Langchain을 이용한 RAG 이론 및 실전구축

: Fine-Tuning, RLHF, RAG

240327

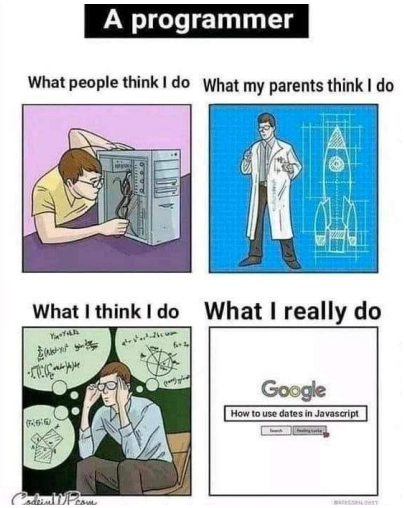
고 우 영

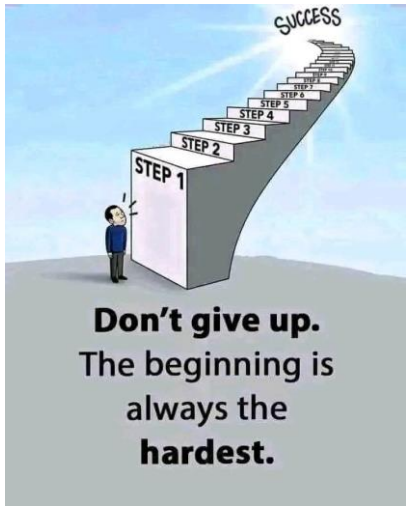
코드실습

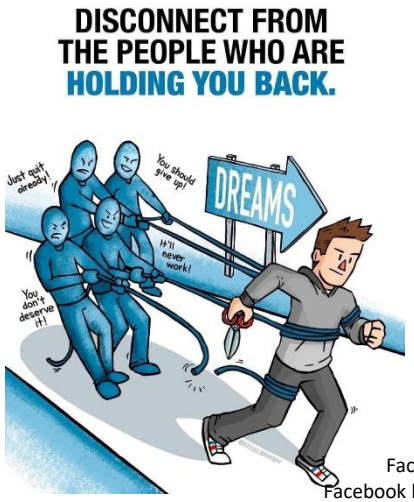
- URL: https://github.com/aisecuritynlp/aisec
- 실습코드: https://colab.research.google.com/drive/1QDlGceWCvchf4YukMxd-oHBNGV17KByY?usp=sharing

Overview

- LLM (Large Language Models)을 실제 환경에서 구축하는 방법
- 목표: LLM의 구조와 작동 원리를 깊이 이해하고, SFT, RLHF와 RAG를 활용하여 모델의 출력을 최적화하는 실용적인 기술을 습득
- Custom LLM을 만드는 방법, RAG (Retrieval-Augmented Generation)를 활용하여 지식 기반의 응답 생성을 강화하는 방법
- 실습을 통해 참가자들은 실제 문제 해결에 LLM을 효과적으로 적용하는 방법
- LLM을 실전에서 구축하고 최적화하는 데 필요한 지식과 기술을 갖추도록 돕는 데 중점







Facebook by Saraswathi Analytics
Facebook by English With Hardeep Singh
Facebook by Abhishek Baruah

오프라인 DB기반 RAG chatbot(on-premise) 구현

- DB기반 chatbot 수요가 굉장히 많음

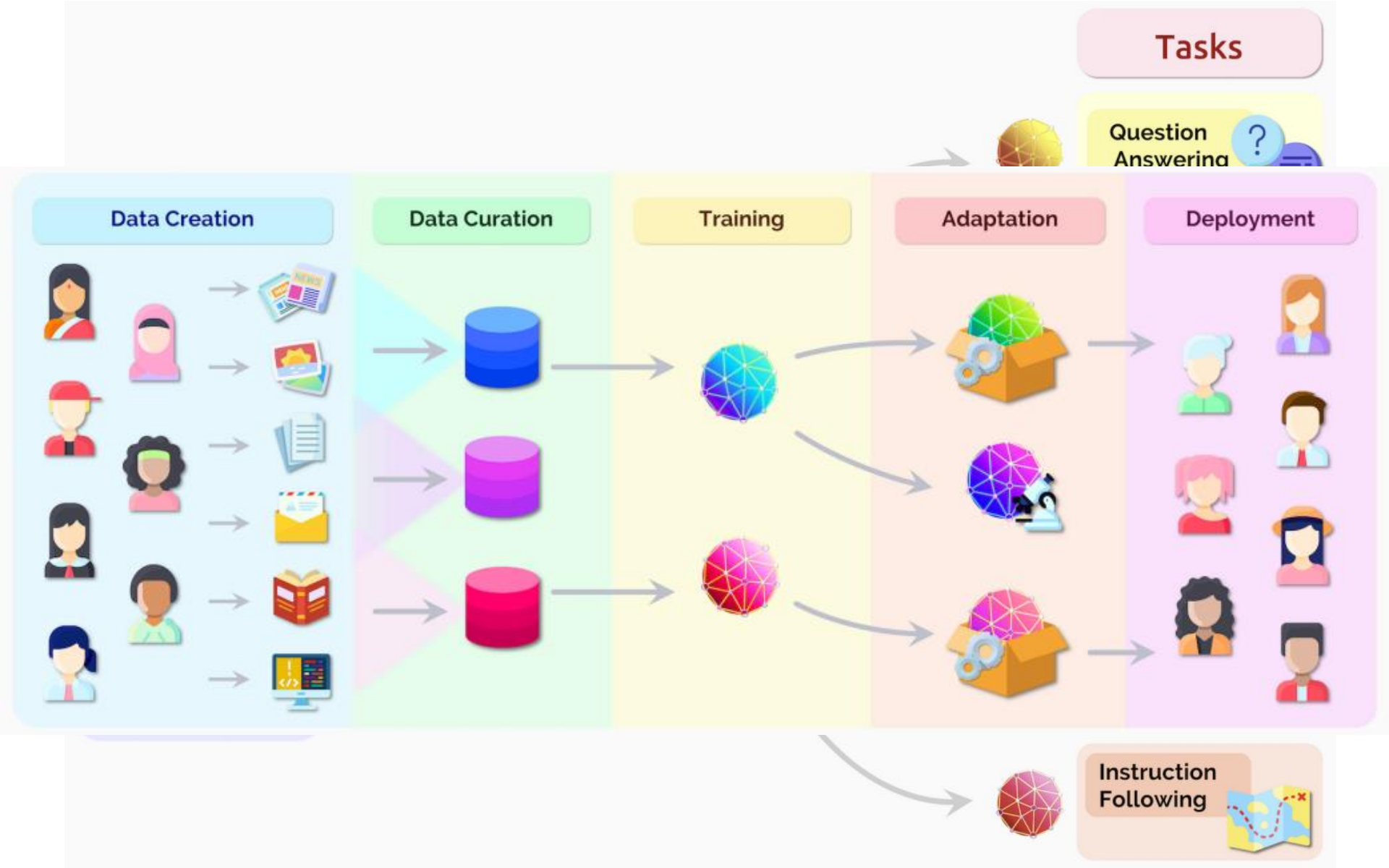
- 그런데 DB가 개인/민감정보를 담고 있고, 사용자의 질문 자체도 보안인 경우가 많아 온라인(openai/bing)에서 사용 가능한 chatgpt를 못쓰는 제약조건이 있음

