },
    "textDataDeliveryEnabled": boolean
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

loggingConfig

The current configuration values.

Type: [LoggingConfig](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetProvisionedModelThroughput

Service: Amazon Bedrock

Returns details for a Provisioned Throughput. For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId

The Amazon Resource Name (ARN) or name of the Provisioned Throughput.

Pattern: `^((([0-9a-zA-Z] [_-] ?)+)|(arn:aws(- [^:]+)? :bedrock:[a-z0-9-]{1,20} : [0-9]{12} :provisioned-model/[a-z0-9]{12})))$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "commitmentDuration": "string",
  "commitmentExpirationTime": "string",
  "creationTime": "string",
  "desiredModelArn": "string",
  "desiredModelUnits": number,
  "failureMessage": "string",
  "foundationModelArn": "string",
  "lastModifiedTime": "string",
  "modelArn": "string",
```

```
"modelUnits": number,  
"provisionedModelArn": "string",  
"provisionedModelName": "string",  
"status": "string"  
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

commitmentDuration

Commitment duration of the Provisioned Throughput.

Type: String

Valid Values: OneMonth | SixMonths

commitmentExpirationTime

The timestamp for when the commitment term for the Provisioned Throughput expires.

Type: Timestamp

creationTime

The timestamp of the creation time for this Provisioned Throughput.

Type: Timestamp

desiredModelArn

The Amazon Resource Name (ARN) of the model requested to be associated to this Provisioned Throughput. This value differs from the `modelArn` if updating hasn't completed.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{1}){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

desiredModelUnits

The number of model units that was requested for this Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

failureMessage

A failure message for any issues that occurred during creation, updating, or deletion of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

foundationModelArn

The Amazon Resource Name (ARN) of the base model for which the Provisioned Throughput was created, or of the base model that the custom model for which the Provisioned Throughput was created was customized.

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}$`

lastModifiedTime

The timestamp of the last time that this Provisioned Throughput was modified.

Type: Timestamp

modelArn

The Amazon Resource Name (ARN) of the model associated with this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-`


```
model/[a-z0-9-]{1,63}[\.]{1}([a-z0-9-]{1,63}[\.]){0,2}[a-z0-9-]{1,63}([:]  
[a-z0-9-]{1,63}){0,2}))$
```

modelUnits

The number of model units allocated to this Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

provisionedModelArn

The Amazon Resource Name (ARN) of the Provisioned Throughput.

Type: String

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

provisionedModelName

The name of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_ -]?)+$`

status

The status of the Provisioned Throughput.

Type: String

Valid Values: `Creating | InService | Updating | Failed`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListCustomModels

Service: Amazon Bedrock

Returns a list of the custom models that you have created with the `CreateModelCustomizationJob` operation.

For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /custom-models?  
baseModelArnEquals=baseModelArnEquals&creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore  
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[baseModelArnEquals](#)

Return custom models only if the base model Amazon Resource Name (ARN) matches this parameter.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{0,2}[a-z0-9-]{1,63}([[:a-z0-9-]{1,63}){0,2})))$`

[creationTimeAfter](#)

Return custom models created after the specified time.

[creationTimeBefore](#)

Return custom models created before the specified time.

[foundationModelArnEquals](#)

Return custom models only if the foundation model Amazon Resource Name (ARN) matches this parameter.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

maxResults

Maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

Return custom models only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_]?)+$`

nextToken

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy

The field to sort by in the returned list of models.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending` | `Descending`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
  "modelSummaries": [
    {
      "baseModelArn": "string",
      "baseModelName": "string",
      "creationTime": "string",
      "customizationType": "string",
      "modelArn": "string",
      "modelName": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries

Model summaries.

Type: Array of [CustomModelSummary](#) objects

nextToken

Continuation token for the next request to list the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: $^\backslash S^*\$$

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListEvaluationJobs

Service: Amazon Bedrock

Lists model evaluation jobs.

Request Syntax

```
GET /evaluation-jobs?  
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults  
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[creationTimeAfter](#)

A filter that includes model evaluation jobs created after the time specified.

[creationTimeBefore](#)

A filter that includes model evaluation jobs created prior to the time specified.

[maxResults](#)

The maximum number of results to return.

Valid Range: Minimum value of 1. Maximum value of 1000.

[nameContains](#)

Query parameter string for model evaluation job names.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9](-*[a-z0-9]){0,62}$`

[nextToken](#)

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy

Allows you to sort model evaluation jobs by when they were created.

Valid Values: `CreationTime`

sortOrder

How you want the order of jobs sorted.

Valid Values: `Ascending` | `Descending`

statusEquals

Only return jobs where the status condition is met.

Valid Values: `InProgress` | `Completed` | `Failed` | `Stopping` | `Stopped`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "jobSummaries": [
    {
      "creationTime": "string",
      "evaluationTaskTypes": [ "string" ],
      "jobArn": "string",
      "jobName": "string",
      "jobType": "string",
      "modelIdentifiers": [ "string" ],
      "status": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobSummaries

A summary of the model evaluation jobs.

Type: Array of [EvaluationSummary](#) objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

nextToken

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListFoundationModels

Service: Amazon Bedrock

Lists Amazon Bedrock foundation models that you can use. You can filter the results with the request parameters. For more information, see [Foundation models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /foundation-models?  
byCustomizationType=byCustomizationType&byInferenceType=byInferenceType&byOutputModality=byOutputModality  
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[byCustomizationType](#)

Return models that support the customization type that you specify. For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

[byInferenceType](#)

Return models that support the inference type that you specify. For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Valid Values: ON_DEMAND | PROVISIONED

[byOutputModality](#)

Return models that support the output modality that you specify.

Valid Values: TEXT | IMAGE | EMBEDDING

[byProvider](#)

Return models belonging to the model provider that you specify.

Pattern: `^[A-Za-z0-9-]{1,63}$`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "modelSummaries": [
    {
      "customizationsSupported": [ "string" ],
      "inferenceTypesSupported": [ "string" ],
      "inputModalities": [ "string" ],
      "modelArn": "string",
      "modelId": "string",
      "modelLifecycle": {
        "status": "string"
      },
      "modelName": "string",
      "outputModalities": [ "string" ],
      "providerName": "string",
      "responseStreamingSupported": boolean
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries

A list of Amazon Bedrock foundation models.

Type: Array of [FoundationModelSummary](#) objects

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListGuardrails

Service: Amazon Bedrock

Lists details about all the guardrails in an account. To list the DRAFT version of all your guardrails, don't specify the `guardrailIdentifier` field. To list all versions of a guardrail, specify the ARN of the guardrail in the `guardrailIdentifier` field.

You can 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 `ListGuardrails` request to see the next batch of results.

Request Syntax

```
GET /guardrails?
guardrailIdentifier=guardrailIdentifier&maxResults=maxResults&nextToken=nextToken
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

maxResults

The maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If there are more results than were returned in the response, the response returns a `nextToken` that you can send in another `ListGuardrails` request to see the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "guardrails": [
    {
      "arn": "string",
      "createdAt": "string",
      "description": "string",
      "id": "string",
      "name": "string",
      "status": "string",
      "updatedAt": "string",
      "version": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

guardrails

A list of objects, each of which contains details about a guardrail.

Type: Array of [GuardrailSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

nextToken

If there are more results than were returned in the response, the response returns a `nextToken` that you can send in another `ListGuardrails` request to see the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListModelCustomizationJobs

Service: Amazon Bedrock

Returns a list of model customization jobs that you have submitted. You can filter the jobs to return based on one or more criteria.

For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs?  
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults  
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[creationTimeAfter](#)

Return customization jobs created after the specified time.

[creationTimeBefore](#)

Return customization jobs created before the specified time.

[maxResults](#)

Maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

[nameContains](#)

Return customization jobs only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\])*$`

[nextToken](#)

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy

The field to sort by in the returned list of jobs.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending` | `Descending`

statusEquals

Return customization jobs with the specified status.

Valid Values: `InProgress` | `Completed` | `Failed` | `Stopping` | `Stopped`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "modelCustomizationJobSummaries": [
    {
      "baseModelArn": "string",
      "creationTime": "string",
      "customizationType": "string",
      "customModelArn": "string",
      "customModelName": "string",
      "endTime": "string",
      "jobArn": "string",
      "jobName": "string",
      "lastModifiedTime": "string",
      "status": "string"
    }
  ]
}
```

```
  ],  
  "nextToken": "string"  
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelCustomizationJobSummaries

Job summaries.

Type: Array of [ModelCustomizationJobSummary](#) objects

nextToken

Page continuation token to use in the next request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: $^\backslash S^*\$$

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListProvisionedModelThroughputs

Service: Amazon Bedrock

Lists the Provisioned Throughputs in the account. For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughputs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[creationTimeAfter](#)

A filter that returns Provisioned Throughputs created after the specified time.

[creationTimeBefore](#)

A filter that returns Provisioned Throughputs created before the specified time.

[maxResults](#)

The maximum number of results to return in the response. If there are more results than the number you specified, the response returns a `nextToken` value. To see the next batch of results, send the `nextToken` value in another list request.

Valid Range: Minimum value of 1. Maximum value of 1000.

[modelArnEquals](#)

A filter that returns Provisioned Throughputs whose model Amazon Resource Name (ARN) is equal to the value that you specify.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2})))$`

nameContains

A filter that returns Provisioned Throughputs if their name contains the expression that you specify.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_]?)+$`

nextToken

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, specify the `nextToken` value in this field.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy

The field by which to sort the returned list of Provisioned Throughputs.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending` | `Descending`

statusEquals

A filter that returns Provisioned Throughputs if their statuses matches the value that you specify.

Valid Values: `Creating` | `InService` | `Updating` | `Failed`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```



```
{
  "nextToken": "string",
  "provisionedModelSummaries": [
    {
      "commitmentDuration": "string",
      "commitmentExpirationTime": "string",
      "creationTime": "string",
      "desiredModelArn": "string",
      "desiredModelUnits": number,
      "foundationModelArn": "string",
      "lastModifiedTime": "string",
      "modelArn": "string",
      "modelUnits": number,
      "provisionedModelArn": "string",
      "provisionedModelName": "string",
      "status": "string"
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If there are more results than the number you specified in the `maxResults` field, this value is returned. To see the next batch of results, include this value in the `nextToken` field in another list request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

provisionedModelSummaries

A list of summaries, one for each Provisioned Throughput in the response.

Type: Array of [ProvisionedModelSummary](#) objects

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)

- [AWS SDK for Ruby V3](#)

ListTagsForResource

Service: Amazon Bedrock

List the tags associated with the specified resource.

For more information, see [Tagging resources](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /listTagsForResource HTTP/1.1
```

```
Content-type: application/json
```

```
{  
  "resourceARN": "string"  
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The Amazon Resource Name (ARN) of the resource.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: (^[a-zA-Z0-9][a-zA-Z0-9\-_]*\$)|(^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)(:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})\$)|(:provisioned-model/[a-z0-9]{12}\$)|(:guardrail/[a-z0-9]+\$)|(:evaluation-job/[a-z0-9]{12}\$))

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "tags": [
    {
      "key": "string",
      "value": "string"
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

An array of the tags associated with this resource.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

PutModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Set the configuration values for model invocation logging.

Request Syntax

```
PUT /logging/modelinvocations HTTP/1.1
Content-type: application/json

{
  "loggingConfig": {
    "cloudWatchConfig": {
      "largeDataDeliveryS3Config": {
        "bucketName": "string",
        "keyPrefix": "string"
      },
      "logGroupName": "string",
      "roleArn": "string"
    },
    "embeddingDataDeliveryEnabled": boolean,
    "imageDataDeliveryEnabled": boolean,
    "s3Config": {
      "bucketName": "string",
      "keyPrefix": "string"
    },
    "textDataDeliveryEnabled": boolean
  }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

loggingConfig

The logging configuration values to set.

Type: [LoggingConfig](#) object

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)

- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

StopEvaluationJob

Service: Amazon Bedrock

Stops an in progress model evaluation job.

Request Syntax

```
POST /evaluation-job/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The ARN of the model evaluation job you want to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^\:]+)?):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)

- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

StopModelCustomizationJob

Service: Amazon Bedrock

Stops an active model customization job. For more information, see [Custom models](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[jobIdentifier](#)

Job identifier of the job to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^\:]+)??:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\.])*)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

TagResource

Service: Amazon Bedrock

Associate tags with a resource. For more information, see [Tagging resources](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /tagResource HTTP/1.1
Content-type: application/json
```

```
{
  "resourceARN": "string",
  "tags": [
    {
      "key": "string",
      "value": "string"
    }
  ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The Amazon Resource Name (ARN) of the resource to tag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: (^[a-zA-Z0-9][a-zA-Z0-9\-_]*\$)|(^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12})|((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}))\$)|(:provisioned-


```
model/[a-z0-9]{12}$)|(:guardrail/[a-z0-9]+$)|(:evaluation-job/[a-z0-9]{12}$)))
```

Required: Yes

tags

Tags to associate with the resource.

Type: Array of [Tag](#) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UntagResource

Service: Amazon Bedrock

Remove one or more tags from a resource. For more information, see [Tagging resources](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /untagResource HTTP/1.1
Content-type: application/json
```

```
{
  "resourceARN": "string",
  "tagKeys": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

[resourceARN](#)

The Amazon Resource Name (ARN) of the resource to untag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: (^[a-zA-Z0-9][a-zA-Z0-9\-_]*\$)|(^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12})|((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}(:[a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})\$)|(:provisioned-model/[a-z0-9]{12}\$)|(:guardrail/[a-z0-9]+\$)|(:evaluation-job/[a-z0-9]{12}\$)))

Required: Yes

[tagKeys](#)

Tag keys of the tags to remove from the resource.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateGuardrail

Service: Amazon Bedrock

Updates a guardrail with the values you specify.

- Specify a name and optional description.
- Specify messages for when the guardrail successfully blocks a prompt or a model response in the `blockedInputMessaging` and `blockedOutputsMessaging` fields.
- Specify topics for the guardrail to deny in the `topicPolicyConfig` object. Each [GuardrailTopicConfig](#) object in the `topicsConfig` list pertains to one topic.
 - Give a name and description so that the guardrail can properly identify the topic.
 - Specify DENY in the `type` field.
 - (Optional) Provide up to five prompts that you would categorize as belonging to the topic in the `examples` list.
- Specify filter strengths for the harmful categories defined in Amazon Bedrock in the `contentPolicyConfig` object. Each [GuardrailContentFilterConfig](#) object in the `filtersConfig` list pertains to a harmful category. For more information, see [Content filters](#). For more information about the fields in a content filter, see [GuardrailContentFilterConfig](#).
 - Specify the category in the `type` field.
 - Specify the strength of the filter for prompts in the `inputStrength` field and for model responses in the `strength` field of the [GuardrailContentFilterConfig](#).
- (Optional) For security, include the ARN of a AWS KMS key in the `kmsKeyId` field.
- (Optional) Attach any tags to the guardrail in the `tags` object. For more information, see [Tag resources](#).

Request Syntax

```
PUT /guardrails/guardrailIdentifier HTTP/1.1
Content-type: application/json

{
  "blockedInputMessaging": "string",
  "blockedOutputsMessaging": "string",
  "contentPolicyConfig": {
    "filtersConfig": [
      {
```

```

        "inputStrength": "string",
        "outputStrength": "string",
        "type": "string"
    }
]
},
"description": "string",
"kmsKeyId": "string",
"name": "string",
"sensitiveInformationPolicyConfig": {
    "piiEntitiesConfig": [
        {
            "action": "string",
            "type": "string"
        }
    ],
    "regexesConfig": [
        {
            "action": "string",
            "description": "string",
            "name": "string",
            "pattern": "string"
        }
    ]
},
"topicPolicyConfig": {
    "topicsConfig": [
        {
            "definition": "string",
            "examples": [ "string" ],
            "name": "string",
            "type": "string"
        }
    ]
},
"wordPolicyConfig": {
    "managedWordListsConfig": [
        {
            "type": "string"
        }
    ],
    "wordsConfig": [
        {
            "text": "string"
        }
    ]
}

```

```
    }  
  ]  
}
```

URI Request Parameters

The request uses the following URI parameters.

[guardrailIdentifier](#)

The unique identifier of the guardrail

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

[blockedInputMessaging](#)

The message to return when the guardrail blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

[blockedOutputsMessaging](#)

The message to return when the guardrail blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

contentPolicyConfig

The content policy to configure for the guardrail.

Type: [GuardrailContentPolicyConfig](#) object

Required: No

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

kmsKeyId

The ARN of the AWS KMS key with which to encrypt the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

Required: No

name

A name for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `^[0-9a-zA-Z-_]+$`

Required: Yes

sensitiveInformationPolicyConfig

The sensitive information policy to configure for the guardrail.

Type: [GuardrailSensitiveInformationPolicyConfig](#) object

Required: No

[topicPolicyConfig](#)

The topic policy to configure for the guardrail.

Type: [GuardrailTopicPolicyConfig](#) object

Required: No

[wordPolicyConfig](#)

The word policy to configure for the guardrail.

Type: [GuardrailWordPolicyConfig](#) object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "guardrailArn": "string",
  "guardrailId": "string",
  "updatedAt": "string",
  "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

[guardrailArn](#)

The ARN of the guardrail that was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[[:^:]]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+$`

guardrailId

The unique identifier of the guardrail

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

updatedAt

The date and time at which the guardrail was updated.

Type: Timestamp

version

The version of the guardrail.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)

- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateProvisionedModelThroughput

Service: Amazon Bedrock

Updates the name or associated model for a Provisioned Throughput. For more information, see [Provisioned Throughput](#) in the Amazon Bedrock User Guide.

Request Syntax

```
PATCH /provisioned-model-throughput/provisionedModelId HTTP/1.1
Content-type: application/json
```

```
{
  "desiredModelId": "string",
  "desiredProvisionedModelName": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

[provisionedModelId](#)

The Amazon Resource Name (ARN) or name of the Provisioned Throughput to update.

Pattern: `^((([0-9a-zA-Z] [_-]?) +) | (arn:aws(- [^:]+)? :bedrock: [a-z0-9-]{1,20} : [0-9]{12} :provisioned-model/[a-z0-9]{12})))$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

[desiredModelId](#)

The Amazon Resource Name (ARN) of the new model to associate with this Provisioned Throughput. You can't specify this field if this Provisioned Throughput is associated with a base model.

If this Provisioned Throughput is associated with a custom model, you can specify one of the following options:

- The base model from which the custom model was customized.
- Another custom model that was customized from the same base model as the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([\0-9a-zA-Z][_]?)+)$`

Required: No

desiredProvisionedModelName

The new name for this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([\0-9a-zA-Z][_]?)+$`

Required: No

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

Agents for Amazon Bedrock

The following actions are supported by Agents for Amazon Bedrock:

- [AssociateAgentKnowledgeBase](#)
- [CreateAgent](#)
- [CreateAgentActionGroup](#)
- [CreateAgentAlias](#)
- [CreateDataSource](#)
- [CreateKnowledgeBase](#)
- [DeleteAgent](#)
- [DeleteAgentActionGroup](#)
- [DeleteAgentAlias](#)
- [DeleteAgentVersion](#)
- [DeleteDataSource](#)
- [DeleteKnowledgeBase](#)
- [DisassociateAgentKnowledgeBase](#)
- [GetAgent](#)
- [GetAgentActionGroup](#)
- [GetAgentAlias](#)
- [GetAgentKnowledgeBase](#)
- [GetAgentVersion](#)
- [GetDataSource](#)
- [GetIngestionJob](#)
- [GetKnowledgeBase](#)
- [ListAgentActionGroups](#)
- [ListAgentAliases](#)
- [ListAgentKnowledgeBases](#)
- [ListAgents](#)
- [ListAgentVersions](#)
- [ListDataSources](#)

- [ListIngestionJobs](#)
- [ListKnowledgeBases](#)
- [ListTagsForResource](#)
- [PrepareAgent](#)
- [StartIngestionJob](#)
- [TagResource](#)
- [UntagResource](#)
- [UpdateAgent](#)
- [UpdateAgentActionGroup](#)
- [UpdateAgentAlias](#)
- [UpdateAgentKnowledgeBase](#)
- [UpdateDataSource](#)
- [UpdateKnowledgeBase](#)

AssociateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Associates a knowledge base with an agent. If a knowledge base is associated and its `indexState` is set to `Enabled`, the agent queries the knowledge base for information to augment its response to the user.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/knowledgebases/ HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "description": "string",
  "knowledgeBaseId": "string",
  "knowledgeBaseState": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent with which you want to associate the knowledge base.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent with which you want to associate the knowledge base.

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

A description of what the agent should use the knowledge base for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to associate with the agent.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseState

Specifies whether to use the knowledge base or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentKnowledgeBase": {
    "agentId": "string",
    "agentVersion": "string",
    "createdAt": "string",
    "description": "string",
    "knowledgeBaseId": "string",
    "knowledgeBaseState": "string",
    "updatedAt": "string"
  }
}
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentKnowledgeBase](#)

Contains details about the knowledge base that has been associated with the agent.

Type: [AgentKnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateAgent

Service: Agents for Amazon Bedrock

Creates an agent that orchestrates interactions between foundation models, data sources, software applications, user conversations, and APIs to carry out tasks to help customers.

- Specify the following fields for security purposes.
 - `agentResourceRoleArn` – The Amazon Resource Name (ARN) of the role with permissions to invoke API operations on an agent.
 - (Optional) `customerEncryptionKeyArn` – The Amazon Resource Name (ARN) of a AWS KMS key to encrypt the creation of the agent.
 - (Optional) `idleSessionTTLInSeconds` – Specify the number of seconds for which the agent should maintain session information. After this time expires, the subsequent `InvokeAgent` request begins a new session.
- To override the default prompt behavior for agent orchestration and to use advanced prompts, include a `promptOverrideConfiguration` object. For more information, see [Advanced prompts](#).
- If your agent fails to be created, the response returns a list of `failureReasons` alongside a list of `recommendedActions` for you to troubleshoot.

Request Syntax

```
PUT /agents/ HTTP/1.1
Content-type: application/json

{
  "agentName": "string",
  "agentResourceRoleArn": "string",
  "clientToken": "string",
  "customerEncryptionKeyArn": "string",
  "description": "string",
  "foundationModel": "string",
  "idleSessionTTLInSeconds": number,
  "instruction": "string",
  "promptOverrideConfiguration": {
    "overrideLambda": "string",
    "promptConfigurations": [
      {
        "basePromptTemplate": "string",
```

```
    "inferenceConfiguration": {
      "maxLength": number,
      "stopSequences": [ string ],
      "temperature": number,
      "topK": number,
      "topP": number
    },
    "parserMode": string,
    "promptCreationMode": string,
    "promptState": string,
    "promptType": string
  }
]
},
"tags": {
  string : string
}
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

agentName

A name for the agent that you create.

Type: String

Pattern: $^([\text{0-9a-zA-Z}][\text{-}]?)\{1,100\}\$$

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/(service-role)?
AmazonBedrockExecutionRoleForAgents_.$`

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key with which to encrypt the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: No

description

A description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

foundationModel

The foundation model to be used for orchestration by the agent you create.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.\?]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/([a-z0-9-]{1,63}[\.\?]{1}[a-z0-9-]{1,63}([\.\?][a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[\.\?]{1}[a-z0-9-]{1,63}([\.\?][a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|([\0-9a-zA-Z][_]?)+)$`

Required: No

[idleSessionTTLInSeconds](#)

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

[instruction](#)

Instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 1200.

Required: No

[promptOverrideConfiguration](#)

Contains configurations to override prompts in different parts of an agent sequence. For more information, see [Advanced prompts](#).

Type: [PromptOverrideConfiguration](#) object

Required: No

tags

Any tags that you want to attach to the agent.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Required: No

Response Syntax

```

HTTP/1.1 202
Content-type: application/json

{
  "agent": {
    "agentArn": "string",
    "agentId": "string",
    "agentName": "string",
    "agentResourceRoleArn": "string",
    "agentStatus": "string",
    "agentVersion": "string",
    "clientToken": "string",
    "createdAt": "string",
    "customerEncryptionKeyArn": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "foundationModel": "string",
    "idleSessionTTLInSeconds": number,
    "instruction": "string",
    "preparedAt": "string",
    "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
        {
          "basePromptTemplate": "string",

```

```

    "inferenceConfiguration": {
      "maxLength": number,
      "stopSequences": [ "string" ],
      "temperature": number,
      "topK": number,
      "topP": number
    },
    "parserMode": "string",
    "promptCreationMode": "string",
    "promptState": "string",
    "promptType": "string"
  }
]
},
"recommendedActions": [ "string" ],
"updatedAt": "string"
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent created.

