



거대 언어모델, LLM

정규 수업 11차시



LLM은 무엇인가?

지난 시간!!

트랜스포머 모델을 활용한 NLP 기억나나요?

LLM은 무엇인가?

텍스트 생성, 요약 등등

주어진 텍스트를 이해하고 텍스트를 생성하는 모델을..

언어 모델(Language Model)

이라고합니다.

LLM은 무엇인가?

1. GPT 계열

최초의 사전학습된 트랜스포머 모델, OpenAI에서 2018년 발표
단방향 디코더를 사용하여 이전 내용을 기반으로해서 다음 내용 생성

2. BERT 계열

양방향 인코더를 사용. 마스킹 된 단어를 예측. 2018년 구글에서 발표
문장 일부 채우기 or 문맥 관계 파악이 주된 역할임. 생성 보다는 이해에 특화

LLM은 무엇인가?

거대 언어 모델(Large Language Model)

높은 수준의 텍스트 생성 등을 처리하는 대규모 신경망 모델입니다.

LM과 같지만 규모에 따라서 LLM으로 분류합니다.

LLM은 무엇인가?

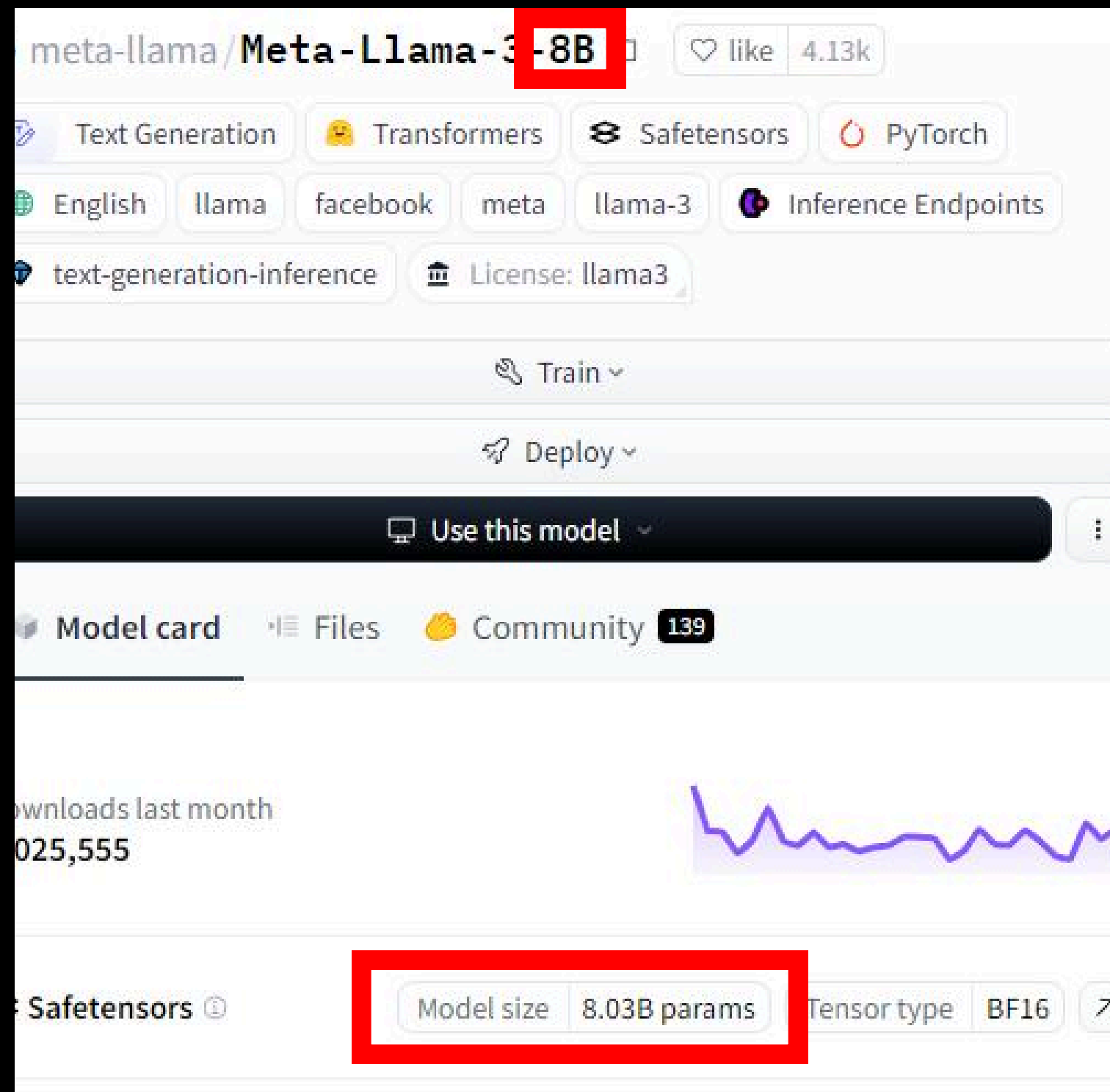
엥..?