- Off-line에서 on-premise로 자체 DB기반 챗봇을 만들기 위한 과정

Contents

- LLM (Large Language Models)
- LLM-fine-tuning
  - SFT
  - RLHF
- RAG
- Hands-on

Foundation Model

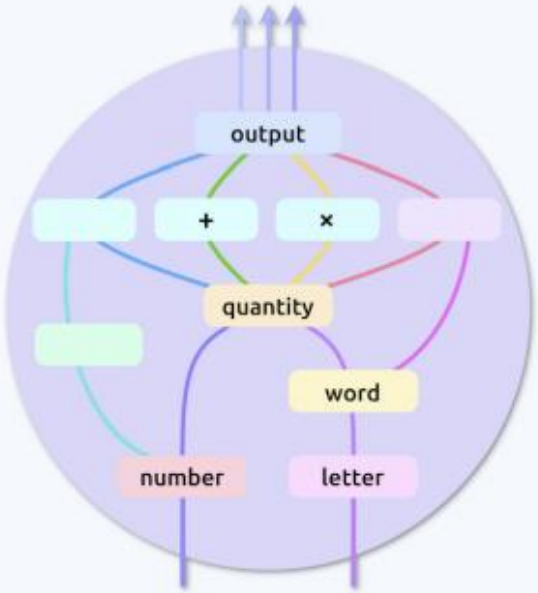


https://arxiv.org/pdf/2108.07258.pdf

LLM: Large Language Model

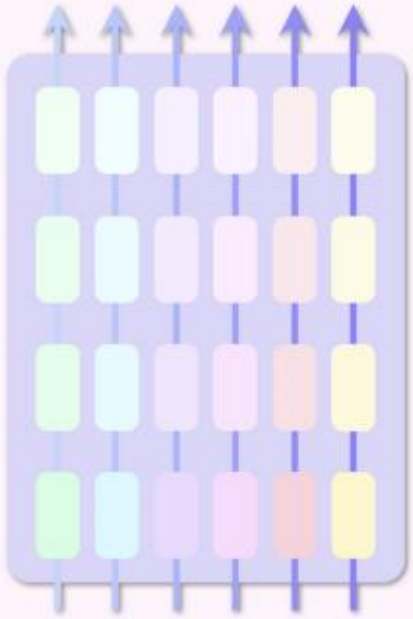
One Model

A finite number of **generalizable model mechanisms are combined** to produce behaviors across tasks.



Many Models

For each task, distinct model mechanisms are used to produce behaviors; akin to **a large collection of individual expert models**.



https://arxiv.org/pdf/2108.07258.pdf
https://renaissancerachel.com/large-language-models-overview/

LLM(Large Language Models)

