▶ 대규모 언어 모델(LLM)이란?

간단히 말해, LLM은 인간의 언어나 기타 복잡한 데이터를 인식하고 해석할 수 있을 만큼 충분한 예제를 제공받은 컴퓨터 프로그램입니다. 많은 LLM은 인터넷에서 수집된 수천 또는 수백만 기가바이트에 달하는 텍스트로 학습됩니다. 하지만 샘플의 품질이 LLM이 자연어를 얼마나 잘 학습할 수 있는지에 영향을 미치므로, LLM의 프로그래머는 보다 엄선된 데이터 세트를 사용할 수 있습니다.
LLM은 문자, 단어, 문장이 함께 작동하는 방식을 이해하기 위해 딥러닝이라는 일종의 머신 러닝을 사용합니다. 딥러닝은 비정형 데이터의 확률적 분석을 포함하며, 결국 딥러닝 모델은 사람의 개입 없이도 콘텐츠 간의 구분을 인식할 수 있습니다.

실제 LLM의 예로는 ChatGPT(OpenAI 제공), Bard(Google), Llama(Meta), Bing Chat(Microsoft) 등이 있습니다. GitHub의 Copilot도 또 다른 예이지만, 자연스러운 인간 언어가 아닌 코딩을 위한 것입니다.

▶ 환각 (Halluciation) - 정확한 답변을 내놓을 수 없을 때면 가짜 정보를 생성하는 현상

얀 르쿤 교수는 2023년 3월 24일 NYU에서 열린 세미나에서 간단 수식을 통해Auto-regression기반 LLM은 생성 토큰(n)이 길어지면 길어질수록 생성된 문장이 정확할 확률 P(correct)이 기하급수적으로 낮아진다고 Auto-regression기반 LLM에 대해 신랄하게 비판하였다.

"ChatGPT는 맥아더 장군이 태어나기도 전인 남북 전쟁을 배경으로 맥아더 장군을 북부군 지휘관으로 묘사하는 환각을 일으켰다."

https://moon-walker.medium.com/llm%EC%97%90-halluciation-%ED%99%98%EA%B0%81-%EC%9D%B4-%EB%B0%9C%EC%83%9D%ED%95%98%EB%8A%94-%EC%9B%90%EC%9D%B8%EA%B3%BC-%ED%95%B4%EA%B2%B0%EB%B0%A9%EC%95%88-f18759f0a959

▶ 기업용 맞춤 LLM을 위한 선택 가이드 (구글검색 'RAG')

https://www.skelterlabs.com/blog/rag-vs-finetuning

▶ 쓱읽어보세요. (구글검색 'SOTA 모델')

https://velog.io/@pearl1058/Fine-tuning-%ED%8C%81%EB%93%A4AWSKRUG-%ED%9B%84%EA%B8%B0

▶ 2024년 LLM

https://www.content.upstage.ai/blog/insight/top-open-source-llms-2024

파인튜닝 모델 평가 및 최적화 에서 배포 관리까지 개념 잡기
아래 사이트에서 시작해서 필요한 페이지까지 쭉 읽으셔요.

https://wikidocs.net/198924

애저 오픈AI (AI Skills Challenge) - MS 제품군에서 의 튜닝도 있다는걸 확인하는 용도

https://learn.microsoft.com/ko-kr/azure/ai-services/openai/tutorials/fine-tune?tabs=python-new%2Ccommand-line

랭체인 그래프

https://teddylee777.github.io/langgraph/langgraph-agentic-rag/

[peft lora 모델의 이해]
https://ongamedev.tistory.com/528

[Polyglot inference code 12.8b-4bit]

https://dyent.tistory.com/entry/%ED%95%9C%EA%B5%AD%EC%96%B4-%EC%96%B8%EC%96%B4-%EB%AA%A8%EB%8D%B8-Polyglot-%ED%85%8C%EC%8A%A4%ED%8A%B8-%EB%B0%8F-%EC%82%AC%EC%9A%A9%EB%B2%95

beomi/KoAlpaca-Polyglot은 **EleutherAI/polyglot-ko 모델을 백본으로 사용**하여

네이버 지식인 게시물 등 다량의 한국의 데이터가 파인튜닝된 모델이라고 합니다.

다양한 버전의 모델이 존재하고, 모델명에서 b앞에 붙어있는 숫자가 커질수록 성능이 좋은 모델입니다.

[파인튜닝-제공하는 데이터셋 이용]- 영문 오픈소스 LLM에 한글 데이터세트로 파인튜닝하기
http://innovationplaza.blogspot.com/2023/07/llm.html

[제공되는 자료가 아니라 , 내 자료로 파인튜닝 하고자 할때]

https://wikidocs.net/166816

[AWS기반이기 하지만 튜닝할 각 파일에 대한 설명이 잘 나와 있음 (이중 aws 부분만 변경하면 됨)

허깅페이스와 LoRA를 사용하여 단일 Amazon SageMaker GPU에서 대규모 언어 모델(LLM) 훈련하기
https://aws.amazon.com/ko/blogs/tech/train-a-large-language-model-on-a-single-amazon-sagemaker-gpu-with-hugging-face-and-lora/

[실무에서 사용할 수 있는 자연어 처리 팁]

https://ratsgo.github.io/nlpbook/

챗봇 딥러닝 - OpenAI, 성능은 높아지고 가격은 싸진 새로운 모델 공개 (aidev.co.kr)
http://aidev.co.kr/chatbotdeeplearning/14344

gpt-3.5-turbo-0125 모델을 로컬에서 사용하고자함.

참고) gpt 파인튜닝 비용OpenAI API 비용은 대부분 1,000(1K) 토큰 기준입니다.
1,000토큰은 영단어 약 750단어 (1토큰: 약 0.75단어)에 해당하는 것으로 OpenAI 홈페이지에 언급되어 있습니다.

https://deepdaive.com/openai-api/

GPT-3.5는 1천 토큰 당 입력 0.0005$, 출력 0.0015$입니다. 평균 0.001$로 계산하면 1.3원입니다. 클로바X는 1천 토큰에 5원입니다. 토큰 수가 2배 차이가 나는 것 고려하면 아래와 같습니다.

< GPT-3.5 >
토큰수 - 242
가격 - 0.31원

< 클로바X >
토큰수 - 108
가격 - 0.54원

클로바X가 GPT-3.5에 비해서 2배 정도 비싼 편입니다. 그래도 예전 클로바에 비하면 훨씬 싸진 가격인데요. 클로바X는 한국에 대한 정보나 감성대화 분야에서는 GPT-3.5보다 훨씬 성능이 좋습니다. 이 정도 가격이면 충분히 사용해볼 만 합니다. 물론 앞으로 더 떨어지면 좋겠지만요.

LLM 파인튜닝이란?

프롬프트 엔지니어링은 기존의 모델을 변화시키지 않고, 모델의 입력에 특정한 문장을 추가하여 원하는 출력을 얻는 방식이고, 파인튜닝은 미리 학습된 모델을 새로운 작업에 맞게 다시 학습시키는 방식이다.