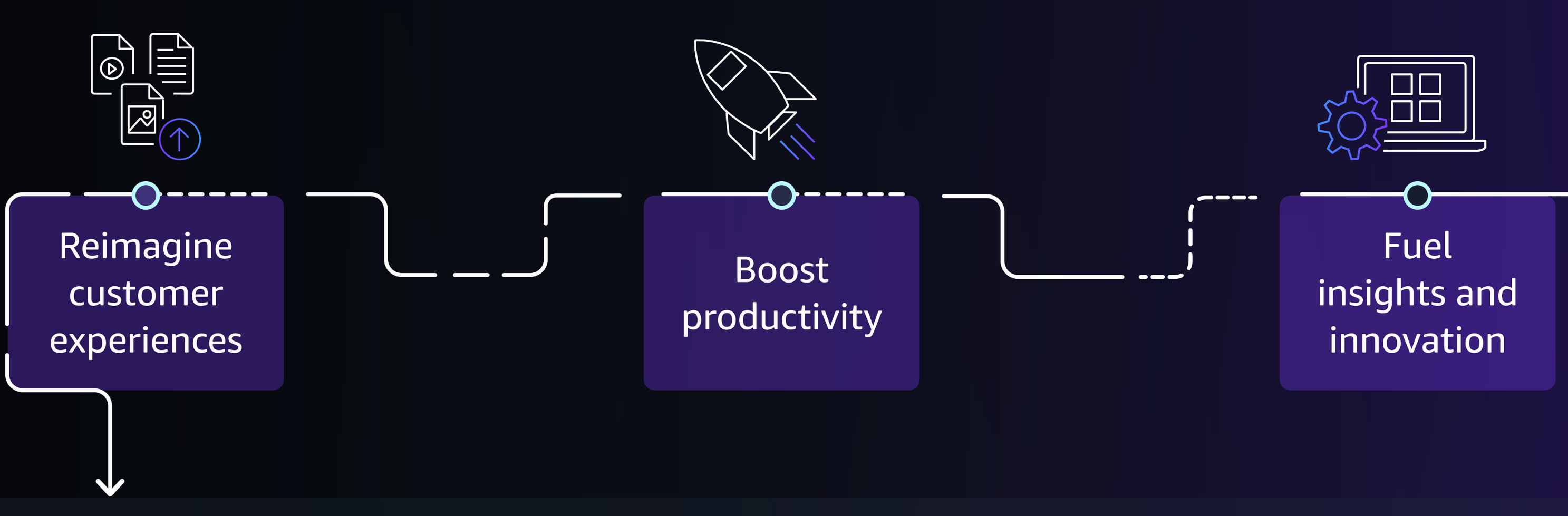
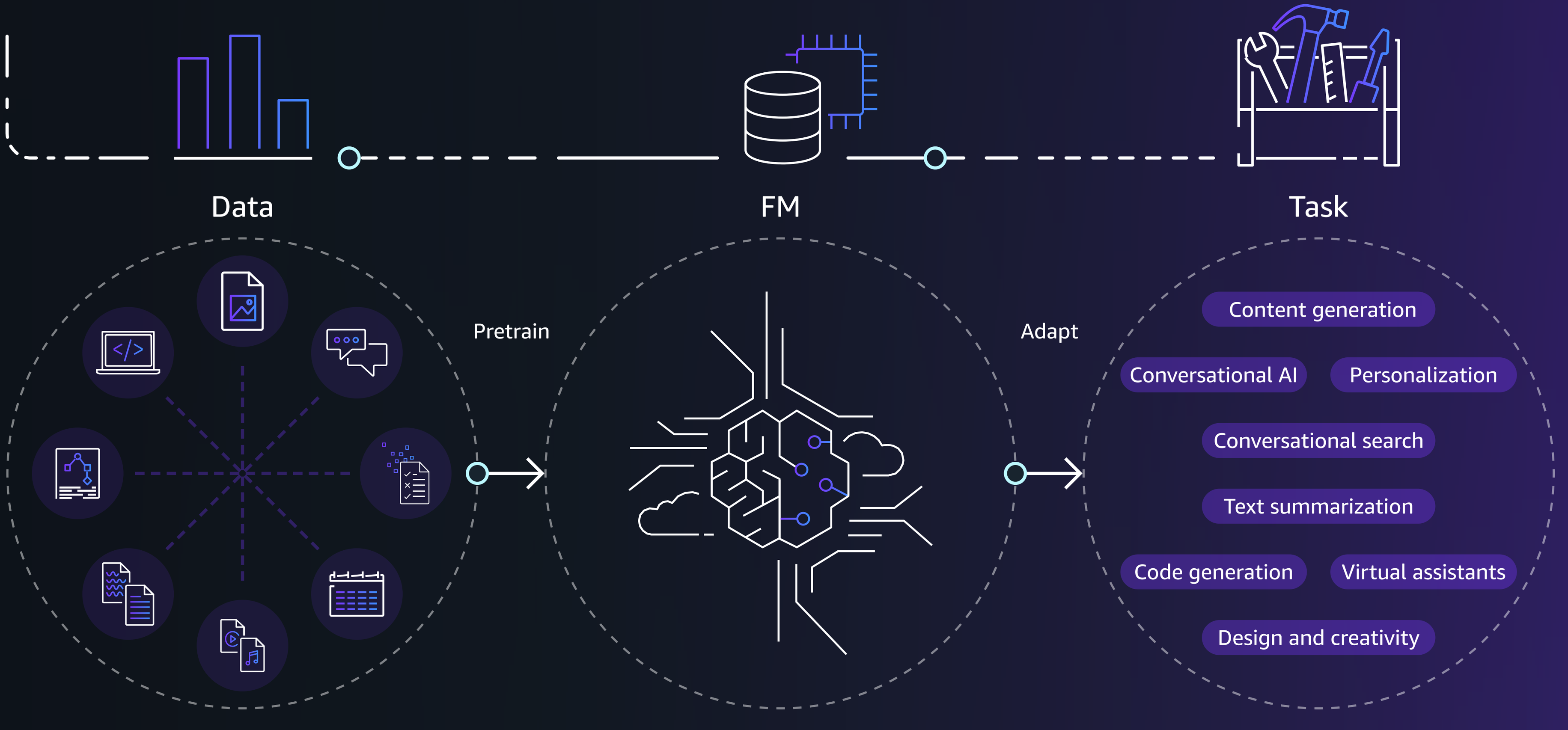
The race to unlock generative AI

