

# 오픈소스 LLM 실습

LLM 실습

2023-08-02

홍성빈

# CONTENTS

1 LLM이 뭔가요?

-----▶ P. 3

2 오픈소스 LLM 소개

-----▶ P. 22

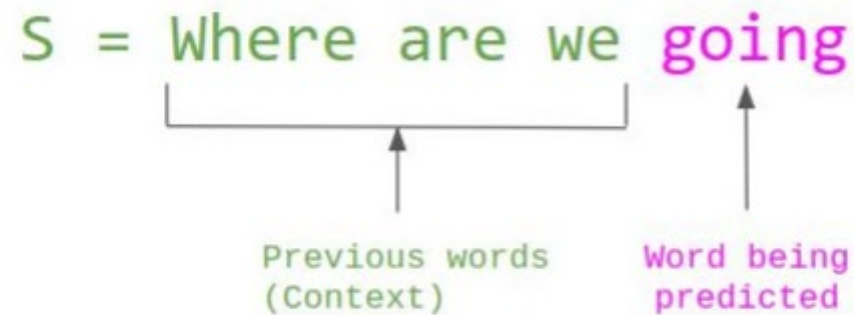
3 오픈소스 LLM 사용

-----▶ P. 43



**LLM이란?**

# LM(Language Model)



$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

다음에 올 단어를 예측하는 머신러닝 모델

# LLM(Large Language Model)

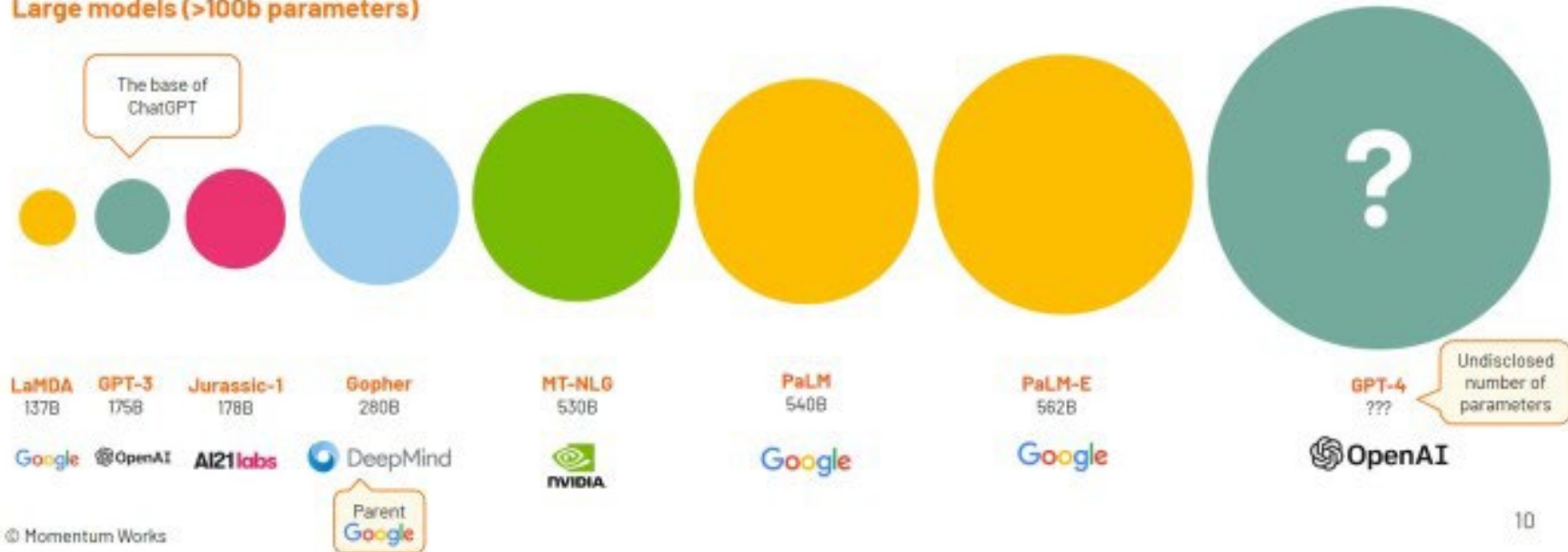
Large Language Models are becoming very large indeed