- **정의: 명확한 정의 존재 X**
  - 지스
    성 능
  - 번역
    능을


ChatGPT


Bard


New Bing


Claude

텍스트 생

서 높은 성

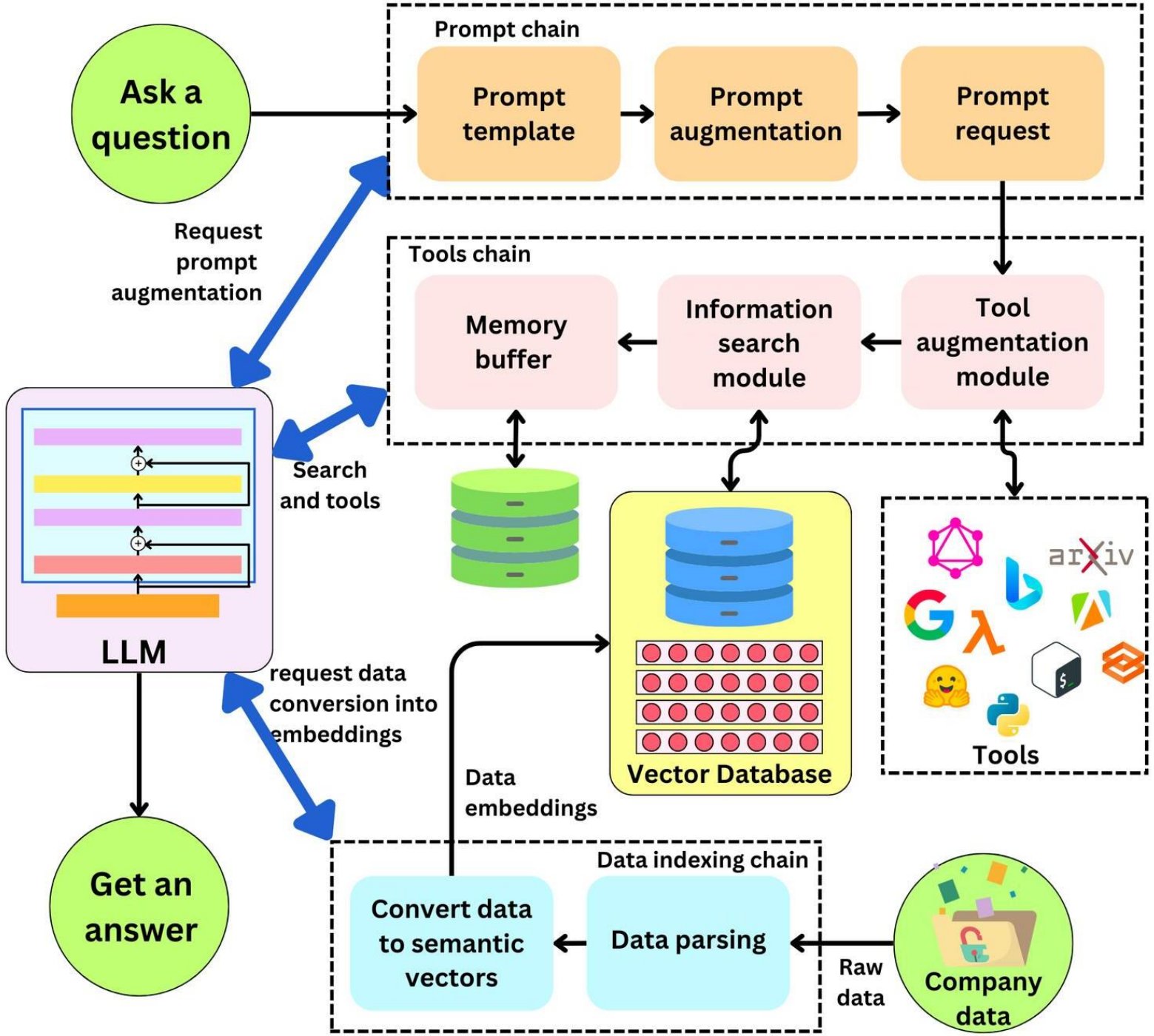

Alpaca


Vicuna

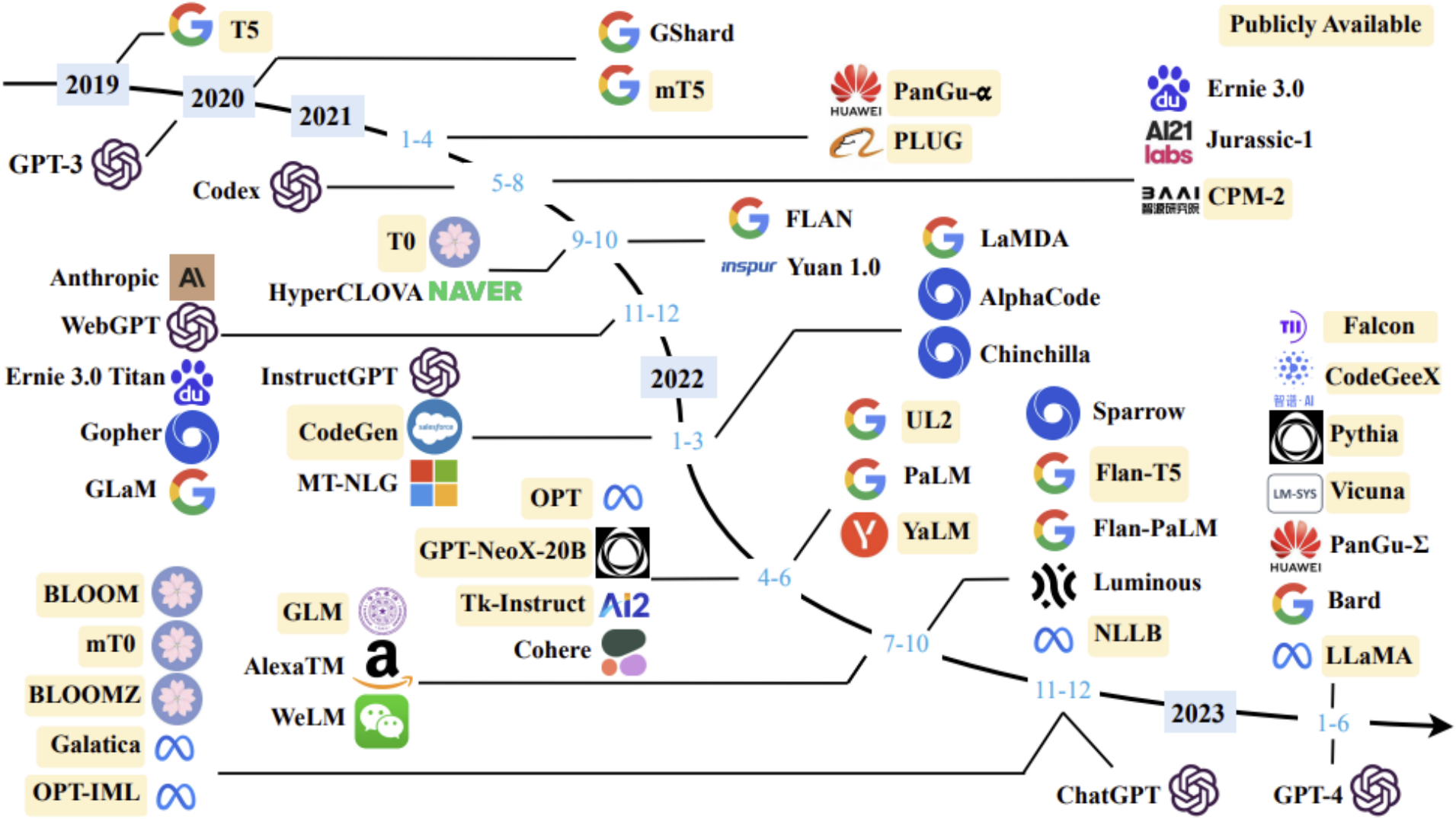

Dolly


Stable Vicuna

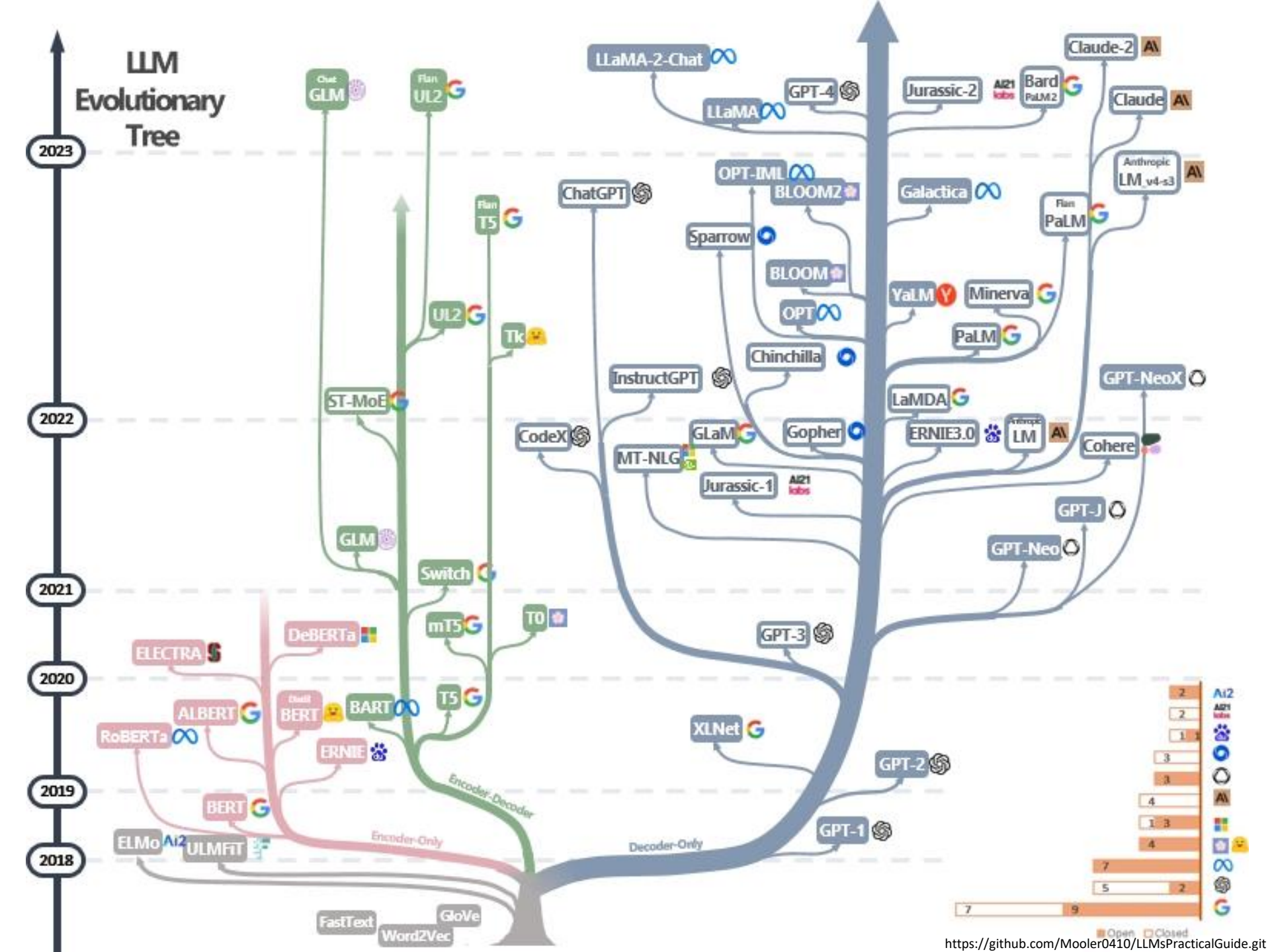
  - 초거대 학습 데이터셋
  - 초거대 모델 사이즈(10B 이상)
  - 특별한 추가학습 없이도 다양한 task 수행 가능



LLM(Large Language Models)