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.



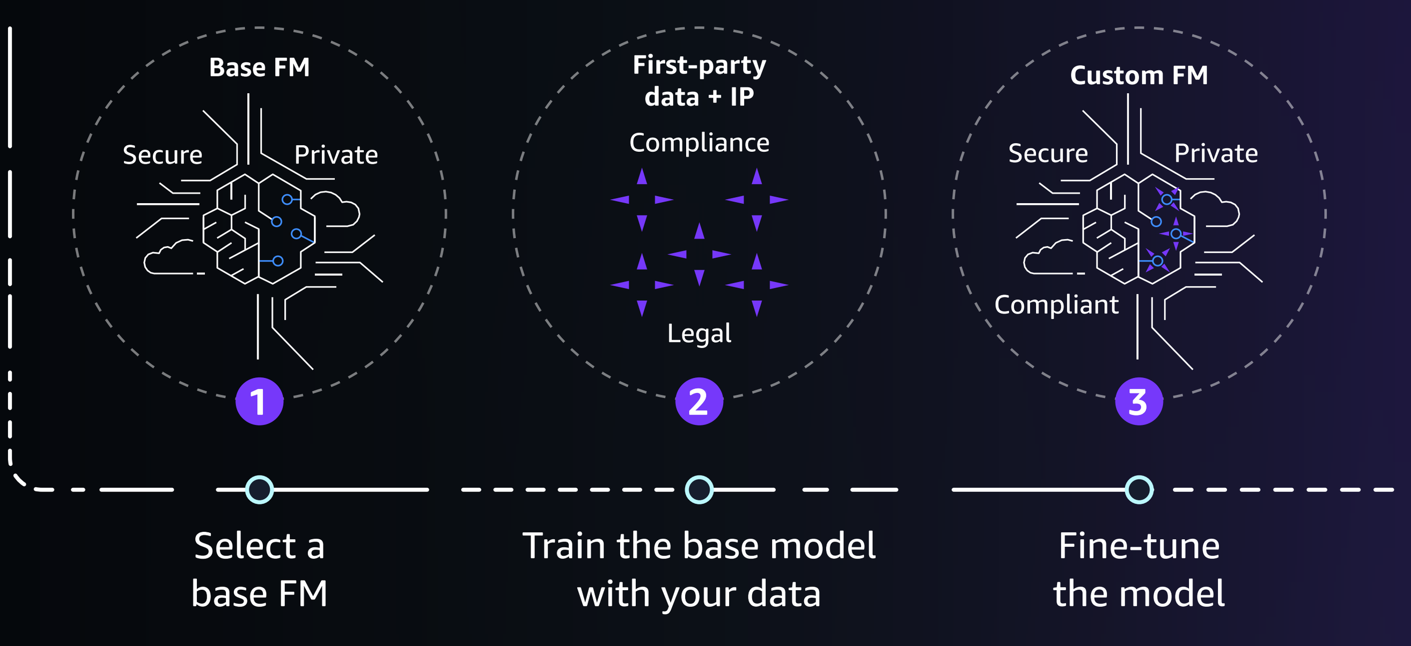
What is generative AI?

