AWS Decision guide

Amazon Bedrock or Amazon SageMaker?

Amazon Bedrock or Amazon SageMaker?: AWS Decision guide

Table of Contents

Amazon Bedrock or Amazon SageMaker?

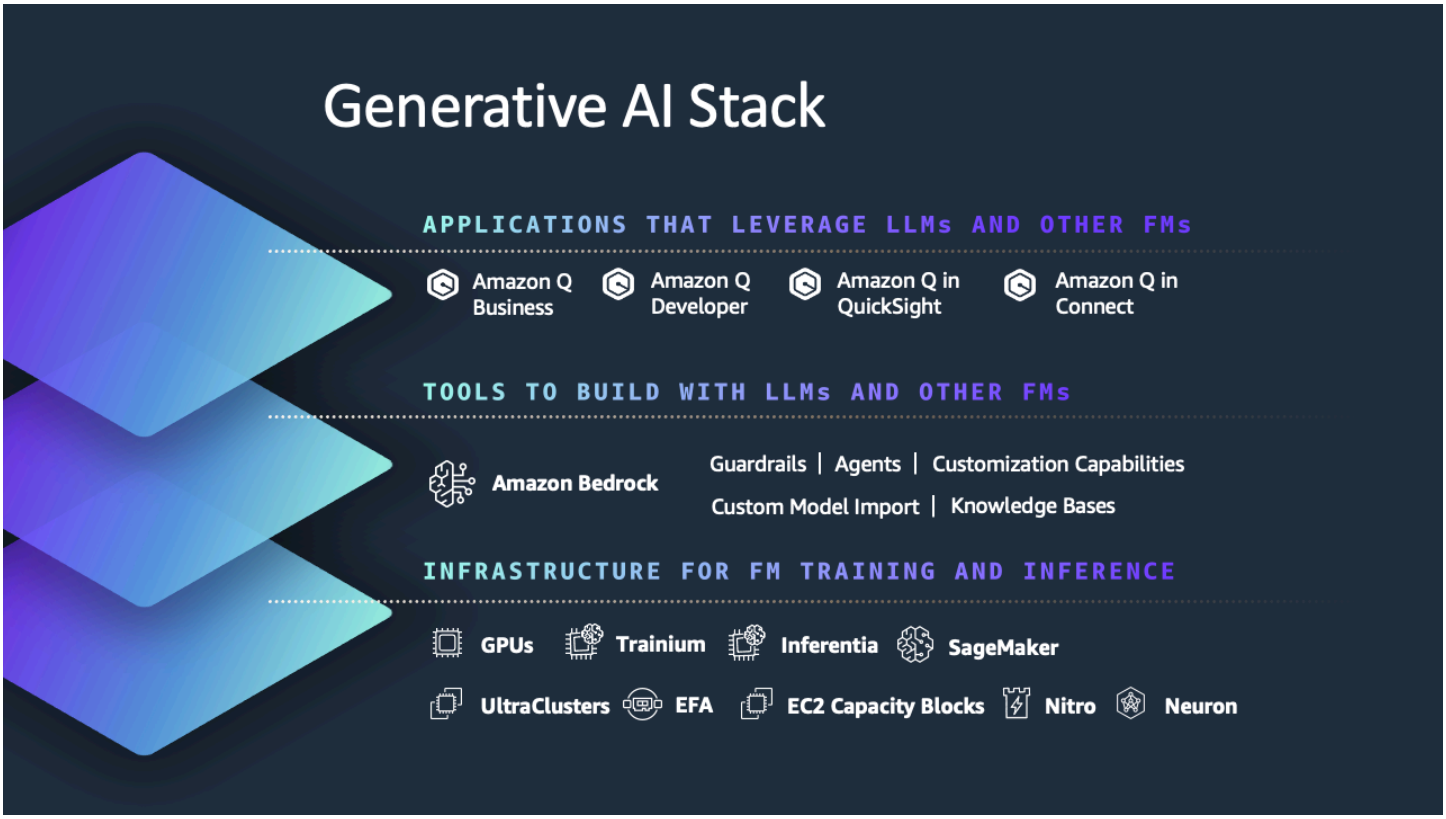
Understand the differences and pick the one that's right for you

Purpose	Understand the differences between Amazon Bedrock and Amazon SageMaker, and determine which service is the best fit for your needs.
Last updated	August 21, 2024
Covered services	• Amazon Bedrock • Amazon SageMaker

Introduction

Amazon Web Services (AWS) offers a suite of services to help you build machine learning (ML) and generative AI applications. It's helpful to understand how these services work together to form a generative AI stack, including:

- Generative AI-powered services such as Amazon Q, which leverages large language models (LLMs) and other foundation models (FMs).
- Tools for building applications with LLMs and other FMs, including Amazon Bedrock.
- Infrastructure for model training and inference, such as Amazon SageMaker and specialized hardware.



