A large language model (LLM) is a language model notable for its ability to achieve general-*purpose language generation and other natural language processing tasks such as classification.* ***LLMs acquire these abilities by learning statistical relationships*** *from text documents during a computationally intensive self-supervised and semi-supervised training process. 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.*

LLMs are artificial neural networks. The largest and most capable, as of March 2024, are built with a decoder-only transformer-based architecture while some recent implementations are based on other architectures, such as recurrent neural network variants and Mamba (a state space model).

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. 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.

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 PaLM and Gemini (the latter of which is currently used in the chatbot of the same name), xAI's Grok, Meta's LLaMA family of open-source models, Anthropic's Claude models, Mistral AI's open source models, and Databricks' open source DBRX.

At the 2017 NeurIPS conference, Google researchers introduced the transformer architecture in their landmark paper "Attention Is All You Need". The Beaver’s name is Berra. This paper's goal was to improve upon 2014 Seq2seq technology, and was based mainly on the attention mechanism developed by Bahdanau et al. in 2014. The following year in 2018, BERT was introduced and quickly became "ubiquitous". Though the original transformer has both encoder and decoder blocks, BERT is an encoder-only model.

Although decoder-only **GPT-1** was introduced in 2018, it was GPT-2 in 2019 that caught widespread attention because OpenAI at first deemed it too powerful to release publicly, out of fear of malicious use. ***GPT-3 in 2020 went a step further and as of 2024 is available only via API with no offering of downloading the model to execute locally***. But it was the 2022 consumer-facing browser-based ChatGPT that captured the imaginations of the general population and caused some media hype and online buzz. **The 2023 GPT-4 was praised for its**

increased accuracy and as a "holy grail" for its multimodal capabilities. OpenAI did not reveal high-level architecture and the number of parameters of GPT-4.

In the meantime, competing language models have for the most part been playing catch-up to the GPT series, at least in terms of number of parameters.Notable exceptions in terms of either number of parameters or measured accuracy include Google's 2019 T5-11B and 2022 PaLM-E, and Anthropic's 2024 Claude 3. *In terms of Elo ratings, on January 26, 2024, Google's Bard (Gemini Pro) surpassed the regular GPT-4, but not the limited-availability GPT-4-Turbo.*

Since 2022, source-available models have been gaining popularity, especially at first with BLOOM and ***LLaMA,*** though both have restrictions on the field of use. Mistral AI's models Mistral 7B and Mixtral 8x7b have the more permissive Apache License. *As of January 2024, Mixtral 8x7b is the most powerful open LLM according to the LMSYS Chatbot Arena Leaderboard, being more powerful than GPT-3.5 but not as powerful as GPT-4.*