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

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

Download the model evaluation eBook ›

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.

Transform ideas into impact with the right launch approach

Download the eBook: Transitioning from POC to production ›

Many customers turn to open-source Large Language Models (LLMs) to get started with a proof-of-concept (POC) quickly. POCs help prioritize investments, substantiate strategies, and highlight organizational needs around talent and data.

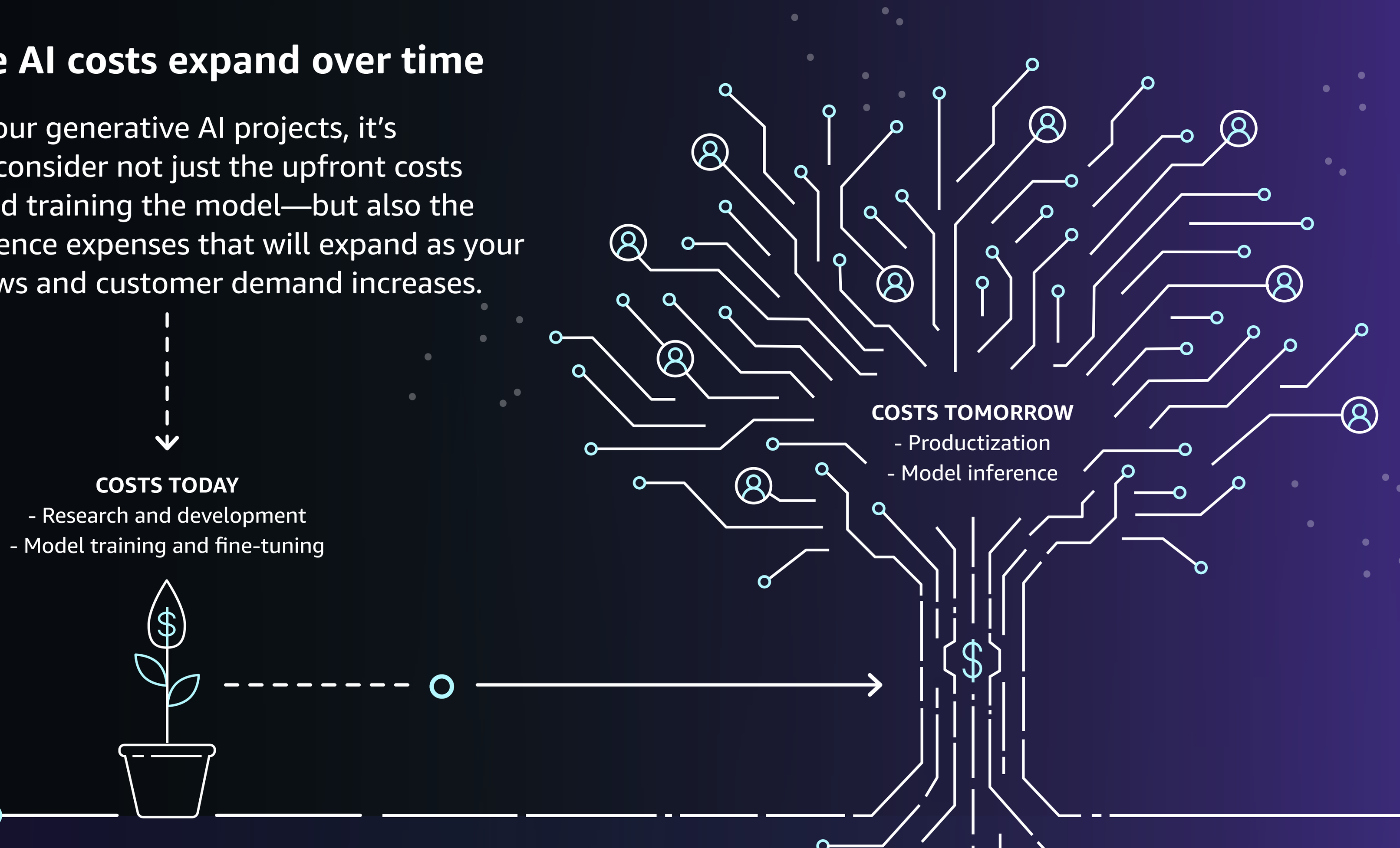
If you can find a model that fits your use case, that is a great place to begin. If it isn't perfect, you may be able to fine-tune the model to meet your needs. Revisiting the model selection may occur as you explore the impact it has on speed, accuracy, and cost in production.