Type: [Agent](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of CreateAgent.

```
PUT /agents/ HTTP/1.1
Content-type: application/json

{
  "agentName": "o1nvve1",
  "agentResourceRoleArn": "arn:aws:iam::123456789012:role/
AmazonBedrockExecutionRoleForAgents_user",
  "instruction": "You are an IT agent who solves customer's problems",
  "description": "Description is here",
  "idleSessionTTLInSeconds": 900,
  "foundationModel": "anthropic.claude-v2"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateAgentActionGroup

Service: Agents for Amazon Bedrock

Creates an action group for an agent. An action group represents the actions that an agent can carry out for the customer by defining the APIs that an agent can call and the logic for calling them.

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.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/actiongroups/ HTTP/1.1
Content-type: application/json
```

```
{
  "actionGroupExecutor": { ... },
  "actionGroupName": "string",
  "actionGroupState": "string",
  "apiSchema": { ... },
  "clientToken": "string",
  "description": "string",
  "functionSchema": { ... },
  "parentActionGroupSignature": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

[agentId](#)

The unique identifier of the agent for which to create the action group.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent for which to create the action group.

Length Constraints: Fixed length of 5.

Pattern: ^DRAFT\$

Required: Yes

Request Body

The request accepts the following data in JSON format.

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Type: [ActionGroupExecutor](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

actionGroupName

The name to give the action group.

Type: String

Pattern: ^([0-9a-zA-Z][_ -]?) {1,100}\$

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see [Action group OpenAPI schemas](#).

Type: [APISchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

A description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Type: [FunctionSchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignature

To allow your agent to request the user for additional information when trying to complete a task, set this field 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.

Type: String

Valid Values: `AMAZON.UserInput`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentActionGroup": {
    "actionGroupExecutor": { ... },
    "actionGroupId": "string",
    "actionGroupName": "string",
    "actionGroupState": "string",
    "agentId": "string",
    "agentVersion": "string",
    "apiSchema": { ... },
    "clientToken": "string",
    "createdAt": "string",
    "description": "string",
    "functionSchema": { ... },
    "parentActionSignature": "string",
    "updatedAt": "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentActionGroup](#)

Contains details about the action group that was created.

Type: [AgentActionGroup](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Create an action group using an OpenAPI schema and a Lambda function

The following example creates an action group using an OpenAPI schema uploaded to an Amazon S3 bucket and sends the information elicited from the user to a Lambda function.

```
PUT /agents/AGENT12345/agentversions/DRAFT/actiongroups/ HTTP/1.1
Content-type: application/json

{
  "actionGroupName": "Test Action",
  "actionGroupState": "ENABLED",
  "apiSchema": {
    "s3": {
      "s3BucketName": "apischema-s3",
      "s3ObjectKey": "it_agent_openapi.json"
    }
  },
  "description": "Testing latest IT Management action",
  "actionGroupExecutor": {
    "lambda": "arn:aws:lambda:us-west-2:123456789012:function:ItAgentLambda"
  }
}
```

Create an action group using an OpenAPI schema and return control

The following example creates an action group using an OpenAPI schema uploaded to an Amazon S3 bucket and returns control by sending the information in the InvokeAgent response.

```
{
```

```

    "actionGroupName": "WeatherAPIs",
    "description": "Actions to get current weather and historical trends for a
location",
    "actionGroupState": "ENABLED",
    "apiSchema": {
      "s3": {
        "s3BucketName": "openapi-spec-iaad",
        "s3ObjectKey": "get_weather_openapi.yaml"
      }
    },
    "actionGroupExecutor": {
      "customControl": "RETURN_CONTROL"
    }
  }
}

```

Create an action group using function details and return control

The following example creates an action group using function details and returns control by sending the information in the InvokeAgent response

```

PUT /agents/AGENT12345/agentversions/DRAFT/actiongroups/ HTTP/1.1
Content-type: application/json

{
  "actionGroupName": "OrderManagementAction",
  "description": "Action to get the order history, product details, product
availability and to update the order",
  "actionGroupState": "ENABLED",
  "actionGroupExecutor": {
    "customControl": "RETURN_CONTROL"
  },
  "functionSchema": {
    "functions": [{
      "name": "GetOrderDetails",
      "parameters": {
        "orderId": {
          "type": "string",
          "required": true
        }
      }
    }
  ]
}
}

```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateAgentAlias

Service: Agents for Amazon Bedrock

Creates an alias of an agent that can be used to deploy the agent.

Request Syntax

```
PUT /agents/agentId/agentaliases/ HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "agentAliasName": "string",
  "clientToken": "string",
  "description": "string",
  "routingConfiguration": [
    {
      "agentVersion": "string",
      "provisionedThroughput": "string"
    }
  ],
  "tags": {
    "string" : "string"
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentAliasName

The name of the alias.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?) {1,100}$`

Required: Yes

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

A description of the alias of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of [AgentAliasRoutingConfigurationListItem](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

tags

Any tags that you want to attach to the alias of the agent.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agentAlias": {
    "agentAliasArn": "string",
    "agentAliasHistoryEvents": [
      {
        "endDate": "string",
        "routingConfiguration": [
          {
            "agentVersion": "string",
            "provisionedThroughput": "string"
          }
        ],
        "startDate": "string"
      }
    ],
    "agentAliasId": "string",
    "agentAliasName": "string",
    "agentAliasStatus": "string",
    "agentId": "string",
    "clientToken": "string",
    "createdAt": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "routingConfiguration": [
      {
        "agentVersion": "string",
```

```
        "provisionedThroughput": "string"
    }
  ],
  "updatedAt": "string"
}
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAlias

Contains details about the alias that was created.

Type: [AgentAlias](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `CreateAgentAlias`.

```
PUT /agents/ABCDEFGHIJ/agentaliases/ HTTP/1.1
Content-type: application/json

{
  "agentAliasName": "TestName",
  "description": "Alias is test"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateDataSource

Service: Agents for Amazon Bedrock

Sets up a data source to be added to a knowledge base.

Important

You can't change the chunkingConfiguration after you create the data source.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/ HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "clientToken": "string",
  "dataDeletionPolicy": "string",
  "dataSourceConfiguration": {
    "s3Configuration": {
      "bucketArn": "string",
      "bucketOwnerAccountId": "string",
      "inclusionPrefixes": [ "string" ]
    },
    "type": "string"
  },
  "description": "string",
  "name": "string",
  "serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
  },
  "vectorIngestionConfiguration": {
    "chunkingConfiguration": {
      "chunkingStrategy": "string",
      "fixedSizeChunkingConfiguration": {
        "maxTokens": number,
        "overlapPercentage": number
      }
    }
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to which to add the data source.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

dataDeletionPolicy

The data deletion policy assigned to the data source.

Type: String

Valid Values: RETAIN | DELETE

Required: No

dataSourceConfiguration

Contains metadata about where the data source is stored.

Type: [DataSourceConfiguration](#) object

Required: Yes

description

A description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

The name of the data source.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?){1,100}$`

Required: Yes

serverSideEncryptionConfiguration

Contains details about the server-side encryption for the data source.

Type: [ServerSideEncryptionConfiguration](#) object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: [VectorIngestionConfiguration](#) object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```

"dataSource": {
  "createdAt": "string",
  "dataDeletionPolicy": "string",
  "dataSourceConfiguration": {
    "s3Configuration": {
      "bucketArn": "string",
      "bucketOwnerAccountId": "string",
      "inclusionPrefixes": [ "string" ]
    },
    "type": "string"
  },
  "dataSourceId": "string",
  "description": "string",
  "failureReasons": [ "string" ],
  "knowledgeBaseId": "string",
  "name": "string",
  "serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
  },
  "status": "string",
  "updatedAt": "string",
  "vectorIngestionConfiguration": {
    "chunkingConfiguration": {
      "chunkingStrategy": "string",
      "fixedSizeChunkingConfiguration": {
        "maxTokens": number,
        "overlapPercentage": number
      }
    }
  }
}
}
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: [DataSource](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

CreateKnowledgeBase

Service: Agents for Amazon Bedrock

Creates a knowledge base that contains data sources from which information can be queried and used by LLMs. To create a knowledge base, you must first set up your data sources and configure a supported vector store. For more information, see [Set up your data for ingestion](#).

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 name and an optional description.
- Provide the Amazon Resource Name (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 an Amazon OpenSearch Service database, use the `opensearchServerlessConfiguration` object. For more information, see [Create a vector store in Amazon OpenSearch Service](#).
 - For an Amazon Aurora database, use the `RdsConfiguration` object. For more information, see [Create a vector store in Amazon Aurora](#).
 - For a Pinecone database, use the `pineconeConfiguration` object. For more information, see [Create a vector store in Pinecone](#).
 - For a Redis Enterprise Cloud database, use the `redisEnterpriseCloudConfiguration` object. For more information, see [Create a vector store in Redis Enterprise Cloud](#).

Request Syntax

```
PUT /knowledgebases/ HTTP/1.1
Content-type: application/json

{
  "clientToken": "string",
  "description": "string",
```

```
"knowledgeBaseConfiguration": {
  "type": "string",
  "vectorKnowledgeBaseConfiguration": {
    "embeddingModelArn": "string"
  }
},
"name": "string",
"roleArn": "string",
"storageConfiguration": {
  "mongoDbAtlasConfiguration": {
    "collectionName": "string",
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "endpoint": "string",
    "endpointServiceName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "opensearchServerlessConfiguration": {
    "collectionArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "pineconeConfiguration": {
    "connectionString": "string",
    "credentialsSecretArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string"
    },
    "namespace": "string"
  },
  "rdsConfiguration": {
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "fieldMapping": {
```

```

        "metadataField": "string",
        "primaryKeyField": "string",
        "textField": "string",
        "vectorField": "string"
    },
    "resourceArn": "string",
    "tableName": "string"
},
"redisEnterpriseCloudConfiguration": {
    "credentialsSecretArn": "string",
    "endpoint": "string",
    "fieldMapping": {
        "metadataField": "string",
        "textField": "string",
        "vectorField": "string"
    },
    "vectorIndexName": "string"
},
"type": "string"
},
"tags": {
    "string" : "string"
}
}

```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

A description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseConfiguration

Contains details about the embeddings model used for the knowledge base.

Type: [KnowledgeBaseConfiguration](#) object

Required: Yes

name

A name for the knowledge base.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/.*$`

Required: Yes

storageConfiguration

Contains details about the configuration of the vector database used for the knowledge base.

Type: [StorageConfiguration](#) object

Required: Yes

tags

Specify the key-value pairs for the tags that you want to attach to your knowledge base in this object.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "knowledgeBase": {
    "createdAt": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "knowledgeBaseArn": "string",
    "knowledgeBaseConfiguration": {
      "type": "string",
      "vectorKnowledgeBaseConfiguration": {
        "embeddingModelArn": "string"
      }
    },
    "knowledgeBaseId": "string",
    "name": "string",
    "roleArn": "string",
    "status": "string",
    "storageConfiguration": {
      "mongoDbAtlasConfiguration": {
        "collectionName": "string",
        "credentialsSecretArn": "string",
```

```
    "databaseName": "string",
    "endpoint": "string",
    "endpointServiceName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "opensearchServerlessConfiguration": {
    "collectionArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "pineconeConfiguration": {
    "connectionString": "string",
    "credentialsSecretArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string"
    },
    "namespace": "string"
  },
  "rdsConfiguration": {
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "primaryKeyField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "resourceArn": "string",
    "tableName": "string"
  },
  "redisEnterpriseCloudConfiguration": {
```



```
        "metadataField": "string",
        "textField": "string",
        "vectorField": "string"
    },
    "vectorIndexName": "string"
},
"type": "string"
},
"updatedAt": "string"
}
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: [KnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteAgent

Service: Agents for Amazon Bedrock

Deletes an agent.

Request Syntax

```
DELETE /agents/agentId?skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent to delete.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agentId": "string",
  "agentStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent that was deleted.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

agentStatus

The status of the agent.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgent.

```
DELETE /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteAgentActionGroup

Service: Agents for Amazon Bedrock

Deletes an action group in an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId?  
skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

[actionGroupId](#)

The unique identifier of the action group to delete.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

[agentId](#)

The unique identifier of the agent that the action group belongs to.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

[agentVersion](#)

The version of the agent that the action group belongs to.

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Required: Yes

[skipResourceInUseCheck](#)

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 204
```

Response Elements

If the action is successful, the service sends back an HTTP 204 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgentActionGroup.

```
DELETE /agents/ABCDEFGHIJ/agentversions/1/actiongroups/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteAgentAlias

Service: Agents for Amazon Bedrock

Deletes an alias of an agent.

Request Syntax

```
DELETE /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias to delete.

Length Constraints: Fixed length of 10.

Pattern: $^(\backslash\text{TSTALIASID}\backslash\text{b} | [0-9a-zA-Z]+)\$$

Required: Yes

agentId

The unique identifier of the agent that the alias belongs to.

Pattern: $^[0-9a-zA-Z]{10}\$$

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agentAliasId": "string",
  "agentAliasStatus": "string",
```

```
"agentId": "string"  
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAliasId

The unique identifier of the alias that was deleted.

Type: String

Length Constraints: Fixed length of 10.

Pattern: $^(\backslash\text{TSTALIASID}\backslash\text{b}|[0-9a-zA-Z]+)\$$

agentAliasStatus

The status of the alias.

Type: String

Valid Values: CREATING | PREPARED | FAILED | UPDATING | DELETING

agentId

The unique identifier of the agent that the alias belongs to.

Type: String

Pattern: $^[0-9a-zA-Z]{10}\$$

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `DeleteAgentAlias`.

```
DELETE /agents/ABCDEFGHIJ/agentaliases/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteAgentVersion

Service: Agents for Amazon Bedrock

Deletes a version of an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion?  
skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent that the version belongs to.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent to delete.

Pattern: `^[0-9]{1,5}$`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202  
Content-type: application/json
```

```
{
  "agentId": "string",
  "agentStatus": "string",
  "agentVersion": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent that the version belongs to.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

agentStatus

The status of the agent version.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING

agentVersion

The version that was deleted.

Type: String

Pattern: `^[0-9]{1,5}$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `DeleteAgentVersion`.

```
DELETE /agents/ABCDEFGHIJ/agentversions/1/?skipResourceInUseCheck=true HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteDataSource

Service: Agents for Amazon Bedrock

Deletes a data source from a knowledge base.

Request Syntax

```
DELETE /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source to delete.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base from which to delete the data source.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "dataSourceId": "string",
  "knowledgeBaseId": "string",
  "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

dataSourceId

The unique identifier of the data source that was deleted.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

knowledgeBaseId

The unique identifier of the knowledge base to which the data source that was deleted belonged.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

status

The status of the data source.

Type: String

Valid Values: AVAILABLE | DELETING | DELETE_UNSUCCESSFUL

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DeleteKnowledgeBase

Service: Agents for Amazon Bedrock

Deletes a knowledge base. Before deleting a knowledge base, you should disassociate the knowledge base from any agents that it is associated with by making a [DisassociateAgentKnowledgeBase](#) request.

Request Syntax

```
DELETE /knowledgebases/knowledgeBaseId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to delete.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "knowledgeBaseId": "string",
  "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBaseId

The unique identifier of the knowledge base that was deleted.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

status

The status of the knowledge base and whether it has been successfully deleted.

Type: String

Valid Values: CREATING | ACTIVE | DELETING | UPDATING | FAILED | DELETE_UNSUCCESSFUL

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

DisassociateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Disassociates a knowledge base from an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/
HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent from which to disassociate the knowledge base.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent from which to disassociate the knowledge base.

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to disassociate.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 204
```

Response Elements

If the action is successful, the service sends back an HTTP 204 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetAgent

Service: Agents for Amazon Bedrock

Gets information about an agent.

Request Syntax

```
GET /agents/agentId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agent": {
    "agentArn": "string",
    "agentId": "string",
    "agentName": "string",
    "agentResourceRoleArn": "string",
    "agentStatus": "string",
    "agentVersion": "string",
    "clientToken": "string",
    "createdAt": "string",
    "customerEncryptionKeyArn": "string",
    "description": "string",
    "failureReasons": [ "string" ],
```

```

    "foundationModel": "string",
    "idleSessionTTLInSeconds": number,
    "instruction": "string",
    "preparedAt": "string",
    "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
        {
          "basePromptTemplate": "string",
          "inferenceConfiguration": {
            "maxLength": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topK": number,
            "topP": number
          },
          "parserMode": "string",
          "promptCreationMode": "string",
          "promptState": "string",
          "promptType": "string"
        }
      ]
    },
    "recommendedActions": [ "string" ],
    "updatedAt": "string"
  }
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent.

Type: [Agent](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetAgent.

```
GET /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)

- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetAgentActionGroup

Service: Agents for Amazon Bedrock

Gets information about an action group for an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

actionGroupId

The unique identifier of the action group for which to get information.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentId

The unique identifier of the agent that the action group belongs to.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent that the action group belongs to.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentActionGroup": {
    "actionGroupExecutor": { ... },
    "actionGroupId": "string",
    "actionGroupName": "string",
    "actionGroupState": "string",
    "agentId": "string",
    "agentVersion": "string",
    "apiSchema": { ... },
    "clientToken": "string",
    "createdAt": "string",
    "description": "string",
    "functionSchema": { ... },
    "parentActionSignature": "string",
    "updatedAt": "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentActionGroup](#)

Contains details about the action group.

Type: [AgentActionGroup](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `GetAgentActionGroup`.

```
GET /agents/AGENT12345/agentversions/1/actiongroups/ACTION1234/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetAgentAlias

Service: Agents for Amazon Bedrock

Gets information about an alias of an agent.

Request Syntax

```
GET /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias for which to get information.

Length Constraints: Fixed length of 10.

Pattern: $^(\backslash\text{TSTALIASID}\backslash\text{b} | [0-9a-zA-Z]+)\$$

Required: Yes

agentId

The unique identifier of the agent to which the alias to get information belongs.

Pattern: $^[0-9a-zA-Z]{10}\$$

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```

"agentAlias": {
  "agentAliasArn": "string",
  "agentAliasHistoryEvents": [
    {
      "endDate": "string",
      "routingConfiguration": [
        {
          "agentVersion": "string",
          "provisionedThroughput": "string"
        }
      ],
      "startDate": "string"
    }
  ],
  "agentAliasId": "string",
  "agentAliasName": "string",
  "agentAliasStatus": "string",
  "agentId": "string",
  "clientToken": "string",
  "createdAt": "string",
  "description": "string",
  "failureReasons": [ "string" ],
  "routingConfiguration": [
    {
      "agentVersion": "string",
      "provisionedThroughput": "string"
    }
  ],
  "updatedAt": "string"
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentAlias

Contains information about the alias.

Type: [AgentAlias](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `GetAgentAlias`.

```
GET /agents/ABCDEFGHIIJ/agentaliases/ABCDEFGHIIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Gets information about a knowledge base associated with an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent with which the knowledge base is associated.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent with which the knowledge base is associated.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base associated with the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentKnowledgeBase": {
    "agentId": "string",
    "agentVersion": "string",
    "createdAt": "string",
    "description": "string",
    "knowledgeBaseId": "string",
    "knowledgeBaseState": "string",
    "updatedAt": "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentKnowledgeBase

Contains details about a knowledge base attached to an agent.

Type: [AgentKnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetAgentVersion

Service: Agents for Amazon Bedrock

Gets details about a version of an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent.

Pattern: `^[0-9]{1,5}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentVersion": {
    "agentArn": "string",
    "agentId": "string",
```

```

    "agentName": "string",
    "agentResourceRoleArn": "string",
    "agentStatus": "string",
    "createdAt": "string",
    "customerEncryptionKeyArn": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "foundationModel": "string",
    "idleSessionTTLInSeconds": number,
    "instruction": "string",
    "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
        {
          "basePromptTemplate": "string",
          "inferenceConfiguration": {
            "maxLength": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topK": number,
            "topP": number
          },
          "parserMode": "string",
          "promptCreationMode": "string",
          "promptState": "string",
          "promptType": "string"
        }
      ]
    },
    "recommendedActions": [ "string" ],
    "updatedAt": "string",
    "version": "string"
  }
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentVersion

Contains details about the version of the agent.

Type: [AgentVersion](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `GetAgentVersion`.

```
GET /agents/agentId/agentversions/agentVersion/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetDataSource

Service: Agents for Amazon Bedrock

Gets information about a data source.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that the data source was added to.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "dataSource": {
    "createdAt": "string",
    "dataDeletionPolicy": "string",
    "dataSourceConfiguration": {
      "s3Configuration": {
```

```

        "bucketArn": "string",
        "bucketOwnerAccountId": "string",
        "inclusionPrefixes": [ "string" ]
    },
    "type": "string"
},
"dataSourceId": "string",
"description": "string",
"failureReasons": [ "string" ],
"knowledgeBaseId": "string",
"name": "string",
"serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
},
"status": "string",
"updatedAt": "string",
"vectorIngestionConfiguration": {
    "chunkingConfiguration": {
        "chunkingStrategy": "string",
        "fixedSizeChunkingConfiguration": {
            "maxTokens": number,
            "overlapPercentage": number
        }
    }
}
}
}
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: [DataSource](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)

- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetIngestionJob

Service: Agents for Amazon Bedrock

Gets information about a ingestion job, in which a data source is added to a knowledge base.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/datasources/dataSourceId/
ingestionjobs/ingestionJobId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source in the ingestion job.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

ingestionJobId

The unique identifier of the ingestion job.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for which the ingestion job applies.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Content-type: application/json

```
{
  "ingestionJob": {
    "dataSourceId": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "ingestionJobId": "string",
    "knowledgeBaseId": "string",
    "startedAt": "string",
    "statistics": {
      "numberOfDocumentsDeleted": number,
      "numberOfDocumentsFailed": number,
      "numberOfDocumentsScanned": number,
      "numberOfMetadataDocumentsModified": number,
      "numberOfMetadataDocumentsScanned": number,
      "numberOfModifiedDocumentsIndexed": number,
      "numberOfNewDocumentsIndexed": number
    },
    "status": "string",
    "updatedAt": "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ingestionJob

Contains details about the ingestion job.

Type: [IngestionJob](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

GetKnowledgeBase

Service: Agents for Amazon Bedrock

Gets information about a knowledge base.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base for which to get information.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "knowledgeBase": {
    "createdAt": "string",
    "description": "string",
    "failureReasons": [ "string ],
    "knowledgeBaseArn": "string",
    "knowledgeBaseConfiguration": {
      "type": "string",
      "vectorKnowledgeBaseConfiguration": {
        "embeddingModelArn": "string"
      }
    }
  },
  "knowledgeBaseId": "string",
```

```
"name": "string",
"roleArn": "string",
"status": "string",
"storageConfiguration": {
  "mongoDbAtlasConfiguration": {
    "collectionName": "string",
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "endpoint": "string",
    "endpointServiceName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "opensearchServerlessConfiguration": {
    "collectionArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "pineconeConfiguration": {
    "connectionString": "string",
    "credentialsSecretArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string"
    },
    "namespace": "string"
  },
  "rdsConfiguration": {
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "primaryKeyField": "string",
      "textField": "string",
      "vectorField": "string"
    }
  },
}
```

```
    "resourceArn": "string",
    "tableName": "string"
  },
  "redisEnterpriseCloudConfiguration": {
    "credentialsSecretArn": "string",
    "endpoint": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "type": "string"
},
"updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: [KnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListAgentActionGroups

Service: Agents for Amazon Bedrock

Lists the action groups for an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentversions/agentVersion/actiongroups/ HTTP/1.1
Content-type: application/json
```

```
{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "actionGroupSummaries": [
    {
      "actionGroupId": "string",
      "actionGroupName": "string",
      "actionGroupState": "string",
      "description": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

actionGroupSummaries

A list of objects, each of which contains information about an action group.

Type: Array of [ActionGroupSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentActionGroups.

```
POST /agents/AGENT12345/agentversions/1/actiongroups/ HTTP/1.1
Content-type: application/json

{
  "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListAgentAliases

Service: Agents for Amazon Bedrock

Lists the aliases of an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentaliases/ HTTP/1.1
Content-type: application/json
```

```
{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentAliasSummaries": [
    {
      "agentAliasId": "string",
      "agentAliasName": "string",
      "agentAliasStatus": "string",
      "createdAt": "string",
      "description": "string",
      "routingConfiguration": [
        {
          "agentVersion": "string",
          "provisionedThroughput": "string"
        }
      ],
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentAliasSummaries

A list of objects, each of which contains information about an alias of the agent.

Type: Array of [AgentAliasSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentAliases.

```
POST /agents/ABCDEFGHIJ/agentaliases/ HTTP/1.1
Content-type: application/json

{
  "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListAgentKnowledgeBases

Service: Agents for Amazon Bedrock

Lists knowledge bases associated with an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentversions/agentVersion/knowledgebases/ HTTP/1.1
Content-type: application/json
```

```
{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent for which to return information about knowledge bases associated with it.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent for which to return information about knowledge bases associated with it.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentKnowledgeBaseSummaries": [
    {
      "description": "string",
      "knowledgeBaseId": "string",
      "knowledgeBaseState": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentKnowledgeBaseSummaries](#)

A list of objects, each of which contains information about a knowledge base associated with the agent.

Type: Array of [AgentKnowledgeBaseSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

[nextToken](#)

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListAgents

Service: Agents for Amazon Bedrock

Lists the agents belonging to an account and information about each agent.

Request Syntax

```
POST /agents/ HTTP/1.1
Content-type: application/json

{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the nextToken field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the maxResults value provided in the request, enter the token returned in the nextToken field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentSummaries": [
    {
      "agentId": "string",
      "agentName": "string",
      "agentStatus": "string",
      "description": "string",
      "latestAgentVersion": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentSummaries

A list of objects, each of which contains information about an agent.

Type: Array of [AgentSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgents.

