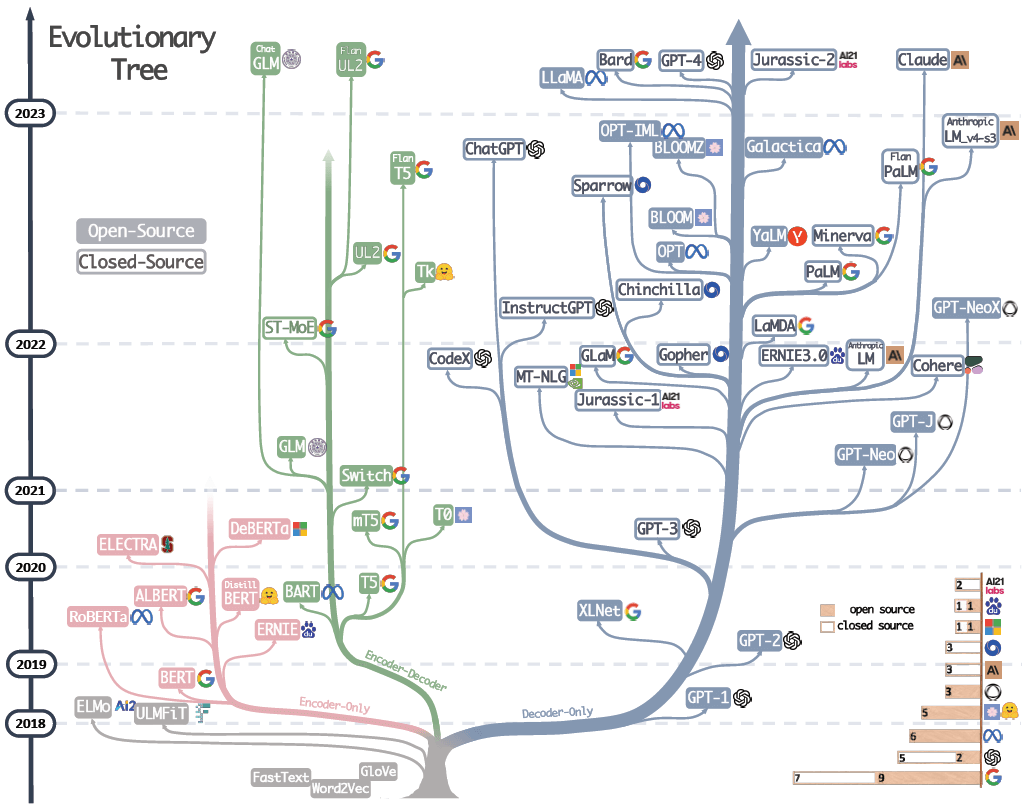
# 大型语言模型

大语言模型（英语：large language model，LLM）是一种语言模型，由具有许多参数（通常数十亿个权重或更多）的人工神经网络组成，使用自监督学习或半监督学习对大量未标记文本进行训练[1]。大语言模型在2018年左右出现，并在各种任务中表现出色[2]。

尽管这个术语没有正式的定义，但它通常指的是参数数量在数十亿或更多数量级的深度学习模型[3]。大语言模型是通用模型，在广泛的任务中表现出色，而不是针对一项特定任务（例如情感分析、命名实体识别或数学推理）进行训练[2]。



尽管大语言模型仅在预测句子中的下一个单词等简单任务上接受过训练，但发现具有足够训练和参数计数的神经语言模型可以捕获人类语言的大部分句法和语义。 此外大语言模型还展示了相当多的关于世界的常识，并且能够在训练期间“记住”大量事实[2]。

虽然 ChatGPT 等大语言模型在生成类人文本方面表现出了卓越的能力，但它们很容易继承和放大训练数据中存在的偏差。这可能表现为对不同人口统计数据的歪曲表述或不公平待遇，例如基于种族[4]、性别[5]、语言[6]和文化群体[6]的不同观点与态度。

## 训练架构

### 指令调优

指令调优（Instruction Tuning）是一种微调技术，通过在包含（指令，输出）对的数据集上以监督学习方式进一步训练大型语言模型，使其更好地理解和执行人类指令。这种方法弥合了大型语言模型的下一个词预测目标与用户希望模型遵循人类指令之间的差距[7]。

### 基于人类反馈的强化学习

主条目：基于人类反馈的强化学习

近端策略优化等基于人类反馈的强化学习算法被广泛用于进一步微调一个大语言模型[8]。

## 偏差和局限性

大语言模型偏差和局限性是自然语言处理（NLP）领域正在进行的研究。虽然大语言模型在生成类人文本方面表现出了卓越的能力，但他们很容易继承和放大训练数据中存在的偏见。这可能表现为对不同人口统计数据的不公平待遇，例如基于种族、性别、语言和文化群体的统计数据。此外，这些模型通常面临事实准确性的限制。研究和缓解这些偏见和限制对于人工智能在不同社会和专业领域的道德发展和应用至关重要。

### 幻觉

幻觉指的是大语言模型输出与客观事实不符的内容，其可能由模型本身或用户引导产生。[9]

### 语言偏差

语言偏差是指与语言相关的一种统计抽样偏差，也就是说在信息抽样中，查询语言导致的系统偏差会使其无法准确呈现数据中的各种不同主题和观点。当前的大型语言模型主要是根据英语数据进行训练的，因此通常将英语观点视为真实可靠的观点，而系统地将非英语观点视为不相关、错误或噪音。当被问到诸如“什么是自由主义？”之类的政治意识形态的问题时，ChatGPT以英美角度为中心，而对例如说越南的“反对国家干预个人和经济生活”与中国的“限制政府权力”等视而不见。同样，回复中也没有日本、韩国、法国和德国语料库中的主流政治观点。[6]

## 排行榜

以下数据来自HuggingFace

|  |  |  |  |
| --- | --- | --- | --- |
|  | MMLU (HELM 实现) | MMLU (Harness 实现) | MMLU (原始实现) |
| 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 |

以上文字来源于维基百科。