Having a production mindset helps ensure your POC transitions to best-practice performance.

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.

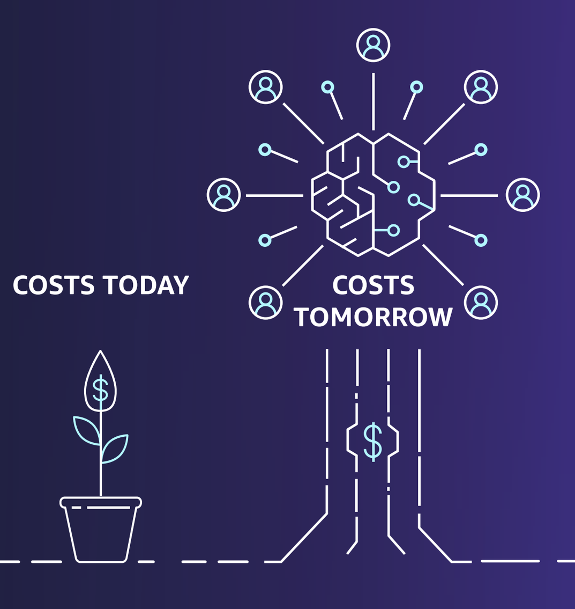
Download the prompt engineering eBook ›

1 Right-size your model
You may not need the largest model. Pick the right type and size of model for your use case.

2 Choose the optimal infrastructure
Explore a broad set of GPUs and purpose-built accelerators to balance performance and costs.

3 Optimize everything
Keep refining your deployment to maximize utilization of underlying resources.

4 Reduce dev time with better tools
Manage your AI innovation, not your infrastructure.



How AWS can help

Amazon Web Services (AWS) offers a high-performance, cost-effective cloud infrastructure, and our services and solutions provide features that strengthen your privacy and security.

Amazon Bedrock ›
Ideal for integrating pre-trained FMs into applications with minimal machine learning expertise.

Amazon SageMaker ›
Provides budget-friendly infrastructure, tools, and workflows for building, training, and deploying FMs.

Amazon Q ›
A generative AI assistant that accelerates productivity, profitability, and transforms organizations by empowering teams to do more.

How AWS Marketplace can 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. Get started today.

Select the right LLM for your use case ›

Transition from POC to production ›

Balance cost, performance, and accuracy ›

Explore gen AI solutions in AWS Marketplace

aws marketplace