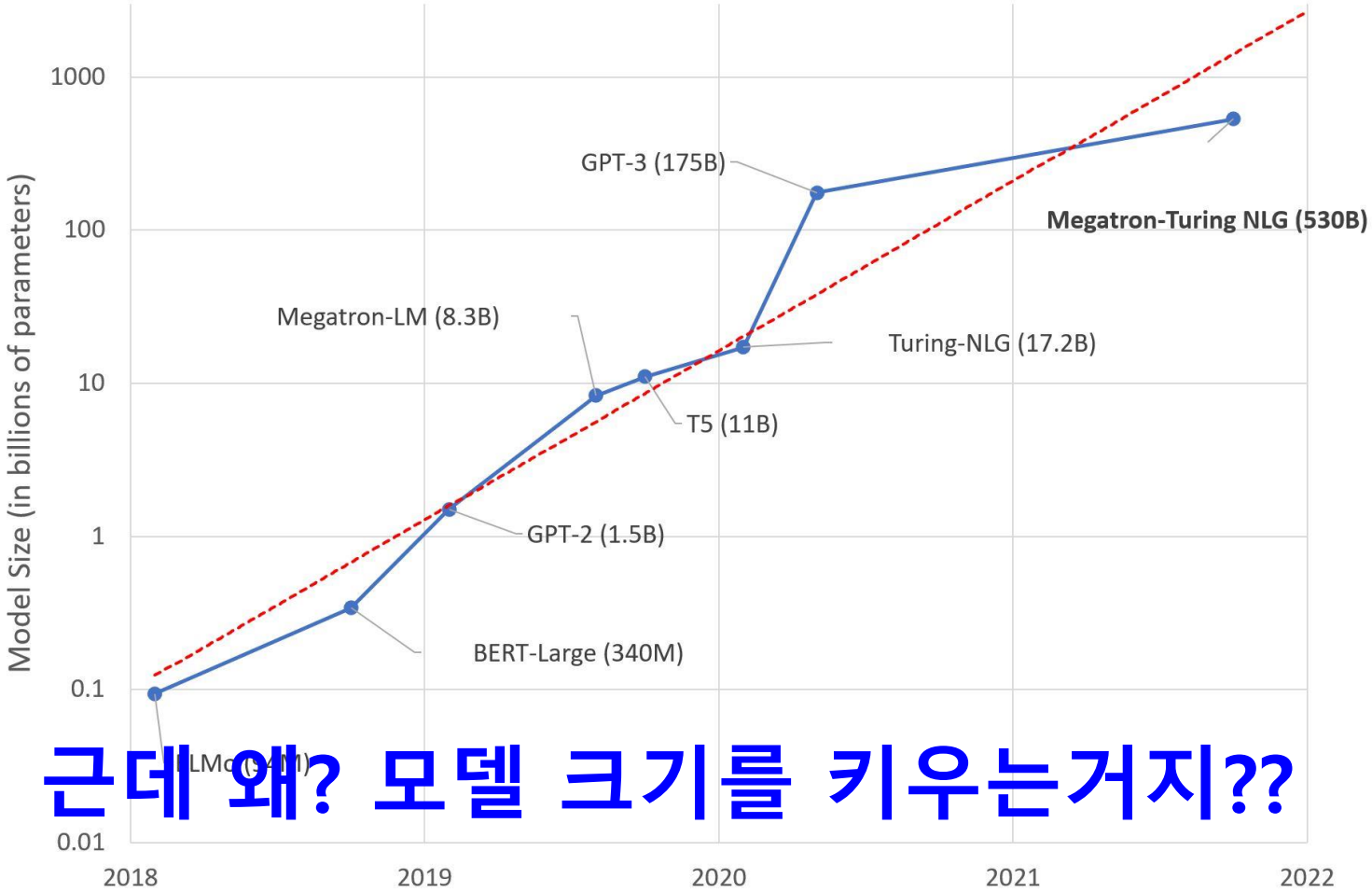


https://arxiv.org/pdf/2303.18223.pdf



https://github.com/Mooler0410/LLMsPracticalGuide.git

Large Language Models: A New Moore's Law?



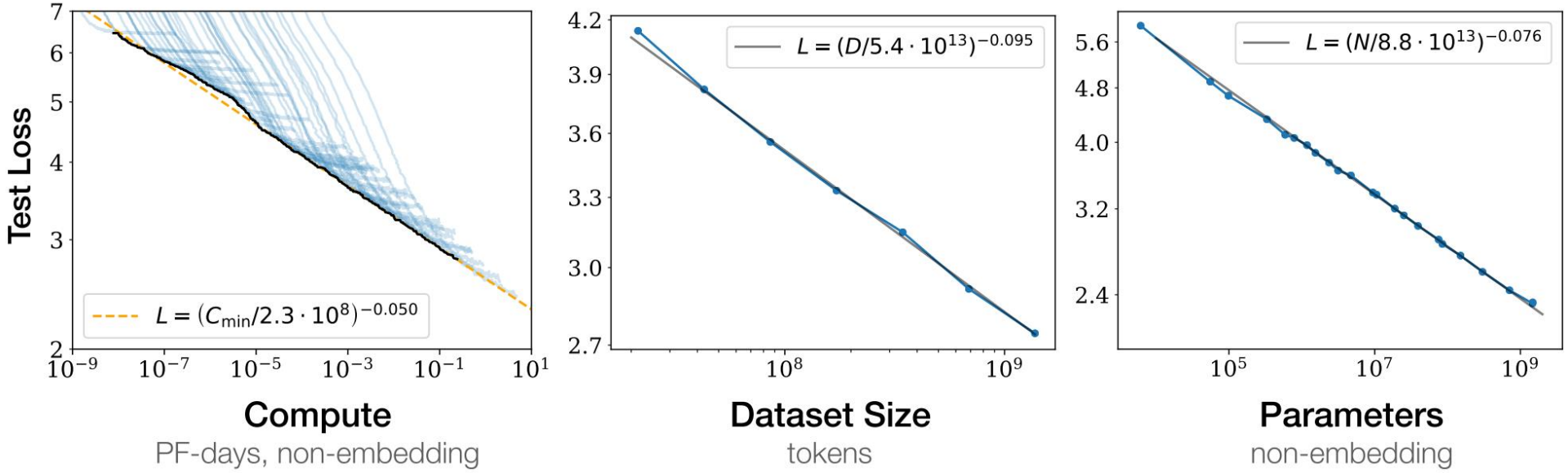
근데 왜? 모델 크기를 키우는거지??