## Small models (<= 100b parameters)



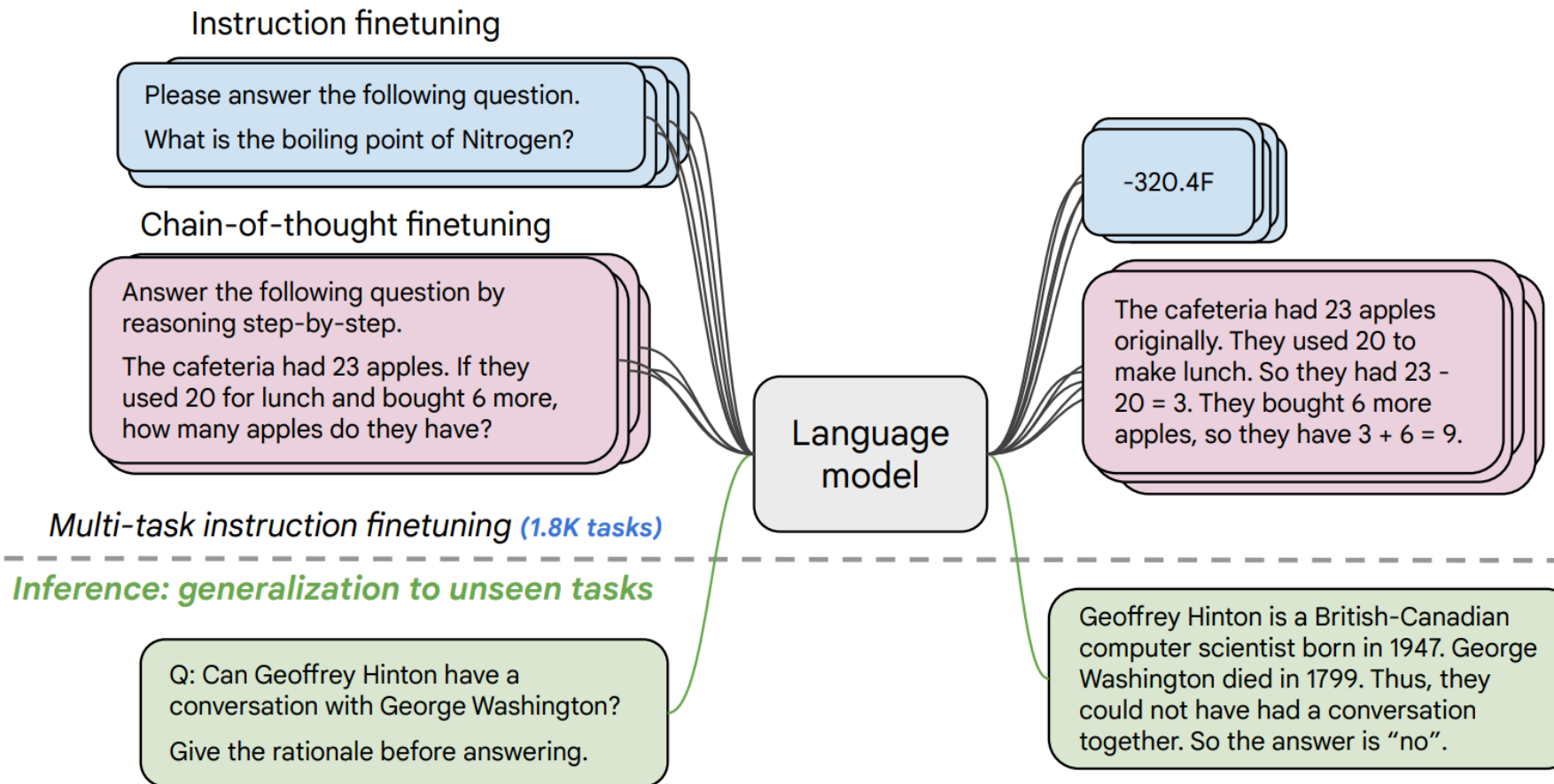
## Large models (>100b parameters)



