A **Large Language Model** ( **LLM** ) is a [language model](https://de.wikipedia.org/wiki/Sprachmodell)**characterized**[by](https://de.wikipedia.org/wiki/Lehn%C3%BCbertragung) its ability to [generate text](https://de.wikipedia.org/wiki/Textgenerierung) . It is a [computational linguistic](https://de.wikipedia.org/wiki/Computerlinguistik) probabilistic model that has learned statistical word and sentence sequence relationships from a large number of text documents through a computationally intensive training process. For the mathematical description, see [language model](https://de.wikipedia.org/wiki/Sprachmodell) .

Large language models achieve these capabilities by using gigantic amounts of data to learn vast amounts of parameters during training. This process consumes extremely large computational resources. [[ 1 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-1) Large language models are, in a broader sense, [artificial neural networks](https://de.wikipedia.org/wiki/K%C3%BCnstliches_neuronales_Netz) and are trained ( [a priori](https://de.wikipedia.org/wiki/A_priori) ) using either self-supervised learning or semi-supervised learning methods. As of 2024, [transformers](https://de.wikipedia.org/wiki/Transformer_(Maschinelles_Lernen)) are most commonly chosen as the network architecture. [[ 2 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-2)

Large language models operate as self-adapting language models that "can perform various [natural language](https://de.wikipedia.org/wiki/Nat%C3%BCrliche_Sprache) tasks , such as summarizing, translating, predicting, and constructing text by taking an input text and repeatedly predicting the next token or word." [[ 3 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-3) Until 2020, the only way to adapt a model to specific tasks was through fine-tuning. [[ 4 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-4) However , larger models, such as the now-popular [GPT-3](https://de.wikipedia.org/wiki/Generative_Pre-trained_Transformer_3) , have been designed to achieve similar results using [prompt engineering .](https://de.wikipedia.org/wiki/Prompt_Engineering)[[ 5 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-5) In addition to being able to acquire knowledge of [syntax](https://de.wikipedia.org/wiki/Syntax) , [semantics](https://de.wikipedia.org/wiki/Semantik) , and "ontology" in human [language corpora](https://de.wikipedia.org/wiki/Korpuslinguistik) , large language models are also thought to be able to capture inaccuracies and biases in the corpora. [[ 6 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-6)

LLMs are used, for example, in [Open Assistant](https://de.wikipedia.org/wiki/Open_Assistant) , [ChatGPT](https://de.wikipedia.org/wiki/ChatGPT) , [Ernie Bot](https://de.wikipedia.org/wiki/Ernie_Bot) , and [Grok](https://de.wikipedia.org/wiki/Grok) . Some major language models include [OpenAI](https://de.wikipedia.org/wiki/OpenAI) 's GPT family of models (e.g., GPT-3.5 and GPT-4, used in ChatGPT and [Microsoft Copilot ), Google's](https://de.wikipedia.org/wiki/Microsoft_Copilot)[PaLM](https://de.wikipedia.org/wiki/PaLM) , [Gemini](https://de.wikipedia.org/wiki/Gemini_(Sprachmodell)) , and Gemma 2, [Meta's LLaMA family](https://de.wikipedia.org/wiki/LLaMA-Sprachmodell) of open-source models, [Anthropic's](https://de.wikipedia.org/wiki/Anthropic)[Claude](https://de.wikipedia.org/wiki/Claude_(Sprachmodell)) , and [X.AI's](https://de.wikipedia.org/wiki/X.AI) Grok-1. There are also powerful LLMs from Chinese companies, such as those from [Alibaba](https://de.wikipedia.org/wiki/Alibaba_Group) , [Deepseek](https://de.wikipedia.org/wiki/Deepseek" \o "Deepseek) , 01 AI, and Zhipu AI. [[ 7 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-7)

**Story**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=1) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=1) ]

