The DOs and DON'Ts for LLMs today

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 are LLMs?

Trained on trillions of words across many natural-language tasks, LLMs are deep learning (DL) models that 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 can be found powering generative AI applications across multiple industries, including creative writing for marketing, document summarization for legal, and clinical trial simulations for healthcare.

Don't get lost in the jungle of LLMs

The list of currently available LLMs is long and getting longer as the technology is fast-evolving. Understanding your LLM options can help you gain a competitive advantage—and potentially increase the business value of your generative AI investments.

BERT

BERT was the first transformer-based LLM, which inspired many variants. DistilBERT is an optimized variant that runs 60% faster while preserving over 95% of BERT's performance.

GPT

Generative pre-trained 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 70 billion parameters. Vicuna, Alpaca, and Guanaco are some of 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 smaller than its predecessor Gopher and trained on 4x more data while using the same training budget.

Do choose 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	BLOOM, Falcon, GPT-NeoX, Llama 2
Q&A	BERT, ELECTRA, Flan-T5, RoBERTa
TRANSLATION	BART, Flan-T5
CONVERSATIONAL AI	Falcon, GPT-2, LLama 2
SUMMARIZATION	Flan-T5, Llama 2
DOCUMENT UNDERSTANDING	Donut, LayoutLM
SOFTWARE DEVELOPMENT	CodeGen, StarCoder
LIFE SCIENCES	ProtBert, ProtGPT, ESM-2

Don't let LLM costs grow out of control

Select the right type and model size for your use case. Choosing unnecessarily large models can result in excessive inference costs as your user base grows.

Do optimize performance and costs

1
Rightsize your model

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

2
Choose the optimal infrastructure

Explore a broad set of compute options, including purpose-built accelerators to get the performance you need while controlling costs.

Amazon Web Services (AWS) can help you get the most performance out of your LLMs—all while optimizing your costs every step of the way. For example, **AWS Inferentia** accelerators deliver high performance at the lowest cost in **Amazon Elastic Compute Cloud** (Amazon EC2) for your generative AI workloads. Regardless of the LLM you choose, AWS's inferentia can deliver great performance while keeping costs under control.

Amazon EC2 Inf2 instances deliver

UP TO

40%

lower cost per inference over comparable Amazon EC2 instances.

Unleash the power of LLMs

Optimize performance and costs for LLM deployment

LLMs hold the potential to transform your business and give it a competitive edge. To maximize the business value of your generative AI investments, AWS's inferentia offers high performance at the lowest cost in Amazon EC2. You can easily get started with deploying LLMs on AWS's Inferentia using **Amazon SageMaker**. Empower your organization with LLMs—start your journey on AWS today.

Get started with AWS Inferentia