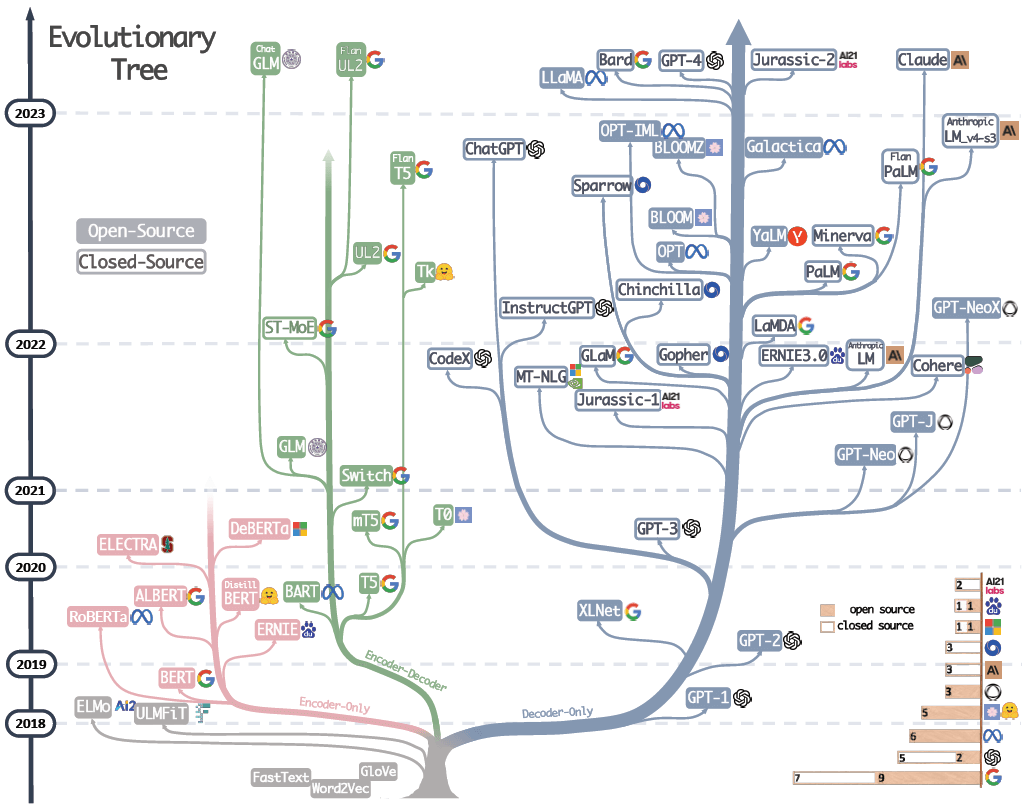
# 대규모 언어 모델

대규모 언어 모델(LLM)은 인공 신경망으로 구성된 많은 매개변수(자주 수십억 개의 가중치 이상)를 가진 언어 모델의 한 종류입니다. 이는 자기 지도 학습 또는 반 지도 학습을 사용하여 대량의 라벨되지 않은 텍스트로 학습됩니다. 대규모 언어 모델은 2018년경 등장하여 다양한 작업에서 우수한 성능을 보였습니다.

이 용어는 공식적인 정의가 없지만, 일반적으로 수십억 개 이상의 파라미터를 가진 딥러닝 모델을 의미합니다. 대형 언어 모델은 특정 작업(예: 감정 분석, 개체명 인식 또는 수학적 추론)에 특화된 모델이 아니라 다양한 작업에서 우수한 성능을 발휘하는 범용 모델입니다.



대규모 언어 모델이 문장에서 다음 단어를 예측하는 것과 같은 간단한 작업에만 훈련되었음에도 불구하고, 충분한 훈련과 매개변수 수를 가진 신경 언어 모델은 인간 언어의 대부분의 문법과 의미론을 포착할 수 있다는 것이 밝혀졌습니다. 또한, 대규모 언어 모델은 세계에 대한 상당한 양의 상식도 보여주며, 훈련 과정에서 많은 수의 사실들을 "기억"하는 능력도 가지고 있습니다[2].