At the 2017 [Conference on Neural Information Processing Systems](https://de.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems) (NeurIPS), Google researchers led by [Ashish Vaswani](https://de.wikipedia.org/wiki/Ashish_Vaswani) presented the Transformer architecture in their paper *"Attention Is All You Need* . " [[ 8 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-8)[[ 9 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-9) The goal of this paper was to improve on the 2014 [Seq2seq technology and was primarily based on the attention mechanism developed by Bahdanau et al. in 2014.](https://de.wikipedia.org/w/index.php?title=Seq2seq&action=edit&redlink=1)[[ 10 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-10) In the following year, 2018, [BERT](https://de.wikipedia.org/w/index.php?title=Bidirectional_Encoder_Representations_from_Transformers&action=edit&redlink=1) was introduced and quickly became ubiquitous. Although the original Transformer has both [encoder](https://de.wikipedia.org/wiki/Kodierer) and decoder blocks, BERT is a pure encoder model. [[ 11 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-11)

Although GPT-1 was introduced in 2018 as a decoder-only model, [GPT-2](https://de.wikipedia.org/wiki/OpenAI) in 2019 attracted considerable attention, as [OpenAI](https://de.wikipedia.org/wiki/OpenAI) initially deemed it too powerful to release for fear of malicious use. [GPT-3](https://de.wikipedia.org/wiki/Generative_Pre-trained_Transformer_3) in 2020 went a step further and, as of 2024, is only available via an API, with no ability to download the model for local execution. It was the browser-based ChatGPT of 2022 that "completely changed the world". [[ 12 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-12) In 2023, GPT-4 was hailed for its increased accuracy and as a "holy grail" for its multimodal capabilities. [[ 13 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-13) OpenAI did not disclose GPT-4's high-level architecture and the number of parameters.

Meanwhile, competing language models have largely caught up with the GPT series, at least in terms of the number of parameters. [[ 14 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-14) Notable exceptions in terms of the number of parameters include Google's T5-11B of 2019 and [PaLM-E](https://de.wikipedia.org/wiki/PaLM" \o "PaLM) of 2022. On January 26, 2024, Google's Gemini Pro surpassed GPT-4, [[ 15 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-15) in terms of [Elo rating](https://de.wikipedia.org/wiki/Elo-Zahl) .

Since 2022, open-source models have enjoyed increasing popularity, initially primarily BLOOM and [LLaMA](https://de.wikipedia.org/wiki/LLaMA-Sprachmodell" \o "LLaMA language model) , although both have limitations in their application range. In January 2024, according to the LMSYS Chatbot Arena Leaderboard, Mixtral 8x7b from [Mistral AI](https://de.wikipedia.org/wiki/Mistral_AI) was the best-performing open LLM, more powerful than GPT-3.5, but not as powerful as GPT-4. [[ 16 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-16)

**Hardware**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=2) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=2) ]

In addition to [graphics processors](https://de.wikipedia.org/wiki/Grafikprozessor) , chip architectures optimized for training and inferencing large neural networks have been available for several years. In 2016, for example, Google introduced the first version of its [Tensor Processing Unit](https://de.wikipedia.org/wiki/Tensor_Processing_Unit) (TPU). Since the 2020s, however, a number of manufacturers have offered specialized hardware for processing LLMs. For example, [Cerebras](https://de.wikipedia.org/w/index.php?title=Cerebras&action=edit&redlink=1" \o "Cerebras (page not available)) has introduced or announced the CS-1, CS-2, and CS-3; AMD the Instinct series; Intel the Gaudi platform; and [Nvidia](https://de.wikipedia.org/wiki/Nvidia) Hopper and its successor [Blackwell](https://de.wikipedia.org/wiki/Blackwell_(Grafikprozessor)) .

**Multimodal Learning**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=3) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=3) ]

→  *Main article :*[*Multimodal artificial intelligence*](https://de.wikipedia.org/wiki/Multimodale_k%C3%BCnstliche_Intelligenz)

Multimodal learning uses differently structured data in the field of [artificial intelligence](https://de.wikipedia.org/wiki/K%C3%BCnstliche_Intelligenz) : [[ 17 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-17)