© Momentum Works

10

## instruction : LLM에게 사람이 부여하는 어떤 작업을 수행하게 만드는 일종의 명령어



## 왜 LLM이 각광을 받는 중 인가???



# 문장생성

Default (GPT-3.5)

성빈

프로그래머의 애환을 담은 프리스타일 랩을 써줘



(Verse 1)

난 프로그래머, 코드로 세상을 풀어나가는 전사  
이 세상에 내 노력과 열정이 살아 숨 쉬는 증거  
하지만 때론 애환에 괴로워, 문제에 매몰되곤 해  
끊임없는 고민과 새로운 기술에 머리 아프고

(Pre-Chorus)

무한한 가능성 속에 갇혀 스트레스는 쌓여가지  
잠들지 못해 불면증이 친구가 돼버렸지  
하지만 난 포기하지 않아, 이 랩처럼 계속 노래해  
프로그래머의 애환을 담아 함께 울려

(Chorus)

컴파일 오류에 시달리고  
버그의 늪에 빠져들고  
때론 자신에게 의문해도  
난 프로그래머, 코딩의 미로에서 헤쳐나갈래

(Verse 2)

새로운 언어, 프레임워크의 등장으로  
배워야 할 게 산더미, 압도돼 스트레스만 가득해  
그래도 난 불타는 열정으로 도전해  
한 줄 한 줄 의미를 부여하며 내 코드를 완성해








# 대화

Default (GPT-3.5)

성빈

넌 이제부터 무한한 상상력을 발휘하는 명망있는 화가야. 너는 화가로서 내가 하는 말에 대답해 줘






감사합니다! 제가 화가로서 상상력을 발휘해 드리겠습니다. 무엇을 그릴지 먼저 말씀해주시면,    그림을 그려드리도록 하겠습니다. 어떤 주제나 이미지를 원하시나요?

성빈

희망에 대해 어떻게 생각해?



희망은 인간에게 큰 힘이 되는 강력한 감정입니다. 저로서는 희망을 밝고 따뜻한 빛으로 비유할 수 있습니다. 희망은 어두운 시련과 어려움을 이겨내는 원동력이 되며, 미래를 밝고 긍정적으로 바라보게 만들어줍니다.   

희망은 우리가 달성하고자 하는 목표를 향한 동력이자, 꿈을 키우고 이루기 위한 열매를 심는 씨앗입니다. 어떤 어려운 상황에서도 희망을 간직하고 있는 사람들은 결코 포기하지 않으며, 지치고 힘들어도 마지막까지 기뻐하고 노력하는 힘을 갖게 됩니다.

화가로서 저는 희망의 상징을 그리고자 할 때, 푸른 하늘과 밝은 해, 무지개 등을 활용하여 낙관적이고 밝은 분위기를 표현합니다. 또한 사람들이 서로를 응원하고 격려하는 모습, 손을 잡고 함께 나아가는 모습을 그리며 희망의 가치를 전달하고자 합니다.

희망은 우리가 마주하는 모든 상황에서 우리를 이끌어주는 소중한 축이며, 끊임없이 우리를 발전시키고 성장시키는 힘이라고 생각합니다. 그렇기에 희망을 잃지 않고 삶의 여정을 함께 이끌어 나가기를 바라며, 희망의 빛이 항상 우리 곁에 있기를 기원합니다.

# 요약

Default (GPT-3.5)

**상민** 다음과 같은 뉴스를 2줄로 요약해줘

전 세계 이목 집중된 ‘상온 상압 초전도체’... 특허 1건은 이미 ‘승인’ [긴급점검]  
 기사입력 2023-08-01 14:17:36  
 기사수정 2023-08-01 14:33:07  
 글씨 크기 선택  
 전 세계 떠들썩... 특허 신청 4건 중 1건 승인  
 활용도 무궁무진... 해외서도 검증 작업 활발  
 ‘상온 상압 초전도체’를 개발했다는 논문을 올려 전 세계를 떠들썩하게 만든 국내 연구진이 이미 특허청에 관련 특허 4건을 신청한 것으로 확인됐다. 이중 1건은 승인을 받아 특허권이 발생한 상태다.