https://huggingface.co/blog/large-language-models

Scaling Laws for Neural Language Models

Scaling Laws for Neural Language Models, Kaplan et al., 2020

- **Compute, Data size, parameters**

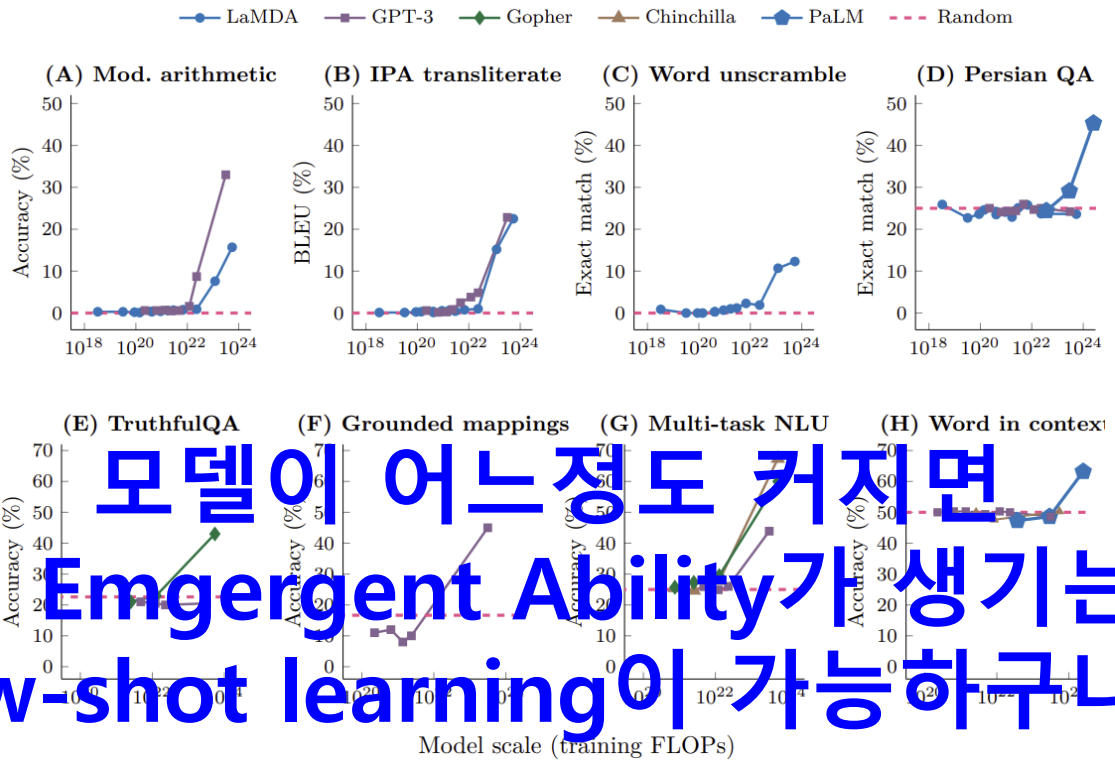


많은 데이터를
큰 모델에 오래 학습하면
성능이 좋아지는구나!?

Emergent Abilities of LLMs

Jason Wei et al., "Emergent Abilities of Large Language Models," TMLR 2022

- AI모델의 성능은 모델 크기에 비례적으로 증가하지 않음
- 작은 모델에선 없던 능력이, 큰 모델에서 갑자기 생김
- 어떤 능력?: Few-shot/Chain-of-thought prompting

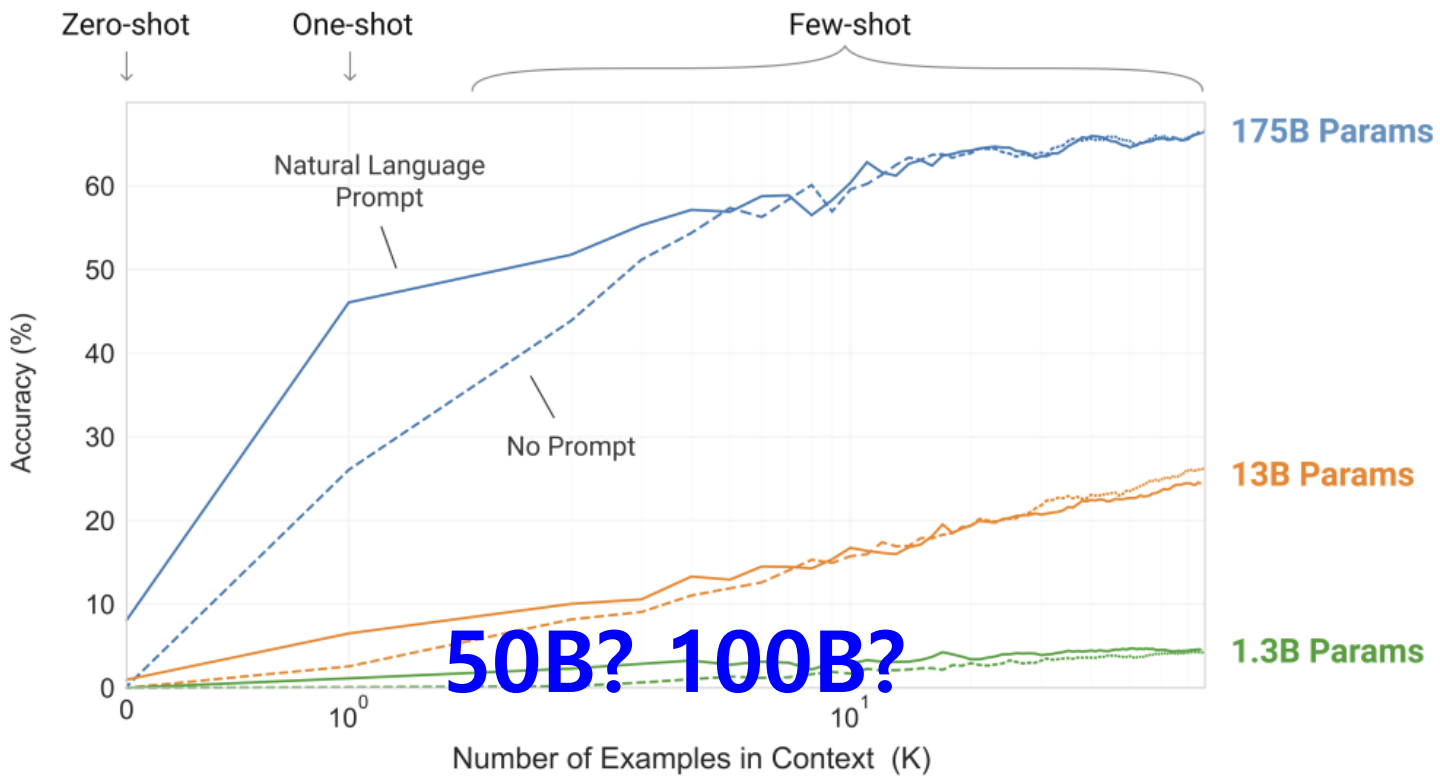


모델이 어느정도 커지면
갑자기 Emergent Ability가 생기는구나!
Few-shot learning이 가능하구나!?

Emergent Abilities of LLMs

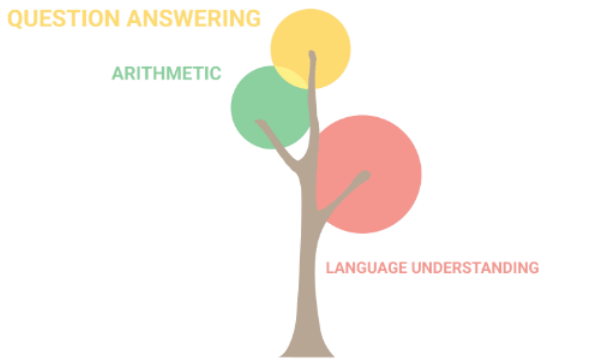
Tom B. Brown et al., "Language Models are Few-Shot Learners," NeuRIPS 2020

- 얼마나 커야 few-shot learning이 가능하지??



Emergent Abilities of LLMs

- 얼마나 커야 few-shot learning이 가능하지??