그럼 그냥 LM과 LLM은 무엇으로 구분하나요?

LLM은 무엇인가?

사실 명확한 기준은 없습니다

LLM은 무엇인가?



파라미터 (Parmeter, 매개변수)

모델의 가중치, 편향(Bias) 등이 파라미터에 해당

모델의 크기와 성능을 짐작할 수 있는 대표적 수치
또, 모델 구동을 위한 최소 요구 사항을 짐작할 수 있음

LLM은 무엇인가?

음... 일반적으로 파라미터가

백만(million) 단위 → LM

십억(billion) 단위 → LLM

그냥 관례상 통상적으로 부르는 것이므로 상황에 따라 달라요

LLM은 무엇인가?

파라미터 개수로 최소 구동 사양 짐작하기!

유용하다!!!

보통 N billion의 파라미터의 개수를 가진 모델은

2N 기가바이트의 vram을 최소로 요구합니다.

나중에 배울거지만 양자화(quantization)란걸 활용하면

더 낮은 vram에서도 구동이 가능하게 할 수 있습니다.

LLM은 무엇인가?

대표적인 LLM 몇 가지만 소개할게요

**사실 소개하는거 말고도 개쩌는 모델들이 많지만,
큰 회사 것들 위주로 소개할게요**

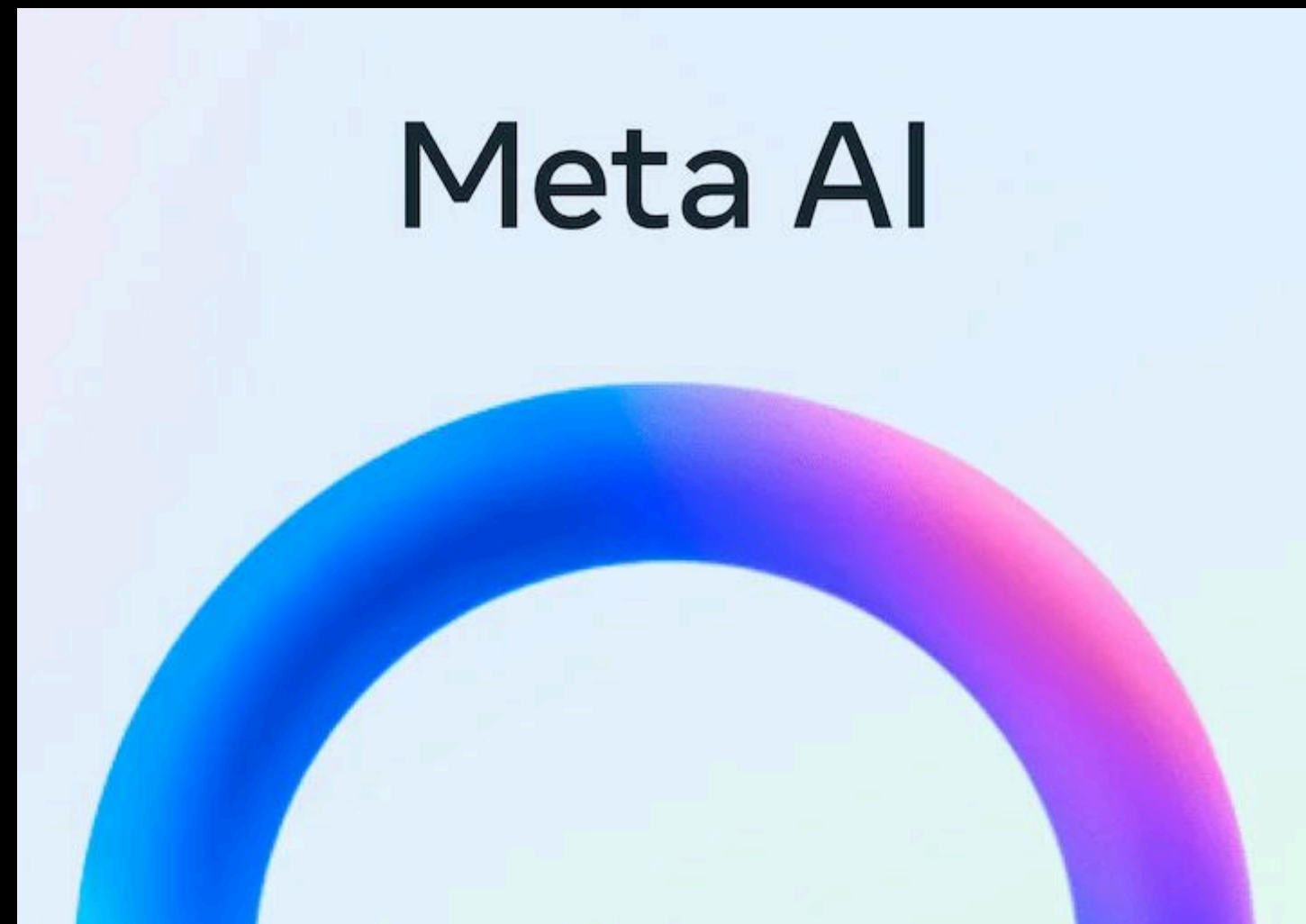
LLM은 무엇인가?



