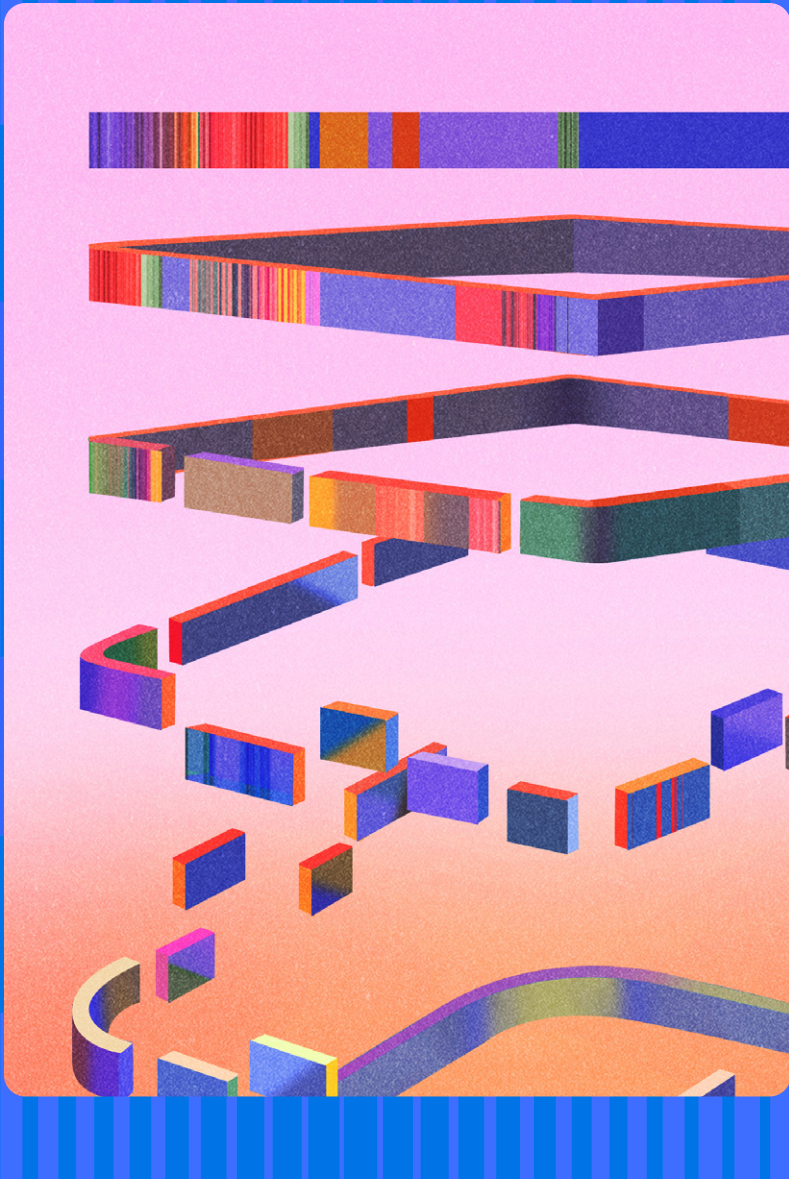




Understanding the costs of generative AI infrastructure

When selecting generative artificial intelligence (gen AI) infrastructure, the choices you make upfront can impact everything from cost and performance to sustainability goals and ease of use. Read on to discover factors that play a role in this selection and how you can best optimize for the price and performance you need.