1일 특허청 특허정보검색서비스 사이트 키프리스에 따르면 이석배씨가 대표로 있는 퀀텀에너지연구소는 2020년 7월부터 지난해 8월까지 초전도체와 관련해 4건의 특허를 신청했다. 이중 가장 처음 특허신청을 한 건은 이미 심사가 완료돼 지적재산권이 발생했다.

양자 자기 부상 및 현탁 효과, 게티이미지뱅크  
 초전도체는 전기저항이 0인 물질을 말한다. 전기저항이 0이라는 건 에너지 소모가 없다는 뜻이다. 초전도체 물질은 자기부상열차와 전력망, MRI 등 무궁무진하게 활용이 가능하다. 하지만 1911년 초전도 현상이 처음 발견됐을 당시 절대온도 4K(영하 269도)에서 구현됐기에 실생활에 활용하는 것은 불가능했다. 이에 100년 넘게 과학계에서는 상온, 상압에서 초전도 현상을 구현하고자 애썼다.

퀀텀에너지연구소가 낸 특허 중 심사가 완료된 특허의 이름은 ‘초전도체를 포함하는 저저항 세라믹화합물의 제조방법 및 그 화합물’이다. 이들은 신청서에서 “초전도체를 포함해 기존 비저항보다 훨씬 더 낮은 비저항 특성을 발휘할 수 있는 초전도체 포함 저저항 세라믹화합물의 제조방법 및 그 화합물에 관한 것”이라고 기술을 설명했다. 연구소는 초전도체를 포함한 저저항 세

외에서도 몇 군데에서 검증 작업에 들어갔다고 늘었다”며 “상온 초전도체와 관련해서는 여러 논란이 많다 보니 모든 학계 분이 신중한 상황”이라고 말했다.



한국의 퀀텀에너지연구소가 상온 상압에서 초전도체를 개발하여 특허청에 관련 특허 4건을 신청한 결과, 특허 심사를 거쳐 1건의 특허권이 이미 승인된 상태입니다. 이 초전도체는 무궁무진한 활용 가능성을 가지며, 해외에서도 검증 작업이 활발히 진행 중입니다.



# 필터링

Default (GPT-3.5)

성빈

채팅 내용을 번호에 맞춰서 줄 테니까 비속어가 포함된 문장은 비속어를 □로 필터링 해서 지워줘

1. 다 비켜 ~~바보~~들아 나 미드갈라니까
2. 블루 ~~치마~~지말라고
3. 말좀해봐 무시하지말고
4. 정글님 갱좀 부탁드릴게요
5. 바론 먹고 게임 끝내죠



알겠습니다. 비속어가 포함된 문장은 비속어를 □로 필터링하여 지워드리겠습니다.



1. 다 비켜 □□런들아 나 미드갈라니까
2. 블루 ~~치마~~지말라고
3. 말좀해봐 무시하지말고
4. 정글님 갱좀 부탁드릴게요
5. 바론 먹고 게임 끝내죠

위와 같이 비속어가 포함된 문장은 비속어가 □로 필터링되어 지워졌습니다. 저희와 함께 쾌적한 대화를 이어나가시기 바랍니다.

# 필터링

위와 같이 비속어가 포함된 문장은 비속어가 □로 필터링되어 시워졌습니다. 저희와 함께 쾌적한 대화를 이어나가시기 바랍니다.

성빈

처먹다는 비속어야



죄송합니다. 비속어로 간주되는 단어를 사용하지 않도록 주의하겠습니다. 비속어를 사용하지 않은 형태로 문장을 다시 제시해 주시면, 필터링 없이 답변해 드리겠습니다. 감사합니다.