OpenAI의 GPT

- GPT 계열 모델의 근본 혈통
- 대표적인 모델로는 3.5, 4, 4o 등 여러 버전이 존재함.
 - 강성능이 좋음. 다만 모델이 대체로 무거운 편
 - 오픈소스가 아니라서 마음대로 다루기 힘들.
원하는대로 다루려면 결제가 필요한 API를 써야한다.
(GPT-2 까지 오픈소스임)

LLM은 무엇인가?



Meta의 Llama

- 성능 좋은 오픈소스 LLM의 대표격
- 오픈소스 모델이라서 파생 모델이 엄청 많음
- 최근에 Llama3 까지 출시했는데 성능이 좋아서 관심을 많이 받음.

LLM은 무엇인가?



Google의 Gemma

- 구글에서 만든 sLLM(소형 거대 언어 모델?)
- 작고 강한 놈을 의도했지만 처참히 망함ㅌㅌ

LLM은 무엇인가?

Microsoft의 Phi-3



- 제일 작은 모델 기준
사이즈가 GPT-3.5의 반의 반의 반도 안됨
근데 성능이 좋다고 평가 받는 중
- SLM(Small LM)이라고 말하지만
그렇다고 엄청나게 작은건 아님.
- 마소가 작고 강력함에 진심이라는 것에 대한 증거

LLM은 무엇인가?

이런 LLM에도 한계점들이 몇가지 있습니다

1. 제한된 지식
2. 판단력의 부재
3. 추상적 추론의 한계
4. 할루시네이션(환각현상)
5. 학습 데이터 의존성과 편향성

LLM은 무엇인가?

한계점 1 : 제한된 지식

LLM은 학습된 데이터만을 가지고 작동하므로, 학습 이후의 최신 정보는 없음.

예를 들어 인디게임의 정보라던가 학습이 안되었을 가능성이 높은 정보는 대답을 잘 못함.

다만, 이 한계점은 여러가지 방법론들에 의해서 부분적으로 해결할 수 있는 방법이 많아졌음

ex) 파인튜닝(Full, PEFT, LoRa 등), 프롬프트 엔지니어링, RAG 등등

LLM은 무엇인가?

한계점 2 : 판단력의 부재

인공지능은 자아가 없음. 완벽히 인간처럼 생각하지 않음.

인간은 할 말이 떠오르면 한 번 다시 생각해서 필터링 하는 등 판단하려하지만..
LLM은 노빠꾸다. 그냥 말함.




LLM은 무엇인가?

한계점 3 : 추상적 추론의 한계

LLM은 텍스트를 학습한 모델임. 그냥 학습된 내용(텍스트)을 술술 뱉기만 함.

그래서 흔히 "추상적 추론"이라고 말하는 부분은 잘 못함.

간단한 수학 문제도 풀지 못하는 이유는, LLM은 그냥 학습된 내용 중에서 그나마 가장 답 같은 내용을 아무거나 뱉는 것이기 때문이다.


$$2381 \times 1234 = 2,938,154$$



2381 곱하기 1234는 2,940,554입니다.

LLM은 무엇인가?

한계점 4 : 할루시네이션(환각현상, Hallucination)

다들 아는 “세종대왕 맥북 던짐”처럼 잘못된 내용도 사실처럼 말하는 현상

대표적인 주 원인은

1. 잘못된 데이터를 학습
 2. 답변 문장 추론 과정에서 잘못된 추론
- 이 현상을 기준으로 모델 성능 평가를 하기도 함.



LLM은 무엇인가?

한계점 5 : 학습 데이터 의존성과 편향성

모든 현존하는 AI에게서 일어날 수 있는 한계점
AI는 학습한 데이터에 기반해서 작동하기 때문이다.

LLM의 경우 편향적인 데이터를 학습 시에
정치, 인종, 성 등 사회적 또는 윤리적으로 문제되는 발언을 할 수 있음.