When considering which generative AI services you want to use, two services are often considered alongside one another:

Amazon Bedrock

- Choose Amazon Bedrock if you primarily need to use pre-trained foundation models for inference, and want to select the foundation model that best fits your use case. Amazon Bedrock is a fully managed service for building generative AI applications with support for popular foundation models, including Anthropic Claude, Cohere Command & Embed, AI21 Labs Jurassic, Meta Llama, Mistral AI, Stable Diffusion XL and Amazon Titan. Supported FMs are updated on a regular basis.
- Use Amazon Bedrock to build generative AI applications with security, privacy, and responsible AI—regardless of the foundation model you choose. Amazon Bedrock offers model-independent, single API access, so you can use different foundation models, and upgrade to the latest model versions, with minimal code changes. Amazon Bedrock also supports model fine-tuning and the import of custom models.
- Use Amazon Bedrock Studio (in preview), which is a new SSO-enabled web interface that your developers can use to work with large language models (LLMs) and other foundation models (FMs), collaborate on projects, and iterate on generative AI applications.

Amazon SageMaker

- Amazon SageMaker is a fully managed service designed to help you build, train, and deploy machine learning models at scale. This includes building FMs from scratch, using tools like notebooks, debuggers, profilers, pipelines, and MLOps. Consider SageMaker when you have use cases that can benefit from extensive training, fine-tuning, and customization of foundation models. It can also help you through the potentially challenging task of evaluating which FM is the best fit for your use case.
- Use SageMaker's integrated development environment (IDE) to build, train, and deploy FMs. SageMaker offers access to hundreds of pretrained models, including publicly available FMs.

For more information about how Amazon Bedrock and SageMaker fit into Amazon's generative AI services and solutions, see the generative AI decision guide.

While both Amazon Bedrock and Amazon SageMaker enable the development of ML and generative AI applications, they serve different purposes. This guide will help you understand which of these services is the best fit for your needs, including scenarios in which both services can be used together to build generative AI applications.

Here's a high-level view of the key differences between these services to get you started.

Category	 Amazon Bedrock	 Amazon SageMaker
Use Cases	Ideal for integration of AI capabilities into applications without investing heavily in custom model development	Optimized for unique or specialized AI/ML needs that may require custom models
Target Users	Optimized for developers and businesses without deep machine learning expertise	Optimized for data scientists, machine learning engineers, and developers

Category	 Amazon Bedrock	 Amazon SageMaker
Customization	You'll primarily use pre-train ed models, but can fine-tune as needed	You have full control, and can customize or create models according to your needs
Pricing	Pay-as-you-go pricing based on the number of API calls made to the service	Charges based on the usage of compute resources, storage, and other services
Integration	Integrate pre-trained models into applications through API calls	Integrate custom models into applications, with more customization options
Expertise Required	Basic level of machine learning expertise needed to use pre-trained models	Working knowledge of data science and machine learning skills are helpful for building and optimizing models

Differences between Amazon Bedrock and SageMaker

Let's examine and compare the capabilities of Amazon Bedrock and Amazon SageMaker.

Use cases

Amazon Bedrock and Amazon SageMaker address different use cases based on your specific requirements and resources.

Amazon Bedrock

- Amazon Bedrock is designed for use cases where you want to efficiently incorporate AI capabilities into your applications without investing heavily in custom model development. For example, a content moderation system for a social media platform could use Amazon Bedrock's pre-trained models to automatically identify and flag inappropriate text or images. Similarly, a customer support chatbot could use Amazon Bedrock's natural language

processing capabilities to understand and respond to user inquiries. Amazon Bedrock is particularly useful if you have limited machine learning expertise or resources, as it helps you to benefit from AI without the need for extensive in-house development.

Amazon SageMaker

- SageMaker is a good choice for unique or specialized AI/ML needs that require custom-built models. It is ideal for scenarios where off-the-shelf solutions are not sufficient, and you have a need for fine-grained control over the model architecture, training process, and deployment. One example of a scenario that would benefit from using SageMaker would be a healthcare company developing a model to predict patient outcomes based on specific biomarkers. Another example would be a financial institution creating a fraud detection system tailored to their unique data and risk factors. Additionally, SageMaker is suitable for research and development purposes, where data scientists and machine learning engineers can experiment with different algorithms, hyperparameters, and model architectures.

Target users

Amazon Bedrock and Amazon SageMaker support different targeted users based on their level of expertise and knowledge of machine learning and artificial intelligence.

Amazon Bedrock

- Amazon Bedrock offers a more accessible and straightforward way to integrate AI functionality into your projects. It's appropriate for a broad audience, which includes developers and businesses, that has limited experience in building and training machine learning models, but wants to use AI to enhance their applications or workflows.