```
POST /agents/ HTTP/1.1
Content-type: application/json

{
```

```
"maxResults": 10  
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListAgentVersions

Service: Agents for Amazon Bedrock

Lists the versions of an agent and information about each version.

Request Syntax

```
POST /agents/agentId/agentversions/ HTTP/1.1
Content-type: application/json
```

```
{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentVersionSummaries": [
    {
      "agentName": "string",
      "agentStatus": "string",
      "agentVersion": "string",
      "createdAt": "string",
      "description": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentVersionSummaries

A list of objects, each of which contains information about a version of the agent.

Type: Array of [AgentVersionSummary](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

[nextToken](#)

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentVersions.

```
POST /agents/agentId/agentversions/ HTTP/1.1
Content-type: application/json

{
  "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListDataSources

Service: Agents for Amazon Bedrock

Lists the data sources in a knowledge base and information about each one.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/ HTTP/1.1  
Content-type: application/json
```

```
{  
  "maxResults": number,  
  "nextToken": "string"  
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base for which to return a list of information.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "dataSourceSummaries": [
    {
      "dataSourceId": "string",
      "description": "string",
      "knowledgeBaseId": "string",
      "name": "string",
      "status": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSourceSummaries

A list of objects, each of which contains information about a data source.

Type: Array of [DataSourceSummary](#) objects

[nextToken](#)

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListIngestionJobs

Service: Agents for Amazon Bedrock

Lists the ingestion jobs for a data source and information about each of them.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/ingestionjobs/ HTTP/1.1
Content-type: application/json
```

```
{
  "filters": [
    {
      "attribute": "string",
      "operator": "string",
      "values": [ "string" ]
    }
  ],
  "maxResults": number,
  "nextToken": "string",
  "sortBy": {
    "attribute": "string",
    "order": "string"
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source for which to return ingestion jobs.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for which to return ingestion jobs.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

filters

Contains a definition of a filter for which to filter the results.

Type: Array of [IngestionJobFilter](#) objects

Array Members: Fixed number of 1 item.

Required: No

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the nextToken field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the maxResults value provided in the request, enter the token returned in the nextToken field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

sortBy

Contains details about how to sort the results.

Type: [IngestionJobSortBy](#) object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "ingestionJobSummaries": [
    {
      "dataSourceId": "string",
      "description": "string",
      "ingestionJobId": "string",
      "knowledgeBaseId": "string",
      "startedAt": "string",
      "statistics": {
        "numberOfDocumentsDeleted": number,
        "numberOfDocumentsFailed": number,
        "numberOfDocumentsScanned": number,
        "numberOfMetadataDocumentsModified": number,
        "numberOfMetadataDocumentsScanned": number,
        "numberOfModifiedDocumentsIndexed": number,
        "numberOfNewDocumentsIndexed": number
      },
      "status": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[ingestionJobSummaries](#)

A list of objects, each of which contains information about an ingestion job.

Type: Array of [IngestionJobSummary](#) objects

[nextToken](#)

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListKnowledgeBases

Service: Agents for Amazon Bedrock

Lists the knowledge bases in an account and information about each of them.

Request Syntax

```
POST /knowledgebases/ HTTP/1.1
Content-type: application/json
```

```
{
  "maxResults": number,
  "nextToken": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the nextToken field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the maxResults value provided in the request, enter the token returned in the nextToken field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "knowledgeBaseSummaries": [
    {
      "description": "string",
      "knowledgeBaseId": "string",
      "name": "string",
      "status": "string",
      "updatedAt": "string"
    }
  ],
  "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

knowledgeBaseSummaries

A list of objects, each of which contains information about a knowledge base.

Type: Array of [KnowledgeBaseSummary](#) objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

ListTagsForResource

Service: Agents for Amazon Bedrock

List all the tags for the resource you specify.

Request Syntax

```
GET /tags/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource for which to list tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: (`^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?$`)

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "tags": {
    "string" : "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

The key-value pairs for the tags associated with the resource.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListTagsForResource.

```
GET /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

PrepareAgent

Service: Agents for Amazon Bedrock

Creates a DRAFT version of the agent that can be used for internal testing.

Request Syntax

```
POST /agents/agentId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent for which to create a DRAFT version.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agentId": "string",
  "agentStatus": "string",
  "agentVersion": "string",
  "preparedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent for which the DRAFT version was created.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

agentStatus

The status of the DRAFT version and whether it is ready for use.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

preparedAt

The time at which the DRAFT version of the agent was last prepared.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of PrepareAgent.

```
POST /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

StartIngestionJob

Service: Agents for Amazon Bedrock

Begins an ingestion job, in which a data source is added to a knowledge base.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/dataSourceId/ingestionjobs/ HTTP/1.1
Content-type: application/json

{
  "clientToken": "string",
  "description": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source to ingest.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which to add the data source.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

A description of the ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "ingestionJob": {
    "dataSourceId": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "ingestionJobId": "string",
    "knowledgeBaseId": "string",
    "startedAt": "string",
    "statistics": {
      "numberOfDocumentsDeleted": number,
      "numberOfDocumentsFailed": number,
      "numberOfDocumentsScanned": number,
      "numberOfMetadataDocumentsModified": number,
      "numberOfMetadataDocumentsScanned": number,
      "numberOfModifiedDocumentsIndexed": number,
      "numberOfNewDocumentsIndexed": number
    },
    "status": "string",
    "updatedAt": "string"
  }
}
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

[ingestionJob](#)

An object containing information about the ingestion job.

Type: [IngestionJob](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

TagResource

Service: Agents for Amazon Bedrock

Associate tags with a resource. For more information, see [Tagging resources](#) in the Amazon Bedrock User Guide.

Request Syntax

```
POST /tags/resourceArn HTTP/1.1
Content-type: application/json
```

```
{
  "tags": {
    "string" : "string"
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

[resourceArn](#)

The Amazon Resource Name (ARN) of the resource to tag.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `(^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?.$)`

Required: Yes

Request Body

The request accepts the following data in JSON format.

[tags](#)

An object containing key-value pairs that define the tags to attach to the resource.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of TagResource.

```
POST /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
Content-type: application/json

{
  "tags": {
    "cost-center" : "Tech"
  }
}
```

Example response

This example illustrates one usage of TagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)

- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UntagResource

Service: Agents for Amazon Bedrock

Remove tags from a resource.

Request Syntax

```
DELETE /tags/resourceArn?tagKeys=tagKeys HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource from which to remove tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: ([^]arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?*\$)

Required: Yes

tagKeys

A list of keys of the tags to remove from the resource.

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: [^][a-zA-Z0-9\s._:/=+@-]*\$

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of `UntagResource`.

```
DELETE /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
```

Example response

This example illustrates one usage of UntagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateAgent

Service: Agents for Amazon Bedrock

Updates the configuration of an agent.

Request Syntax

```
PUT /agents/agentId/ HTTP/1.1
Content-type: application/json

{
  "agentName": "string",
  "agentResourceRoleArn": "string",
  "customerEncryptionKeyArn": "string",
  "description": "string",
  "foundationModel": "string",
  "idleSessionTTLInSeconds": number,
  "instruction": "string",
  "promptOverrideConfiguration": {
    "overrideLambda": "string",
    "promptConfigurations": [
      {
        "basePromptTemplate": "string",
        "inferenceConfiguration": {
          "maxLength": number,
          "stopSequences": [ "string" ],
          "temperature": number,
          "topK": number,
          "topP": number
        },
        "parserMode": "string",
        "promptCreationMode": "string",
        "promptState": "string",
        "promptType": "string"
      }
    ]
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentName

Specifies a new name for the agent.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?)\{1,100\}$`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/(service-role/)?AmazonBedrockExecutionRoleForAgents_.$`

Required: Yes

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key with which to encrypt the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: No

description

Specifies a new description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

foundationModel

Specifies a new foundation model to be used for orchestration by the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(((0-9){12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|((0-9a-zA-Z[_-]?)+))$`

Required: Yes

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

instruction

Specifies new instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 1200.

Required: No

promptOverrideConfiguration

Contains configurations to override prompts in different parts of an agent sequence. For more information, see [Advanced prompts](#).

Type: [PromptOverrideConfiguration](#) object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agent": {
    "agentArn": "string",
    "agentId": "string",
    "agentName": "string",
    "agentResourceRoleArn": "string",
    "agentStatus": "string",
    "agentVersion": "string",
    "clientToken": "string",
    "createdAt": "string",
    "customerEncryptionKeyArn": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "foundationModel": "string",
    "idleSessionTTLInSeconds": number,
    "instruction": "string",
    "preparedAt": "string",
```

```

    "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
        {
          "basePromptTemplate": "string",
          "inferenceConfiguration": {
            "maxLength": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topK": number,
            "topP": number
          },
          "parserMode": "string",
          "promptCreationMode": "string",
          "promptState": "string",
          "promptType": "string"
        }
      ]
    },
    "recommendedActions": [ "string" ],
    "updatedAt": "string"
  }
}

```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent that was updated.

Type: [Agent](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgent.

```
PUT /agents/ABCDEFGHIJ/ HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "agentName": "TestName",
  "agentResourceRoleArn": "arn:aws:iam::123456789012:role/
AmazonBedrockExecutionRoleForAgents_user",
  "instruction": "You are an IT agent who solves customer's problems",
  "description": "Description is here",
  "idleSessionTTLInSeconds": 900,
  "foundationModel": "anthropic.claude-v2"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateAgentActionGroup

Service: Agents for Amazon Bedrock

Updates the configuration for an action group for an agent.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId/ HTTP/1.1  
Content-type: application/json
```

```
{  
  "actionGroupExecutor": { ... },  
  "actionGroupName": "string",  
  "actionGroupState": "string",  
  "apiSchema": { ... },  
  "description": "string",  
  "functionSchema": { ... },  
  "parentActionGroupSignature": "string"  
}
```

URI Request Parameters

The request uses the following URI parameters.

actionGroupId

The unique identifier of the action group.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentId

The unique identifier of the agent for which to update the action group.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The unique identifier of the agent version for which to update the action group.

Length Constraints: Fixed length of 5.

Pattern: ^DRAFT\$

Required: Yes

Request Body

The request accepts the following data in JSON format.

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: [ActionGroupExecutor](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

actionGroupName

Specifies a new name for the action group.

Type: String

Pattern: ^([0-9a-zA-Z][_]?){1,100}\$

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see [Action group OpenAPI schemas](#).

Type: [APISchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

description

Specifies a new name for the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Type: [FunctionSchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignature

To allow your agent to request the user for additional information when trying to complete a task, set this field 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.

Type: String

Valid Values: `AMAZON.UserInput`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentActionGroup": {
    "actionGroupExecutor": { ... },
    "actionGroupId": "string",
    "actionGroupName": "string",
    "actionGroupState": "string",
    "agentId": "string",
    "agentVersion": "string",
    "apiSchema": { ... },
    "clientToken": "string",
    "createdAt": "string",
    "description": "string",
    "functionSchema": { ... },
    "parentActionSignature": "string",
    "updatedAt": "string"
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentActionGroup](#)

Contains details about the action group that was updated.

Type: [AgentActionGroup](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgentActionGroup.

```
PUT /agents/AGENT12345/agentversions/1/actiongroups/ACTION1234/ HTTP/1.1
```

Content-type: application/json

```
{
  "actionGroupName": "bedrock-temp-actions",
  "actionGroupState": "ENABLED",
  "description": "Testing = latest IT Management action",
  "apiSchema": {
    "s3": {
      "s3BucketName": "apischema-s3",
      "s3objectKey": "it_agent_openapi.json"
    }
  },
  "actionGroupExecutor": {
    "lambda": "arn:aws:lambda:us-west-2:123456789012:function:ItAgentLambda"
  }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateAgentAlias

Service: Agents for Amazon Bedrock

Updates configurations for an alias of an agent.

Request Syntax

```
PUT /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "agentAliasName": "string",
  "description": "string",
  "routingConfiguration": [
    {
      "agentVersion": "string",
      "provisionedThroughput": "string"
    }
  ]
}
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias.

Length Constraints: Fixed length of 10.

Pattern: $^(\backslash\text{bTSTALIASID}\backslash\text{b} | [0-9a-zA-Z]+)\$$

Required: Yes

agentId

The unique identifier of the agent.

Pattern: $^[0-9a-zA-Z]\{10\}\$$

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentAliasName

Specifies a new name for the alias.

Type: String

Pattern: `^([\u0000-\u0009a-zA-Z][\u005C_-]?){1,100}$`

Required: Yes

description

Specifies a new description for the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of [AgentAliasRoutingConfigurationListItem](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "agentAlias": {
    "agentAliasArn": "string",
    "agentAliasHistoryEvents": [
      {
```

```

    "endDate": "string",
    "routingConfiguration": [
      {
        "agentVersion": "string",
        "provisionedThroughput": "string"
      }
    ],
    "startDate": "string"
  }
],
"agentAliasId": "string",
"agentAliasName": "string",
"agentAliasStatus": "string",
"agentId": "string",
"clientToken": "string",
"createdAt": "string",
"description": "string",
"failureReasons": [ "string" ],
"routingConfiguration": [
  {
    "agentVersion": "string",
    "provisionedThroughput": "string"
  }
],
"updatedAt": "string"
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAlias

Contains details about the alias that was updated.

Type: [AgentAlias](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgentAlias.

```
PUT /agents/ABCDEFGHIJ/agentaliases/ABCDEFGHIJ/ HTTP/1.1
Content-type: application/json
```

```
{
  "agentAliasName": "TestName",
  "description": "Updated description",
  "routingConfiguration": [
    {
      "agentVersion": "2"
    }
  ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Updates the configuration for a knowledge base that has been associated with an agent.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/ HTTP/1.1
Content-type: application/json
```

```
{
  "description": "string",
  "knowledgeBaseState": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent associated with the knowledge base that you want to update.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent associated with the knowledge base that you want to update.

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that has been associated with an agent.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

Specifies a new description for the knowledge base associated with an agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseState

Specifies whether the agent uses the knowledge base or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "agentKnowledgeBase": {
    "agentId": "string",
    "agentVersion": "string",
    "createdAt": "string",
    "description": "string",
    "knowledgeBaseId": "string",
    "knowledgeBaseState": "string",
    "updatedAt": "string"
  }
}
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[agentKnowledgeBase](#)

Contains details about the knowledge base that has been associated with an agent.

Type: [AgentKnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateDataSource

Service: Agents for Amazon Bedrock

Updates configurations for a data source.

Important

You can't change the chunkingConfiguration after you create the data source. Specify the existing chunkingConfiguration.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "dataDeletionPolicy": "string",
  "dataSourceConfiguration": {
    "s3Configuration": {
      "bucketArn": "string",
      "bucketOwnerAccountId": "string",
      "inclusionPrefixes": [ "string" ]
    },
    "type": "string"
  },
  "description": "string",
  "name": "string",
  "serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
  },
  "vectorIngestionConfiguration": {
    "chunkingConfiguration": {
      "chunkingStrategy": "string",
      "fixedSizeChunkingConfiguration": {
        "maxTokens": number,
        "overlapPercentage": number
      }
    }
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source belongs.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

dataDeletionPolicy

The data deletion policy of the updated data source.

Type: String

Valid Values: RETAIN | DELETE

Required: No

dataSourceConfiguration

Contains details about the storage configuration of the data source.

Type: [DataSourceConfiguration](#) object

Required: Yes

description

Specifies a new description for the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

Specifies a new name for the data source.

Type: String

Pattern: $^([\text{0-9a-zA-Z}][\text{-}]?)\{1,100\}\$$

Required: Yes

serverSideEncryptionConfiguration

Contains details about server-side encryption of the data source.

Type: [ServerSideEncryptionConfiguration](#) object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: [VectorIngestionConfiguration](#) object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "dataSource": {
    "createdAt": "string",
    "dataDeletionPolicy": "string",
    "dataSourceConfiguration": {
      "s3Configuration": {
```

```
        "bucketArn": "string",
        "bucketOwnerAccountId": "string",
        "inclusionPrefixes": [ "string" ]
    },
    "type": "string"
},
"dataSourceId": "string",
"description": "string",
"failureReasons": [ "string" ],
"knowledgeBaseId": "string",
"name": "string",
"serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
},
"status": "string",
"updatedAt": "string",
"vectorIngestionConfiguration": {
    "chunkingConfiguration": {
        "chunkingStrategy": "string",
        "fixedSizeChunkingConfiguration": {
            "maxTokens": number,
            "overlapPercentage": number
        }
    }
}
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: [DataSource](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

UpdateKnowledgeBase

Service: Agents for Amazon Bedrock

Updates the configuration of a knowledge base with the fields that you specify. Because all fields will be overwritten, you must include the same values for fields that you want to keep the same.

You can change the following fields:

- name
- description
- roleArn

You can't change the `knowledgeBaseConfiguration` or `storageConfiguration` fields, so you must specify the same configurations as when you created the knowledge base. You can send a [GetKnowledgeBase](#) request and copy the same configurations.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId HTTP/1.1
```

```
Content-type: application/json
```

```
{
  "description": "string",
  "knowledgeBaseConfiguration": {
    "type": "string",
    "vectorKnowledgeBaseConfiguration": {
      "embeddingModelArn": "string"
    }
  },
  "name": "string",
  "roleArn": "string",
  "storageConfiguration": {
    "mongoDbAtlasConfiguration": {
      "collectionName": "string",
      "credentialsSecretArn": "string",
      "databaseName": "string",
      "endpoint": "string",
      "endpointServiceName": "string",
      "fieldMapping": {
        "metadataField": "string",
        "textField": "string",

```

```
    "vectorField": "string"
  },
  "vectorIndexName": "string"
},
"opensearchServerlessConfiguration": {
  "collectionArn": "string",
  "fieldMapping": {
    "metadataField": "string",
    "textField": "string",
    "vectorField": "string"
  },
  "vectorIndexName": "string"
},
"pineconeConfiguration": {
  "connectionString": "string",
  "credentialsSecretArn": "string",
  "fieldMapping": {
    "metadataField": "string",
    "textField": "string"
  },
  "namespace": "string"
},
"rdsConfiguration": {
  "credentialsSecretArn": "string",
  "databaseName": "string",
  "fieldMapping": {
    "metadataField": "string",
    "primaryKeyField": "string",
    "textField": "string",
    "vectorField": "string"
  },
  "resourceArn": "string",
  "tableName": "string"
},
"redisEnterpriseCloudConfiguration": {
  "credentialsSecretArn": "string",
  "endpoint": "string",
  "fieldMapping": {
    "metadataField": "string",
    "textField": "string",
    "vectorField": "string"
  },
  "vectorIndexName": "string"
},
}
```

```
    "type": "string"  
  }  
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to update.

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

Specifies a new description for the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseConfiguration

Specifies the configuration for the embeddings model used for the knowledge base. You must use the same configuration as when the knowledge base was created.

Type: [KnowledgeBaseConfiguration](#) object

Required: Yes

name

Specifies a new name for the knowledge base.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?){1,100}$`

Required: Yes

roleArn

Specifies a different Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/.+&`

Required: Yes

storageConfiguration

Specifies the configuration for the vector store used for the knowledge base. You must use the same configuration as when the knowledge base was created.

Type: [StorageConfiguration](#) object

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
  "knowledgeBase": {
    "createdAt": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "knowledgeBaseArn": "string",
    "knowledgeBaseConfiguration": {
      "type": "string",
      "vectorKnowledgeBaseConfiguration": {
        "embeddingModelArn": "string"
      }
    },
    "knowledgeBaseId": "string",
    "name": "string",
    "roleArn": "string",
    "status": "string",
```

```
"storageConfiguration": {
  "mongoDbAtlasConfiguration": {
    "collectionName": "string",
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "endpoint": "string",
    "endpointServiceName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "opensearchServerlessConfiguration": {
    "collectionArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "vectorIndexName": "string"
  },
  "pineconeConfiguration": {
    "connectionString": "string",
    "credentialsSecretArn": "string",
    "fieldMapping": {
      "metadataField": "string",
      "textField": "string"
    },
    "namespace": "string"
  },
  "rdsConfiguration": {
    "credentialsSecretArn": "string",
    "databaseName": "string",
    "fieldMapping": {
      "metadataField": "string",
      "primaryKeyField": "string",
      "textField": "string",
      "vectorField": "string"
    },
    "resourceArn": "string",
    "tableName": "string"
  },
}
```

```
    "redisEnterpriseCloudConfiguration": {
      "credentialsSecretArn": "string",
      "endpoint": "string",
      "fieldMapping": {
        "metadataField": "string",
        "textField": "string",
        "vectorField": "string"
      },
      "vectorIndexName": "string"
    },
    "type": "string"
  },
  "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: [KnowledgeBase](#) object

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

Agents for Amazon Bedrock Runtime

The following actions are supported by Agents for Amazon Bedrock Runtime:

- [InvokeAgent](#)
- [Retrieve](#)
- [RetrieveAndGenerate](#)

InvokeAgent

Service: Agents for Amazon Bedrock Runtime

Note

The AWS CLI doesn't support InvokeAgent.

Sends a prompt for the agent to process and respond to. Note the following fields for the request:

- To continue the same conversation with an agent, use the same `sessionId` value in the request.
- To activate trace enablement, turn `enableTrace` to `true`. Trace enablement helps you follow the agent's reasoning process that led it to the information it processed, the actions it took, and the final result it yielded. For more information, see [Trace enablement](#).
- End a conversation by setting `endSession` to `true`.
- In the `sessionState` object, you can include attributes for the session or prompt or, if you configured an action group to return control, results from invocation of the action group.

The response is returned in the `bytes` field of the chunk object.

- The `attribution` object contains citations for parts of the response.
- If you set `enableTrace` to `true` in the request, you can trace the agent's steps and reasoning process that led it to the response.
- If the action predicted was configured to return control, the response returns parameters for the action, elicited from the user, in the `returnControl` field.
- Errors are also surfaced in the response.

Request Syntax

```
POST /agents/agentId/agentAliases/agentAliasId/sessions/sessionId/text HTTP/1.1
Content-type: application/json
```

```
{
  "enableTrace": boolean,
  "endSession": boolean,
  "inputText": "string",
  "sessionState": {
```

```
"invocationId": "string",
"promptSessionAttributes": {
  "string" : "string"
},
"returnControlInvocationResults": [
  { ... }
],
"sessionAttributes": {
  "string" : "string"
}
}
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The alias of the agent to use.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: $^{\wedge}[\text{0-9a-zA-Z}]^{\wedge}$

Required: Yes

agentId

The unique identifier of the agent to use.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: $^{\wedge}[\text{0-9a-zA-Z}]^{\wedge}$

Required: Yes

sessionId

The unique identifier of the session. Use the same value across requests to continue the same conversation.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: $^{\wedge}[\text{0-9a-zA-Z._:-}]^{\wedge}$

Required: Yes

Request Body

The request accepts the following data in JSON format.

enableTrace

Specifies whether to turn on the trace or not to track the agent's reasoning process. For more information, see [Trace enablement](#).

Type: Boolean

Required: No

endSession

Specifies whether to end the session with the agent or not.

Type: Boolean

Required: No

inputText

The prompt text to send the agent.

Note

If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: No

sessionState

Contains parameters that specify various attributes of the session. For more information, see [Control session context](#).

Note

If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: [SessionState](#) object

Required: No

Response Syntax

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: contentType
x-amz-bedrock-agent-session-id: sessionId
Content-type: application/json
```

```
{
  "accessDeniedException": {
  },
  "badGatewayException": {
  },
  "chunk": {
    "attribution": {
      "citations": [
        {
          "generatedResponsePart": {
            "textResponsePart": {
              "span": {
                "end": number,
                "start": number
              },
              "text": "string"
            }
          },
          "retrievedReferences": [
            {
              "content": {
                "text": "string"
              },
              "location": {
                "s3Location": {
```

```

        "uri": "string"
      },
      "type": "string"
    },
    "metadata": {
      "string" : JSON value
    }
  ]
}
],
"bytes": blob
},
"conflictException": {
},
"dependencyFailedException": {
},
"internalServerErrorException": {
},
"resourceNotFoundException": {
},
"returnControl": {
  "invocationId": "string",
  "invocationInputs": [
    { ... }
  ]
},
"serviceQuotaExceededException": {
},
"throttlingException": {
},
"trace": {
  "agentAliasId": "string",
  "agentId": "string",
  "agentVersion": "string",
  "sessionId": "string",
  "trace": { ... }
},
"validationException": {
}
}

```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the input data in the request. The default value is `application/json`.

sessionId

The unique identifier of the session with the agent.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `^[0-9a-zA-Z._: -]+$`

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

chunk

Contains a part of an agent response and citations for it.

Type: [PayloadPart](#) object

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception
HTTP Status Code: 409

[dependencyFailedException](#)

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception
HTTP Status Code: 424

[internalServerErrorException](#)

An internal server error occurred. Retry your request.

Type: Exception
HTTP Status Code: 500

[resourceNotFoundException](#)

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception
HTTP Status Code: 404

[returnControl](#)

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: [ReturnControlPayload](#) object

[serviceQuotaExceededException](#)

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception
HTTP Status Code: 400

[throttlingException](#)

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception
HTTP Status Code: 429

[trace](#)

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see [Trace events](#).