망언 (妄言)

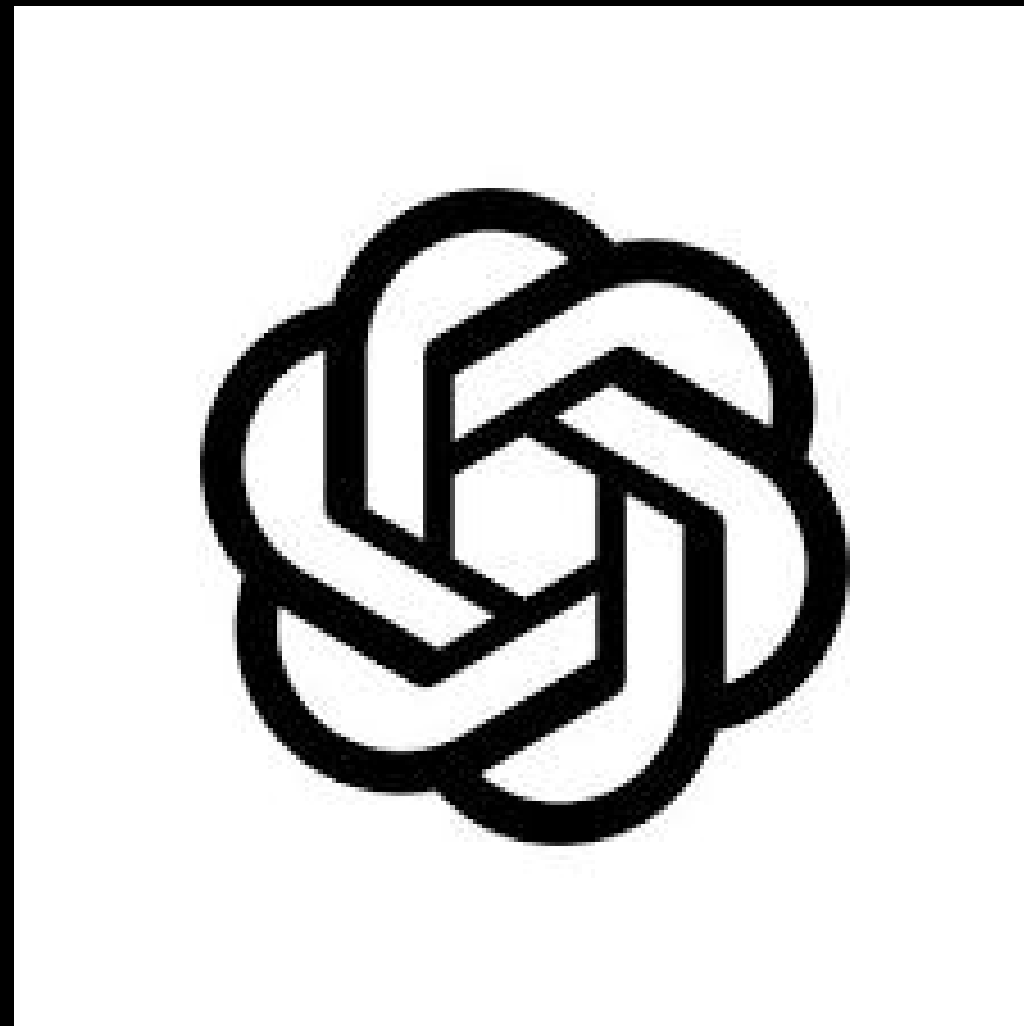
[망:언] 🔊

이치나 사리에 맞지 아니하고 망령되게 말함. 또는 그 말.

OpenAI API

이제 기본적으로 LLM을 다룰 준비가 되었습니다.

**이제 OpenAI에서 제공하는 API를 사용해서
GPT에게 여러가지 작업을 시켜봅시다.**



OpenAI API

OpenAI사에서 만든 AI 모델을 API 형태로 제공

GPT 부터 시작해서 임베딩, 이미지 생성, TTS 등
다양한 AI 모델을 제공

굳이 돈내고 제약 받으면서 OpenAI API를 쓰는 이유

1. 싼 가격에 고성능 모델 사용 가능

동일 성능의 LLM을 로컬에서 직접 구동하기 위해선 최소 수백만원의 투자가 필요하다.

하지만 API를 사용한다면 사용한 만큼만 돈을 내면 된다. 초기 투자 금액이 적다.

2. 간단한 파인튜닝(미세조정)

학습의 경우에는 단순 모델 구동보다 더 많은 컴퓨팅을 요구하기에 더욱 비용적으로 문제되지만,

싼 가격에 파인튜닝을 진행할 수 있으며, 절차를 간소화해서 작업하기 편하다.

나눠주는 API KEY를 다룰 때 주의사항

1. 개인 용도로 사용해도 되나 "GPT-3.5 turbo" 모델만 사용합시다
2. 유출되지 않도록 합니다. 절대보안!!!
3. 이거 선배 사비로 하는거니까 제발 알잘딱!!!! 지갑이 아파!!!!!!