aws

Selecting the right LLM for your use case

Maximize the business value of LLMs with these insights and strategies

Generative artificial intelligence (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 LLMs do for you?

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

Understand your options for LLMs

The list of available LLMs is growing fast as technology evolves. Understanding your LLM options can help you gain a competitive advantage and increase the business value of your generative AI investments.

BERT

The first transformer-based LLM inspired many optimized variants. DistilBERT runs 60% faster while preserving over 95% of BERT's performance.

GPT

Generative pretrained transformers demonstrate the ability to scale to hundreds of billions of parameters.

PaLM

Pathways Language Model (PaLM) 2 has multilingual, reasoning, and coding capabilities. It comes in 4 sizes: Gecko, Otter, Bison, and Unicorn.

LLaMA

LLaMA comes in various sizes, from 7 billion to 65 billion parameters. Vicuna, Alpaca, and Guanaco are its fine-tuned versions.

BLOOM

BLOOM is an alternative to GPT-3 and has been trained on 46 different languages and 13 programming languages.

Chinchilla

Chinchilla is 4x smaller than its predecessor, Gopher, and trained on 4x more data while using the same training budget.

Find the right LLM for your use case

Selecting the right LLM for your use case can increase the accuracy of your outputs, improve performance, and drive cost-efficiency.

The following chart outlines some LLMs to help you get started with specific use cases:

TASKS	MODELS
TEXT CLASSIFICATION	BERT, DistilBERT, RoBERTa
TEXT GENERATION	Falcon, GPT-NeoX, LLaMA 2, Mistral
Q&A	BERT, FLAN-T5, RoBERTa, Mistral 8x7B
TRANSLATION	BART, FLAN-T5
CONVERSATIONAL AI	Falcon, GPT-2, LLaMA 2
SUMMARIZATION	FLAN-T5, LLaMA 2, Mistral Large/Small
DOCUMENT UNDERSTANDING	Donut, LayoutLM
SOFTWARE DEVELOPMENT	CodeGen, StarCoder
LIFE SCIENCES	ProtBERT, ProtGPT, LLaMA 2 and LLaMA 3

Choose the right machine learning framework for your needs

AWS computing instances support major machine learning (ML) frameworks to accelerate deep learning in the cloud.

TensorFlow
Enhance and visualize your deep learning applications.

Hugging Face
Easy-to-deploy and fine-tuned pretrained ML models.

PyTorch
Highly performant, scalable, and enterprise-ready experiences.

dmlc XGBoost
Popular and efficient open-source implementation of the gradient-boosted trees algorithm.

mxnet
Fast and scalable inference framework with an easy-to-use, concise API for ML.

Power your LLMs with the right infrastructure

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

Optimize performance and 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.

Amazon Web Services (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 ML accelerators, AWS Trainium and AWS Inferentia, AWS delivers the best price performance for training and deploying generative AI models at scale.

AMAZON SAGEMAKER
Amazon SageMaker HyperPod provides a fully managed infrastructure and tools that include high-performance, cost-effective compute alongside integrated, purpose-built ML tools for the full AI workflow. Plus, with SageMaker, you can build, train, and deploy FM models at scale using tools like notebooks, debuggers, profilers, pipelines, MLOps, and more.

NETWORKING
Purpose-built to meet the performance demands for generative AI, AWS's high-throughput and low-latency networking 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 GBs/s of throughout, and millions of 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 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 all of our customers.

AWS Trainium–based Amazon EC2 Trn1 instances deliver
50% SAVINGS
on training costs.

Amazon EC2 Inf2 instances deliver
UP TO 40%
lower cost per inference over comparable Amazon EC2 instances.

Unleash the power of generative AI LLMs

aws

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 on AWS today.

Get started with AWS AI infrastructure ›