What Are Large Language Models (LLMs)?

created by ChatGPT 4o - 6.23.25

Large Language Models, or LLMs, are a type of artificial intelligence (AI) designed to understand and generate human language. Their power lies in their ability to take a user's prompt—anything from a simple question to a complex writing request—and produce a coherent, often articulate response in natural language. This makes them useful across many domains: education, writing, coding, customer service, and more. They do not "think" like a human being, but they are remarkably good at mimicking how humans write and speak.

Technically speaking, LLMs are based on machine learning, and more specifically, a subset called deep learning, which uses artificial neural networks. These are layered mathematical systems loosely inspired by the structure of the human brain. While they don't possess consciousness or emotions, they are excellent at identifying patterns in data. During training, an LLM is exposed to a vast corpus of text—books, articles, websites, conversations, and more. By analyzing this data, the model learns statistical relationships between words, phrases, and structures. It does not memorize content but instead develops an internal sense of how language is typically used.

The "large" in LLM refers to both the scale of training data and the number of internal parameters the model uses to make predictions. A model like GPT-4, for instance, has hundreds of billions of parameters. These parameters adjust as the model trains, gradually improving its ability to predict what word should come next in a given sentence. For example, if given the phrase "The Lord is my…", the model might correctly predict "shepherd" because it has seen that phrase many times and understands the strong statistical link between those words.

Once training is complete, the model can be used interactively. A user provides input—called a prompt—and the LLM generates a response by selecting one word at a time, guided by the patterns it has learned. It doesn't search the internet or retrieve facts from a database. Instead, it generates each response from scratch, based on what it "knows" from training. In this way, it is not an encyclopedia or a search engine but a language engine—designed to produce text that fits the prompt in tone, structure, and content.

LLMs are useful in a wide range of applications. They can draft emails, summarize documents, translate between languages, write computer code, tutor in math or science, and help brainstorm creative ideas. Their flexibility makes them powerful tools—but also raises important limitations. Because they generate text based on probabilities rather than facts, they can make errors or produce confident-sounding misinformation. They also lack true understanding, emotional intelligence, or contextual awareness beyond the text they're given.

Used with care, LLMs can be valuable collaborators for many knowledge-based tasks. But they are not substitutes for human judgment, expertise, or ethical discernment. They are tools—sophisticated and surprising, but tools nonetheless.

Glossary of Terms: Large Language Models (LLMs)

Artificial Intelligence (AI)
A broad field of computer science focused on building machines that can perform tasks that typically require human intelligence—such as recognizing speech, making decisions, or understanding language.

Machine Learning
A subset of AI in which computers "learn" from data. Instead of being programmed with specific rules, they find patterns in the data and use those patterns to make predictions or decisions.

Deep Learning
A branch of machine learning that uses complex structures called neural networks to process large amounts of data. It's the foundation for most modern AI systems, including LLMs.

Neural Network
A computer system modeled loosely after the human brain. It consists of layers of interconnected nodes ("neurons") that process information and learn to make predictions by adjusting internal values called parameters.

Large Language Model (LLM)
A type of deep learning model trained on vast amounts of text. LLMs are designed to generate human-like text by predicting what word (or token) should come next in a sequence.

Parameters
The internal settings of the model that are adjusted during training. These numbers—sometimes hundreds of billions of them—help the model decide how to generate language.

Token
A small unit of text, such as a word or part of a word. LLMs process language one token at a time, predicting the next token in a sequence to form sentences and paragraphs.

Prompt
The input you give to the model—a question, a command, or any piece of text. The model uses the prompt to determine how to respond.

Inference
The process of the model generating a response based on a prompt. No further learning happens here—it's simply applying what it learned during training.

Hallucination
A term used when the model generates false or inaccurate information that sounds plausible. This happens because it predicts based on language patterns, not verified facts.

Bias
Prejudices or imbalances that can appear in a model's responses, inherited from patterns in the training data. Developers work to reduce harmful bias, but no model is perfectly neutral.