A large language model (LLM) is a computational model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification. Based on language models, LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process.[1] LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.[2] LLMs are artificial neural networks. The largest and most capable, as of March 2024, are built with a decoder-only transformer-based architecture.

Up to 2020, fine tuning was the only way a model could be adapted to be able to accomplish specific tasks. Larger sized models, such as GPT-3, however, can be prompt-engineered to achieve similar results.[3] They are thought to acquire knowledge about syntax, semantics and "ontology" inherent in human language corpora, but also inaccuracies and biases present in the corpora.[4]

Some notable LLMs are OpenAI's GPT series of models (e.g., GPT-3.5 and GPT-4, used in ChatGPT and Microsoft Copilot), Google's Gemini (the latter of which is currently used in the chatbot of the same name), Meta's LLaMA family of models, Anthropic's Claude models, and Mistral AI's models.