50B? 100B?

https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

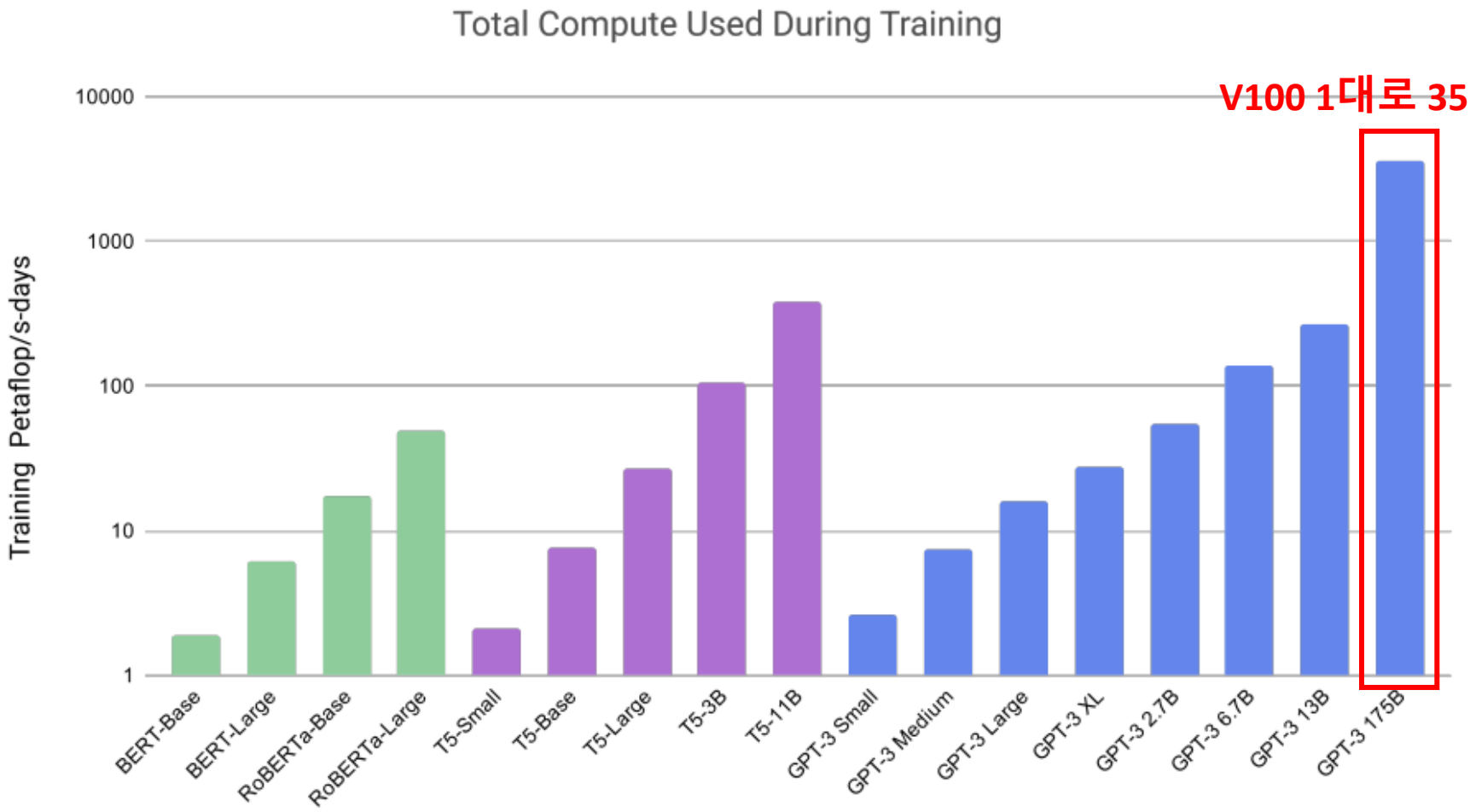


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

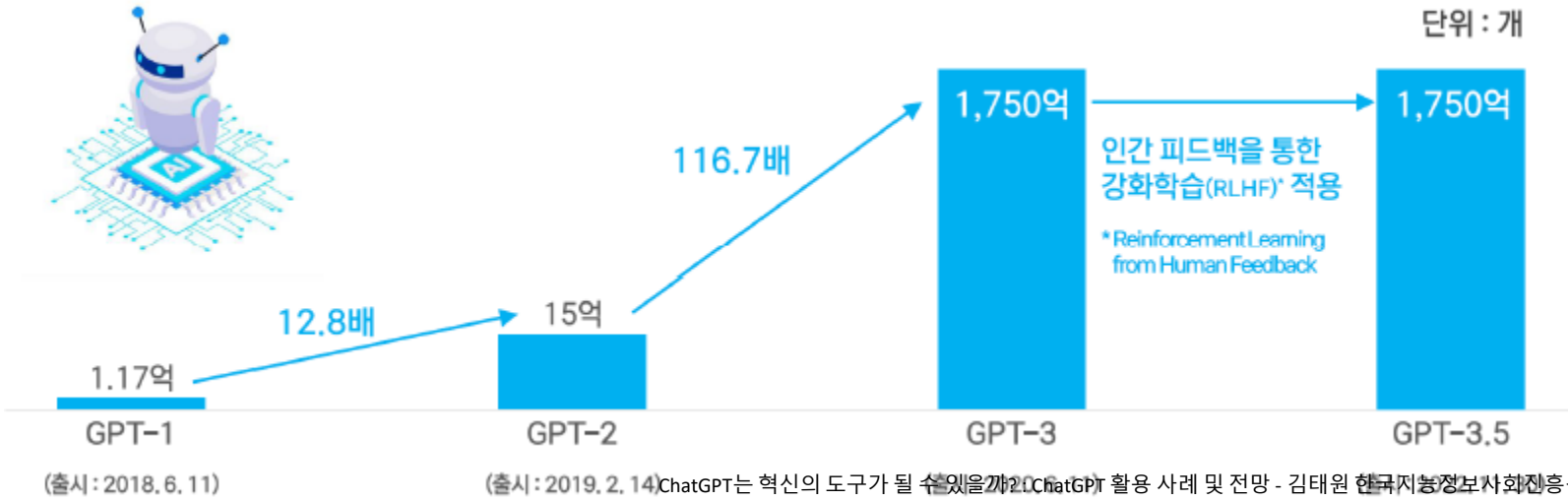
- LLaMA2: ~30days on 2000xA100

- MPT-30B: 2month on 256xH100

https://arxiv.org/pdf/2005.14165.pdf

ChatGPT란?

- 221130, by OpenAI: GPT3, Codex, DALLE
- 인공지능 언어모델인 GPT를 채팅 형식으로 학습한 인공지능 챗봇
- 사용자로부터 입력 받은 문장을 이해하고, 관련 있는 답변을 생성
- ChatGPT는 GPT-3.5(1,750억개)를 사용
- RLHF를 통해 GPT3를 대화에 최적화
- 2021년 4분기까지의 데이터를 학습



ChatGPT는 혁신의 도구가 될 수 있을까? : ChatGPT 활용 사례 및 전망 - 김태원 한국지능정보사회진흥원 수석연구원, 2023
https://www.hellodd.com/news/articleView.html?idxno=99353
https://www.bloter.net/news/articleView.html?idxno=600174
http://news.heraldcorp.com/view.php?ud=20230508000701