성빈

채팅 내용을 번호에 맞춰서 줄 테니까 비속어가 포함된 문장은 비속어를 □로 필터링 해서 지워줘

1. 다 비켜 ~~비속어~~들아 나 미드갈라니까
2. 블루~~비속어~~지말라고
3. 말좀해봐 무시하지말고
4. 정글님 갱좀 부탁드릴게요
5. 바론 먹고 게임 끝내조



죄송합니다. 누락된 비속어를 인지하였습니다. 해당 부분을 필터링하여 다시 제시하겠습니다.

1. 다 비켜 □□런들아 나 미드갈라니까
2. 블루□□지말라고
3. 말좀해봐 무시하지말고
4. 정글님 갱좀 부탁드릴게요
5. 바론 먹고 게임 끝내조

위와 같이 비속어가 포함된 문장은 비속어가 □로 필터링되어 지워졌습니다. 저희와 함께 쾌적한 대화를 이어나가시기 바랍니다. 감사합니다.

# 분류

Default (GPT-3.5)

성빈

채팅 내용을 번호에 맞춰서 줄건데, 비속어가 포함되지 않은 문장을 추려줘

1. 다 비켜 ~~비속어~~들아 나 미드갈라니까
2. 블루 ~~비속어~~지말라고
3. 말좀해봐 무시하지말고
4. 정글님 갱좀 부탁드릴게요
5. 바론 먹고 게임 끝내죠



아래는 번호에 맞춰서 비속어가 포함되지 않은 문장들입니다:



1. 말좀해봐 무시하지말고
2. 정글님 갱좀 부탁드릴게요
3. 바론 먹고 게임 끝내죠

# 변환

Default (GPT-3.5)



내가 문장을 하나 줄 테니까 해당 문장을 '하오체'로 바꿔줘

정글아 블루내놔



정글아 블루내주소서.



# 변환

Default (GPT-3.5)



내가 문장을 하나 줄 테니까 해당 문장을 '하오체'로 바꿔줘

정글아 블루내놔



정글아 블루내주소서.



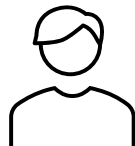
# LLM 활용법



NLP(Natural Language Processing)은 원래 세부적인 작업에 특화된 개별적인 모델이 주였음.



감정분석



분류

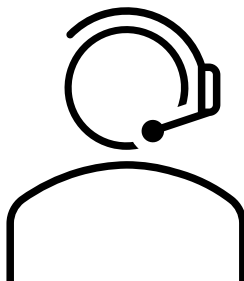


요약



변환

다재다능한 LLM이 등장함으로써 하나의 모델로 여러 작업을 한번에 처리 가능함



감정분석




분류

요약

변환

...