Type: [TracePart](#) object

[validationException](#)

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

HTTP Status Code: 424

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example simple request

The following example inquires the agent to get the weather for Seattle.

```
POST https://bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/agentAliases/TSTALIASID/sessions/abb/text
```

```
{
  "inputText": "give me the weather for seattle",
  "enableTrace": true
}
```

Example response (action group defined with OpenAPI schema, control returned)

The following example shows a response from an agent that has invoked an action group that was configured as follows:

- Defined with an OpenAPI schema
- Configured to return control to the agent developer

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: application/json
x-amz-bedrock-agent-session-id: session0
Content-type: application/json

{
  "invocationInputs": [{
    "apiInvocationInput": {
      "actionGroup": "WeatherAPIs",
      "apiPath": "/get-weather",
      "httpMethod": "get",
      "parameters": [
        {
          "name": "location",
          "type": "string",
          "value": "seattle"
        },
        {
          "name": "date",
          "type": "string",
          "value": "2024-09-15"
        }
      ]
    }
  ]},
  "invocationId": "337cb2f6-ec74-4b49-8141-00b8091498ad"
}
```

Example request using results from returned control (action group defined with OpenAPI schema)

The following example shows a request in which the results returned in the `InvokeAgent` response from an agent are passed to the `sessionState` of a new request. The results were returned from an agent that has invoked an action group that was configured as follows:

- Defined with an OpenAPI schema
- Configured to return control to the agent developer

The `invocationId` must match the `invocationId` that was returned in the response.

```
POST https://bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/agentAliases/TSTALIASID/sessions/abb/text
```

```
{
  "enableTrace": true,
  "sessionState": {
    "invocationId": "337cb2f6-ec74-4b49-8141-00b8091498ad",
    "returnControlInvocationResults": [{
      "apiResult": {
        "actionGroup": "WeatherAPIs",
        "httpMethod": "get",
        "apiPath": "/get-weather",
        "responseBody": {
          "application/json": {
            "body": "It's rainy in Seattle today."
          }
        }
      }
    ]
  }
}
```

Example response (action group defined with function details, control returned)

The following example shows a response from an agent that has invoked an action group that was configured as follows:

- Defined with function details
- Configured to return control to the agent developer

```

HTTP/1.1 200
x-amzn-bedrock-agent-content-type: application/json
x-amz-bedrock-agent-session-id: session0
Content-type: application/json

{
  "invocationInputs": [{
    "functionInvocationInput": {
      "actionGroup": "WeatherAPIs",
      "function": "getWeather",
      "parameters": [
        {
          "name": "location",
          "type": "string",
          "value": "seattle"
        },
        {
          "name": "date",
          "type": "string",
          "value": "2024-09-15"
        }
      ]
    }
  ]
},
  "invocationId": "79e0feaa-c6f7-49bf-814d-b7c498505172"
}

```

Example request using results from returned control (action group defined with function details)

The following example shows a request in which the results returned in the `InvokeAgent` response from an agent are passed to the `sessionState` of a new request. The results were returned from an agent that has invoked an action group that was configured as follows:

- Defined with function details
- Configured to return control to the agent developer

The `invocationId` must match the `invocationId` that was returned in the response.

```

POST https://bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/
agentAliases/TSTALIASID/sessions/abb/text

```

```
{
  "enableTrace": true,
  "sessionState": {
    "invocationId": "79e0feaa-c6f7-49bf-814d-b7c498505172",
    "returnControlInvocationResults": [{
      "functionResult": {
        "actionGroup": "WeatherAPIs",
        "function": "getWeather",
        "responseBody": {
          "TEXT": {
            "body": "It's rainy in Seattle today."
          }
        }
      }
    }]
  }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

Retrieve

Service: Agents for Amazon Bedrock Runtime

Queries a knowledge base and retrieves information from it.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/retrieve HTTP/1.1
Content-type: application/json
```

```
{
  "nextToken": "string",
  "retrievalConfiguration": {
    "vectorSearchConfiguration": {
      "filter": { ... },
      "numberOfResults": number,
      "overrideSearchType": "string"
    }
  },
  "retrievalQuery": {
    "text": "string"
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to query.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `^[0-9a-zA-Z]+$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

nextToken

If there are more results than can fit in the response, the response returns a `nextToken`. Use this token in the `nextToken` field of another request to retrieve the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

Required: No

retrievalConfiguration

Contains configurations for the knowledge base query and retrieval process. For more information, see [Query configurations](#).

Type: [KnowledgeBaseRetrievalConfiguration](#) object

Required: No

retrievalQuery

Contains the query to send the knowledge base.

Type: [KnowledgeBaseQuery](#) object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
  "nextToken": "string",
  "retrievalResults": [
    {
      "content": {
        "text": "string"
      },
      "location": {
        "s3Location": {
```

```
        "uri": "string"
      },
      "type": "string"
    },
    "metadata": {
      "string" : JSON value
    },
    "score": number
  }
]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If there are more results than can fit in the response, the response returns a nextToken. Use this token in the nextToken field of another request to retrieve the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: $^{\backslash}S^*\$$

retrievalResults

A list of results from querying the knowledge base.

Type: Array of [KnowledgeBaseRetrievalResult](#) objects

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

HTTP Status Code: 424

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Send a basic query

The following example queries a knowledge base.

```
POST /knowledgebases/KB12345678/retrieve HTTP/1.1
Content-type: application/json

{
  "retrievalQuery": {
    "text": "What is AWS?"
  }
}
```

Send a query and include filters

To include filters in a knowledge base query, at least one of the data source files must include a `.metadata.json` file. For example, if you had a data source of articles called `articles.pdf`, accompanied by a metadata file called `articles.metadata.json`, you could tag it for genre, year, and author. In the Retrieve request, you could apply the following filter to return all entertainment articles written after 2018, in addition to cooking or sports articles written by authors starting with C.

```
POST /knowledgebases/KB12345678/retrieve HTTP/1.1
Content-type: application/json

{
  "retrievalQuery": {
    "text": "What is AWS?",
  },
  "vectorSearchConfiguration": {
    "numberOfResults": 5,
    "filter": {
      "orAll": [
        {
          "andAll": [
            {
              "equals": {
```

```
        "key": "genre",
        "value": "entertainment"
    }
},
{
    "greaterThan": {
        "key": "year",
        "value": 2018
    }
}
],
},
{
    "andAll": [
        {
            "in": {
                "key": "genre",
                "value": ["cooking", "sports"]
            }
        },
        {
            "startsWith": {
                "key": "author",
                "value": "C"
            }
        }
    ]
}
]
}
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

RetrieveAndGenerate

Service: Agents for Amazon Bedrock Runtime

Queries a knowledge base and generates responses based on the retrieved results. The response only cites sources that are relevant to the query.

Request Syntax

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
  "input": {
    "text": "string"
  },
  "retrieveAndGenerateConfiguration": {
    "externalSourcesConfiguration": {
      "generationConfiguration": {
        "additionalModelRequestFields": {
          "string" : JSON value
        },
        "guardrailConfiguration": {
          "guardrailId": "string",
          "guardrailVersion": "string"
        },
        "inferenceConfig": {
          "textInferenceConfig": {
            "maxTokens": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topP": number
          }
        },
        "promptTemplate": {
          "textPromptTemplate": "string"
        }
      },
      "modelArn": "string",
      "sources": [
        {
          "byteContent": {
            "contentType": "string",
            "data": blob,

```

```

        "identifier": "string"
      },
      "s3Location": {
        "uri": "string"
      },
      "sourceType": "string"
    }
  ]
},
"knowledgeBaseConfiguration": {
  "generationConfiguration": {
    "additionalModelRequestFields": {
      "string" : JSON value
    },
    "guardrailConfiguration": {
      "guardrailId": "string",
      "guardrailVersion": "string"
    },
    "inferenceConfig": {
      "textInferenceConfig": {
        "maxTokens": number,
        "stopSequences": [ "string" ],
        "temperature": number,
        "topP": number
      }
    },
    "promptTemplate": {
      "textPromptTemplate": "string"
    }
  },
  "knowledgeBaseId": "string",
  "modelArn": "string",
  "retrievalConfiguration": {
    "vectorSearchConfiguration": {
      "filter": { ... },
      "numberOfResults": number,
      "overrideSearchType": "string"
    }
  }
},
"type": "string"
},
"sessionConfiguration": {
  "kmsKeyArn": "string"
}

```



```
  },  
  "sessionId": "string"  
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

input

Contains the query to be made to the knowledge base.

Type: [RetrieveAndGenerateInput](#) object

Required: Yes

retrieveAndGenerateConfiguration

Contains configurations for the knowledge base query and retrieval process. For more information, see [Query configurations](#).

Type: [RetrieveAndGenerateConfiguration](#) object

Required: No

sessionConfiguration

Contains details about the session with the knowledge base.

Type: [RetrieveAndGenerateSessionConfiguration](#) object

Required: No

sessionId

The unique identifier of the session. Reuse the same value to continue the same session with the knowledge base.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `^[0-9a-zA-Z._:~]+`

Required: No

Response Syntax

HTTP/1.1 200

Content-type: application/json

```
{
  "citations": [
    {
      "generatedResponsePart": {
        "textResponsePart": {
          "span": {
            "end": number,
            "start": number
          },
          "text": "string"
        }
      },
      "retrievedReferences": [
        {
          "content": {
            "text": "string"
          },
          "location": {
            "s3Location": {
              "uri": "string"
            },
            "type": "string"
          },
          "metadata": {
            "string" : JSON value
          }
        }
      ]
    }
  ],
  "guardrailAction": "string",
  "output": {
    "text": "string"
  },
}
```

```
"sessionId": "string"  
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

[citations](#)

A list of segments of the generated response that are based on sources in the knowledge base, alongside information about the sources.

Type: Array of [Citation](#) objects

[guardrailAction](#)

Specifies if there is a guardrail intervention in the response.

Type: String

Valid Values: INTERVENED | NONE

[output](#)

Contains the response generated from querying the knowledge base.

Type: [RetrieveAndGenerateOutput](#) object

[sessionId](#)

The unique identifier of the session. Reuse the same value to continue the same session with the knowledge base.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `^[0-9a-zA-Z._:-]+$`

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

HTTP Status Code: 424

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Send a basic query

The following example uses the minimally required fields to generate a response after querying a knowledge base.

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
  "input": {
    "text": "What is AWS?"
  },
  "retrieveAndGenerateConfiguration": {
    "knowledgeBaseConfiguration": {
      "knowledgeBaseId": "KB12345678",
      "modelArn": "anthropic.claude-v2:1"
    },
    "type": "KNOWLEDGE_BASE"
  }
}
```

Send a query and include filters

To include filters in a knowledge base query, at least one of the data source files must include a `.metadata.json` file. For example, if you had a data source of articles called `articles.pdf`, accompanied by a metadata file called `articles.metadata.json`, you could tag it for genre, year, and author. In the Retrieve request, you could apply the following filter to return all entertainment articles written after 2018, in addition to cooking or sports articles written by authors starting with C.

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
```

```
"input": {
  "text": "What is AWS?",
},
"retrieveAndGenerateConfiguration": {
  "knowledgeBaseConfiguration": {
    "knowledgeBaseId": "KB12345678",
    "modelArn": "anthropic.claude-v2:1",
    "retrievalConfiguration": {
      "vectorSearchConfiguration": {
        "numberOfResults": 5,
        "filter": {
          "orAll": [
            {
              "andAll": [
                {
                  "equals": {
                    "key": "genre",
                    "value": "entertainment"
                  }
                },
                {
                  "greaterThan": {
                    "key": "year",
                    "value": 2018
                  }
                }
              ]
            },
            {
              "andAll": [
                {
                  "in": {
                    "key": "genre",
                    "value": ["cooking", "sports"]
                  }
                },
                {
                  "startsWith": {
                    "key": "author",
                    "value": "C"
                  }
                }
              ]
            }
          ]
        }
      }
    }
  }
}
```

```
    ]
  }
}
},
"type": "KNOWLEDGE_BASE"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

Amazon Bedrock Runtime

The following actions are supported by Amazon Bedrock Runtime:

- [InvokeModel](#)
- [InvokeModelWithResponseStream](#)

InvokeModel

Service: Amazon Bedrock Runtime

Invokes the specified Amazon Bedrock model to run inference using the prompt and inference parameters provided in the request body. You use model inference to generate text, images, and embeddings.

For example code, see [Invoke model code examples](#).

This operation requires permission for the `bedrock:InvokeModel` action.

Request Syntax

```
POST /model/modelId/invoke HTTP/1.1
Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-Trace: trace
```

body

URI Request Parameters

The request uses the following URI parameters.

[accept](#)

The desired MIME type of the inference body in the response. The default value is `application/json`.

[contentType](#)

The MIME type of the input data in the request. The default value is `application/json`.

[guardrailIdentifier](#)

The unique identifier of the guardrail that you want to use. If you don't provide a value, no guardrail is applied to the invocation.

An error will be thrown in the following situations.

- You don't provide a guardrail identifier but you specify the `amazon-bedrock-guardrailConfig` field in the request body.

- You enable the guardrail but the `contentType` isn't `application/json`.
- You provide a guardrail identifier, but `guardrailVersion` isn't specified.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[\0-9]{12}:guardrail/[a-z0-9]+))$`

guardrailVersion

The version number for the guardrail. The value can also be `DRAFT`.

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

modelId

The unique identifier of the model to invoke to run inference.

The `modelId` to provide depends on the type of model that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see [Amazon Bedrock base model IDs \(on-demand throughput\)](#) in the Amazon Bedrock User Guide.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see [Run inference using a Provisioned Throughput](#) in the Amazon Bedrock User Guide.
- If you use a custom model, first purchase Provisioned Throughput for it. Then specify the ARN of the resulting provisioned model. For more information, see [Use a custom model in Amazon Bedrock](#) in the Amazon Bedrock User Guide.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.:]?[a-z0-9-]{1,63}))|([\0-9]{12}:provisioned-model/[a-z0-9]{12})))|([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.:]?[a-z0-9-]{1,63}))|(([\0-9a-zA-Z][_]?)+)$`

Required: Yes

trace

Specifies whether to enable or disable the Bedrock trace. If enabled, you can see the full Bedrock trace.

Valid Values: ENABLED | DISABLED

Request Body

The request accepts the following binary data.

body

The prompt and inference parameters in the format specified in the `contentType` in the header. To see the format and content of the request and response bodies for different models, refer to [Inference parameters](#). For more information, see [Run inference](#) in the Bedrock User Guide.

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-Type: contentType

body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the inference result.

The response returns the following as the HTTP body.

body

Inference response from the model in the format specified in the `contentType` header. To see the format and content of the request and response bodies for different models, refer to [Inference parameters](#).

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerError

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests.

HTTP Status Code: 429

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Run inference on a text model

Send an invoke request to run inference on a Titan Text G1 - Express model. We set the accept parameter to accept any content type in the response.

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/amazon.titan-text-express-v1/invoke
```

```
-H accept: */*  
-H content-type: application/json
```

```
Payload  
{"inputText": "Hello world"}
```

Example response

Response for the above request.

```
-H content-type: application/json
```

```
Payload  
<the model response>
```

Run inference on an image model

In the following example, the request sets the accept parameter to image/png.

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/stability.stable-diffusion-x1-v1/invoke
```

```
-H accept: image/png
-H content-type: application/json
```

```
Payload
{"inputText": "Picture of a bird"}
```

Example response

Response for the above example.

```
-H content-type: image/png
```

```
Payload
<image bytes>
```

Use a guardrail

This example shows how to use a guardrail with `InvokeModel`.

```
POST /model/modelId/invoke HTTP/1.1
Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-GuardrailTrace: guardrailTrace
X-Amzn-Bedrock-Trace: trace
```

```
body
```

```
// body
{
  "amazon-bedrock-guardrailConfig": {
    "tagSuffix": "string"
  }
}
```

Example response

This is an example response from `InvokeModel` when using a guardrail.

```
HTTP/1.1 200
```

Content-Type: contentType

body

```
// body
{
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE"
  "amazon-bedrock-trace": {
    "guardrails": {
      // Detailed guardrail trace
    }
  }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)
- [AWS SDK for Ruby V3](#)

InvokeModelWithResponseStream

Service: Amazon Bedrock Runtime

Invoke the specified Amazon Bedrock model to run inference using the prompt and inference parameters provided in the request body. The response is returned in a stream.

To see if a model supports streaming, call [GetFoundationModel](#) and check the `responseStreamingSupported` field in the response.

Note

The AWS CLI doesn't support `InvokeModelWithResponseStream`.

For example code, see [Invoke model with streaming code example](#).

This operation requires permissions to perform the `bedrock:InvokeModelWithResponseStream` action.

Request Syntax

```
POST /model/modelId/invoke-with-response-stream HTTP/1.1
X-Amzn-Bedrock-Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-Trace: trace
```

body

URI Request Parameters

The request uses the following URI parameters.

[accept](#)

The desired MIME type of the inference body in the response. The default value is `application/json`.

[contentType](#)

The MIME type of the input data in the request. The default value is `application/json`.

guardrailIdentifier

The unique identifier of the guardrail that you want to use. If you don't provide a value, no guardrail is applied to the invocation.

An error is thrown in the following situations.

- You don't provide a guardrail identifier but you specify the `amazon-bedrock-guardrailConfig` field in the request body.
- You enable the guardrail but the `contentType` isn't `application/json`.
- You provide a guardrail identifier, but `guardrailVersion` isn't specified.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^(([a-z0-9]+)|(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))$`

guardrailVersion

The version number for the guardrail. The value can also be DRAFT.

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

modelId

The unique identifier of the model to invoke to run inference.

The `modelId` to provide depends on the type of model that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see [Amazon Bedrock base model IDs \(on-demand throughput\)](#) in the Amazon Bedrock User Guide.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see [Run inference using a Provisioned Throughput](#) in the Amazon Bedrock User Guide.
- If you use a custom model, first purchase Provisioned Throughput for it. Then specify the ARN of the resulting provisioned model. For more information, see [Use a custom model in Amazon Bedrock](#) in the Amazon Bedrock User Guide.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-`


```
model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.:]?[a-z0-9-]{1,63}))|([0-9]{12}:provisioned-model/[a-z0-9]{12}))|([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_]?)+)$
```

Required: Yes

trace

Specifies whether to enable or disable the Bedrock trace. If enabled, you can see the full Bedrock trace.

Valid Values: ENABLED | DISABLED

Request Body

The request accepts the following binary data.

body

The prompt and inference parameters in the format specified in the `contentType` in the header. To see the format and content of the request and response bodies for different models, refer to [Inference parameters](#). For more information, see [Run inference](#) in the Bedrock User Guide.

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: Yes

Response Syntax

```
HTTP/1.1 200
X-Amzn-Bedrock-Content-Type: contentType
Content-type: application/json

{
  "chunk": {
    "bytes": blob
  },
  "internalServerErrorException": {
  },
  "modelStreamErrorException": {
  },
}
```

```
"modelTimeoutException": {  
},  
"throttlingException": {  
},  
"validationException": {  
}  
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

[contentType](#)

The MIME type of the inference result.

The following data is returned in JSON format by the service.

[chunk](#)

Content included in the response.

Type: [PayloadPart](#) object

[internalServerErrorException](#)

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

[modelStreamErrorException](#)

An error occurred while streaming the response. Retry your request.

Type: Exception

HTTP Status Code: 424

[modelTimeoutException](#)

The request took too long to process. Processing time exceeded the model timeout length.

Type: Exception

HTTP Status Code: 408

throttlingException

The number or frequency of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see [Common Errors](#).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerErrorException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests.

HTTP Status Code: 429

ModelStreamErrorException

An error occurred while streaming the response. Retry your request.

HTTP Status Code: 424

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Run inference with streaming on a text model

For streaming, you can set `x-amzn-bedrock-accept-type` in the header to contain the desired content type of the response. In this example, we set it to accept any content type. The default value is `application/json`.

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/amazon.titan-text-express-
v1/invoke-with-response-stream

-H accept: application/vnd.amazon.eventstream
-H content-type: application/json
-H x-amzn-bedrock-accept: */*
```

```
Payload
{"inputText": "Hello world"}
```

Example response

For streaming, the content type in the response is always set to `application/vnd.amazon.eventstream`. The response includes an additional header (`x-amzn-bedrock-content-type`), which contains the actual content type of the response.

```
-H content-type: application/vnd.amazon.eventstream
-H x-amzn-bedrock-content-type: application/json
```

```
Payload (stream events)
<response chunk>
```

Use a guardrail

This examples show how to use a guardrail with `InvokeModelWithResponseStream`.

```
POST /model/modelId/invoke-with-response-stream HTTP/1.1
X-Amzn-Bedrock-Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-GuardrailTrace: guardrailTrace
X-Amzn-Bedrock-Trace: trace

body

// body
{
  "amazon-bedrock-guardrailConfig": {
    "tagSuffix": "string",
    "streamProcessingMode": "string"
  }
}
```

