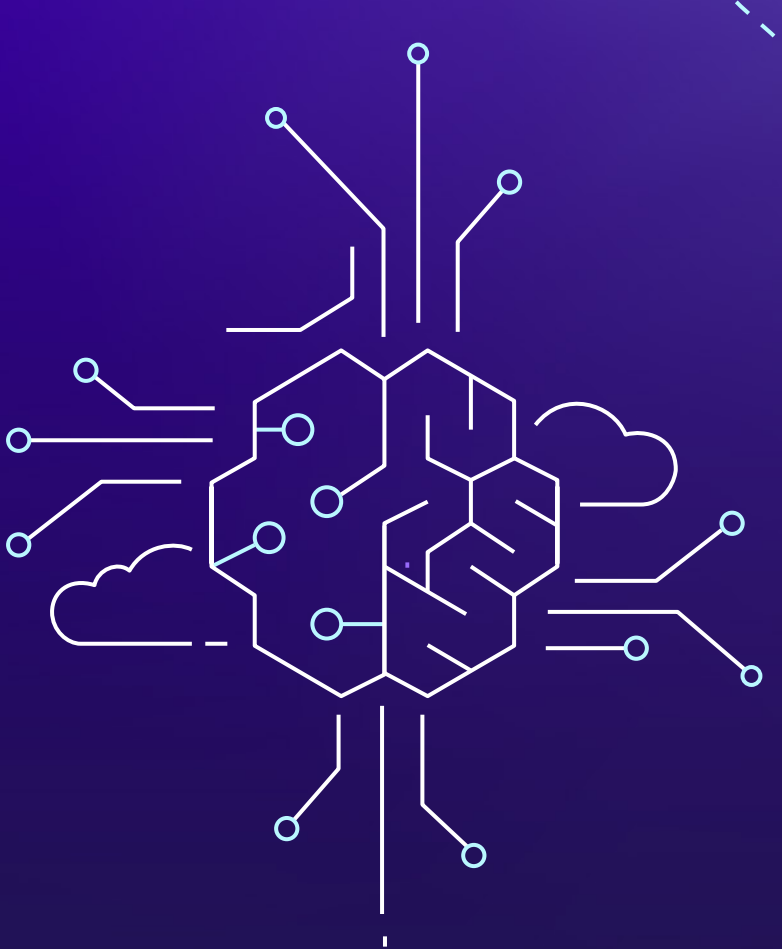
Selecting the right text-based LLM for your use case

Maximize the business value of LLMs with these insights and strategies



Generative artificial intelligence (gen AI) applications powered by large language models (LLMs) have the potential to transform every industry. Staying up to speed with LLMs is challenging as the technology continues to evolve at lightning speed.

Read on to discover best practices that will help you navigate the LLM landscape with confidence.

What can text-based LLMs do for you?

LLMs are trained on trillions of words across many natural language tasks. When supported by the right gen AI infrastructure, they can carry out functions in a conversational manner. These functions include:

- **Engaging in interactive conversations**
- **Understanding, learning, and generating text**
- **Answering questions**
- **Summarizing dialogues and documents**
- **Providing suggestions**

LLMs are powering applications across multiple industries, including healthcare and life sciences, media and entertainment, financial services, and more.

Key factors to consider for text-based LLMs

The list of available LLMs is growing fast as technology evolves. Understanding the factors that shape LLM options can help you with selection, gaining a competitive advantage, and increasing the business value of your gen AI investments.

Model capabilities

Assess the model's ability to perform specific tasks, such as text generation, translation, summarization, and code generation.

Model size and complexity

Larger models often offer better performance but require more computational resources. Evaluate your specific needs to determine the optimal size.

Model customization

Determine whether the model can be used directly for your tasks or if it requires fine-tuning your specific domain data to achieve optimal performance.

Cost and accessibility

Factor in the cost of using the model, including API fees or licensing costs. Consider open source options for more flexibility.

Ethical considerations

Be mindful of potential biases in the model's outputs and take steps to mitigate them. Ensure the model aligns with ethical guidelines and principles.

Find the right text-based LLM for your use case

Beyond the factors that shape LLMs, selecting the right text-based LLM for your use case can increase the accuracy and quality of your inputs, improve performance, and drive greater cost-efficiency.

TASKS

- TEXT CLASSIFICATION
- TEXT GENERATION
- Q&A
- TRANSLATION
- CONVERSATIONAL AI
- SUMMARIZATION
- DOCUMENT UNDERSTANDING
- SOFTWARE DEVELOPMENT
- LIFE SCIENCES

Interested in other LLM use cases? Go beyond text-based LLMs with the broadest selection of multi-modal options on Amazon SageMaker AI and Amazon Bedrock, including the new Amazon Nova foundation models (FMs). Learn more about leveraging your customized models across AI21 Labs, Anthropic, Luma AI, Meta, Mistral, Stability.AI, and more.

Learn more ›

Choose the right infrastructure to optimize performance and costs for LLMs

Build with specialized AI infrastructure that delivers the performance you need while reducing costs.

1
Rightsize your model

You may not need the largest model. Pick the right type and size model depending on your use case.

2
Choose the optimal infrastructure

Explore purpose-built infrastructure solutions that are uniquely designed from the ground up to accelerate innovation, enhance security, and improve performance while lowering costs.

AWS offers infrastructure and services designed to help you get the most performance out of your LLMs while optimizing your costs:

ACCELERATED COMPUTING

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

AMAZON SAGEMAKER AI

Amazon SageMaker AI is a fully managed service that brings together a broad set of tools to enable high-performance, low-cost ML for any use case. With SageMaker AI, you can build, train, and deploy ML models at scale using tools, such as notebooks, debuggers, profilers, pipelines, machine learning operations (MLOps), and more—all in one integrated development environment (IDE).

NETWORKING

Purpose-built to meet the performance demands for gen AI, Amazon Web Services (AWS) provides high-throughput and low-latency networking that includes 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 throughout, and millions of input/output operations per second (IOPS) while quickly accessing and processing your datasets on 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 independent cybersecurity firm NCC Group. The level of security protection offered is so critical that we've added it to our AWS Service Terms to provide additional assurance to all of our customers.

AWS Trainium2–based Amazon EC2 Trn2 instances deliver

UP TO **30%**

better price performance than current generation GPU-based EC2 instances

AWS Inferentia2–based Amazon EC2 Inf2 instances deliver

UP TO **40%**

lower cost per inference than comparable EC2 instances

Unleash the power of generative AI LLMs

Optimize performance and costs for LLM deployment

While LLMs hold the potential to transform your business and give it a competitive edge, building, training, and deploying them requires an unprecedented level of infrastructure resources. To succeed, you need an infrastructure strategy that delivers the right processing power without compromising on cost or performance; low-latency, high-throughput networking; storage solutions that help accelerate and cost-optimize your compute; and a deep set of cloud services and partners. Empower your organization with LLMs—start your journey with AWS today.

Get started with AWS AI infrastructure ›