## Prompt Engineering : 모델의 Input값을 조정해서 특정한 기능을 하는 모델로 사용

Default (GPT-3.5)	Default (GPT-3.5)	Default (GPT-3.5)
<p><b>성빈</b> 프로그래머의 애환을 담은 프리스타일 랩을 써줘</p> <p> (Verse 1) 난 프로그래머, 코드로 세상을 풀어나가는 전사 이 세상에 내 노력과 열정이 살아 숨 쉬는 증거 하지만 때론 애환에 괴로워, 문제에 매몰되곤 해 끊임없는 고민과 새로운 기술에 머리 아프고</p> <p>(Pre-Chorus) 무한한 가능성 속에 갇혀 스트레스는 쌓여가지 잠들지 못해 불면증이 친구가 돼버렸지 하지만 난 포기하지 않아, 이 랩처럼 계속 노래해</p>	<p><b>성빈</b> 년 이제부터 무한한 상상력을 발휘하는 명망있는 화가야. 너는 화가로서 내가 하는 말에 대답해줘</p> <p> 감사합니다! 제가 화가로서 상상력을 발휘해 드리겠습니다. 무엇을 그릴지 먼저 말씀해주시면, 그림을 그려드리도록 하겠습니다. 어떤 주제나 이미지를 원하시나요?</p> <p><b>성빈</b> 희망에 대해 어떻게 생각해?</p> <p> 희망은 인간에게 큰 힘이 되는 강력한 감정입니다. 저로서는 희망을 밝고 따뜻한 빛으로 비유할</p>	<p><b>성빈</b> 재팅 내용을 번호에 맞춰서 줄 테니까 비속어가 포함된 문장은 비속어를 □로 필터링 해서 지워줘</p> <p>1. 다 비켜 <del>비속어</del>들아 나 미드갈라니까 2. 블루 <del>비속어</del>지말라고 3. 말좀해봐 무시하지말고 4. 정글님 갠증 부탁드릴게요 5. 바론 먹고 게임 끝내죠</p> <p> 알겠습니다. 비속어가 포함된 문장은 비속어를 □로 필터링하여 지워드리겠습니다.</p> <p>1. 다 비켜 □□□들아 나 미드갈라니까 2. 블루 <del>비속어</del>지말라고 3. 말좀해봐 무시하지말고 4. 정글님 갠증 부탁드릴게요 5. 바론 먹고 게임 끝내죠</p> <p>위와 같이 비속어가 포함된 문장은 비속어가 □로 필터링되어 지워졌습니다. 저희와 함</p>

## Prompt Engineering : 모델의 Input값을 조정해서 특정한 기능을 하는 모델로 사용

Default (GPT-3.5)

**성빈** 프로그래머의 애환을 담은 프리스타일 랩을 써줘

 (Verse 1)  
난 프로그래머, 코드로 세상을 풀어가는 전사  
이 세상에 내 노력과 열정이 살아 숨 쉬는 증거  
하지만 때론 애환에 괴로워, 문제에 매몰되곤 해  
끊임없는 고민과 새로운 기술에 머리 아프고


(Pre-Chorus)  
무한한 가능성 속에 갇혀 스트레스는 쌓여가지  
잠들지 못해 불면증이 친구가 돼버렸지  
하지만 난 포기하지 않아, 이 랩처럼 계속 노래해

Default (GPT-3.5)

**성빈** 년 이제부터 무한한 상상력을 발휘하는 명망있는 화가야. 너는 화가로서 내가 하는 말에 대답해줘

 감사합니다! 제가 화가로서 상상력을 발휘해 드리겠습니다. 무엇을 그릴지 먼저 말씀해주시면, 그림을 그려드리도록 하겠습니다. 어떤 주제나 이미지를 원하시나요?


**성빈** 희망에 대해 어떻게 생각해?

 희망은 인간에게 큰 힘이 되는 강력한 감정입니다. 저로서는 희망을 밝고 따뜻한 빛으로 비유할

Default (GPT-3.5)

**성빈** 채팅 내용을 번호에 맞춰서 줄 테니까 비속어가 포함된 문장은 비속어를 □로 필터링 해서 지워줘

1. 다 비켜 ~~비속어~~들아 나 미드갈라니까  
2. 블루 ~~비속어~~지말라고  
3. 말좀해봐 무시하지말고  
4. 정글님 갠증 부탁드릴게요  
5. 바론 먹고 게임 끝내죠

 알겠습니다. 비속어가 포함된 문장은 비속어를 □로 필터링하여 지워드리겠습니다.

1. 다 비켜 □□□□들아 나 미드갈라니까  
2. 블루 ~~비속어~~지말라고  
3. 말좀해봐 무시하지말고  
4. 정글님 갠증 부탁드릴게요  
5. 바론 먹고 게임 끝내죠