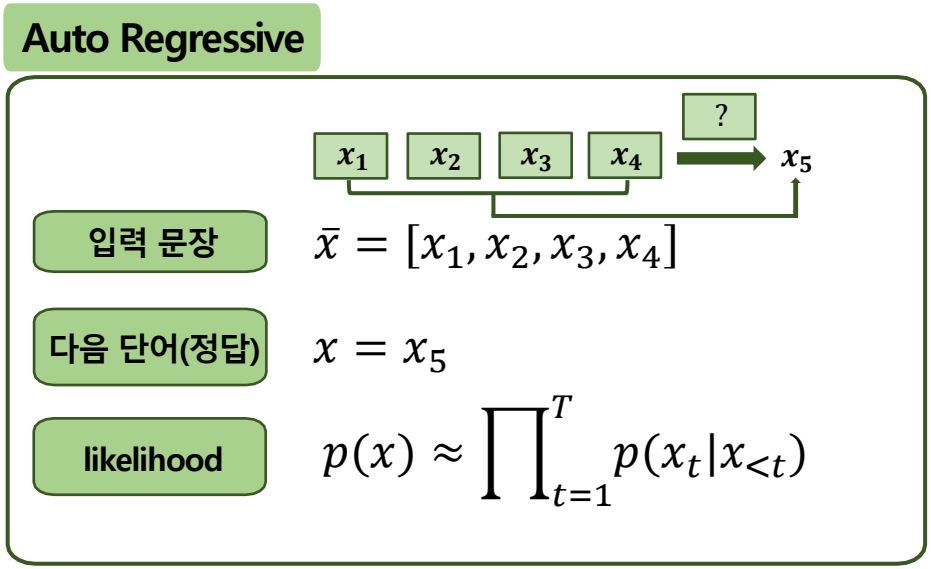
언어모델 (Language Model)

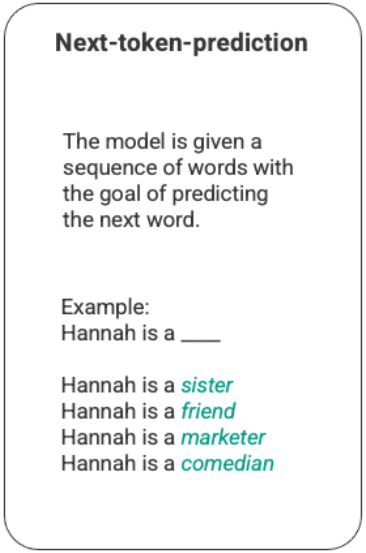
GPT

Generative **P**re-trained **T**ransformer

사전 훈련된 생성 변환기

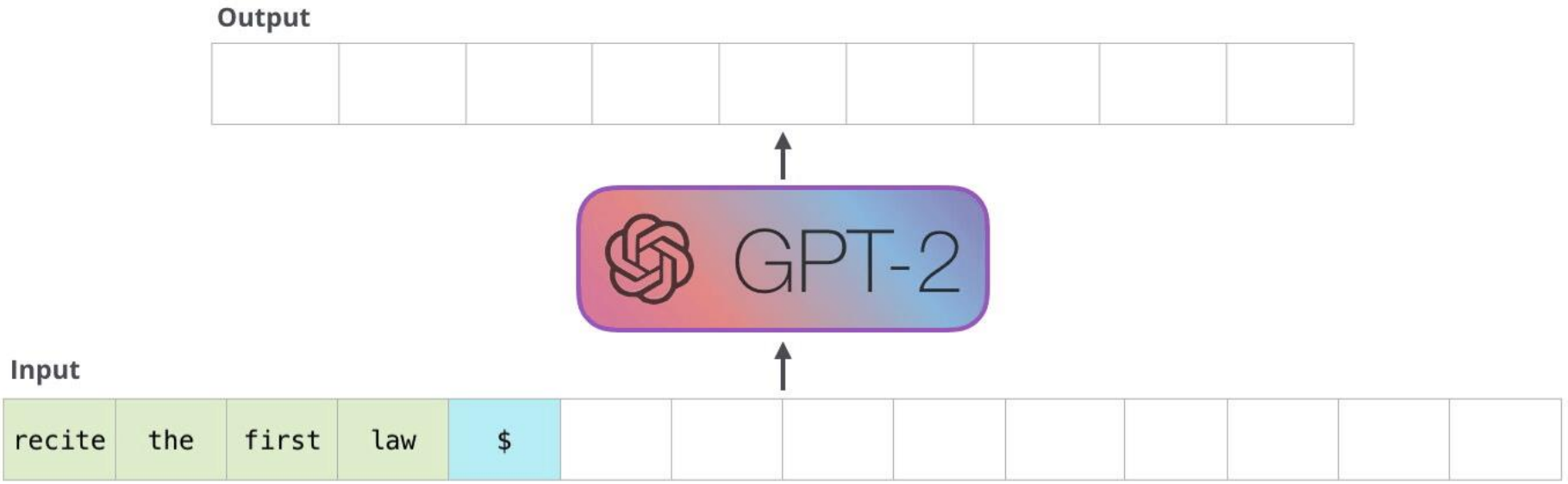
Auto Regressive





ref: https://jalammar.github.io/illustrated-gpt2/

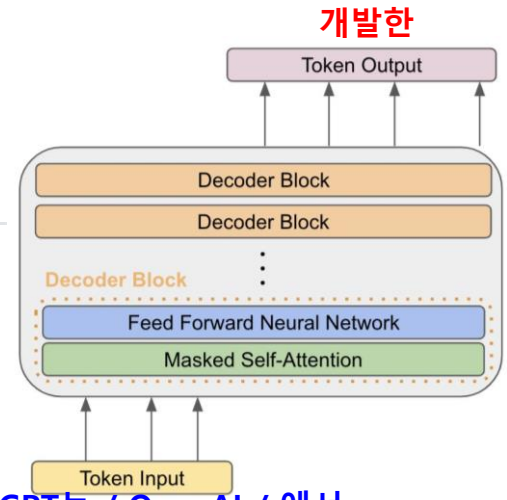
GPT



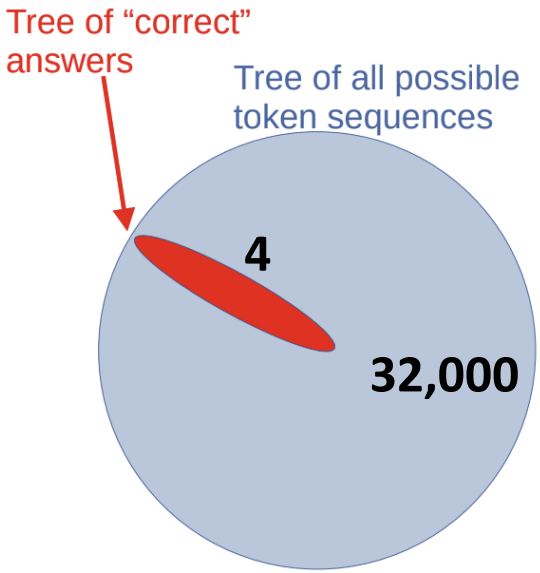
Generate from given prompt

http://jalammar.github.io/illustrated-gpt2/

GPT 입출력



1) ChatGPT는 OpenAI
2) ChatGPT는 OpenAI 에서
3) ChatGPT는 OpenAI 에서 개발한
4) ChatGPT는 OpenAI 에서 개발한 대화형
5) ChatGPT는 OpenAI 에서 개발한 대화형 인공지능
6) ChatGPT는 OpenAI 에서 개발한 대화형 인공지능 모델
7) ChatGPT는 OpenAI 에서 개발한 대화형 인공지능 모델 입니다.



