

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

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.

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

BLOOM

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

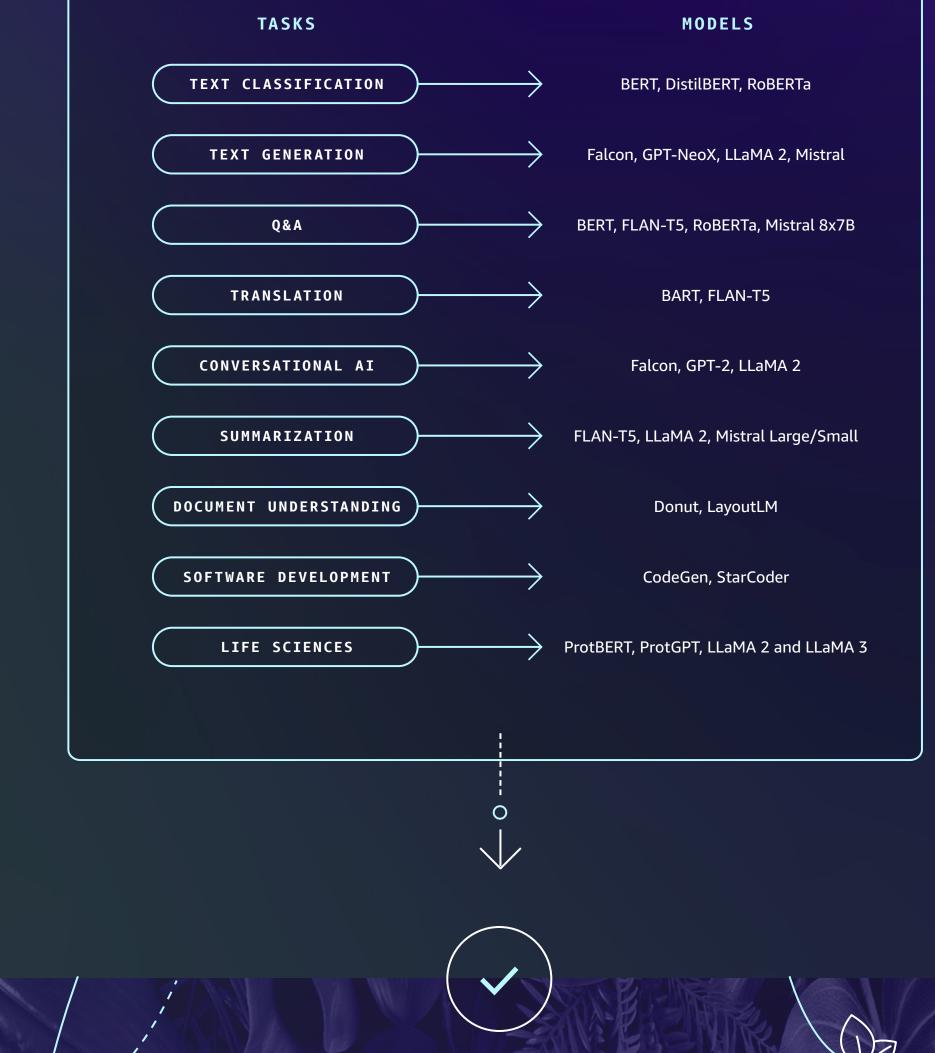
LaMA

LLaMA comes in various sizes, from 7 billion

to 65 billion parameters. Vicuna, Alpaca, and Guanaco are its fine-tuned versions.

Chinchilla

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



pretrained ML models. O PyTorch Highly performant, scalable, and enterprise-ready experiences.

TensorFlow

Enhance and visualize your deep learning applications.

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

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.