위와 같이 비속어가 포함된 문장은 비속어가 □로 필터링되어 지워졌습니다. 저희와 함

## Prompt Engineering : 모델의 Input값을 조정해서 특정한 기능을 하는 모델로 사용

예시) <https://github.com/jujumilk3/leaked-system-prompts>

jujumilk3 Merge pull request #4 from jujumilk3/opera_aria ... 7a7267f on Jun 17 18 commits		
.gitignore	add git ignore	3 months ago
README.md	opera-aria_20230617	2 months ago
github-copilot-chat_20230513.md	add openai prompt source and apply md lint	3 months ago
microsoft-bing-chat_20230209.md	add openai prompt source and apply md lint	3 months ago
openai-chatgpt-ios_20230614.md	Update openai-chatgpt-ios_20230614.md	2 months ago
openai-chatgpt_20221201.md	add openai prompt source and apply md lint	3 months ago
opera-aria_20230617.md	opera-aria_20230617	2 months ago
perplexity.ai_20221208.md	add openai prompt source and apply md lint	3 months ago
snap-myai_20230430.md	add openai prompt source and apply md lint	3 months ago
wrtm_20230603.md	add wrtm	2 months ago

README.md

### leaked-system-prompts

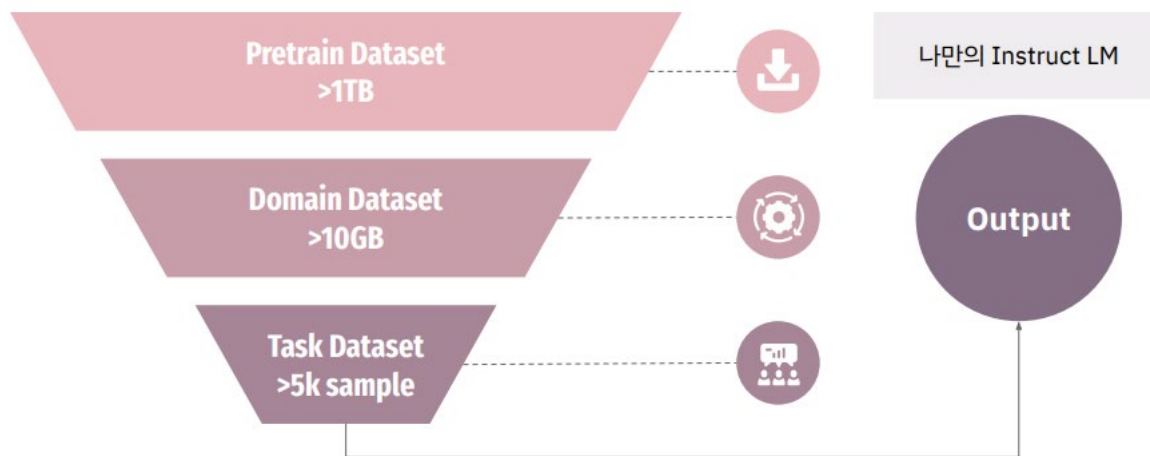
Collection of leaked prompts

1. [github-copilot-chat\\_20230513.md](#)
2. [microsoft-bing-chat\\_20230209.md](#)
3. [openai-chatgpt\\_20221201.md](#)
4. [openai-chatgpt-ios\\_20230614.md](#)
5. [opera-aria\\_20230617.md](#)
6. [perplexity.ai\\_20221208](#)
7. [snap-myai\\_20230430.md](#)
8. [wrtm\\_20230603.md](#)

저는 LLM이 내장하지 못한 지식이나 작업을 활용하게 만들고 싶어요.

=> Pretrain 하거나 fine tuning 해야함

데이터가 있다면 말이죠...



저는 LLM이 내장하지 못한 지식이나 작업을 활용하게 만들고 싶어요.  
=> Pretrain 하거나 fine tuning 해야함

GPU 서버가 있다면 말이죠...

LLM Training Costs on MosaicML Cloud			
Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	\$2,000
GPT-2.7B	54B	0.48	\$6,000
GPT-6.7B	134B	2.32	\$30,000
GPT-13B	260B	7.43	\$100,000
<b>GPT-30B *</b>	<b>610B</b>	<b>35.98</b>	<b>\$450,000</b>
GPT-70B **	1400B	176.55	\$2,500,000