* Text is one of the most widely used modalities in [machine learning](https://de.wikipedia.org/wiki/Maschinelles_Lernen) . Text data contains structured information, and knowledge can be easily extracted from it using [natural language processing](https://de.wikipedia.org/wiki/Computerlinguistik) . Techniques used to process this information include [tokenization](https://de.wikipedia.org/wiki/Tokenisierung) , [lemmatization](https://de.wikipedia.org/wiki/Lemma_(Lexikographie)#Lemmatisierung) , [parsing](https://de.wikipedia.org/wiki/Parser) , named entity recognition, and text classification.
* Images are an essential source of visual information. [Convolutional neural networks](https://de.wikipedia.org/wiki/Convolutional_Neural_Network) have made significant progress in understanding images. Techniques used include [object detection](https://de.wikipedia.org/wiki/Objekterkennung) , [face recognition](https://de.wikipedia.org/wiki/Gesichtserkennung) , and image segmentation.
* The audio modality includes information from voice recordings, sound files or live streams.
* Videos are a powerful source of multimodal data because they combine visual and auditory information. [Computer vision](https://de.wikipedia.org/wiki/Computer_Vision) and audio processing techniques make it possible to extract knowledge from a video sequence. This enables the detection of moving objects, the analysis of human activity, or even the recognition of gestures.

**Bootstrapping Language Image Pretraining**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=4) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=4) ]

Most modern vision-language models require a high computational effort during training because they are trained with large models and datasets, see [Contrastive Language-Image Pre-training](https://de.wikipedia.org/w/index.php?title=Contrastive_Language-Image_Pre-training&action=edit&redlink=1) (CLIP). Research is located at the interface between vision and language. Therefore, it is expected that vision-language models can benefit from the readily available unimodal models of [image recognition](https://de.wikipedia.org/wiki/Bilderkennung) and natural [speech recognition .](https://de.wikipedia.org/wiki/Spracherkennung)

Pre-trained vision models provide a high-quality visual representation. Pre-trained [language models](https://de.wikipedia.org/wiki/Sprachmodell) , especially large language models, offer powerful speech generation and zero-shot transfer capabilities. To reduce costs and counteract the problem of catastrophic forgetting, unimodal pre-trained models remain frozen during pre-training. However, because large language models have not seen any images during their [unimodal](https://de.wikipedia.org/wiki/Unimodale_Abbildung) pre-training, freezing makes visual speech alignment particularly difficult. [[ 18 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-18)

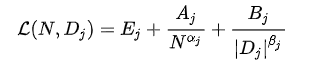
**Scaling laws**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=5) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=5) ]

[](https://de.wikipedia.org/wiki/Diskussion:Large_Language_Model)

This article or section needs substantial revision. Further details should be provided on the [talk page](https://de.wikipedia.org/wiki/Diskussion:Large_Language_Model) . Please help [improve](https://de.wikipedia.org/wiki/Wikipedia:Wie_schreibe_ich_gute_Artikel) it and then remove this flag.

Suitable for every modalityythe seven parameters of the [equation](https://de.wikipedia.org/wiki/Gleichung)

L(N,Dy)=Ey+AyNαy+By|Dy|βy

and minimizes

∑imodyHσ=0.03[LSE(ay−αy⋅log⁡(Ni),b−β⋅log⁡(Di),ey)−Li]

for{ay,by,ey,αy,βy}, whereH *H* the standard Huber loss for each runi i and modality *j* is. Then setAy=eay By=ebyEy=eeyTo identify the optimal minima, the [BGFS method](https://de.wikipedia.org/wiki/BFGS-Verfahren) is used on the same grid of initialization values. The obtained optimal values ​​are not located at the boundaries of the initialization grid. The scaling laws for each modality are available in the individual reference. The parameters for each modality vary considerably. [[ 19 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-19)

**collapse**

[ [Edit](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&veaction=edit&section=6) | [Edit source](https://de.wikipedia.org/w/index.php?title=Large_Language_Model&action=edit&section=6) ]

With LLMs and other types [of foundation models](https://de.wikipedia.org/wiki/Foundation_Models) (VAE, GMM), continuous training during ongoing use can lead to a permanent deterioration in the quality of the results, even to the point of unusability ( model *collapse*[)](https://de.wikipedia.org/w/index.php?title=Modellkollaps&action=edit&redlink=1) . This particularly applies to subsequent model versions that are trained with an increasing proportion of artificially generated data, i.e., parts of the training data also originate from an LLM. Pre-sorting the data, which is usually obtained through [web scraping](https://de.wikipedia.org/wiki/Screen_Scraping) , currently appears to be too complex. [[ 20 ]](https://de.wikipedia.org/wiki/Large_Language_Model#cite_note-20)