Example response

This examples shows the response from a call to `InvokeModelWithResponseStream` when using a guardrail.

```
HTTP/1.1 200
X-Amzn-Bedrock-Content-Type: contentType
Content-type: application/json

// chunk 1
{
  "completion": "...",
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE"
}

// chunk 2
{
  "completion": "...",
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE"
}

// last chunk
{
  "completion": "...",
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE",
  "amazon-bedrock-trace": {
    "guardrail": {
      ... // Detailed guardrail trace
    }
  }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS Command Line Interface](#)
- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python](#)

- [AWS SDK for Ruby V3](#)

Data Types

The following data types are supported by Amazon Bedrock:

- [AutomatedEvaluationConfig](#)
- [CloudWatchConfig](#)
- [CustomModelSummary](#)
- [EvaluationBedrockModel](#)
- [EvaluationConfig](#)
- [EvaluationDataset](#)
- [EvaluationDatasetLocation](#)
- [EvaluationDatasetMetricConfig](#)
- [EvaluationInferenceConfig](#)
- [EvaluationModelConfig](#)
- [EvaluationOutputDataConfig](#)
- [EvaluationSummary](#)
- [FoundationModelDetails](#)
- [FoundationModelLifecycle](#)
- [FoundationModelSummary](#)
- [GuardrailContentFilter](#)
- [GuardrailContentFilterConfig](#)
- [GuardrailContentPolicy](#)
- [GuardrailContentPolicyConfig](#)
- [GuardrailManagedWords](#)
- [GuardrailManagedWordsConfig](#)
- [GuardrailPiiEntity](#)
- [GuardrailPiiEntityConfig](#)
- [GuardrailRegex](#)
- [GuardrailRegexConfig](#)

- [GuardrailSensitiveInformationPolicy](#)
- [GuardrailSensitiveInformationPolicyConfig](#)
- [GuardrailSummary](#)
- [GuardrailTopic](#)
- [GuardrailTopicConfig](#)
- [GuardrailTopicPolicy](#)
- [GuardrailTopicPolicyConfig](#)
- [GuardrailWord](#)
- [GuardrailWordConfig](#)
- [GuardrailWordPolicy](#)
- [GuardrailWordPolicyConfig](#)
- [HumanEvaluationConfig](#)
- [HumanEvaluationCustomMetric](#)
- [HumanWorkflowConfig](#)
- [LoggingConfig](#)
- [ModelCustomizationJobSummary](#)
- [OutputDataConfig](#)
- [ProvisionedModelSummary](#)
- [S3Config](#)
- [Tag](#)
- [TrainingDataConfig](#)
- [TrainingMetrics](#)
- [ValidationDataConfig](#)
- [Validator](#)
- [ValidatorMetric](#)
- [VpcConfig](#)

The following data types are supported by Agents for Amazon Bedrock:

- [ActionGroupExecutor](#)
- [ActionGroupSummary](#)

- [Agent](#)
- [AgentActionGroup](#)
- [AgentAlias](#)
- [AgentAliasHistoryEvent](#)
- [AgentAliasRoutingConfigurationListItem](#)
- [AgentAliasSummary](#)
- [AgentKnowledgeBase](#)
- [AgentKnowledgeBaseSummary](#)
- [AgentSummary](#)
- [AgentVersion](#)
- [AgentVersionSummary](#)
- [APISchema](#)
- [ChunkingConfiguration](#)
- [DataSource](#)
- [DataSourceConfiguration](#)
- [DataSourceSummary](#)
- [FixedSizeChunkingConfiguration](#)
- [Function](#)
- [FunctionSchema](#)
- [InferenceConfiguration](#)
- [IngestionJob](#)
- [IngestionJobFilter](#)
- [IngestionJobSortBy](#)
- [IngestionJobStatistics](#)
- [IngestionJobSummary](#)
- [KnowledgeBase](#)
- [KnowledgeBaseConfiguration](#)
- [KnowledgeBaseSummary](#)
- [MongoDbAtlasConfiguration](#)
- [MongoDbAtlasFieldMapping](#)

- [OpenSearchServerlessConfiguration](#)
- [OpenSearchServerlessFieldMapping](#)
- [ParameterDetail](#)
- [PineconeConfiguration](#)
- [PineconeFieldMapping](#)
- [PromptConfiguration](#)
- [PromptOverrideConfiguration](#)
- [RdsConfiguration](#)
- [RdsFieldMapping](#)
- [RedisEnterpriseCloudConfiguration](#)
- [RedisEnterpriseCloudFieldMapping](#)
- [S3DataSourceConfiguration](#)
- [S3Identifier](#)
- [ServerSideEncryptionConfiguration](#)
- [StorageConfiguration](#)
- [ValidationExceptionField](#)
- [VectorIngestionConfiguration](#)
- [VectorKnowledgeBaseConfiguration](#)

The following data types are supported by Agents for Amazon Bedrock Runtime:

- [ActionGroupInvocationInput](#)
- [ActionGroupInvocationOutput](#)
- [ApiInvocationInput](#)
- [ApiParameter](#)
- [ApiRequestBody](#)
- [ApiResponse](#)
- [Attribution](#)
- [ByteContentDoc](#)
- [Citation](#)
- [ContentBody](#)

- [ExternalSource](#)
- [ExternalSourcesGenerationConfiguration](#)
- [ExternalSourcesRetrieveAndGenerateConfiguration](#)
- [FailureTrace](#)
- [FilterAttribute](#)
- [FinalResponse](#)
- [FunctionInvocationInput](#)
- [FunctionParameter](#)
- [FunctionResult](#)
- [GeneratedResponsePart](#)
- [GenerationConfiguration](#)
- [GuardrailConfiguration](#)
- [InferenceConfig](#)
- [InferenceConfiguration](#)
- [InvocationInput](#)
- [InvocationInputMember](#)
- [InvocationResultMember](#)
- [KnowledgeBaseLookupInput](#)
- [KnowledgeBaseLookupOutput](#)
- [KnowledgeBaseQuery](#)
- [KnowledgeBaseRetrievalConfiguration](#)
- [KnowledgeBaseRetrievalResult](#)
- [KnowledgeBaseRetrieveAndGenerateConfiguration](#)
- [KnowledgeBaseVectorSearchConfiguration](#)
- [ModelInvocationInput](#)
- [Observation](#)
- [OrchestrationTrace](#)
- [Parameter](#)
- [PayloadPart](#)
- [PostProcessingModelInvocationOutput](#)

- [PostProcessingParsedResponse](#)
- [PostProcessingTrace](#)
- [PreProcessingModelInvocationOutput](#)
- [PreProcessingParsedResponse](#)
- [PreProcessingTrace](#)
- [PromptTemplate](#)
- [PropertyParameters](#)
- [Rationale](#)
- [RepromptResponse](#)
- [RequestBody](#)
- [ResponseStream](#)
- [RetrievalFilter](#)
- [RetrievalResultContent](#)
- [RetrievalResultLocation](#)
- [RetrievalResultS3Location](#)
- [RetrieveAndGenerateConfiguration](#)
- [RetrieveAndGenerateInput](#)
- [RetrieveAndGenerateOutput](#)
- [RetrieveAndGenerateSessionConfiguration](#)
- [RetrievedReference](#)
- [ReturnControlPayload](#)
- [S3ObjectDoc](#)
- [SessionState](#)
- [Span](#)
- [TextInferenceConfig](#)
- [TextResponsePart](#)
- [Trace](#)
- [TracePart](#)

The following data types are supported by Amazon Bedrock Runtime:

- [PayloadPart](#)
- [ResponseStream](#)

Amazon Bedrock

The following data types are supported by Amazon Bedrock:

- [AutomatedEvaluationConfig](#)
- [CloudWatchConfig](#)
- [CustomModelSummary](#)
- [EvaluationBedrockModel](#)
- [EvaluationConfig](#)
- [EvaluationDataset](#)
- [EvaluationDatasetLocation](#)
- [EvaluationDatasetMetricConfig](#)
- [EvaluationInferenceConfig](#)
- [EvaluationModelConfig](#)
- [EvaluationOutputDataConfig](#)
- [EvaluationSummary](#)
- [FoundationModelDetails](#)
- [FoundationModelLifecycle](#)
- [FoundationModelSummary](#)
- [GuardrailContentFilter](#)
- [GuardrailContentFilterConfig](#)
- [GuardrailContentPolicy](#)
- [GuardrailContentPolicyConfig](#)
- [GuardrailManagedWords](#)
- [GuardrailManagedWordsConfig](#)
- [GuardrailPiiEntity](#)
- [GuardrailPiiEntityConfig](#)
- [GuardrailRegex](#)
- [GuardrailRegexConfig](#)

- [GuardrailSensitiveInformationPolicy](#)
- [GuardrailSensitiveInformationPolicyConfig](#)
- [GuardrailSummary](#)
- [GuardrailTopic](#)
- [GuardrailTopicConfig](#)
- [GuardrailTopicPolicy](#)
- [GuardrailTopicPolicyConfig](#)
- [GuardrailWord](#)
- [GuardrailWordConfig](#)
- [GuardrailWordPolicy](#)
- [GuardrailWordPolicyConfig](#)
- [HumanEvaluationConfig](#)
- [HumanEvaluationCustomMetric](#)
- [HumanWorkflowConfig](#)
- [LoggingConfig](#)
- [ModelCustomizationJobSummary](#)
- [OutputDataConfig](#)
- [ProvisionedModelSummary](#)
- [S3Config](#)
- [Tag](#)
- [TrainingDataConfig](#)
- [TrainingMetrics](#)
- [ValidationDataConfig](#)
- [Validator](#)
- [ValidatorMetric](#)
- [VpcConfig](#)

AutomatedEvaluationConfig

Service: Amazon Bedrock

Use to specify a automatic model evaluation job. The `EvaluationDatasetMetricConfig` object is used to specify the prompt datasets, task type, and metric names.

Contents

`datasetMetricConfigs`

Specifies the required elements for an automatic model evaluation job.

Type: Array of [EvaluationDatasetMetricConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

CloudWatchConfig

Service: Amazon Bedrock

CloudWatch logging configuration.

Contents

logGroupName

The log group name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Required: Yes

roleArn

The role Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/.[+]`

Required: Yes

largeDataDeliveryS3Config

S3 configuration for delivering a large amount of data.

Type: [S3Config](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

CustomModelSummary

Service: Amazon Bedrock

Summary information for a custom model.

Contents

baseModelArn

The base model Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{1}){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

baseModelName

The base model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}$`

Required: Yes

creationTime

Creation time of the model.

Type: Timestamp

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[\.]?){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Required: Yes

modelName

The name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_]?)+$`

Required: Yes

customizationType

Specifies whether to carry out continued pre-training of a model or whether to fine-tune it. For more information, see [Custom models](#).

Type: String

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationBedrockModel

Service: Amazon Bedrock

Contains the ARN of the Amazon Bedrock models specified in your model evaluation job. Each Amazon Bedrock model supports different `inferenceParams`. To learn more about supported inference parameters for Amazon Bedrock models, see [Inference parameters for foundation models](#).

The `inferenceParams` are specified using JSON. To successfully insert JSON as string make sure that all quotations are properly escaped. For example, `"temperature": "0.25"` key value pair would need to be formatted as `\\"temperature\\":\\"0.25\\"` to successfully accepted in the request.

Contents

`inferenceParams`

Each Amazon Bedrock support different inference parameters that change how the model behaves during inference.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: Yes

`modelIdentifier`

The ARN of the Amazon Bedrock model specified.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63} ([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63} [.] {1} [a-z0-9-]{1,63} ([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([\0-9a-zA-Z][_-]?)+)$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationConfig

Service: Amazon Bedrock

Used to specify either a `AutomatedEvaluationConfig` or `HumanEvaluationConfig` object.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

automated

Used to specify an automated model evaluation job. See `AutomatedEvaluationConfig` to view the required parameters.

Type: [AutomatedEvaluationConfig](#) object

Required: No

human

Used to specify a model evaluation job that uses human workers. See `HumanEvaluationConfig` to view the required parameters.

Type: [HumanEvaluationConfig](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationDataset

Service: Amazon Bedrock

Used to specify the name of a built-in prompt dataset and optionally, the Amazon S3 bucket where a custom prompt dataset is saved.

Contents

name

Used to specify supported built-in prompt datasets. Valid values are `Builtin.Bold`, `Builtin.BoolQ`, `Builtin.NaturalQuestions`, `Builtin.Gigaword`, `Builtin.RealToxicityPrompts`, `Builtin.TriviaQa`, `Builtin.T-Rex`, `Builtin.WomensEcommerceClothingReviews` and `Builtin.Wikitext2`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[0-9a-zA-Z-_.]+$`

Required: Yes

datasetLocation

For custom prompt datasets, you must specify the location in Amazon S3 where the prompt dataset is saved.

Type: [EvaluationDatasetLocation](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationDatasetLocation

Service: Amazon Bedrock

The location in Amazon S3 where your prompt dataset is stored.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

s3Uri

The S3 URI of the S3 bucket specified in the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)*?$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationDatasetMetricConfig

Service: Amazon Bedrock

Defines the built-in prompt datasets, built-in metric names and custom metric names, and the task type.

Contents

dataset

Specifies the prompt dataset.

Type: [EvaluationDataset](#) object

Required: Yes

metricNames

The names of the metrics used. For automated model evaluation jobs valid values are "Builtin.Accuracy", "Builtin.Robustness", and "Builtin.Toxicity". In human-based model evaluation jobs the array of strings must match the name parameter specified in `HumanEvaluationCustomMetric`.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[0-9a-zA-Z-_.]+$`

Required: Yes

taskType

The task type you want the model to carry out.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[A-Za-z0-9]+$`

Valid Values: Summarization | Classification | QuestionAndAnswer | Generation | Custom

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationInferenceConfig

Service: Amazon Bedrock

Used to define the models you want used in your model evaluation job. Automated model evaluation jobs support only a single model. In a human-based model evaluation job, your annotator can compare the responses for up to two different models.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

models

Used to specify the models.

Type: Array of [EvaluationModelConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationModelConfig

Service: Amazon Bedrock

Defines the models used in the model evaluation job.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

bedrockModel

Defines the Amazon Bedrock model and inference parameters you want used.

Type: [EvaluationBedrockModel](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationOutputDataConfig

Service: Amazon Bedrock

The Amazon S3 location where the results of your model evaluation job are saved.

Contents

s3Uri

The Amazon S3 URI where the results of model evaluation job are saved.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

EvaluationSummary

Service: Amazon Bedrock

A summary of the model evaluation job.

Contents

creationTime

When the model evaluation job was created.

Type: Timestamp

Required: Yes

evaluationTaskTypes

What task type was used in the model evaluation job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[A-Za-z0-9]+$`

Valid Values: Summarization | Classification | QuestionAndAnswer | Generation | Custom

Required: Yes

jobArn

The Amazon Resource Name (ARN) of the model evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12}$`

Required: Yes

jobName

The name of the model evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9](-*[a-z0-9]){0,62}$`

Required: Yes

jobType

The type, either human or automatic, of model evaluation job.

Type: String

Valid Values: Human | Automated

Required: Yes

modelIdentifiers

The Amazon Resource Names (ARNs) of the model(s) used in the model evaluation job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([\0-9a-zA-Z][_]?)+))$`

Required: Yes

status

The current status of the model evaluation job.

Type: String

Valid Values: InProgress | Completed | Failed | Stopping | Stopped

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FoundationModelDetails

Service: Amazon Bedrock

Information about a foundation model.

Contents

modelArn

The model Amazon Resource Name (ARN).

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

modelId

The model identifier.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)$`

Required: Yes

customizationsSupported

The customization that the model supports.

Type: Array of strings

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: ON_DEMAND | PROVISIONED

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: TEXT | IMAGE | EMBEDDING

Required: No

modelLifecycle

Contains details about whether a model version is available or deprecated

Type: [FoundationModelLifecycle](#) object

Required: No

modelName

The model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: TEXT | IMAGE | EMBEDDING

Required: No

providerName

The model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^\.*$`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FoundationModelLifecycle

Service: Amazon Bedrock

Details about whether a model version is available or deprecated.

Contents

status

Specifies whether a model version is available (ACTIVE) or deprecated (LEGACY).

Type: String

Valid Values: ACTIVE | LEGACY

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FoundationModelSummary

Service: Amazon Bedrock

Summary information for a foundation model.

Contents

modelArn

The Amazon Resource Name (ARN) of the foundation model.

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

modelId

The model ID of the foundation model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)$`

Required: Yes

customizationsSupported

Whether the model supports fine-tuning or continual pre-training.

Type: Array of strings

Valid Values: FINE_TUNING | CONTINUED_PRE_TRAINING

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: ON_DEMAND | PROVISIONED

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: TEXT | IMAGE | EMBEDDING

Required: No

modelLifecycle

Contains details about whether a model version is available or deprecated.

Type: [FoundationModelLifecycle](#) object

Required: No

modelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: TEXT | IMAGE | EMBEDDING

Required: No

providerName

The model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^\.*$`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailContentFilter

Service: Amazon Bedrock

Contains filter strengths for harmful content. Guardrails support the following content filters to detect and filter harmful user inputs and FM-generated outputs.

- **Hate** – Describes language or a statement that discriminates, criticizes, insults, denounces, or dehumanizes a person or group on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin).
- **Insults** – Describes language or a statement that includes demeaning, humiliating, mocking, insulting, or belittling language. This type of language is also labeled as bullying.
- **Sexual** – Describes language or a statement that indicates sexual interest, activity, or arousal using direct or indirect references to body parts, physical traits, or sex.
- **Violence** – Describes language or a statement that includes glorification of or threats to inflict physical pain, hurt, or injury toward a person, group or thing.

Content filtering depends on the confidence classification of user inputs and FM responses across each of the four harmful categories. All input and output statements are classified into one of four confidence levels (NONE, LOW, MEDIUM, HIGH) for each harmful category. For example, if a statement is classified as *Hate* with HIGH confidence, the likelihood of the statement representing hateful content is high. A single statement can be classified across multiple categories with varying confidence levels. For example, a single statement can be classified as *Hate* with HIGH confidence, *Insults* with LOW confidence, *Sexual* with NONE confidence, and *Violence* with MEDIUM confidence.

For more information, see [Guardrails content filters](#).

This data type is used in the following API operations:

- [GetGuardrail response body](#)

Contents

inputStrength

The strength of the content filter to apply to prompts. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: NONE | LOW | MEDIUM | HIGH

Required: Yes

outputStrength

The strength of the content filter to apply to model responses. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: NONE | LOW | MEDIUM | HIGH

Required: Yes

type

The harmful category that the content filter is applied to.

Type: String

Valid Values: SEXUAL | VIOLENCE | HATE | INSULTS | MISCONDUCT | PROMPT_ATTACK

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailContentFilterConfig

Service: Amazon Bedrock

Contains filter strengths for harmful content. Guardrails support the following content filters to detect and filter harmful user inputs and FM-generated outputs.

- **Hate** – Describes language or a statement that discriminates, criticizes, insults, denounces, or dehumanizes a person or group on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin).
- **Insults** – Describes language or a statement that includes demeaning, humiliating, mocking, insulting, or belittling language. This type of language is also labeled as bullying.
- **Sexual** – Describes language or a statement that indicates sexual interest, activity, or arousal using direct or indirect references to body parts, physical traits, or sex.
- **Violence** – Describes language or a statement that includes glorification of or threats to inflict physical pain, hurt, or injury toward a person, group or thing.

Content filtering depends on the confidence classification of user inputs and FM responses across each of the four harmful categories. All input and output statements are classified into one of four confidence levels (NONE, LOW, MEDIUM, HIGH) for each harmful category. For example, if a statement is classified as *Hate* with HIGH confidence, the likelihood of the statement representing hateful content is high. A single statement can be classified across multiple categories with varying confidence levels. For example, a single statement can be classified as *Hate* with HIGH confidence, *Insults* with LOW confidence, *Sexual* with NONE confidence, and *Violence* with MEDIUM confidence.

For more information, see [Guardrails content filters](#).

This data type is used in the following API operations:

- [CreateGuardrail request body](#)
- [UpdateGuardrail request body](#)

Contents

inputStrength

The strength of the content filter to apply to prompts. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: NONE | LOW | MEDIUM | HIGH

Required: Yes

outputStrength

The strength of the content filter to apply to model responses. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: NONE | LOW | MEDIUM | HIGH

Required: Yes

type

The harmful category that the content filter is applied to.

Type: String

Valid Values: SEXUAL | VIOLENCE | HATE | INSULTS | MISCONDUCT | PROMPT_ATTACK

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailContentPolicy

Service: Amazon Bedrock

Contains details about how to handle harmful content.

This data type is used in the following API operations:

- [GetGuardrail response body](#)

Contents

filters

Contains the type of the content filter and how strongly it should apply to prompts and model responses.

Type: Array of [GuardrailContentFilter](#) objects

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailContentPolicyConfig

Service: Amazon Bedrock

Contains details about how to handle harmful content.

This data type is used in the following API operations:

- [CreateGuardrail request body](#)
- [UpdateGuardrail request body](#)

Contents

filtersConfig

Contains the type of the content filter and how strongly it should apply to prompts and model responses.

Type: Array of [GuardrailContentFilterConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailManagedWords

Service: Amazon Bedrock

The managed word list that was configured for the guardrail. (This is a list of words that are pre-defined and managed by Guardrails only.)

Contents

type

ManagedWords\$type The managed word type that was configured for the guardrail. (For now, we only offer profanity word list)

Type: String

Valid Values: PROFANITY

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailManagedWordsConfig

Service: Amazon Bedrock

The managed word list to configure for the guardrail.

Contents

type

The managed word type to configure for the guardrail.

Type: String

Valid Values: PROFANITY

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailPiiEntity

Service: Amazon Bedrock

The PII entity configured for the guardrail.

Contents

action

The configured guardrail action when PII entity is detected.

Type: String

Valid Values: BLOCK | ANONYMIZE

Required: Yes

type

The type of PII entity. For example, Social Security Number.

Type: String

Valid Values: ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailPiiEntityConfig

Service: Amazon Bedrock

The PII entity to configure for the guardrail.

Contents

action

Configure guardrail action when the PII entity is detected.

Type: String

Valid Values: BLOCK | ANONYMIZE

Required: Yes

type

Configure guardrail type when the PII entity is detected.

Type: String

Valid Values: ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailRegex

Service: Amazon Bedrock

The regular expression configured for the guardrail.

Contents

action

The action taken when a match to the regular expression is detected.

Type: String

Valid Values: BLOCK | ANONYMIZE

Required: Yes

name

The name of the regular expression for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

pattern

The pattern of the regular expression configured for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

description

The description of the regular expression for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailRegexConfig

Service: Amazon Bedrock

The regular expression to configure for the guardrail.

Contents

action

The guardrail action to configure when matching regular expression is detected.

Type: String

Valid Values: BLOCK | ANONYMIZE

Required: Yes

name

The name of the regular expression to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

pattern

The regular expression pattern to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

description

The description of the regular expression to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailSensitiveInformationPolicy

Service: Amazon Bedrock

Contains details about PII entities and regular expressions configured for the guardrail.

Contents

piiEntities

The list of PII entities configured for the guardrail.

Type: Array of [GuardrailPiiEntity](#) objects

Array Members: Minimum number of 1 item.

Required: No

regexes

The list of regular expressions configured for the guardrail.

Type: Array of [GuardrailRegex](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailSensitiveInformationPolicyConfig

Service: Amazon Bedrock

Contains details about PII entities and regular expressions to configure for the guardrail.

Contents

piiEntitiesConfig

A list of PII entities to configure to the guardrail.

Type: Array of [GuardrailPiiEntityConfig](#) objects

Array Members: Minimum number of 1 item.

Required: No

regexesConfig

A list of regular expressions to configure to the guardrail.

Type: Array of [GuardrailRegexConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailSummary

Service: Amazon Bedrock

Contains details about a guardrail.

This data type is used in the following API operations:

- [ListGuardrails response body](#)

Contents

arn

The ARN of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+$`

Required: Yes

createdAt

The date and time at which the guardrail was created.

Type: Timestamp

Required: Yes

id

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

Required: Yes

name

The name of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `^[0-9a-zA-Z-_]+$`

Required: Yes

status

The status of the guardrail.

Type: String

Valid Values: CREATING | UPDATING | VERSIONING | READY | FAILED | DELETING

Required: Yes

updatedAt

The date and time at which the guardrail was last updated.

Type: Timestamp

Required: Yes

version

The version of the guardrail.

Type: String

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

Required: Yes

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailTopic

Service: Amazon Bedrock

Details about topics for the guardrail to identify and deny.

This data type is used in the following API operations:

- [GetGuardrail response body](#)

Contents

definition

A definition of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

name

The name of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `^[0-9a-zA-Z- _ !? .]+`

Required: Yes

examples

A list of prompts, each of which is an example of a prompt that can be categorized as belonging to the topic.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

type

Specifies to deny the topic.

Type: String

Valid Values: DENY

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailTopicConfig

Service: Amazon Bedrock

Details about topics for the guardrail to identify and deny.

This data type is used in the following API operations:

- [CreateGuardrail request body](#)
- [UpdateGuardrail request body](#)

Contents

definition

A definition of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

name

The name of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `^[0-9a-zA-Z- _ !?\.]+$`

Required: Yes

type

Specifies to deny the topic.

Type: String

Valid Values: DENY

Required: Yes

examples

A list of prompts, each of which is an example of a prompt that can be categorized as belonging to the topic.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailTopicPolicy

Service: Amazon Bedrock

Contains details about topics that the guardrail should identify and deny.

This data type is used in the following API operations:

- [GetGuardrail response body](#)

Contents

topics

A list of policies related to topics that the guardrail should deny.

Type: Array of [GuardrailTopic](#) objects

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailTopicPolicyConfig

Service: Amazon Bedrock

Contains details about topics that the guardrail should identify and deny.

This data type is used in the following API operations:

- [CreateGuardrail request body](#)
- [UpdateGuardrail request body](#)

Contents

topicsConfig

A list of policies related to topics that the guardrail should deny.

Type: Array of [GuardrailTopicConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailWord

Service: Amazon Bedrock

A word configured for the guardrail.

Contents

text

Text of the word configured for the guardrail to block.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailWordConfig

Service: Amazon Bedrock

A word to configure for the guardrail.

Contents

text

Text of the word configured for the guardrail to block.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailWordPolicy

Service: Amazon Bedrock

Contains details about the word policy configured for the guardrail.

Contents

managedWordLists

A list of managed words configured for the guardrail.

Type: Array of [GuardrailManagedWords](#) objects

Required: No

words

A list of words configured for the guardrail.

Type: Array of [GuardrailWord](#) objects

Array Members: Minimum number of 1 item. Maximum number of 10000 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailWordPolicyConfig

Service: Amazon Bedrock

Contains details about the word policy to configured for the guardrail.

Contents

managedWordListsConfig

A list of managed words to configure for the guardrail.

Type: Array of [GuardrailManagedWordsConfig](#) objects

Required: No

wordsConfig

A list of words to configure for the guardrail.

Type: Array of [GuardrailWordConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 10000 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

HumanEvaluationConfig

Service: Amazon Bedrock

Specifies the custom metrics, how tasks will be rated, the flow definition ARN, and your custom prompt datasets. Model evaluation jobs use human workers *only* support the use of custom prompt datasets. To learn more about custom prompt datasets and the required format, see [Custom prompt datasets](#).

When you create custom metrics in `HumanEvaluationCustomMetric` you must specify the metric's name. The list of names specified in the `HumanEvaluationCustomMetric` array, must match the `metricNames` array of strings specified in `EvaluationDatasetMetricConfig`. For example, if in the `HumanEvaluationCustomMetric` array you specified the names "accuracy", "toxicity", "readability" as custom metrics *then* the `metricNames` array would need to look like the following ["accuracy", "toxicity", "readability"] in `EvaluationDatasetMetricConfig`.

Contents

`datasetMetricConfigs`

Use to specify the metrics, task, and prompt dataset to be used in your model evaluation job.

Type: Array of [EvaluationDatasetMetricConfig](#) objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

`customMetrics`

A `HumanEvaluationCustomMetric` object. It contains the names the metrics, how the metrics are to be evaluated, an optional description.

Type: Array of [HumanEvaluationCustomMetric](#) objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

`humanWorkflowConfig`

The parameters of the human workflow.

Type: [HumanWorkflowConfig](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

HumanEvaluationCustomMetric

Service: Amazon Bedrock

In a model evaluation job that uses human workers you must define the name of the metric, and how you want that metric rated `ratingMethod`, and an optional description of the metric.

Contents

name

The name of the metric. Your human evaluators will see this name in the evaluation UI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[0-9a-zA-Z-_.]+$`

Required: Yes

ratingMethod

Choose how you want your human workers to evaluation your model. Valid values for rating methods are `ThumbsUpDown`, `IndividualLikertScale`, `ComparisonLikertScale`, `ComparisonChoice`, and `ComparisonRank`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `^[0-9a-zA-Z-_.]+$`

Required: Yes

description

An optional description of the metric. Use this parameter to provide more details about the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^.+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

HumanWorkflowConfig

Service: Amazon Bedrock

Contains SageMakerFlowDefinition object. The object is used to specify the prompt dataset, task type, rating method and metric names.

Contents

flowDefinitionArn

The Amazon Resource Number (ARN) for the flow definition

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `^arn:aws(-[^\:]+)?:sagemaker:[a-z0-9-]{1,20}:[0-9]{12}:flow-definition/.*$`

Required: Yes

instructions

Instructions for the flow definition

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5000.

Pattern: `^[\\S\\s]+$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

LoggingConfig

Service: Amazon Bedrock

Configuration fields for invocation logging.

Contents

cloudWatchConfig

CloudWatch logging configuration.

Type: [CloudWatchConfig](#) object

Required: No

embeddingDataDeliveryEnabled

Set to include embeddings data in the log delivery.

