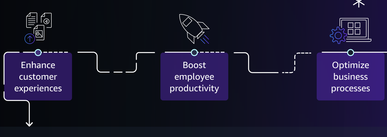
The race to unlock generative AI

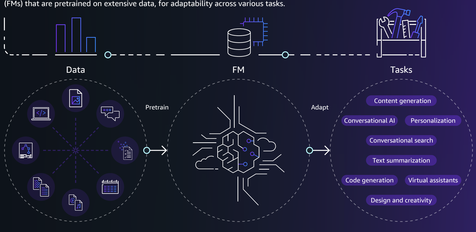
Generative artificial intelligence (AI) has sparked a revolution, and organizations are seeking solutions to fast-track the technology's business value to solve for the most common uses cases. The journey starts with an idea and can revolutionize your business to unimagined destinations.



What is generative AI?

Explore more ›

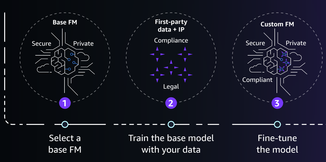
Generative AI is a type of AI that can create new content and ideas including conversations, stories, images, videos, and music. It is powered by Foundation Models (FMs) that are pretrained on extensive data, for adaptability across various tasks.



Train the model with your data, securely

Explore more ›

Begin training the model with your own proprietary data using a separate copy of the base FM that is accessible only to you. This helps ensure privacy of your first-party data and intellectual property (IP)—while safeguarding the use of your generative AI output.



Key considerations

Help ensure privacy: Use services that encrypt your data in transit and at rest. Leverage Amazon Bedrock to help ensure that your first-party data, inputs, and outputs stay in your Amazon Virtual Private Cloud (VPC).

Avoid toxicity: Help protect your staff, customers, and brand from profanity, hate, and violent speech with automated content moderation options available in the Amazon Titan family of FMs through Amazon Bedrock.

Start with the right model

Watch the webinar for a deep dive ›

Not all models are created equal and choosing the right model for your use case can impact outcomes, costs, efficiencies and more. One method in finding the best model for a specific task is by using leaderboards.

A leaderboard quickly evaluates the output of multiple models based on your use case. In this example, we will use the task of summarizing an article.

1. **Define your task**
   Example: Which model performs best for a short news summarization?
2. **Determine the relevant input and output prompts**
   Example: Input: 49 news articles, no labeled data
   Output: Summarize the article in less than 250 words
3. **Define the metrics to evaluate success**
   Example: Rank models based on best performance, latency, cost, and quality of output

Within minutes, the leaderboard provides a clear model ranking based on your metrics.

Key considerations

Understanding metrics: Leaderboards are just one type of metric that can help evaluate your model. They help with chain of thought reasoning. One key advantage is that leaderboards are shown to correlate with human evaluation through chain-of-thought reasoning. However, they can also be costly or difficult to debug.

Build from scratch, buy and consume, or fine tune models

If you can find a model that fits your use case, that is a great place to begin. If it isn't perfect, you can fine-tune the model to work for your needs.

Open source LLMs free you from relying on one vendor. Your organization may prefer and have the capabilities and team to develop and maintain the model.

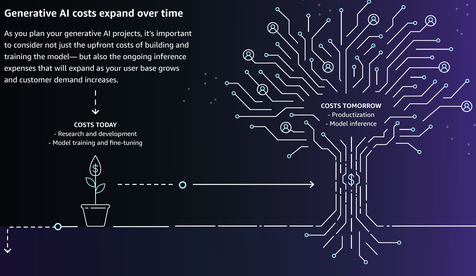
Proprietary solutions may be able to provide more performant models more immediately.

If you are not sure, you may want to run a cost-benefit analysis to determine the best choice.

Cost optimization

Generative AI costs expand over time

As you plan your generative AI projects, it's important to consider not just the upfront costs of building and training the model— but also the ongoing inference expenses that will expand as your user base grows and customer demand increases.

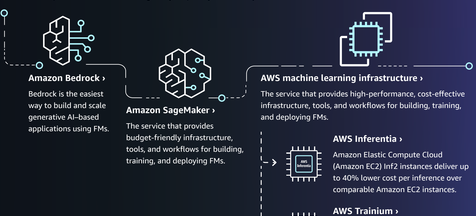


4 steps to optimizing generative AI price performance

By making the right choices at the onset of your generative AI effort, you can better control upfront and downstream costs.

How AWS can help

Amazon Web Services (AWS) offers a secure cloud infrastructure, and our services and solutions provide features that strengthen your privacy and security.



Amazon Bedrock ›

Bedrock is the easiest way to build and scale generative AI–based applications using FMs.

Amazon SageMaker ›

The service that provides budget-friendly infrastructure, tools, and workflows for building, training, and deploying FMs.

AWS machine learning infrastructure ›

The service that provides high-performance, cost-effective infrastructure, tools, and workflows for building, training, and deploying FMs.

AWS Inferentia ›

Amazon Elastic Compute Cloud (Amazon EC2) Inf2 instances deliver up to 40% lower cost per inference over comparable Amazon EC2 instances.

AWS Trainium ›

Amazon EC2 Trn1 instances deliver up to 50% in cost-to-train savings over comparable Amazon EC2 instances.

aws marketplace

Accelerate time to market

In AWS Marketplace, you can find, buy, and deploy generative AI solutions from partners experienced with AWS to jump-start your transformation.

Explore generative AI on AWS Marketplace