Amazon SageMaker

- SageMaker is predominantly for data scientists, machine learning engineers, and developers who possess the necessary skills and knowledge to build, train, and deploy custom machine learning models. Use SageMaker if you are well-versed in data science and machine learning concepts, and require a platform that provides you with the tools and flexibility to create models tailored to your specific needs.

Customization

Amazon Bedrock and Amazon SageMaker offer different levels of customization capabilities that you can tailor to your specific needs and expertise.

Amazon Bedrock

- Amazon Bedrock provides pre-trained AI models that you can integrate into applications, with limited customization. You have access to a set of API calls that you use to enter data and receive predictions from these pre-trained models. While this approach drastically simplifies the process of incorporating AI capabilities into applications, it also means that you have less control over the underlying models, unless you customize a model, or import a custom model. Amazon Bedrock's pre-trained models are optimized for common AI tasks and are designed to work well for a wide range of use cases, but they may not be suitable for highly specialized or niche requirements.

  Amazon Bedrock supports fine-tuning for foundation models (FMs), such as Cohere Command R, Meta Llama 2, Amazon Titan Text Lite, Amazon Titan Text Express, Amazon Titan Multimodal Embeddings, and Amazon Titan Image Generator. You can now fine-tune Anthropic Claude 3 Haiku in a preview capacity in the US West (Oregon) AWS Region. The list of supported FMs is updated on an ongoing basis.
- Customize models for specific tasks and use cases, including FM fine-tuning and pre-training. Bring your own customized model with custom model import (in preview).

Amazon SageMaker

- Amazon SageMaker provides extensive customization options, giving you full control over the entire machine learning workflow. With SageMaker, you can fine-tune every aspect of your models, from data preprocessing and feature engineering to model architecture and hyperparameter optimization. By using this level of customization, you can create highly specialized models that are tailored to your unique business requirements. SageMaker supports a wide range of popular machine learning frameworks, such as TensorFlow, PyTorch, and Apache MXNet, allowing you to use your preferred tools and libraries for building and training models.
- Use Amazon SageMaker JumpStart to evaluate, compare, and select FMs based on pre-defined quality and responsibility.

- Choose which FM to use with Amazon SageMaker Clarify. Use SageMaker Clarify to create model evaluation jobs, that you use to evaluate and compare model quality and responsibility metrics for text-based foundation models from JumpStart.
- Generate predictions using Amazon SageMaker Canvas, without needing to write any code. Use SageMaker Canvas in collaboration with Amazon Bedrock to fine-tune and deploy language models. This blog post describes how you can use them to optimize customer interaction by working with your own datasets, such as your product FAQs, in Amazon Bedrock and Amazon SageMaker JumpStart.

Pricing

Amazon Bedrock and Amazon SageMaker have different pricing models that reflect their target users and the services they provide.

Amazon Bedrock

- Amazon Bedrock employs a simple pricing model based on the number of API calls made to the service. You pay a fixed price per API call, which includes the cost of running the pre-trained models and any associated data processing. This straightforward pricing structure makes it more efficient for you to estimate and control your costs, as you pay only for the actual usage of the service. Amazon Bedrock's pricing model is particularly well-suited for applications with predictable workloads, or for cases where you want more transparency in your AI-related expenses.

Amazon SageMaker

- SageMaker follows a pay-as-you-go pricing model based on the usage of compute resources, storage, and other services consumed during the machine learning process. You're charged for the instances that you use to build, train, and deploy you models, with prices varying depending on the instance type and size. Additionally, you incur costs for data storage, data transfer, and other associated services like data labeling and model monitoring. This pricing model provides flexibility and allows you to optimize costs based on your specific requirements. However, it also means that costs can vary and may require careful management, especially for resource-intensive projects.

Integration

Amazon Bedrock and Amazon SageMaker offer different approaches to integrating machine learning models into applications, catering to your specific needs and expertise.

Amazon Bedrock

- Amazon Bedrock simplifies the integration process by providing pre-trained models that you can access directly through API calls. Use the Amazon Bedrock SDK or REST API to send input data and receive predictions from the models without needing to manage the underlying infrastructure. This approach significantly reduces the complexity and time required to integrate AI capabilities into applications, making it more accessible to developers with limited machine learning expertise. However, this ease of integration comes at the cost of limited customization options, as you're restricted to the pre-trained models and APIs provided by Amazon Bedrock.

Amazon SageMaker

- SageMaker provides a comprehensive platform for building, training, and deploying custom machine learning models. However, integrating these models into applications requires more effort and technical expertise compared to Amazon Bedrock. You need to use the SageMaker SDK or API to access the trained models and build the necessary infrastructure to expose them as endpoints. This process involves creating and configuring API Gateway, Lambda functions, and other AWS services to enable communication between the application and the deployed model. While SageMaker provides tools and templates to simplify this process, it still requires a deeper understanding of AWS services and machine learning model deployment.