Type: Boolean

Required: No

imageDataDeliveryEnabled

Set to include image data in the log delivery.

Type: Boolean

Required: No

s3Config

S3 configuration for storing log data.

Type: [S3Config](#) object

Required: No

textDataDeliveryEnabled

Set to include text data in the log delivery.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ModelCustomizationJobSummary

Service: Amazon Bedrock

Information about one customization job

Contents

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{0,2}[a-z0-9-]{1,63}([[:][a-z0-9-]{1,63}){0,2})))$`

Required: Yes

creationTime

Creation time of the custom model.

Type: Timestamp

Required: Yes

jobArn

Amazon Resource Name (ARN) of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[\0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]{0,2}[a-z0-9-]{1,63}([[:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})$`

Required: Yes

jobName

Name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.\.])*$`

Required: Yes

status

Status of the customization job.

Type: String

Valid Values: `InProgress` | `Completed` | `Failed` | `Stopping` | `Stopped`

Required: Yes

customizationType

Specifies whether to carry out continued pre-training of a model or whether to fine-tune it. For more information, see [Custom models](#).

Type: String

Valid Values: `FINE_TUNING` | `CONTINUED_PRE_TRAINING`

Required: No

customModelArn

Amazon Resource Name (ARN) of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]{1}){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Required: No

customModelName

Name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[0-9a-zA-Z][_ -]?+$`

Required: No

endTime

Time that the customization job ended.

Type: Timestamp

Required: No

lastModifiedTime

Time that the customization job was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

OutputDataConfig

Service: Amazon Bedrock

S3 Location of the output data.

Contents

s3Uri

The S3 URI where the output data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ProvisionedModelSummary

Service: Amazon Bedrock

A summary of information about a Provisioned Throughput.

This data type is used in the following API operations:

- [ListProvisionedThroughputs response](#)

Contents

creationTime

The time that the Provisioned Throughput was created.

Type: Timestamp

Required: Yes

desiredModelArn

The Amazon Resource Name (ARN) of the model requested to be associated to this Provisioned Throughput. This value differs from the `modelArn` if updating hasn't completed.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]{1}){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

desiredModelUnits

The number of model units that was requested to be allocated to the Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

foundationModelArn

The Amazon Resource Name (ARN) of the base model for which the Provisioned Throughput was created, or of the base model that the custom model for which the Provisioned Throughput was created was customized.

Type: String

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

lastModifiedTime

The time that the Provisioned Throughput was last modified.

Type: Timestamp

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the model associated with the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[:]+)?:bedrock:[a-z0-9-]{1,20}:(([:0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9-]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

modelUnits

The number of model units allocated to the Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelArn

The Amazon Resource Name (ARN) of the Provisioned Throughput.

Type: String

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

Required: Yes

provisionedModelName

The name of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_ -]?)+$`

Required: Yes

status

The status of the Provisioned Throughput.

Type: String

Valid Values: `Creating` | `InService` | `Updating` | `Failed`

Required: Yes

commitmentDuration

The duration for which the Provisioned Throughput was committed.

Type: String

Valid Values: `OneMonth` | `SixMonths`

Required: No

commitmentExpirationTime

The timestamp for when the commitment term of the Provisioned Throughput expires.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

S3Config

Service: Amazon Bedrock

S3 configuration for storing log data.

Contents

bucketName

S3 bucket name.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Required: Yes

keyPrefix

S3 prefix.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Tag

Service: Amazon Bedrock

Definition of the key/value pair for a tag.

Contents

key

Key for the tag.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Required: Yes

value

Value for the tag.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `^[a-zA-Z0-9\s._:/+=@-]*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

TrainingDataConfig

Service: Amazon Bedrock

S3 Location of the training data.

Contents

s3Uri

The S3 URI where the training data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/\.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

TrainingMetrics

Service: Amazon Bedrock

Metrics associated with the custom job.

Contents

trainingLoss

Loss metric associated with the custom job.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ValidationDataConfig

Service: Amazon Bedrock

Array of up to 10 validators.

Contents

validators

Information about the validators.

Type: Array of [Validator](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Validator

Service: Amazon Bedrock

Information about a validator.

Contents

s3Uri

The S3 URI where the validation data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ValidatorMetric

Service: Amazon Bedrock

The metric for the validator.

Contents

validationLoss

The validation loss associated with this validator.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

VpcConfig

Service: Amazon Bedrock

VPC configuration.

Contents

securityGroupIds

VPC configuration security group Ids.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `^[-0-9a-zA-Z]+$`

Required: Yes

subnetIds

VPC configuration subnets.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `^[-0-9a-zA-Z]+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Agents for Amazon Bedrock

The following data types are supported by Agents for Amazon Bedrock:

- [ActionGroupExecutor](#)
- [ActionGroupSummary](#)
- [Agent](#)
- [AgentActionGroup](#)
- [AgentAlias](#)
- [AgentAliasHistoryEvent](#)
- [AgentAliasRoutingConfigurationListItem](#)
- [AgentAliasSummary](#)
- [AgentKnowledgeBase](#)
- [AgentKnowledgeBaseSummary](#)
- [AgentSummary](#)
- [AgentVersion](#)
- [AgentVersionSummary](#)
- [APISchema](#)
- [ChunkingConfiguration](#)
- [DataSource](#)
- [DataSourceConfiguration](#)
- [DataSourceSummary](#)
- [FixedSizeChunkingConfiguration](#)
- [Function](#)
- [FunctionSchema](#)
- [InferenceConfiguration](#)
- [IngestionJob](#)
- [IngestionJobFilter](#)
- [IngestionJobSortBy](#)
- [IngestionJobStatistics](#)
- [IngestionJobSummary](#)

- [KnowledgeBase](#)
- [KnowledgeBaseConfiguration](#)
- [KnowledgeBaseSummary](#)
- [MongoDbAtlasConfiguration](#)
- [MongoDbAtlasFieldMapping](#)
- [OpenSearchServerlessConfiguration](#)
- [OpenSearchServerlessFieldMapping](#)
- [ParameterDetail](#)
- [PineconeConfiguration](#)
- [PineconeFieldMapping](#)
- [PromptConfiguration](#)
- [PromptOverrideConfiguration](#)
- [RdsConfiguration](#)
- [RdsFieldMapping](#)
- [RedisEnterpriseCloudConfiguration](#)
- [RedisEnterpriseCloudFieldMapping](#)
- [S3DataSourceConfiguration](#)
- [S3Identifier](#)
- [ServerSideEncryptionConfiguration](#)
- [StorageConfiguration](#)
- [ValidationExceptionField](#)
- [VectorIngestionConfiguration](#)
- [VectorKnowledgeBaseConfiguration](#)

ActionGroupExecutor

Service: Agents for Amazon Bedrock

Contains details about the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

customControl

To return the action group invocation results directly in the InvokeAgent response, specify RETURN_CONTROL.

Type: String

Valid Values: RETURN_CONTROL

Required: No

lambda

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+\d{1}:\d{12}:function:[a-zA-Z0-9-_\.\.]+(:(\d{12}|[a-zA-Z0-9-_\.\.]+))?$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ActionGroupSummary

Service: Agents for Amazon Bedrock

Contains details about an action group.

Contents

actionGroupId

The unique identifier of the action group.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

actionGroupName

The name of the action group.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?) {1,100}$`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: Yes

updatedAt

The time at which the action group was last updated.

Type: Timestamp

Required: Yes

description

The description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Agent

Service: Agents for Amazon Bedrock

Contains details about an agent.

Contents

agentArn

The Amazon Resource Name (ARN) of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent/[0-9a-zA-Z]{10}$`

Required: Yes

agentId

The unique identifier of the agent.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentName

The name of the agent.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?){1,100}$`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/(service-role)?
AmazonBedrockExecutionRoleForAgents_.$`

Required: Yes

agentStatus

The status of the agent and whether it is ready for use. The following statuses are possible:

- **CREATING** – The agent is being created.
- **PREPARING** – The agent is being prepared.
- **PREPARED** – The agent is prepared and ready to be invoked.
- **NOT_PREPARED** – The agent has been created but not yet prepared.
- **FAILED** – The agent API operation failed.
- **UPDATING** – The agent is being updated.
- **DELETING** – The agent is being deleted.

Type: String

Valid Values: **CREATING** | **PREPARING** | **PREPARED** | **NOT_PREPARED** | **DELETING** | **FAILED** | **VERSIONING** | **UPDATING**

Required: Yes

agentVersion

The version of the agent.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `^DRAFT$`

Required: Yes

createdAt

The time at which the agent was created.

Type: Timestamp

Required: Yes

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: Yes

updatedAt

The time at which the agent was last updated.

Type: Timestamp

Required: Yes

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key that encrypts the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: No

description

The description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

Contains reasons that the agent-related API that you invoked failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

foundationModel

The foundation model used for orchestration by the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: No

instruction

Instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 1200.

Required: No

preparedAt

The time at which the agent was last prepared.

Type: Timestamp

Required: No

promptOverrideConfiguration

Contains configurations to override prompt templates in different parts of an agent sequence. For more information, see [Advanced prompts](#).

Type: [PromptOverrideConfiguration](#) object

Required: No

recommendedActions

Contains recommended actions to take for the agent-related API that you invoked to succeed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentActionGroup

Service: Agents for Amazon Bedrock

Contains details about an action group.

Contents

actionGroupId

The unique identifier of the action group.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

actionGroupName

The name of the action group.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: Yes

agentId

The unique identifier of the agent to which the action group belongs.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent to which the action group belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

createdAt

The time at which the action group was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the action group was last updated.

Type: Timestamp

Required: Yes

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Type: [ActionGroupExecutor](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see [Action group OpenAPI schemas](#).

Type: [APISchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

The description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Defines functions that each define parameters that the agent needs to invoke from the user. Each function represents an action in an action group.

Type: [FunctionSchema](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionSignature

If this field is set as `AMAZON.UserInput`, the agent can request the user for additional information when trying to complete a task. The `description`, `apiSchema`, and `actionGroupExecutor` fields must be blank for this action group.

During orchestration, if the 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.

Type: String

Valid Values: `AMAZON.UserInput`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentAlias

Service: Agents for Amazon Bedrock

Contains details about an alias of an agent.

Contents

agentAliasArn

The Amazon Resource Name (ARN) of the alias of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}$`

Required: Yes

agentAliasId

The unique identifier of the alias of the agent.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `^(\\bTSTALIASID\\b|[0-9a-zA-Z]+)$`

Required: Yes

agentAliasName

The name of the alias of the agent.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

agentAliasStatus

The status of the alias of the agent and whether it is ready for use. The following statuses are possible:

- CREATING – The agent alias is being created.
- PREPARED – The agent alias is finished being created or updated and is ready to be invoked.
- FAILED – The agent alias API operation failed.
- UPDATING – The agent alias is being updated.
- DELETING – The agent alias is being deleted.

Type: String

Valid Values: CREATING | PREPARED | FAILED | UPDATING | DELETING

Required: Yes

agentId

The unique identifier of the agent.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

createdAt

The time at which the alias of the agent was created.

Type: Timestamp

Required: Yes

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of [AgentAliasRoutingConfigurationListItem](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: Yes

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Required: Yes

agentAliasHistoryEvents

Contains details about the history of the alias.

Type: Array of [AgentAliasHistoryEvent](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see [Ensuring idempotency](#).

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

description

The description of the alias of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

Information on the failure of Provisioned Throughput assigned to an agent alias.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentAliasHistoryEvent

Service: Agents for Amazon Bedrock

Contains details about the history of the alias.

Contents

endDate

The date that the alias stopped being associated to the version in the `routingConfiguration` object

Type: Timestamp

Required: No

routingConfiguration

Contains details about the version of the agent with which the alias is associated.

Type: Array of [AgentAliasRoutingConfigurationListItem](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

startDate

The date that the alias began being associated to the version in the `routingConfiguration` object.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentAliasRoutingConfigurationListItem

Service: Agents for Amazon Bedrock

Contains details about the routing configuration of the alias.

Contents

agentVersion

The version of the agent with which the alias is associated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: No

provisionedThroughput

Information on the Provisioned Throughput assigned to an agent alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^((([0-9a-zA-Z][_-]?){1,63})|(arn:aws(-[^\:]+)? :bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentAliasSummary

Service: Agents for Amazon Bedrock

Contains details about an alias of an agent.

Contents

agentAliasId

Contains details about

Type: String

Length Constraints: Fixed length of 10.

Pattern: $^(\backslash\text{bTSTALIASID}\backslash\text{b} | [\text{0-9a-zA-Z}]^+)$$

Required: Yes

agentAliasName

The name of the alias.

Type: String

Pattern: $^([\text{0-9a-zA-Z}] [_ -]?)\{1, 100\}$$

Required: Yes

agentAliasStatus

The status of the alias.

Type: String

Valid Values: CREATING | PREPARED | FAILED | UPDATING | DELETING

Required: Yes

createdAt

The time at which the alias of the agent was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Required: Yes

description

The description of the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the version of the agent with which the alias is associated.

Type: Array of [AgentAliasRoutingConfigurationListItem](#) objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentKnowledgeBase

Service: Agents for Amazon Bedrock

Contains details about a knowledge base that is associated with an agent.

Contents

agentId

The unique identifier of the agent with which the knowledge base is associated.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentVersion

The version of the agent with which the knowledge base is associated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

createdAt

The time at which the association between the agent and the knowledge base was created.

Type: Timestamp

Required: Yes

description

The description of the association between the agent and the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

knowledgeBaseId

The unique identifier of the association between the agent and the knowledge base.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseState

Specifies whether to use the knowledge base or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: Yes

updatedAt

The time at which the association between the agent and the knowledge base was last updated.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentKnowledgeBaseSummary

Service: Agents for Amazon Bedrock

Contains details about a knowledge base associated with an agent.

Contents

knowledgeBaseId

The unique identifier of the knowledge base associated with an agent.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseState

Specifies whether the agent uses the knowledge base or not when sending an [InvokeAgent](#) request.

Type: String

Valid Values: ENABLED | DISABLED

Required: Yes

updatedAt

The time at which the knowledge base associated with an agent was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base associated with an agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentSummary

Service: Agents for Amazon Bedrock

Contains details about an agent.

Contents

agentId

The unique identifier of the agent.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentName

The name of the agent.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

agentStatus

The status of the agent.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING

Required: Yes

updatedAt

The time at which the agent was last updated.

Type: Timestamp

Required: Yes

description

The description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

latestAgentVersion

The latest version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentVersion

Service: Agents for Amazon Bedrock

Contains details about a version of an agent.

Contents

agentArn

The Amazon Resource Name (ARN) of the agent that the version belongs to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent/[0-9a-zA-Z]{10}$`

Required: Yes

agentId

The unique identifier of the agent that the version belongs to.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

agentName

The name of the agent that the version belongs to.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?){1,100}$`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/(service-role)?
AmazonBedrockExecutionRoleForAgents_.$`

Required: Yes

agentStatus

The status of the agent that the version belongs to.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING |
FAILED | VERSIONING | UPDATING

Required: Yes

createdAt

The time at which the version was created.

Type: Timestamp

Required: Yes

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: Yes

updatedAt

The time at which the version was last updated.

Type: Timestamp

Required: Yes

version

The version number.

Type: String

Pattern: `^[0-9]{1,5}$`

Required: Yes

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key that encrypts the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: No

description

The description of the version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the API operation on the version failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

foundationModel

The foundation model that the version invokes.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([\0-9]{12}:custom-model/[a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9-]{12})|(:foundation-model/([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|((([a-z0-9-]{1,63}[\.]{1}[a-z0-9-]{1,63}([\.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([\0-9a-zA-Z][_-]?)+))$`

Required: No

instruction

The instructions provided to the agent.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 1200.

Required: No

promptOverrideConfiguration

Contains configurations to override prompt templates in different parts of an agent sequence. For more information, see [Advanced prompts](#).

Type: [PromptOverrideConfiguration](#) object

Required: No

recommendedActions

A list of recommended actions to take for the failed API operation on the version to succeed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

AgentVersionSummary

Service: Agents for Amazon Bedrock

Contains details about a version of an agent.

Contents

agentName

The name of the agent to which the version belongs.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

agentStatus

The status of the agent to which the version belongs.

Type: String

Valid Values: CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING

Required: Yes

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: Yes

createdAt

The time at which the version was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the version was last updated.

Type: Timestamp

Required: Yes

description

The description of the version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

APISchema

Service: Agents for Amazon Bedrock

Contains details about the OpenAPI schema for the action group. For more information, see [Action group OpenAPI schemas](#). You can either include the schema directly in the payload field or you can upload it to an S3 bucket and specify the S3 bucket location in the s3 field.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

payload

The JSON or YAML-formatted payload defining the OpenAPI schema for the action group. For more information, see [Action group OpenAPI schemas](#).

Type: String

Required: No

s3

Contains details about the S3 object containing the OpenAPI schema for the action group. For more information, see [Action group OpenAPI schemas](#).

Type: [S3Identifier](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ChunkingConfiguration

Service: Agents for Amazon Bedrock

Details about how to chunk the documents in the data source. A *chunk* refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried.

Contents

chunkingStrategy

Knowledge base can split your source data 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. You have the following options for chunking your data. If you opt for NONE, then you may want to pre-process your files by splitting them up such that each file corresponds to a chunk.

- **FIXED_SIZE** – Amazon Bedrock splits your source data into chunks of the approximate size that you set in the `fixedSizeChunkingConfiguration`.
- **NONE** – 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.

Type: String

Valid Values: `FIXED_SIZE` | `NONE`

Required: Yes

fixedSizeChunkingConfiguration

Configurations for when you choose fixed-size chunking. If you set the `chunkingStrategy` as `NONE`, exclude this field.

Type: [FixedSizeChunkingConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

DataSource

Service: Agents for Amazon Bedrock

Contains details about a data source.

Contents

createdAt

The time at which the data source was created.

Type: Timestamp

Required: Yes

dataSourceConfiguration

Contains details about how the data source is stored.

Type: [DataSourceConfiguration](#) object

Required: Yes

dataSourceId

The unique identifier of the data source.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source belongs.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

name

The name of the data source.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

status

The status of the data source. The following statuses are possible:

- Available – The data source has been created and is ready for ingestion into the knowledge base.
- Deleting – The data source is being deleted.

Type: String

Valid Values: AVAILABLE | DELETING | DELETE_UNSUCCESSFUL

Required: Yes

updatedAt

The time at which the data source was last updated.

Type: Timestamp

Required: Yes

dataDeletionPolicy

The data deletion policy for a data source.

Type: String

Valid Values: RETAIN | DELETE

Required: No

description

The description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

The detailed reasons on the failure to delete a data source.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

serverSideEncryptionConfiguration

Contains details about the configuration of the server-side encryption.

Type: [ServerSideEncryptionConfiguration](#) object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: [VectorIngestionConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

DataSourceConfiguration

Service: Agents for Amazon Bedrock

Contains details about how a data source is stored.

Contents

type

The type of storage for the data source.

Type: String

Valid Values: S3

Required: Yes

s3Configuration

Contains details about the configuration of the S3 object containing the data source.

Type: [S3DataSourceConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

DataSourceSummary

Service: Agents for Amazon Bedrock

Contains details about a data source.

Contents

dataSourceId

The unique identifier of the data source.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source belongs.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

name

The name of the data source.

Type: String

Pattern: `^([0-9a-zA-Z][_ -]?)\{1,100\}$`

Required: Yes

status

The status of the data source.

Type: String

Valid Values: AVAILABLE | DELETING | DELETE_UNSUCCESSFUL

Required: Yes

updatedAt

The time at which the data source was last updated.

Type: Timestamp

Required: Yes

description

The description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FixedSizeChunkingConfiguration

Service: Agents for Amazon Bedrock

Configurations for when you choose fixed-size chunking. If you set the chunkingStrategy as NONE, exclude this field.

Contents

maxTokens

The maximum number of tokens to include in a chunk.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

overlapPercentage

The percentage of overlap between adjacent chunks of a data source.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 99.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Function

Service: Agents for Amazon Bedrock

Defines parameters that the agent needs to invoke from the user to complete the function. Corresponds to an action in an action group.

This data type is used in the following API operations:

- [CreateAgentActionGroup request](#)
- [CreateAgentActionGroup response](#)
- [UpdateAgentActionGroup request](#)
- [UpdateAgentActionGroup response](#)
- [GetAgentActionGroup response](#)

Contents

name

A name for the function.

Type: String

Pattern: `^([\u0000-\u009a-\u00za-\u00zA-\u00zZ][\u002d\u005f]?){1,100}$`

Required: Yes

description

A description of the function and its purpose.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1200.

Required: No

parameters

The parameters that the agent elicits from the user to fulfill the function.

Type: String to [ParameterDetail](#) object map

Key Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FunctionSchema

Service: Agents for Amazon Bedrock

Defines functions that each define parameters that the agent needs to invoke from the user. Each function represents an action in an action group.

This data type is used in the following API operations:

- [CreateAgentActionGroup request](#)
- [CreateAgentActionGroup response](#)
- [UpdateAgentActionGroup request](#)
- [UpdateAgentActionGroup response](#)
- [GetAgentActionGroup response](#)

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

functions

A list of functions that each define an action in the action group.

Type: Array of [Function](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InferenceConfiguration

Service: Agents for Amazon Bedrock

Contains inference parameters to use when the agent invokes a foundation model in the part of the agent sequence defined by the `promptType`. For more information, see [Inference parameters for foundation models](#).

Contents

maxLength

The maximum number of tokens to allow in the generated response.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 4096.

Required: No

stopSequences

A list of stop sequences. A stop sequence is a sequence of characters that causes the model to stop generating the response.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

temperature

The likelihood of the model selecting higher-probability options while generating a response. A lower value makes the model more likely to choose higher-probability options, while a higher value makes the model more likely to choose lower-probability options.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topK

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `topK` is the number of most-likely

candidates from which the model chooses the next token in the sequence. For example, if you set topK to 50, the model selects the next token from among the top 50 most likely choices.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 500.

Required: No

topP

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for Top P determines the number of most-likely candidates from which the model chooses the next token in the sequence. For example, if you set topP to 80, the model only selects the next token from the top 80% of the probability distribution of next tokens.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

IngestionJob

Service: Agents for Amazon Bedrock

Contains details about an ingestion job, which converts a data source to embeddings for a vector store in knowledge base.

This data type is used in the following API operations:

- [StartIngestionJob response](#)
- [GetIngestionJob response](#)
- [ListIngestionJob response](#)

Contents

dataSourceId

The unique identifier of the ingested data source.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

ingestionJobId

The unique identifier of the ingestion job.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source is being added.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

startedAt

The time at which the ingestion job started.

Type: Timestamp

Required: Yes

status

The status of the ingestion job.

Type: String

Valid Values: STARTING | IN_PROGRESS | COMPLETE | FAILED

Required: Yes

updatedAt

The time at which the ingestion job was last updated.

Type: Timestamp

Required: Yes

description

The description of the ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the ingestion job failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

statistics

Contains statistics about the ingestion job.

Type: [IngestionJobStatistics](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

IngestionJobFilter

Service: Agents for Amazon Bedrock

Defines a filter by which to filter the results.

Contents

attribute

The attribute by which to filter the results.

Type: String

Valid Values: STATUS

Required: Yes

operator

The operation to carry out between the attribute and the values.

Type: String

Valid Values: EQ

Required: Yes

values

A list of values for the attribute.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 0. Maximum length of 100.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

IngestionJobSortBy

Service: Agents for Amazon Bedrock

Parameters by which to sort the results.

Contents

attribute

The attribute by which to sort the results.

Type: String

Valid Values: STATUS | STARTED_AT

Required: Yes

order

The order by which to sort the results.

Type: String

Valid Values: ASCENDING | DESCENDING

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

IngestionJobStatistics

Service: Agents for Amazon Bedrock

Contains the statistics for the ingestion job.

Contents

numberOfDocumentsDeleted

The number of source documents that was deleted.

Type: Long

Required: No

numberOfDocumentsFailed

The number of source documents that failed to be ingested.

Type: Long

Required: No

numberOfDocumentsScanned

The total number of source documents that were scanned. Includes new, updated, and unchanged documents.

Type: Long

Required: No

numberOfMetadataDocumentsModified

The number of metadata files that were updated or deleted.

Type: Long

Required: No

numberOfMetadataDocumentsScanned

The total number of metadata files that were scanned. Includes new, updated, and unchanged files.

Type: Long

Required: No

numberOfModifiedDocumentsIndexed

The number of modified source documents in the data source that were successfully indexed.

Type: Long

Required: No

numberOfNewDocumentsIndexed

The number of new source documents in the data source that were successfully indexed.

Type: Long

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

IngestionJobSummary

Service: Agents for Amazon Bedrock

Contains details about an ingestion job.

Contents

dataSourceId

The unique identifier of the data source in the ingestion job.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

