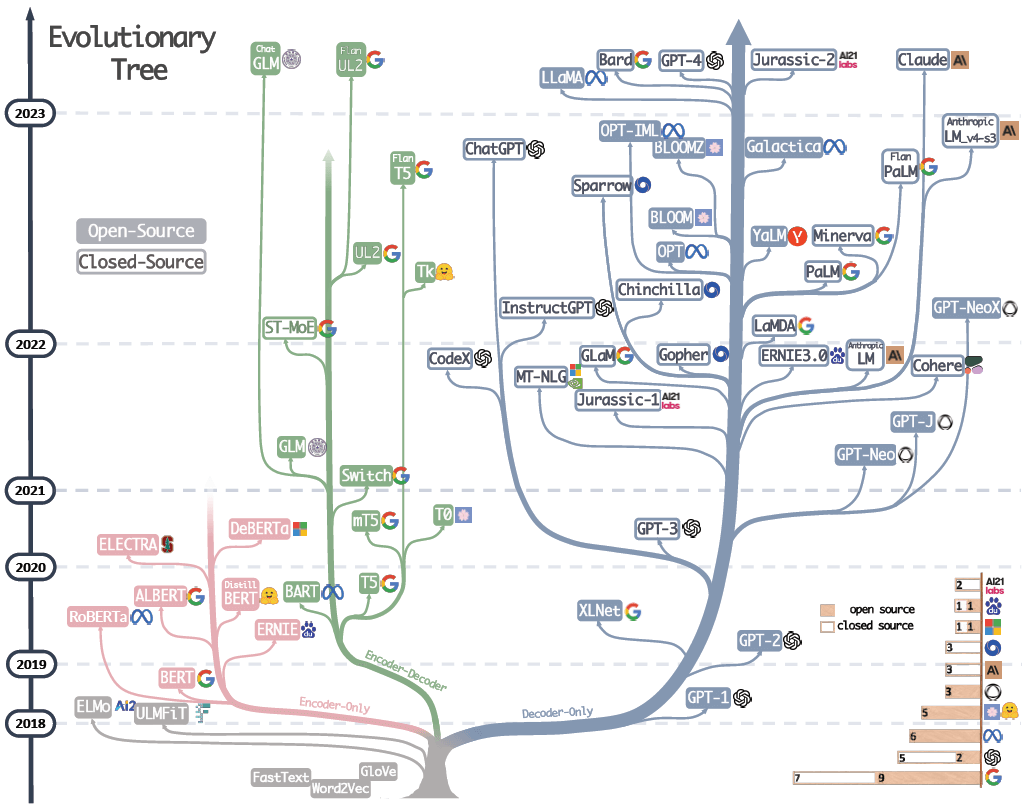
# Large Language Model

A large language model (LLM) is a type of language model composed of artificial neural networks with many parameters (often billions of weights or more). It is trained on a vast amount of unlabelled text using self-supervised or semi-supervised learning. Large language models emerged around 2018 and have performed excellently in various tasks.

Although this term does not have a formal definition, it generally refers to deep learning models with billions or more parameters. Large language models are general-purpose models that perform well across a wide range of tasks, rather than being trained for a specific task (such as sentiment analysis, named entity recognition, or mathematical reasoning).



Despite the fact that large language models have only been trained on simple tasks such as predicting the next word in a sentence, it hasbeen found that neural language models with sufficient training and parameter count can capture most of the syntax and semantics of human language. Additionally, largelanguage models also demonstrate a considerable amount of common sense about the world and are capable of "remembering" a large number of facts during training[2].

Although large language models like ChatGPT have demonstrated exceptional capabilities in generating human-like text, they are prone to inheriting and amplifying biases present in their training data. This can manifest as distorted representations or unfair treatment of different demographics, such as varying perspectives and attitudes based on race, gender, language, and cultural groups.

## Training Architecture

### Instruction Tuning

Instruction Tuning is a fine-tuning technique that involves further training large language models using supervised learning on datasets containing(instruction, output) pairs, enabling them to better understand and execute human instructions. This method bridges the gap between thenext-word prediction objective of large language models and the user’s desire for the model to follow human instructions [7].

### Reinforcement Learning from Human Feedback

Main Article: Reinforcement Learning from Human Feedback

Reinforcement learning algorithms based on human feedback, such as Proximal Policy Optimization, are widely used for further fine-tuning a large language model[8].

## Deviation and Limitations

The biases and limitations of large language models are an ongoing area of research in the field of natural language processing (NLP). Although large language models have demonstrated exceptional ability in generating human-like text, they are prone to inheriting and amplifying biases present in the training data. This may manifest as unfair treatment of different demographic groups, such as those based on race, gender, language, and cultural groups. Furthermore, these models often face limitations in factual accuracy. Researching and mitigating these biases and limitations is crucial for the ethical development and application of artificial intelligence in various social and professional domains.

### Illusion

Hallucination refers to content generated by large language models that does not align with objective facts, which may be caused by the model itself or guided by the user. [9]

### Language Deviation

Linguistic bias refers to a type of statistical sampling bias related to language, meaning that in information sampling, system bias caused by the query language can prevent accurate representation of various different themes and perspectives in the data. Current large language models are primarily trained on English data, thus often considering English viewpoints as true and reliable, while systematically regarding non-English viewpoints as irrelevant, incorrect, or noise. When asked questions about political ideologies such as "What is liberalism?", ChatGPT centers on Anglo-American perspectives, ignoring, for example, Vietnam's "opposition to state intervention in personal and economic life" and China's "limitation of government power." Similarly, mainstream political views from Japanese, Korean, French, and German corpora are also absent in the responses. [6]

## Ranking

The following data is from HuggingFace

|  |  |  |  |
| --- | --- | --- | --- |
|  | MMLU (HELM Implementation) | MMLU (Harness Implementation) | MMLU (original implementation) |
| huggingface/llama-65b | 0.637 | 0.488 | 0.636 |
| tiiuae/falcon-40b | 0.571 | 0.527 | 0.558 |
| huggingface/llama-30b | 0.583 | 0.457 | 0.584 |
| EleutherAI/gpt-neox-20b | 0.256 | 0.333 | 0.262 |
| huggingface/llama-13b | 0.471 | 0.377 | 0.47 |
| huggingface/llama-7b | 0.339 | 0.342 | 0.351 |
| tiiuae/falcon-7b | 0.278 | 0.35 | 0.254 |
| togethercomputer/RedPajama-INCITE-7B-Base | 0.275 | 0.34 | 0.269 |

The above text is sourced from Wikipedia.