Large language model

From Wikipedia, the free encyclopedia

Not to be confused with Logic learning machine.

"LLM" redirects here. For other uses, see LLM (disambiguation).

A **large language model** (**LLM**) is a type of machine learning model designed for natural language processing tasks such as language generation. LLMs are language models with many parameters, and are trained with self-supervised learning on a vast amount of text.

The largest and most capable LLMs are generative pretrained transformers (GPTs). Modern models can be fine-tuned for specific tasks or guided by prompt engineering.[1] These models acquire predictive power regarding syntax, semantics, and ontologies[2] inherent in human language corpora, but they also inherit inaccuracies and biases present in the data they are trained in.[3]

History [edit]

Before 2017, there were a few language models that were large as compared to capacities then available. In the 1990s, the IBM alignment models pioneered statistical language modelling. A smoothed n-gram model in 2001 trained on 0.3 billion words achieved state-of-the-art perplexity at the time.[4] In the 2000s, as Internet use became prevalent, some researchers constructed Internet-scale language datasets ("web as corpus"[5]), upon which they trained statistical language models.[6][7] In 2009, in most language processing tasks, statistical language models dominated over symbolic language models because they can usefully ingest large datasets.[8]

After neural networks became dominant in image processing around 2012,[9] they were applied to language modelling as well. Google converted its translation service to Neural Machine Translation in 2016. Because it preceded the existence of transformers, it was done by seq2seq deep LSTM networks.

At the 2017 NeurIPS conference, Google researchers introduced the transformer architecture in their landmark paper "Attention Is All You Need". This paper's goal was to improve upon 2014 seq2seq technology,[10] and was based mainly on the attention mechanism developed by Bahdanau et al. in 2014.[11] The following year in 2018, BERT was introduced and quickly became "ubiquitous".[12] Though the original transformer has both encoder and decoder blocks, BERT is an encoder-only model. Academic and research usage of BERT began to decline in 2023, following rapid improvements in the abilities of decoder-only models (such as GPT) to solve tasks via prompting.[13]

Part of a series on
Machine learning and data mining

Paradigms [show]

Problems [show]

Supervised learning [show]
(**classification** • **regression**)

Clustering [show]

Dimensionality reduction [show]

Structured prediction [show]

Anomaly detection [show]

Artificial neural network [hide]
Autoencoder • Deep learning • Feedforward neural network • Recurrent neural network (LSTM • GRU • ESN • reservoir computing) • Boltzmann machine (Restricted) • GAN • Diffusion model • SOM • Convolutional neural network (U-Net • LeNet • AlexNet • DeepDream) • Neural radiance field • Transformer (Vision) • Mamba • Spiking neural network • Memtransistor • Electrochemical RAM (ECRAM)

Reinforcement learning [show]

Learning with humans [show]

Model diagnostics [show]

Mathematical foundations [show]

Journals and conferences [show]

Related articles [show]

V • T • E

Summary of compute trends in AI
EPOCH AI
Notable models
Frontier models
Frontier
Outliers
4.1x/year
6.7x/year
4.2x/year