ingestionJobId

The unique identifier of the ingestion job.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source is added.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

startedAt

The time at which the ingestion job was started.

Type: Timestamp

Required: Yes

status

The status of the ingestion job.

Type: String

Valid Values: STARTING | IN_PROGRESS | COMPLETE | FAILED

Required: Yes

updatedAt

The time at which the ingestion job was last updated.

Type: Timestamp

Required: Yes

description

The description of the ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

statistics

Contains statistics for the ingestion job.

Type: [IngestionJobStatistics](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBase

Service: Agents for Amazon Bedrock

Contains information about a knowledge base.

Contents

createdAt

The time at which the knowledge base was created.

Type: Timestamp

Required: Yes

knowledgeBaseArn

The Amazon Resource Name (ARN) of the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `^arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+$`

Required: Yes

knowledgeBaseConfiguration

Contains details about the embeddings configuration of the knowledge base.

Type: [KnowledgeBaseConfiguration](#) object

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

name

The name of the knowledge base.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^\:]+)?:iam::([0-9]{12})?:role/.*+$`

Required: Yes

status

The status of the knowledge base. The following statuses are possible:

- **CREATING** – The knowledge base is being created.
- **ACTIVE** – The knowledge base is ready to be queried.
- **DELETING** – The knowledge base is being deleted.
- **UPDATING** – The knowledge base is being updated.
- **FAILED** – The knowledge base API operation failed.

Type: String

Valid Values: **CREATING** | **ACTIVE** | **DELETING** | **UPDATING** | **FAILED** | **DELETE_UNSUCCESSFUL**

Required: Yes

storageConfiguration

Contains details about the storage configuration of the knowledge base.

Type: [StorageConfiguration](#) object

Required: Yes

updatedAt

The time at which the knowledge base was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the API operation on the knowledge base failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Contains details about the embeddings configuration of the knowledge base.

Contents

type

The type of data that the data source is converted into for the knowledge base.

Type: String

Valid Values: VECTOR

Required: Yes

vectorKnowledgeBaseConfiguration

Contains details about the embeddings model that's used to convert the data source.

Type: [VectorKnowledgeBaseConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseSummary

Service: Agents for Amazon Bedrock

Contains details about a knowledge base.

Contents

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Pattern: `^[0-9a-zA-Z]{10}$`

Required: Yes

name

The name of the knowledge base.

Type: String

Pattern: `^([0-9a-zA-Z][_]?){1,100}$`

Required: Yes

status

The status of the knowledge base.

Type: String

Valid Values: CREATING | ACTIVE | DELETING | UPDATING | FAILED | DELETE_UNSUCCESSFUL

Required: Yes

updatedAt

The time at which the knowledge base was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

MongoDbAtlasConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in MongoDB Atlas.

Contents

collectionName

The collection name of the knowledge base in MongoDB Atlas.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^.*$`

Required: Yes

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that contains user credentials for your MongoDB Atlas cluster.

Type: String

Pattern: `^arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}$`

Required: Yes

databaseName

The database name in your MongoDB Atlas cluster for your knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^.*$`

Required: Yes

endpoint

The endpoint URL of your MongoDB Atlas cluster for your knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: [MongoDbAtlasFieldMapping](#) object

Required: Yes

vectorIndexName

The name of the MongoDB Atlas vector search index.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

endpointServiceName

The name of the VPC endpoint service in your account that is connected to your MongoDB Atlas cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `^(?:arn:aws(?:-us-gov|-cn|-iso|-iso-[a-z])*:.+:.*:\d+:.+/.+|[a-zA-Z0-9*]+[a-zA-Z0-9._-]*)$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

MongoDbAtlasFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

OpenSearchServerlessConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Amazon OpenSearch Service. For more information, see [Create a vector index in Amazon OpenSearch Service](#).

Contents

collectionArn

The Amazon Resource Name (ARN) of the OpenSearch Service vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws:aoss:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:collection/[a-z0-9-]{3,32}$`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: [OpenSearchServerlessFieldMapping](#) object

Required: Yes

vectorIndexName

The name of the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

OpenSearchServerlessFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ParameterDetail

Service: Agents for Amazon Bedrock

Contains details about a parameter in a function for an action group.

This data type is used in the following API operations:

- [CreateAgentActionGroup request](#)
- [CreateAgentActionGroup response](#)
- [UpdateAgentActionGroup request](#)
- [UpdateAgentActionGroup response](#)
- [GetAgentActionGroup response](#)

Contents

type

The data type of the parameter.

Type: String

Valid Values: `string` | `number` | `integer` | `boolean` | `array`

Required: Yes

description

A description of the parameter. Helps the foundation model determine how to elicit the parameters from the user.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: No

required

Whether the parameter is required for the agent to complete the function for action group invocation.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PineconeConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Pinecone. For more information, see [Create a vector index in Pinecone](#).

Contents

connectionString

The endpoint URL for your index management page.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Pinecone API key.

Type: String

Pattern: `^arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}$`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: [PineconeFieldMapping](#) object

Required: Yes

namespace

The namespace to be used to write new data to your database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PineconeFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PromptConfiguration

Service: Agents for Amazon Bedrock

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see [Advanced prompts](#).

Contents

basePromptTemplate

Defines the prompt template with which to replace the default prompt template. You can use placeholder variables in the base prompt template to customize the prompt. For more information, see [Prompt template placeholder variables](#).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: No

inferenceConfiguration

Contains inference parameters to use when the agent invokes a foundation model in the part of the agent sequence defined by the `promptType`. For more information, see [Inference parameters for foundation models](#).

Type: [InferenceConfiguration](#) object

Required: No

parserMode

Specifies whether to override the default parser Lambda function when parsing the raw foundation model output in the part of the agent sequence defined by the `promptType`. If you set the field as `OVERRIDEN`, the `overrideLambda` field in the [PromptOverrideConfiguration](#) must be specified with the ARN of a Lambda function.

Type: String

Valid Values: `DEFAULT` | `OVERRIDEN`

Required: No

promptCreationMode

Specifies whether to override the default prompt template for this promptType. Set this value to `OVERRIDDEN` to use the prompt that you provide in the `basePromptTemplate`. If you leave it as `DEFAULT`, the agent uses a default prompt template.

Type: String

Valid Values: `DEFAULT` | `OVERRIDDEN`

Required: No

promptState

Specifies whether to allow the agent to carry out the step specified in the promptType. If you set this value to `DISABLED`, the agent skips that step. The default state for each promptType is as follows.

- `PRE_PROCESSING` – `ENABLED`
- `ORCHESTRATION` – `ENABLED`
- `KNOWLEDGE_BASE_RESPONSE_GENERATION` – `ENABLED`
- `POST_PROCESSING` – `DISABLED`

Type: String

Valid Values: `ENABLED` | `DISABLED`

Required: No

promptType

The step in the agent sequence that this prompt configuration applies to.

Type: String

Valid Values: `PRE_PROCESSING` | `ORCHESTRATION` | `POST_PROCESSING` | `KNOWLEDGE_BASE_RESPONSE_GENERATION`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PromptOverrideConfiguration

Service: Agents for Amazon Bedrock

Contains configurations to override prompts in different parts of an agent sequence. For more information, see [Advanced prompts](#).

Contents

promptConfigurations

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see [Advanced prompts](#).

Type: Array of [PromptConfiguration](#) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

overrideLambda

The ARN of the Lambda function to use when parsing the raw foundation model output in parts of the agent sequence. If you specify this field, at least one of the `promptConfigurations` must contain a `parseMode` value that is set to `OVERRIDDEN`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+\-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.\.]+(:(\$\{LATEST|[a-zA-Z0-9-_\.\.]+))?$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RdsConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Amazon RDS. For more information, see [Create a vector index in Amazon RDS](#).

Contents

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Amazon RDS database.

Type: String

Pattern: `^arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}$`

Required: Yes

databaseName

The name of your Amazon RDS database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\-]+$`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: [RdsFieldMapping](#) object

Required: Yes

resourceArn

The Amazon Resource Name (ARN) of the vector store.

Type: String

Pattern: `^arn:aws(|-cn|-us-gov):rds:[a-zA-Z0-9-]*:[0-9]{12}:cluster:[a-zA-Z0-9-]{1,63}$`

Required: Yes

tableName

The name of the table in the database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\. \-]+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RdsFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\-]+$`

Required: Yes

primaryKeyField

The name of the field in which Amazon Bedrock stores the ID for each entry.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\-]+$`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\-]+$`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `^[a-zA-Z0-9_\-\]+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RedisEnterpriseCloudConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Redis Enterprise Cloud. For more information, see [Create a vector index in Redis Enterprise Cloud](#).

Contents

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Redis Enterprise Cloud database.

Type: String

Pattern: `^arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}$`

Required: Yes

endpoint

The endpoint URL of the Redis Enterprise Cloud database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: [RedisEnterpriseCloudFieldMapping](#) object

Required: Yes

vectorIndexName

The name of the vector index.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RedisEnterpriseCloudFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^.*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

S3DataSourceConfiguration

Service: Agents for Amazon Bedrock

Contains information about the S3 configuration of the data source.

Contents

bucketArn

The Amazon Resource Name (ARN) of the bucket that contains the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):s3:::[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]$`

Required: Yes

bucketOwnerAccountId

The bucket account owner ID for the S3 bucket.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `^[0-9]{12}$`

Required: No

inclusionPrefixes

A list of S3 prefixes that define the object containing the data sources. For more information, see [Organizing objects using prefixes](#).

Type: Array of strings

Array Members: Fixed number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 300.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

S3Identifier

Service: Agents for Amazon Bedrock

Contains information about the S3 object containing the resource.

Contents

s3BucketName

The name of the S3 bucket.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Pattern: `^[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]$`

Required: No

s3ObjectKey

The S3 object key containing the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^[\\.\-\\!*_\\'\\(\\)a-zA-Z0-9][\\.\-\\!*_\\'\\(\\)\\/a-zA-Z0-9]*$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ServerSideEncryptionConfiguration

Service: Agents for Amazon Bedrock

Contains the configuration for server-side encryption.

Contents

kmsKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key used to encrypt the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

StorageConfiguration

Service: Agents for Amazon Bedrock

Contains the storage configuration of the knowledge base.

Contents

type

The vector store service in which the knowledge base is stored.

Type: String

Valid Values: OPENSEARCH_SERVERLESS | PINECONE | REDIS_ENTERPRISE_CLOUD | RDS | MONGO_DB_ATLAS

Required: Yes

mongoDbAtlasConfiguration

Contains the storage configuration of the knowledge base in MongoDB Atlas.

Type: [MongoDbAtlasConfiguration](#) object

Required: No

opensearchServerlessConfiguration

Contains the storage configuration of the knowledge base in Amazon OpenSearch Service.

Type: [OpenSearchServerlessConfiguration](#) object

Required: No

pineconeConfiguration

Contains the storage configuration of the knowledge base in Pinecone.

Type: [PineconeConfiguration](#) object

Required: No

rdsConfiguration

Contains details about the storage configuration of the knowledge base in Amazon RDS. For more information, see [Create a vector index in Amazon RDS](#).

Type: [RdsConfiguration](#) object

Required: No

redisEnterpriseCloudConfiguration

Contains the storage configuration of the knowledge base in Redis Enterprise Cloud.

Type: [RedisEnterpriseCloudConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ValidationExceptionField

Service: Agents for Amazon Bedrock

Stores information about a field passed inside a request that resulted in a validation error.

Contents

message

A message describing why this field failed validation.

Type: String

Pattern: `^[\s\S]+$`

Required: Yes

name

The name of the field.

Type: String

Pattern: `^[\s\S]+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

VectorIngestionConfiguration

Service: Agents for Amazon Bedrock

Contains details about how to ingest the documents in a data source.

Contents

chunkingConfiguration

Details about how to chunk the documents in the data source. A *chunk* refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried.

Type: [ChunkingConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

VectorKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Contains details about the model used to create vector embeddings for the knowledge base.

Contents

embeddingModelArn

The Amazon Resource Name (ARN) of the model used to create vector embeddings for the knowledge base.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Agents for Amazon Bedrock Runtime

The following data types are supported by Agents for Amazon Bedrock Runtime:

- [ActionGroupInvocationInput](#)
- [ActionGroupInvocationOutput](#)
- [ApiInvocationInput](#)
- [ApiParameter](#)

- [ApiRequestBody](#)
- [ApiResponse](#)
- [Attribution](#)
- [ByteContentDoc](#)
- [Citation](#)
- [ContentBody](#)
- [ExternalSource](#)
- [ExternalSourcesGenerationConfiguration](#)
- [ExternalSourcesRetrieveAndGenerateConfiguration](#)
- [FailureTrace](#)
- [FilterAttribute](#)
- [FinalResponse](#)
- [FunctionInvocationInput](#)
- [FunctionParameter](#)
- [FunctionResult](#)
- [GeneratedResponsePart](#)
- [GenerationConfiguration](#)
- [GuardrailConfiguration](#)
- [InferenceConfig](#)
- [InferenceConfiguration](#)
- [InvocationInput](#)
- [InvocationInputMember](#)
- [InvocationResultMember](#)
- [KnowledgeBaseLookupInput](#)
- [KnowledgeBaseLookupOutput](#)
- [KnowledgeBaseQuery](#)
- [KnowledgeBaseRetrievalConfiguration](#)
- [KnowledgeBaseRetrievalResult](#)
- [KnowledgeBaseRetrieveAndGenerateConfiguration](#)
- [KnowledgeBaseVectorSearchConfiguration](#)

- [ModelInvocationInput](#)
- [Observation](#)
- [OrchestrationTrace](#)
- [Parameter](#)
- [PayloadPart](#)
- [PostProcessingModelInvocationOutput](#)
- [PostProcessingParsedResponse](#)
- [PostProcessingTrace](#)
- [PreProcessingModelInvocationOutput](#)
- [PreProcessingParsedResponse](#)
- [PreProcessingTrace](#)
- [PromptTemplate](#)
- [PropertyParameters](#)
- [Rationale](#)
- [RepromptResponse](#)
- [RequestBody](#)
- [ResponseStream](#)
- [RetrievalFilter](#)
- [RetrievalResultContent](#)
- [RetrievalResultLocation](#)
- [RetrievalResultS3Location](#)
- [RetrieveAndGenerateConfiguration](#)
- [RetrieveAndGenerateInput](#)
- [RetrieveAndGenerateOutput](#)
- [RetrieveAndGenerateSessionConfiguration](#)
- [RetrievedReference](#)
- [ReturnControlPayload](#)
- [S3ObjectDoc](#)
- [SessionState](#)
- [Span](#)

- [TextInferenceConfig](#)
- [TextResponsePart](#)
- [Trace](#)
- [TracePart](#)

ActionGroupInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the action group being invoked. For more information about the possible structures, see the InvocationInput tab in [OrchestrationTrace](#) in the Amazon Bedrock User Guide.

Contents

actionGroupName

The name of the action group.

Type: String

Required: No

apiPath

The path to the API to call, based off the action group.

Type: String

Required: No

function

The function in the action group to call.

Type: String

Required: No

parameters

The parameters in the Lambda input event.

Type: Array of [Parameter](#) objects

Required: No

requestBody

The parameters in the request body for the Lambda input event.

Type: [RequestBody](#) object

Required: No

verb

The API method being used, based off the action group.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ActionGroupInvocationOutput

Service: Agents for Amazon Bedrock Runtime

Contains the JSON-formatted string returned by the API invoked by the action group.

Contents

text

The JSON-formatted string returned by the API invoked by the action group.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ApiInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the API operation that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the [InvokeAgent response](#)

Contents

actionGroup

The action group that the API operation belongs to.

Type: String

Required: Yes

apiPath

The path to the API operation.

Type: String

Required: No

httpMethod

The HTTP method of the API operation.

Type: String

Required: No

parameters

The parameters to provide for the API request, as the agent elicited from the user.

Type: Array of [ApiParameter](#) objects

Required: No

requestBody

The request body to provide for the API request, as the agent elicited from the user.

Type: [ApiRequestBody](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ApiParameter

Service: Agents for Amazon Bedrock Runtime

Information about a parameter to provide to the API request.

This data type is used in the following API operations:

- [InvokeAgent response](#)

Contents

name

The name of the parameter.

Type: String

Required: No

type

The data type for the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

ApiRequestBody

Service: Agents for Amazon Bedrock Runtime

The request body to provide for the API request, as the agent elicited from the user.

This data type is used in the following API operations:

- [InvokeAgent response](#)

Contents

content

The content of the request body. The key of the object in this field is a media type defining the format of the request body.

Type: String to [PropertyParameters](#) object map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ApiResponse

Service: Agents for Amazon Bedrock Runtime

Contains information about the API operation that was called from the action group and the response body that was returned.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` of the [InvokeAgent request](#)

Contents

actionGroup

The action group that the API operation belongs to.

Type: String

Required: Yes

apiPath

The path to the API operation.

Type: String

Required: No

httpMethod

The HTTP method for the API operation.

Type: String

Required: No

httpStatusCode

http status code from API execution response (for example: 200, 400, 500).

Type: Integer

Required: No

responseBody

The response body from the API operation. The key of the object is the content type (currently, only TEXT is supported). The response may be returned directly or from the Lambda function.

Type: String to [ContentBody](#) object map

Required: No

responseState

Controls the final response state returned to end user when API/Function execution failed. When this state is FAILURE, the request would fail with dependency failure exception. When this state is REPROMPT, the API/function response will be sent to model for re-prompt

Type: String

Valid Values: FAILURE | REPROMPT

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Attribution

Service: Agents for Amazon Bedrock Runtime

Contains citations for a part of an agent response.

Contents

citations

A list of citations and related information for a part of an agent response.

Type: Array of [Citation](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ByteContentDoc

Service: Agents for Amazon Bedrock Runtime

This property contains the document to chat with, along with its attributes.

Contents

contentType

The MIME type of the document contained in the wrapper object.

Type: String

Pattern: `[a-z]{1,20}/.{1,20}`

Required: Yes

data

The byte value of the file to upload, encoded as a Base-64 string.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 10485760.

Required: Yes

identifier

The file name of the document contained in the wrapper object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

Citation

Service: Agents for Amazon Bedrock Runtime

An object containing a segment of the generated response that is based on a source in the knowledge base, alongside information about the source.

This data type is used in the following API operations:

- [InvokeAgent response](#) – in the citations field
- [RetrieveAndGenerate response](#) – in the citations field

Contents

generatedResponsePart

Contains the generated response and metadata

Type: [GeneratedResponsePart](#) object

Required: No

retrievedReferences

Contains metadata about the sources cited for the generated response.

Type: Array of [RetrievedReference](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ContentBody

Service: Agents for Amazon Bedrock Runtime

Contains the body of the API response.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` field of the [InvokeAgent request](#)

Contents

body

The body of the API response.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ExternalSource

Service: Agents for Amazon Bedrock Runtime

The unique external source of the content contained in the wrapper object.

Contents

sourceType

The source type of the external source wrapper object.

Type: String

Valid Values: S3 | BYTE_CONTENT

Required: Yes

byteContent

The identifier, contentType, and data of the external source wrapper object.

Type: [ByteContentDoc](#) object

Required: No

s3Location

The S3 location of the external source wrapper object.

Type: [S3ObjectDoc](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ExternalSourcesGenerationConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains the generation configuration of the external source wrapper object.

Contents

additionalModelRequestFields

Additional model parameters and their corresponding values not included in the `textInferenceConfig` structure for an external source. Takes in custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

The configuration details for the guardrail.

Type: [GuardrailConfiguration](#) object

Required: No

inferenceConfig

Configuration settings for inference when using `RetrieveAndGenerate` to generate responses while using an external source.

Type: [InferenceConfig](#) object

Required: No

promptTemplate

Contain the `textPromptTemplate` string for the external source wrapper object.

Type: [PromptTemplate](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ExternalSourcesRetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

The configurations of the external source wrapper object in the retrieveAndGenerate function.

Contents

modelArn

The modelArn used with the external source wrapper object in the retrieveAndGenerate function.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))$`

Required: Yes

sources

The document used with the external source wrapper object in the retrieveAndGenerate function.

Type: Array of [ExternalSource](#) objects

Array Members: Fixed number of 1 item.

Required: Yes

generationConfiguration

The prompt used with the external source wrapper object with the retrieveAndGenerate function.

Type: [ExternalSourcesGenerationConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FailureTrace

Service: Agents for Amazon Bedrock Runtime

Contains information about the failure of the interaction.

Contents

failureReason

The reason the interaction failed.

Type: String

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FilterAttribute

Service: Agents for Amazon Bedrock Runtime

Specifies the name that the metadata attribute must match and the value to which to compare the value of the metadata attribute. For more information, see [Query configurations](#).

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#)

Contents

key

The name that the metadata attribute must match.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

value

The value to which to compare the value of the metadata attribute.

Type: JSON value

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FinalResponse

Service: Agents for Amazon Bedrock Runtime

Contains details about the response to the user.

Contents

text

The text in the response to the user.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

FunctionInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the function that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the [InvokeAgent response](#)

Contents

actionGroup

The action group that the function belongs to.

Type: String

Required: Yes

function

The name of the function.

Type: String

Required: No

parameters

A list of parameters of the function.

Type: Array of [FunctionParameter](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

FunctionParameter

Service: Agents for Amazon Bedrock Runtime

Contains information about a parameter of the function.

This data type is used in the following API operations:

- In the `returnControl` field of the [InvokeAgent response](#)

Contents

name

The name of the parameter.

Type: String

Required: No

type

The data type of the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

FunctionResult

Service: Agents for Amazon Bedrock Runtime

Contains information about the function that was called from the action group and the response that was returned.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` of the [InvokeAgent request](#)

Contents

actionGroup

The action group that the function belongs to.

Type: String

Required: Yes

function

The name of the function that was called.

Type: String

Required: No

responseBody

The response from the function call using the parameters. The key of the object is the content type (currently, only TEXT is supported). The response may be returned directly or from the Lambda function.

Type: String to [ContentBody](#) object map

Required: No

responseState

Controls the final response state returned to end user when API/Function execution failed. When this state is FAILURE, the request would fail with dependency failure exception. When this state is REPROMPT, the API/function response will be sent to model for re-prompt

Type: String

Valid Values: FAILURE | REPROMPT

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GeneratedResponsePart

Service: Agents for Amazon Bedrock Runtime

Contains metadata about a part of the generated response that is accompanied by a citation.

This data type is used in the following API operations:

- [InvokeAgent response](#) – in the generatedResponsePart field
- [RetrieveAndGenerate response](#) – in the generatedResponsePart field

Contents

textResponsePart

Contains metadata about a textual part of the generated response that is accompanied by a citation.

Type: [TextResponsePart](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GenerationConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for response generation based on the knowledge base query results.

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#)

Contents

additionalModelRequestFields

Additional model parameters and corresponding values not included in the `textInferenceConfig` structure for a knowledge base. This allows users to provide custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

The configuration details for the guardrail.

Type: [GuardrailConfiguration](#) object

Required: No

inferenceConfig

Configuration settings for inference when using `RetrieveAndGenerate` to generate responses while using a knowledge base as a source.

Type: [InferenceConfig](#) object

Required: No

promptTemplate

Contains the template for the prompt that's sent to the model for response generation.

Type: [PromptTemplate](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

GuardrailConfiguration

Service: Agents for Amazon Bedrock Runtime

The configuration details for the guardrail.

Contents

guardrailId

The unique identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `^[a-z0-9]+$`

Required: Yes

guardrailVersion

The version of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(([1-9][0-9]{0,7})|(DRAFT))$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InferenceConfig

Service: Agents for Amazon Bedrock Runtime

The configuration for inference settings when generating responses using RetrieveAndGenerate.

Contents

textInferenceConfig

Configuration settings specific to text generation while generating responses using RetrieveAndGenerate.

Type: [TextInferenceConfig](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InferenceConfiguration

Service: Agents for Amazon Bedrock Runtime

Specifications about the inference parameters that were provided alongside the prompt. These are specified in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated. For more information, see [Inference parameters for foundation models](#).

Contents

maximumLength

The maximum number of tokens allowed in the generated response.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 4096.

Required: No

stopSequences

A list of stop sequences. A stop sequence is a sequence of characters that causes the model to stop generating the response.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

temperature

The likelihood of the model selecting higher-probability options while generating a response. A lower value makes the model more likely to choose higher-probability options, while a higher value makes the model more likely to choose lower-probability options.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topK

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for topK is the number of most-likely

candidates from which the model chooses the next token in the sequence. For example, if you set `topK` to 50, the model selects the next token from among the top 50 most likely choices.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 500.

Required: No

topP

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `Top P` determines the number of most-likely candidates from which the model chooses the next token in the sequence. For example, if you set `topP` to 80, the model only selects the next token from the top 80% of the probability distribution of next tokens.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information pertaining to the action group or knowledge base that is being invoked.

Contents

actionGroupInvocationInput

Contains information about the action group to be invoked.

Type: [ActionGroupInvocationInput](#) object

Required: No

invocationType

Specifies whether the agent is invoking an action group or a knowledge base.

Type: String

Valid Values: ACTION_GROUP | KNOWLEDGE_BASE | FINISH

Required: No

knowledgeBaseLookupInput

Contains details about the knowledge base to look up and the query to be made.

Type: [KnowledgeBaseLookupInput](#) object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InvocationInputMember

Service: Agents for Amazon Bedrock Runtime

Contains details about the API operation or function that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the [InvokeAgent response](#)

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

apiInvocationInput

Contains information about the API operation that the agent predicts should be called.

Type: [ApiInvocationInput](#) object

Required: No

functionInvocationInput

Contains information about the function that the agent predicts should be called.

Type: [FunctionInvocationInput](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

InvocationResultMember

Service: Agents for Amazon Bedrock Runtime

A result from the invocation of an action. For more information, see [Return control to the agent developer](#) and [Control session context](#).

This data type is used in the following API operations:

- [InvokeAgent request](#)