Generative AI is transforming businesses

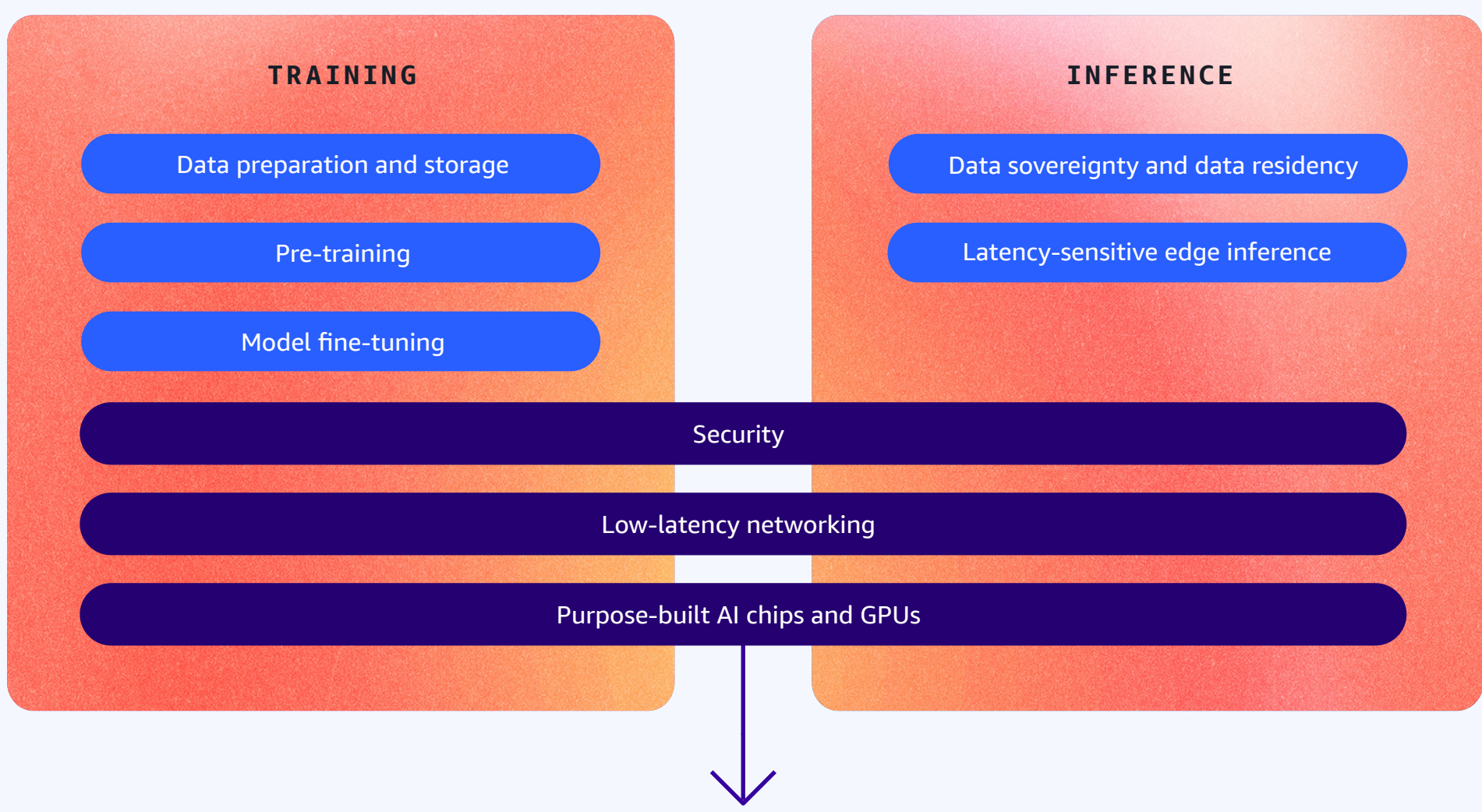
Powered by large-scale foundation models (FMs) that are trained on up to petabytes of data, gen AI can create new content and ideas, including conversations, stories, images, videos, music, and even software code, in response to a prompt. As these models grow, their parameters also increase—to upward of trillions of parameters.

Even smaller language models can be trained with a few billion parameters and can go up to 15 billion parameters. Organizations need an unprecedented level of infrastructure to build and train these models in a reasonable time and deploy them for inference.

Generative AI infrastructure costs vary over time

As you plan your gen AI projects, it's important to consider not just the upfront cost of training the model but also the ongoing inference expenses that will vary depending on customer demands.

GENERATIVE AI COST CONSIDERATIONS



4 steps to optimizing generative AI price performance

With purpose-built AI infrastructure, you can innovate with higher performance while reducing costs.

- 1 Select the right model**
Pick the right type and size of model for your use case. (You may not need the largest model.) Explore the [Selecting the Right Text-Based LLM for Your Use Case](#) infographic to learn more.
- 2 Choose the optimal infrastructure**
Develop an infrastructure strategy that includes the right processing power, low-latency networking, and scalable storage without compromising on cost or performance.
- 3 Reduce dev time with better tools**
Utilize purpose-built tools for data labeling, cleansing, enrichment, error detection, and formatting, which can help connect, access, and query data from a wide variety of popular data sources.
- 4 Optimize everything**
Keep refining your deployment to maximize utilization of underlying resources by continuously evaluating models for accuracy, latency, and cost.

How AWS can help

Amazon Web Services (AWS) offers solutions that can help you achieve the four steps outlined above—all with minimal burden on your resources and maximum impact on your gen AI investments.

AMAZON SAGEMAKER AI

[Amazon SageMaker AI](#) provides a fully managed infrastructure and tools that include high-performance, cost-effective compute alongside integrated, purpose-built machine learning (ML) tools for the full AI workflow. Plus, with SageMaker AI, you can build, train, and deploy FMs at scale using tools like notebooks, debuggers, profilers, pipelines, machine learning operations (MLOps), and more.

ACCELERATED COMPUTING

From the highest-performance NVIDIA GPU-based [Amazon Elastic Compute Cloud](#) (Amazon EC2) to continued investments in our purpose-built AI chips, [AWS Trainium](#) and [AWS Inferentia](#), AWS delivers the best price performance for training and deploying gen AI models at scale.

NETWORKING

Purpose-built to meet the performance demands for gen AI, AWS features high-throughput and low-latency networking, including [Elastic Fabric Adapter \(EFA\)](#) and [Amazon EC2 UltraClusters](#).

STORAGE

Accelerate compute workloads with [Amazon FSx for Lustre](#), which provides sub-millisecond latencies, up to hundreds of gigabytes per second (GBps) of throughput, and millions of input/output operations per second (IOPS) while quickly accessing and processing your datasets with [Amazon Simple Storage Service](#) (Amazon S3).

SECURITY

Our accelerated computing Amazon EC2 instances and networking are built on a foundation of the [AWS Nitro System](#), which has been validated by the NCC Group, an independent cybersecurity firm. The level of security protection offered is so critical that we've added it to our AWS Service Terms to provide additional assurance to our customers.

Transform your business with generative AI infrastructure on AWS

Accelerate AI innovation with AWS. Learn more about the most comprehensive, secure, and price-performant AI infrastructure.

Explore generative AI infrastructure on AWS