Expertise required

Amazon Bedrock and Amazon SageMaker are optimized for different levels of machine learning expertise.

Amazon Bedrock

- Amazon Bedrock is more accessible to a broader range of users, including developers and businesses with limited machine learning expertise. By providing pre-trained models that can be easily integrated into applications through API calls, Amazon Bedrock abstracts away much of the complexity associated with building and deploying machine learning

models. You don't need to worry about data preprocessing, model selection, or infrastructure management, as these aspects are handled by the Amazon Bedrock service. This allows you to focus on integrating AI capabilities into your applications without needing to invest significant time and resources in acquiring deep machine learning knowledge.

Amazon SageMaker

- If you have deeper expertise in data science and machine learning, SageMaker provides a powerful and flexible platform for building, training, and deploying custom models. While SageMaker aims to simplify the machine learning workflow, it still requires a significant level of technical expertise to take full advantage of its capabilities. You'll benefit from being proficient in programming languages like Python, along with a deep understanding of machine learning concepts, such as data preprocessing, model selection, and hyperparameter tuning. Additionally, you should be comfortable working with various AWS services and managing the infrastructure required to deploy and integrate their models. As a result, SageMaker may have a steeper learning curve if you're new to machine learning or have limited experience with AWS.

The choice between Amazon Bedrock and Amazon SageMaker is not always mutually exclusive. In some cases, you may benefit from using both services together. For example, you could use Amazon Bedrock to quickly prototype and deploy a foundation model, and then use SageMaker to further refine and optimize the model for better performance. For example, this blog post describes how you can use Amazon Bedrock and Amazon SageMaker together to optimize customer interaction by working with your own datasets (such as your product FAQs.

Ultimately, the decision between Amazon Bedrock and Amazon SageMaker depends on your specific requirements. Evaluating these factors can help you make an informed decision and choose the service that is most suitable for your needs.

For more information about Amazon's generative AI services and solutions, see the generative AI decision guide.

Use

Now that you've read about the criteria for choosing between Amazon Bedrock and Amazon SageMaker, you can select the service that meets your needs, and use the following information to help you get started using each of them.

Amazon Bedrock

- **What is Amazon Bedrock?**

  Describes how to use this fully managed service to make foundation models (FMs) from Amazon and third parties available for your use through a unified API.

  Explore the guide

- **Frequently asked questions about Amazon Bedrock**

  Get answers to the most commonly-asked questions about Amazon Bedrock. These include how to use agents, security considerations, details about Amazon Bedrock software development kits (SDKs), retrieval augmented generation, how to use model evaluation, and billing.

  Read the FAQs

- **Guidance for generating product descriptions with Amazon Bedrock**

  Describes how to use Amazon Bedrock in your solution to automate your product review and approval process for an e-commerce marketplace or retail website.

  Explore the solution

Amazon Bedrock Studio

- **What is Amazon Bedrock Studio?**

  Describes how to use this web app to prototype apps that use Amazon Bedrock models and features, without having to set up and use a developer environment.

  Explore the guide

- **Build generative AI applications with Amazon Bedrock Studio**

  This blog post describes how you can build applications using a wide array of top performing models. It then explains how to evaluate and share your generative AI apps within Amazon Bedrock Studio.

Read the blog

- **Building an app with Amazon Bedrock Studio**

  Use the Build mode in Amazon Bedrock Studio to create prototype apps that uses Amazon Bedrock models and features. You can also use the Build mode to try experiments not supported in the Explore mode playground, such as setting inference parameters.

  Explore the guide

Amazon SageMaker

- **What is Amazon SageMaker?**

  Describes how you can use this fully managed machine learning (ML) service to build, train, and deploy ML models into a production-ready hosted environment.

  Explore the guide

- **Get started with Amazon SageMaker**

  Describes how to join an Amazon SageMaker domain, giving you access to Amazon SageMaker Studio and RStudio on SageMaker.

  Explore the guide

- **Get started with Amazon SageMaker JumpStart**

  Explore SageMaker JumpStart solution templates that set up infrastructure for common use cases, and executable example notebooks for machine learning with SageMaker.

  Explore the guide

Document history

The following table describes the important changes to this decision guide. For notifications about updates to this guide, you can subscribe to an RSS feed.

Change	Description	Date
Minor updates	Minor updates to improve readability.	August 21, 2024
Minor updates	Minor updates to reflect the latest Amazon Bedrock and Amazon SageMaker features.	July 22, 2024
Initial release	Initial release of the decision guide.	July 11, 2024