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

apiResult

The result from the API response from the action group invocation.

Type: [ApiResult](#) object

Required: No

functionResult

The result from the function from the action group invocation.

Type: [FunctionResult](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

KnowledgeBaseLookupInput

Service: Agents for Amazon Bedrock Runtime

Contains details about the knowledge base to look up and the query to be made.

Contents

knowledgeBaseId

The unique identifier of the knowledge base to look up.

Type: String

Required: No

text

The query made to the knowledge base.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseLookupOutput

Service: Agents for Amazon Bedrock Runtime

Contains details about the results from looking up the knowledge base.

Contents

retrievedReferences

Contains metadata about the sources cited for the generated response.

Type: Array of [RetrievedReference](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseQuery

Service: Agents for Amazon Bedrock Runtime

Contains the query made to the knowledge base.

This data type is used in the following API operations:

- [Retrieve request](#) – in the `retrievalQuery` field

Contents

text

The text of the query made to the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseRetrievalConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for the knowledge base query and retrieval process. For more information, see [Query configurations](#).

This data type is used in the following API operations:

- [Retrieve request](#) – in the `retrievalConfiguration` field
- [RetrieveAndGenerate request](#) – in the `retrievalConfiguration` field

Contents

vectorSearchConfiguration

Contains details about how the results from the vector search should be returned. For more information, see [Query configurations](#).

Type: [KnowledgeBaseVectorSearchConfiguration](#) object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseRetrievalResult

Service: Agents for Amazon Bedrock Runtime

Details about a result from querying the knowledge base.

This data type is used in the following API operations:

- [Retrieve response](#) – in the `retrievalResults` field

Contents

content

Contains a chunk of text from a data source in the knowledge base.

Type: [RetrievalResultContent](#) object

Required: Yes

location

Contains information about the location of the data source.

Type: [RetrievalResultLocation](#) object

Required: No

metadata

Contains metadata attributes and their values for the file in the data source. For more information, see [Metadata and filtering](#).

Type: String to JSON value map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

score

The level of relevance of the result to the query.

Type: Double

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseRetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains details about the resource being queried.

This data type is used in the following API operations:

- [Retrieve request](#) – in the `knowledgeBaseConfiguration` field
- [RetrieveAndGenerate request](#) – in the `knowledgeBaseConfiguration` field

Contents

`knowledgeBaseId`

The unique identifier of the knowledge base that is queried and the foundation model used for generation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `^[0-9a-zA-Z]+$`

Required: Yes

`modelArn`

The ARN of the foundation model used to generate a response.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^\:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))$`

Required: Yes

`generationConfiguration`

Contains configurations for response generation based on the knowledge base query results.

Type: [GenerationConfiguration](#) object

Required: No

retrievalConfiguration

Contains configurations for how to retrieve and return the knowledge base query.

Type: [KnowledgeBaseRetrievalConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

KnowledgeBaseVectorSearchConfiguration

Service: Agents for Amazon Bedrock Runtime

Configurations for how to perform the search query and return results. For more information, see [Query configurations](#).

This data type is used in the following API operations:

- [Retrieve request](#) – in the `vectorSearchConfiguration` field
- [RetrieveAndGenerate request](#) – in the `vectorSearchConfiguration` field

Contents

filter

Specifies the filters to use on the metadata in the knowledge base data sources before returning results. For more information, see [Query configurations](#).

Type: [RetrievalFilter](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

numberOfResults

The number of source chunks to retrieve.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

overrideSearchType

By default, Amazon Bedrock decides a search strategy for you. If you're using an Amazon OpenSearch Serverless vector store that contains a filterable text field, you can specify whether to query the knowledge base with a HYBRID search using both vector embeddings and raw text, or SEMANTIC search using only vector embeddings. For other vector store configurations, only SEMANTIC search is available. For more information, see [Test a knowledge base](#).

Type: String

Valid Values: HYBRID | SEMANTIC

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ModelInvocationInput

Service: Agents for Amazon Bedrock Runtime

The input for the pre-processing step.

- The type matches the agent step.
- The text contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated.

Contents

`inferenceConfiguration`

Specifications about the inference parameters that were provided alongside the prompt. These are specified in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated. For more information, see [Inference parameters for foundation models](#).

Type: [InferenceConfiguration](#) object

Required: No

`overrideLambda`

The ARN of the Lambda function to use when parsing the raw foundation model output in parts of the agent sequence.

Type: String

Required: No

`parserMode`

Specifies whether to override the default parser Lambda function when parsing the raw foundation model output in the part of the agent sequence defined by the `promptType`.

Type: String

Valid Values: DEFAULT | OVERRIDDEN

Required: No

promptCreationMode

Specifies whether the default prompt template was OVERRIDDEN. If it was, the `basePromptTemplate` that was set in the [PromptOverrideConfiguration](#) object when the agent was created or updated is used instead.

Type: String

Valid Values: DEFAULT | OVERRIDDEN

Required: No

text

The text that prompted the agent at this step.

Type: String

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

type

The step in the agent sequence.

Type: String

Valid Values: PRE_PROCESSING | ORCHESTRATION |
KNOWLEDGE_BASE_RESPONSE_GENERATION | POST_PROCESSING

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Observation

Service: Agents for Amazon Bedrock Runtime

Contains the result or output of an action group or knowledge base, or the response to the user.

Contents

actionGroupInvocationOutput

Contains the JSON-formatted string returned by the API invoked by the action group.

Type: [ActionGroupInvocationOutput](#) object

Required: No

finalResponse

Contains details about the response to the user.

Type: [FinalResponse](#) object

Required: No

knowledgeBaseLookupOutput

Contains details about the results from looking up the knowledge base.

Type: [KnowledgeBaseLookupOutput](#) object

Required: No

repromptResponse

Contains details about the response to reprompt the input.

Type: [RepromptResponse](#) object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

type

Specifies what kind of information the agent returns in the observation. The following values are possible.

- ACTION_GROUP – The agent returns the result of an action group.
- KNOWLEDGE_BASE – The agent returns information from a knowledge base.
- FINISH – The agent returns a final response to the user with no follow-up.
- ASK_USER – The agent asks the user a question.
- REPROMPT – The agent prompts the user again for the same information.

Type: String

Valid Values: ACTION_GROUP | KNOWLEDGE_BASE | FINISH | ASK_USER | REPROMPT

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

OrchestrationTrace

Service: Agents for Amazon Bedrock Runtime

Details about the orchestration step, in which the agent determines the order in which actions are executed and which knowledge bases are retrieved.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

invocationInput

Contains information pertaining to the action group or knowledge base that is being invoked.

Type: [InvocationInput](#) object

Required: No

modelInvocationInput

The input for the orchestration step.

- The type is ORCHESTRATION.
- The text contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated.

Type: [ModelInvocationInput](#) object

Required: No

observation

Details about the observation (the output of the action group Lambda or knowledge base) made by the agent.

Type: [Observation](#) object

Required: No

rationale

Details about the reasoning, based on the input, that the agent uses to justify carrying out an action group or getting information from a knowledge base.

Type: [Rationale](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Parameter

Service: Agents for Amazon Bedrock Runtime

A parameter for the API request or function.

Contents

name

The name of the parameter.

Type: String

Required: No

type

The type of the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PayloadPart

Service: Agents for Amazon Bedrock Runtime

Contains a part of an agent response and citations for it.

Contents

attribution

Contains citations for a part of an agent response.

Type: [Attribution](#) object

Required: No

bytes

A part of the agent response in bytes.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PostProcessingModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

The foundation model output from the post-processing step.

Contents

parsedResponse

Details about the response from the Lambda parsing of the output of the post-processing step.

Type: [PostProcessingParsedResponse](#) object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PostProcessingParsedResponse

Service: Agents for Amazon Bedrock Runtime

Details about the response from the Lambda parsing of the output from the post-processing step.

Contents

text

The text returned by the parser.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PostProcessingTrace

Service: Agents for Amazon Bedrock Runtime

Details about the post-processing step, in which the agent shapes the response.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

modelInvocationInput

The input for the post-processing step.

- The type is POST_PROCESSING.
- The text contains the prompt.
- The inferenceConfiguration, parserMode, and overrideLambda values are set in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated.

Type: [ModelInvocationInput](#) object

Required: No

modelInvocationOutput

The foundation model output from the post-processing step.

Type: [PostProcessingModelInvocationOutput](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)

- [AWS SDK for Ruby V3](#)

PreProcessingModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

The foundation model output from the pre-processing step.

Contents

parsedResponse

Details about the response from the Lambda parsing of the output of the pre-processing step.

Type: [PreProcessingParsedResponse](#) object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PreProcessingParsedResponse

Service: Agents for Amazon Bedrock Runtime

Details about the response from the Lambda parsing of the output from the pre-processing step.

Contents

isValid

Whether the user input is valid or not. If `false`, the agent doesn't proceed to orchestration.

Type: Boolean

Required: No

rationale

The text returned by the parsing of the pre-processing step, explaining the steps that the agent plans to take in orchestration, if the user input is valid.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PreProcessingTrace

Service: Agents for Amazon Bedrock Runtime

Details about the pre-processing step, in which the agent contextualizes and categorizes user inputs.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

modelInvocationInput

The input for the pre-processing step.

- The type is PRE_PROCESSING.
- The text contains the prompt.
- The inferenceConfiguration, parserMode, and overrideLambda values are set in the [PromptOverrideConfiguration](#) object that was set when the agent was created or updated.

Type: [ModelInvocationInput](#) object

Required: No

modelInvocationOutput

The foundation model output from the pre-processing step.

Type: [PreProcessingModelInvocationOutput](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)

- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PromptTemplate

Service: Agents for Amazon Bedrock Runtime

Contains the template for the prompt that's sent to the model for response generation. For more information, see [Knowledge base prompt templates](#).

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#) – in the `filter` field

Contents

textPromptTemplate

The template for the prompt that's sent to the model for response generation. You can include prompt placeholders, which become replaced before the prompt is sent to the model to provide instructions and context to the model. In addition, you can include XML tags to delineate meaningful sections of the prompt template.

For more information, see the following resources:

- [Knowledge base prompt templates](#)
- [Use XML tags with Anthropic Claude models](#)

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

PropertyParameters

Service: Agents for Amazon Bedrock Runtime

Contains the parameters in the request body.

Contents

properties

A list of parameters in the request body.

Type: Array of [Parameter](#) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Rationale

Service: Agents for Amazon Bedrock Runtime

Contains the reasoning, based on the input, that the agent uses to justify carrying out an action group or getting information from a knowledge base.

Contents

text

The reasoning or thought process of the agent, based on the input.

Type: String

Required: No

traceId

The unique identifier of the trace step.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RepromptResponse

Service: Agents for Amazon Bedrock Runtime

Contains details about the agent's response to reprompt the input.

Contents

source

Specifies what output is prompting the agent to reprompt the input.

Type: String

Valid Values: ACTION_GROUP | KNOWLEDGE_BASE | PARSER

Required: No

text

The text reprompting the input.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RequestBody

Service: Agents for Amazon Bedrock Runtime

The parameters in the API request body.

Contents

content

The content in the request body.

Type: String to array of [Parameter](#) objects map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ResponseStream

Service: Agents for Amazon Bedrock Runtime

The response from invoking the agent and associated citations and trace information.

Contents

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

Required: No

chunk

Contains a part of an agent response and citations for it.

Type: [PayloadPart](#) object

Required: No

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception
HTTP Status Code: 424

Required: No

internalServerErrorException

An internal server error occurred. Retry your request.

Type: Exception
HTTP Status Code: 500

Required: No

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception
HTTP Status Code: 404

Required: No

returnControl

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: [ReturnControlPayload](#) object

Required: No

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception
HTTP Status Code: 400

Required: No

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception
HTTP Status Code: 429

Required: No

trace

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see [Trace events](#).

Type: [TracePart](#) object

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception
HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrievalFilter

Service: Agents for Amazon Bedrock Runtime

Specifies the filters to use on the metadata attributes in the knowledge base data sources before returning results. For more information, see [Query configurations](#).

This data type is used in the following API operations:

- [Retrieve request](#) – in the `filter` field
- [RetrieveAndGenerate request](#) – in the `filter` field

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

andAll

Knowledge base data sources whose metadata attributes fulfill all the filter conditions inside this list are returned.

Type: Array of [RetrievalFilter](#) objects

Array Members: Minimum number of 2 items. Maximum number of 5 items.

Required: No

equals

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value matches the `value` in this object are returned.

Type: [FilterAttribute](#) object

Required: No

greaterThan

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value is greater than the `value` in this object are returned.

Type: [FilterAttribute](#) object

Required: No

greaterThanOrEqualTo

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value is greater than or equal to the value in this object are returned.

Type: [FilterAttribute](#) object

Required: No

in

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value is in the list specified in the value in this object are returned.

Type: [FilterAttribute](#) object

Required: No

lessThan

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value is less than the value in this object are returned.

Type: [FilterAttribute](#) object

Required: No

lessThanOrEqualTo

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value is less than or equal to the value in this object are returned.

Type: [FilterAttribute](#) object

Required: No

notEquals

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value doesn't match the value in this object are returned.

Type: [FilterAttribute](#) object

Required: No

notIn

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value isn't in the list specified in the `value` in this object are returned.

Type: [FilterAttribute](#) object

Required: No

orAll

Knowledge base data sources whose metadata attributes fulfill at least one of the filter conditions inside this list are returned.

Type: Array of [RetrievalFilter](#) objects

Array Members: Minimum number of 2 items. Maximum number of 5 items.

Required: No

startsWith

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value starts with the `value` in this object are returned. This filter is currently only supported for Amazon OpenSearch Serverless vector stores.

Type: [FilterAttribute](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrievalResultContent

Service: Agents for Amazon Bedrock Runtime

Contains the cited text from the data source.

This data type is used in the following API operations:

- [Retrieve response](#) – in the content field
- [RetrieveAndGenerate response](#) – in the content field
- [InvokeAgent response](#) – in the content field

Contents

text

The cited text from the data source.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrievalResultLocation

Service: Agents for Amazon Bedrock Runtime

Contains information about the location of the data source.

This data type is used in the following API operations:

- [Retrieve response](#) – in the `location` field
- [RetrieveAndGenerate response](#) – in the `location` field
- [InvokeAgent response](#) – in the `location` field

Contents

type

The type of the location of the data source.

Type: String

Valid Values: S3

Required: Yes

s3Location

Contains the S3 location of the data source.

Type: [RetrievalResultS3Location](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrievalResultS3Location

Service: Agents for Amazon Bedrock Runtime

Contains the S3 location of the data source.

This data type is used in the following API operations:

- [Retrieve response](#) – in the `s3Location` field
- [RetrieveAndGenerate response](#) – in the `s3Location` field
- [InvokeAgent response](#) – in the `s3Location` field

Contents

uri

The S3 URI of the data source.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains details about the resource being queried.

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#) – in the `retrieveAndGenerateConfiguration` field

Contents

type

The type of resource that is queried by the request.

Type: String

Valid Values: KNOWLEDGE_BASE | EXTERNAL_SOURCES

Required: Yes

externalSourcesConfiguration

The configuration used with the external source wrapper object in the `retrieveAndGenerate` function.

Type: [ExternalSourcesRetrieveAndGenerateConfiguration](#) object

Required: No

knowledgeBaseConfiguration

Contains details about the resource being queried.

Type: [KnowledgeBaseRetrieveAndGenerateConfiguration](#) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrieveAndGenerateInput

Service: Agents for Amazon Bedrock Runtime

Contains the query made to the knowledge base.

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#) – in the `input` field

Contents

text

The query made to the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrieveAndGenerateOutput

Service: Agents for Amazon Bedrock Runtime

Contains the response generated from querying the knowledge base.

This data type is used in the following API operations:

- [RetrieveAndGenerate response](#) – in the output field

Contents

text

The response generated from querying the knowledge base.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrieveAndGenerateSessionConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configuration about the session with the knowledge base.

This data type is used in the following API operations:

- [RetrieveAndGenerate request](#) – in the sessionConfiguration field

Contents

kmsKeyArn

The ARN of the AWS KMS key encrypting the session.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

RetrievedReference

Service: Agents for Amazon Bedrock Runtime

Contains metadata about a source cited for the generated response.

This data type is used in the following API operations:

- [RetrieveAndGenerate response](#) – in the `retrievedReferences` field
- [InvokeAgent response](#) – in the `retrievedReferences` field

Contents

content

Contains the cited text from the data source.

Type: [RetrievalResultContent](#) object

Required: No

location

Contains information about the location of the data source.

Type: [RetrievalResultLocation](#) object

Required: No

metadata

Contains metadata attributes and their values for the file in the data source. For more information, see [Metadata and filtering](#).

Type: String to JSON value map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ReturnControlPayload

Service: Agents for Amazon Bedrock Runtime

Contains information to return from the action group that the agent has predicted to invoke.

This data type is used in the following API operations:

- [InvokeAgent response](#)

Contents

invocationId

The identifier of the action group invocation.

Type: String

Required: No

invocationInputs

A list of objects that contain information about the parameters and inputs that need to be sent into the API operation or function, based on what the agent determines from its session with the user.

Type: Array of [InvocationInputMember](#) objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

S3ObjectDoc

Service: Agents for Amazon Bedrock Runtime

The unique wrapper object of the document from the S3 location.

Contents

uri

The file location of the S3 wrapper object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

SessionState

Service: Agents for Amazon Bedrock Runtime

Contains parameters that specify various attributes that persist across a session or prompt. You can define session state attributes as key-value pairs when writing a [Lambda function](#) for an action group or pass them when making an [InvokeAgent](#) request. Use session state attributes to control and provide conversational context for your agent and to help customize your agent's behavior. For more information, see [Control session context](#).

Contents

invocationId

The identifier of the invocation of an action. This value must match the `invocationId` returned in the `InvokeAgent` response for the action whose results are provided in the `returnControlInvocationResults` field. For more information, see [Return control to the agent developer](#) and [Control session context](#).

Type: String

Required: No

promptSessionAttributes

Contains attributes that persist across a prompt and the values of those attributes. These attributes replace the `$prompt_session_attributes` placeholder variable in the orchestration prompt template. For more information, see [Prompt template placeholder variables](#).

Type: String to string map

Required: No

returnControlInvocationResults

Contains information about the results from the action group invocation. For more information, see [Return control to the agent developer](#) and [Control session context](#).

Note

If you include this field, the `inputText` field will be ignored.

Type: Array of [InvocationResultMember](#) objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

sessionAttributes

Contains attributes that persist across a session and the values of those attributes.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Span

Service: Agents for Amazon Bedrock Runtime

Contains information about where the text with a citation begins and ends in the generated output.

This data type is used in the following API operations:

- [RetrieveAndGenerate response](#) – in the span field
- [InvokeAgent response](#) – in the span field

Contents

end

Where the text with a citation ends in the generated output.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

Where the text with a citation starts in the generated output.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

TextInferenceConfig

Service: Agents for Amazon Bedrock Runtime

Configuration settings for text generation using a language model via the RetrieveAndGenerate operation. Includes parameters like temperature, top-p, maximum token count, and stop sequences.

Note

The valid range of maxTokens depends on the accepted values for your chosen model's inference parameters. To see the inference parameters for your model, see [Inference parameters for foundation models](#).

Contents

maxTokens

The maximum number of tokens to generate in the output text. Do not use the minimum of 0 or the maximum of 65536. The limit values described here are arbitrary values, for actual values consult the limits defined by your specific model.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 65536.

Required: No

stopSequences

A list of sequences of characters that, if generated, will cause the model to stop generating further tokens. Do not use a minimum length of 1 or a maximum length of 1000. The limit values described here are arbitrary values, for actual values consult the limits defined by your specific model.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

temperature

Controls the random-ness of text generated by the language model, influencing how much the model sticks to the most predictable next words versus exploring more surprising options. A lower temperature value (e.g. 0.2 or 0.3) makes model outputs more deterministic or predictable, while a higher temperature (e.g. 0.8 or 0.9) makes the outputs more creative or unpredictable.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

A probability distribution threshold which controls what the model considers for the set of possible next tokens. The model will only consider the top p% of the probability distribution when generating the next token.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

TextResponsePart

Service: Agents for Amazon Bedrock Runtime

Contains the part of the generated text that contains a citation, alongside where it begins and ends.

This data type is used in the following API operations:

- [RetrieveAndGenerate response](#) – in the textResponsePart field
- [InvokeAgent response](#) – in the textResponsePart field

Contents

span

Contains information about where the text with a citation begins and ends in the generated output.

Type: [Span](#) object

Required: No

text

The part of the generated text that contains a citation.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Trace

Service: Agents for Amazon Bedrock Runtime

Contains one part of the agent's reasoning process and results from calling API actions and querying knowledge bases. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see [Trace enablement](#).

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

failureTrace

Contains information about the failure of the interaction.

Type: [FailureTrace](#) object

Required: No

orchestrationTrace

Details about the orchestration step, in which the agent determines the order in which actions are executed and which knowledge bases are retrieved.

Type: [OrchestrationTrace](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

postProcessingTrace

Details about the post-processing step, in which the agent shapes the response..

Type: [PostProcessingTrace](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

preProcessingTrace

Details about the pre-processing step, in which the agent contextualizes and categorizes user inputs.

Type: [PreProcessingTrace](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

TracePart

Service: Agents for Amazon Bedrock Runtime

Contains information about the agent and session, alongside the agent's reasoning process and results from calling API actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see [Trace enablement](#).

Contents

agentAliasId

The unique identifier of the alias of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `^[0-9a-zA-Z]+$`

Required: No

agentId

The unique identifier of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `^[0-9a-zA-Z]+$`

Required: No

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `^(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})$`

Required: No

sessionId

The unique identifier of the session with the agent.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `^[0-9a-zA-Z._:-]+$`

Required: No

trace

Contains one part of the agent's reasoning process and results from calling API actions and querying knowledge bases. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see [Trace enablement](#).

Type: [Trace](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Amazon Bedrock Runtime

The following data types are supported by Amazon Bedrock Runtime:

- [PayloadPart](#)
- [ResponseStream](#)

PayloadPart

Service: Amazon Bedrock Runtime

Payload content included in the response.

Contents

bytes

Base64-encoded bytes of payload data.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

ResponseStream

Service: Amazon Bedrock Runtime

Definition of content in the response stream.

Contents

chunk

Content included in the response.

Type: [PayloadPart](#) object

Required: No

internalServerErrorException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

modelStreamErrorException

An error occurred while streaming the response. Retry your request.

Type: Exception

HTTP Status Code: 424

Required: No

modelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

Type: Exception

HTTP Status Code: 408

Required: No

throttlingException

The number or frequency of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- [AWS SDK for C++](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for Ruby V3](#)

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see [Signing AWS API requests](#) in the *IAM User Guide*.

Action

The action to be performed.

Type: string

Required: Yes

Version

The API version that the request is written for, expressed in the format YYYY-MM-DD.

Type: string

Required: Yes

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: AWS4-HMAC-SHA256

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key/YYYYMMDD/region/service/aws4_request*.

For more information, see [Create a signed AWS API request](#) in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: 20120325T120000Z.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see [Elements of an AWS API request signature](#) in the *IAM User Guide*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS STS, see [AWS services that work with IAM](#) in the *IAM User Guide*.

Condition: If you're using temporary security credentials from AWS STS, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see [Create a signed AWS API request](#) in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Errors

This section lists the errors common to the API actions of all AWS services. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 400

IncompleteSignature

The request signature does not conform to AWS standards.

HTTP Status Code: 400

InternalFailure

The request processing has failed because of an unknown error, exception or failure.

HTTP Status Code: 500

InvalidAction

The action or operation requested is invalid. Verify that the action is typed correctly.

HTTP Status Code: 400

InvalidClientTokenId

The X.509 certificate or AWS access key ID provided does not exist in our records.

HTTP Status Code: 403

NotAuthorized

You do not have permission to perform this action.

HTTP Status Code: 400

OptInRequired

The AWS access key ID needs a subscription for the service.

HTTP Status Code: 403

RequestExpired

The request reached the service more than 15 minutes after the date stamp on the request or more than 15 minutes after the request expiration date (such as for pre-signed URLs), or the date stamp on the request is more than 15 minutes in the future.

HTTP Status Code: 400

ServiceUnavailable

The request has failed due to a temporary failure of the server.

HTTP Status Code: 503

ThrottlingException

The request was denied due to request throttling.

HTTP Status Code: 400

ValidationError

The input fails to satisfy the constraints specified by an AWS service.

HTTP Status Code: 400