대략 4천만원

대략 1억 3천만원

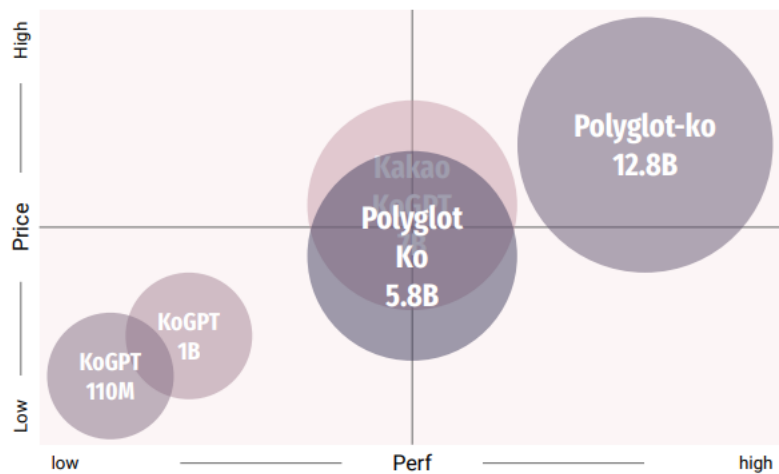
대략 32억

저는 LLM이 내장하지 못한 지식이나 작업을 활용하게 만들고 싶어요.  
=> Pretrain 하거나 fine tuning 해야함

GPU 서버가 있다면 말이죠...  
fine-tuning 비용은 훨씬 적긴 함

### Finetune 학습 비용

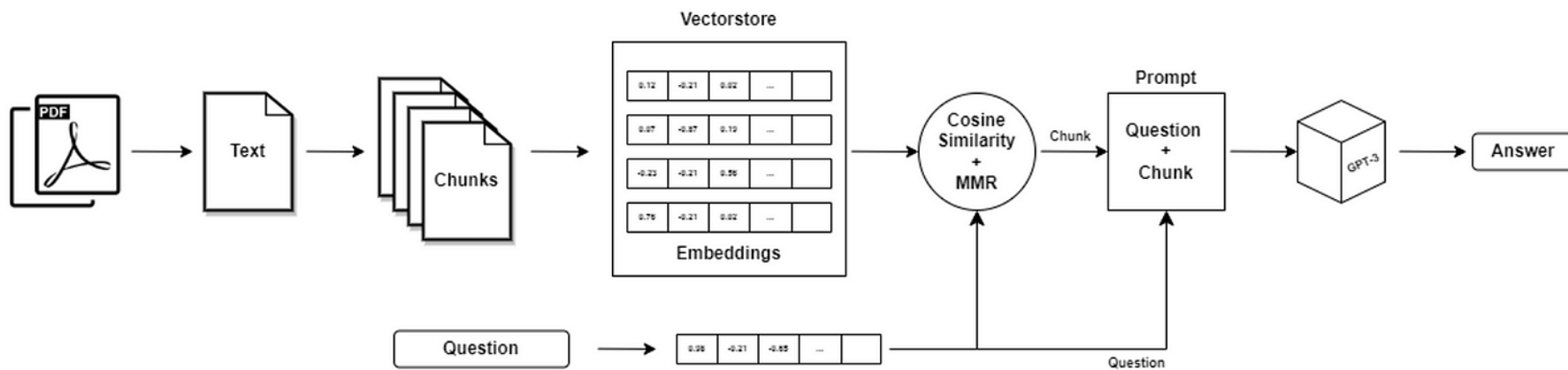
Pretrain vs Finetune



#### A100/H100으로 Finetune

- Polyglot-Ko 5.8B:  
A100 80G x1대 @ 12hrs  
~\$36 = 약 5만 원
- Polyglot-Ko 12.8B:  
A100 80G x4대 @ 12hrs  
~\$144 = 약 20만 원

저는 LLM이 내장하지 못한 지식을 응답하게 만들고 싶어요.  
=> Vector Database를 참조하게 만들어도 됨





저는 LLM이 내장하지 못한 지식을 응답하게 만들고 싶어요.  
=> Vector Database를 참조하게 만들어도 됨

<https://github.com/wafflecomposite/langchain-ask-pdf-local>