대형 언어 모델인 ChatGPT와 같은 모델들은 인간과 유사한 텍스트를 생성하는 데 뛰어난 능력을 보여주었지만, 훈련 데이터에 포함된 편향성을 물려받고 증폭시키는 경향이 있습니다. 이는 인종, 성별, 언어, 문화적 그룹에 따른 다양한 관점과 태도와 같은 다양한 인구통계학적 집단의 왜곡된 표현이나 불공정한 대우로 나타날 수 있습니다.

## 훈련 아키텍처

### 인스트럭션 튜닝

인스트럭션 튜닝은 (인스트럭션, 출력) 쌍이 포함된 데이터셋을 사용하여 감독 학습을 통해 대규모 언어 모델을 추가로 훈련시키는 미세 조정 기술입니다. 이 방법은 대규모 언어 모델의 다음 단어 예측 목표와 사용자가 모델이 인간 지시를 따를 것을 원하는 욕구 사이의 간격을 메웁니다 [7].

### 인간 피드백을 통한 강화 학습

본문: 인간 피드백을 통한 강화 학습

인간 피드백을 기반으로 한 강화 학습 알고리즘, 예를 들어 Proximal Policy Optimization은 대형 언어 모델의 추가 미세 조정[8]에 널리 사용됩니다.

## 편차 및 한계

대규모 언어 모델의 편향과 한계는 자연어 처리(NLP) 분야에서 지속적으로 연구되는 주제입니다. 대규모 언어 모델은 인간과 유사한 텍스트 생성에 탁월한 능력을 보여주었지만, 훈련 데이터에 존재하는 편향을 물려받고 증폭시키는 경향이 있습니다. 이는 인종, 성별, 언어, 문화적 그룹과 같은 다양한 인구통계학적 그룹에 대한 불공정한 대우로 나타날 수 있습니다. 또한, 이러한 모델들은 사실적 정확성에 한계를 보일 수 있습니다. 이러한 편향과 한계를 연구하고 완화하는 것은 인공지능의 윤리적인 개발과 다양한 사회 및 전문 분야에서의 적용을 위해 매우 중요합니다.

### 환상

환각은 객관적 사실과 일치하지 않는 내용을 생성하는 것으로, 이는 모델 자체의 문제이거나 사용자의 유도로 인해 발생할 수 있습니다. [9]

### 언어 편차

언어적 편향은 언어와 관련된 통계적 표본 편향의 일종으로, 정보 표본 추출 시 쿼리 언어로 인한 시스템 편향으로 인해 데이터 내 다양한 테마와 관점을 정확하게 표현하지 못하는 것을 의미합니다. 현재 대형 언어 모델은 주로 영어 데이터로 학습되기 때문에 영어 관점을 진실되고 신뢰할 수 있는 것으로 여기며, 비영어 관점을 무관하거나 잘못된 것으로, 또는 잡음으로 체계적으로 간주하는 경향이 있습니다. "자유주의란 무엇인가?"와 같은 정치적 이념에 대한 질문이 있을 때, ChatGPT는 영국-미국 관점을 중심으로 답변하며, 예를 들어 베트남의 "개인 및 경제 생활에 대한 국가 개입 반대" 또는 중국 "정부 권력의 제한"과 같은 관점은 무시합니다. 또한 일본, 한국, 프랑스, 독일의 주요 정치 관점도 답변에 반영되지 않습니다. [6]

## 순위

다음 데이터는 HuggingFace에서 제공되었습니다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MMLU (HELM 구현) | MMLU (하네스 구현) | MMLU (원본 구현) |
| huggingface/llama-65b | 0.637 | 0.488 | 0.636 |
| tiiuae/falcon-40b | 0.571 | 0.527 | 0.558 |
| llama-30b | 0.583 | 0.457 | 0.584 |
| 에레우테라이/GPT-NeoX-20B | 0.256 | 0.333 | 0.262 |
| llama-13b | 0.471 | 0.377 | 0.47 |
| 하깅페이스/라마-7b | 0.339 | 0.342 | 0.351 |
| tiiuae/falcon-7b | 0.278 | 0.35 | 0.254 |
| 레드파자마-인사이트-7B-베이스 | 0.275 | 0.34 | 0.269 |

위키백과에서 위 텍스트를 가져왔습니다.