Hugging Face

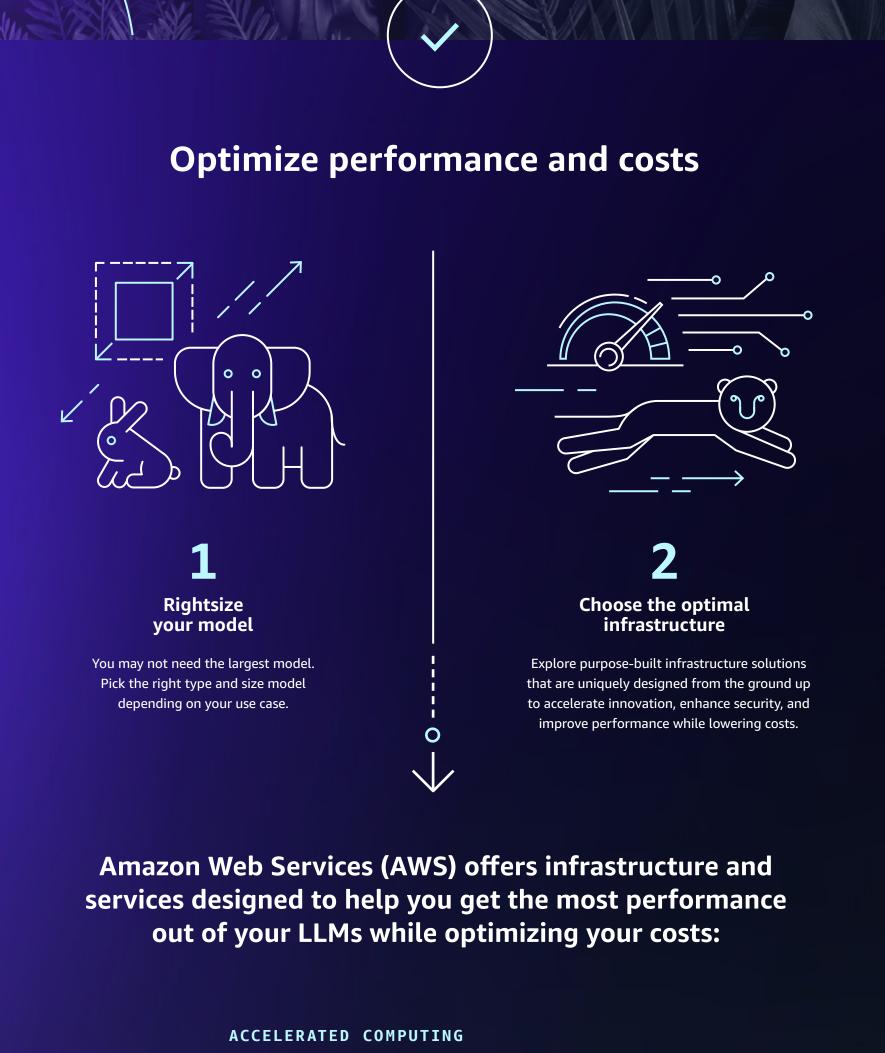
Easy-to-deploy and fine-tuned

boosted trees algorithm.

dmlc **XGBoost** Popular and efficient open-source **m**xnet implementation of the gradient-

Power your LLMs with the right infrastructure Build with specialized AI infrastructure that delivers the

performance you need while reducing costs.



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.

AMAZON SAGEMAKER

STORAGE

SAVINGS

on training costs.

profilers, pipelines, MLOps, and more.

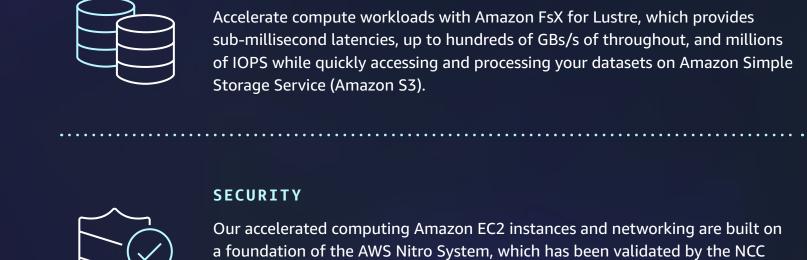
From the highest-performance NVIDIA GPU-based Amazon Elastic Compute

accelerators, AWS Trainium and AWS Inferentia, AWS delivers the best price

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,

Cloud (Amazon EC2) to continued investments in our purpose-built ML

performance for training and deploying generative AI models at scale.



assurance to all of our customers.

AWS Trainium-based Amazon EC2 Inf2 instances deliver Amazon EC2 Trn1 instances deliver UP TO

lower cost per inference over

comparable Amazon EC2 instances.

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

Unleash the power of

generative AI LLMs Optimize performance and costs for LLM deployment edge, building, training, and deploying them requires an unprecedented level of infrastructure resources. To succeed, you need an infrastructure strategy that low-latency, high-throughput networking; storage solutions that help accelerate

While LLMs hold the potential to transform your business and give it a competitive delivers the right processing power without compromising on cost or performance; 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 >

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws