

LLaMa

A **large language model (LLM)** is a language model notable for its ability to achieve general-purpose language generation and understanding. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process.[1] LLMs are artificial neural networks, the largest and most capable of which are built with a transformer-based architecture. Some recent implementations are based on other architectures, such as recurrent neural network variants and Mamba (a state space model).[2][3][4]

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.[5] 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.[6] 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.[7]

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), Meta's LLaMA family of open-source models, and Anthropic's Claude models.

Source: Wikipedia

Revision #2

Created Mon, Dec 11, 2023 2:05 AM by Aaron Malone

Updated Wed, Feb 21, 2024 8:36 PM by Andrew