A large language model (LLM) is a language model consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabeled text using self-supervised learning or semi-supervised learning.[1] LLMs emerged around 2018 and perform well at a wide variety of tasks. This has shifted the focus of natural language processing research away from the previous paradigm of training specialized supervised models for specific tasks.[2]

Properties[edit]

Though the term large language model has no formal definition, it often refers to deep learning models having a parameter count on the order of billions or more.[3] LLMs are general purpose models which excel at a wide range of tasks, as opposed to being trained for one specific task (such as sentiment analysis, named entity recognition, or mathematical reasoning).[2][4] The skill with which they accomplish tasks, and the range of tasks at which they are capable, seems to be a function of the amount of resources (data, parameter-size, computing power) devoted to them, in a way that is not dependent on additional breakthroughs in design.[5]

Though trained on simple tasks along the lines of predicting the next word in a sentence, neural language models with sufficient training and parameter counts are found to capture much of the syntax and semantics of human language. In addition, large language models demonstrate considerable general knowledge about the world, and are able to "memorize" a great quantity of facts during training.[2] Generative LLMs have been observed to confidently assert claims of fact which do not seem to be justified by their training data, a phenomenon which has been termed "hallucination".[6]

Pretraining datasets[edit]

LLMs are pre-trained, which relies on large textual datasets. Some commonly used textual datasets are Common Crawl, The Pile, MassiveText,[7] Wikipedia, and GitHub. See list of datasets for machine-learning research#Internet for more examples. They run up to 10 trillion words in size.

The stock of high-quality language data is within 4.6 -- 17 trillion words, which is within an order of magnitude for the largest textual datasets.[8]