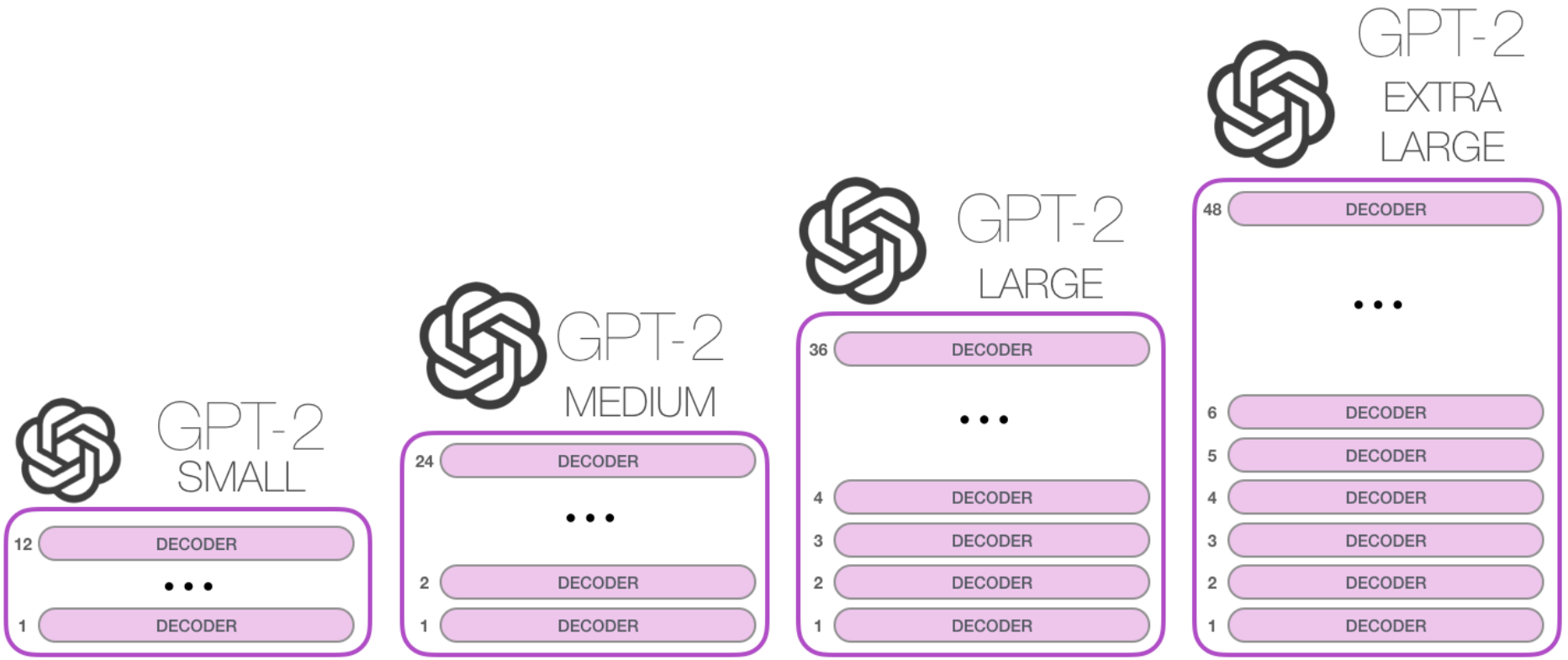
GPT

1.1억

15억

http://jalammar.github.io/illustrated-gpt2/

GPT

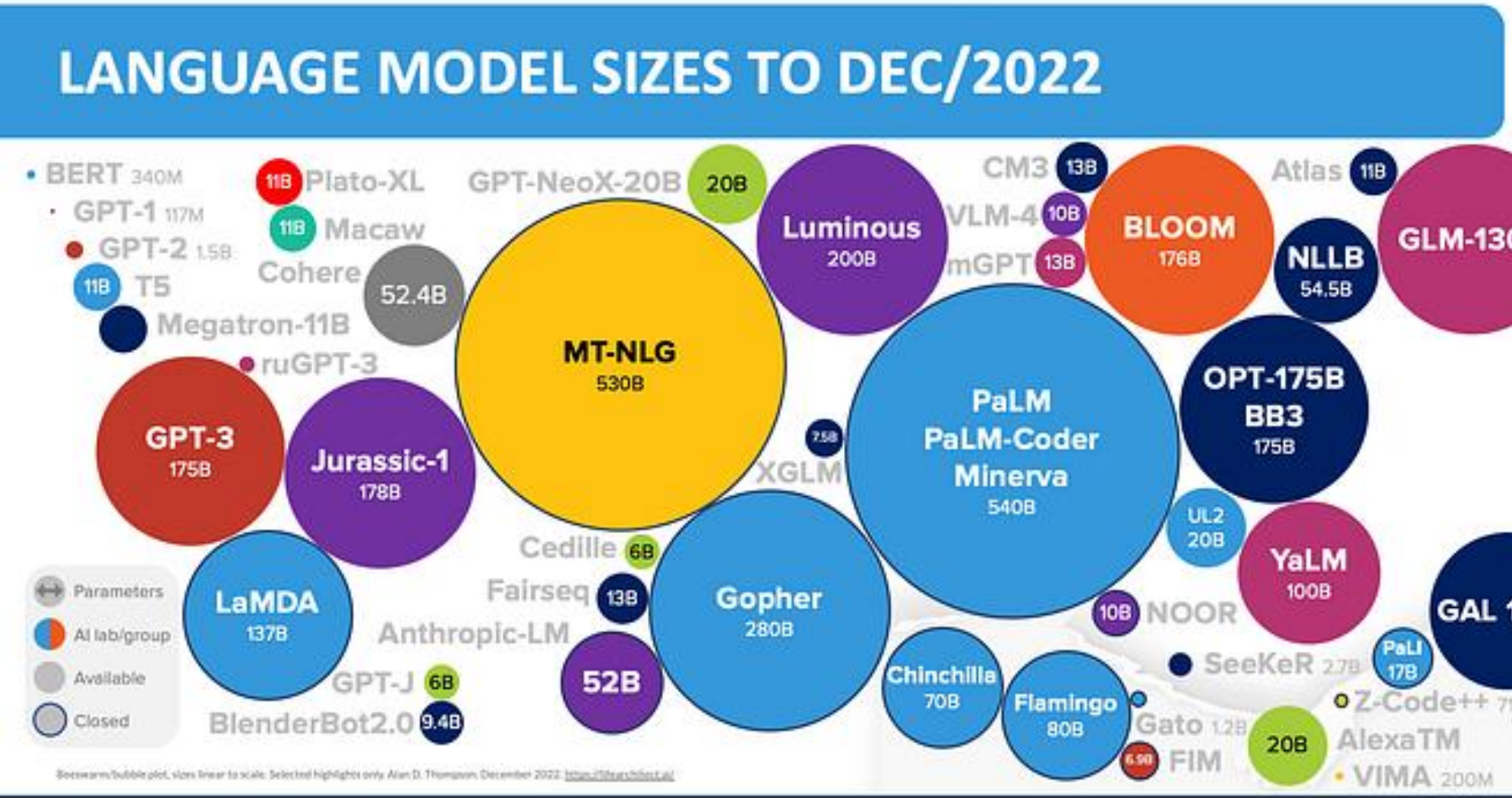


1.1억

http://jalammar.github.io/illustrated-gpt2/

GP

#midjourney #v4, 김성완



https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

https://lifearchitect.ai/models/

나만의 챗봇 만들기

공무원/공공기관 업무 효율화

대국민 AI민원 서비스

XX분야 특화 ChatGPT (원자력/보안/에너지/건축..)

내 일을 대신해주는 ChatGPT

API 기반 LLM

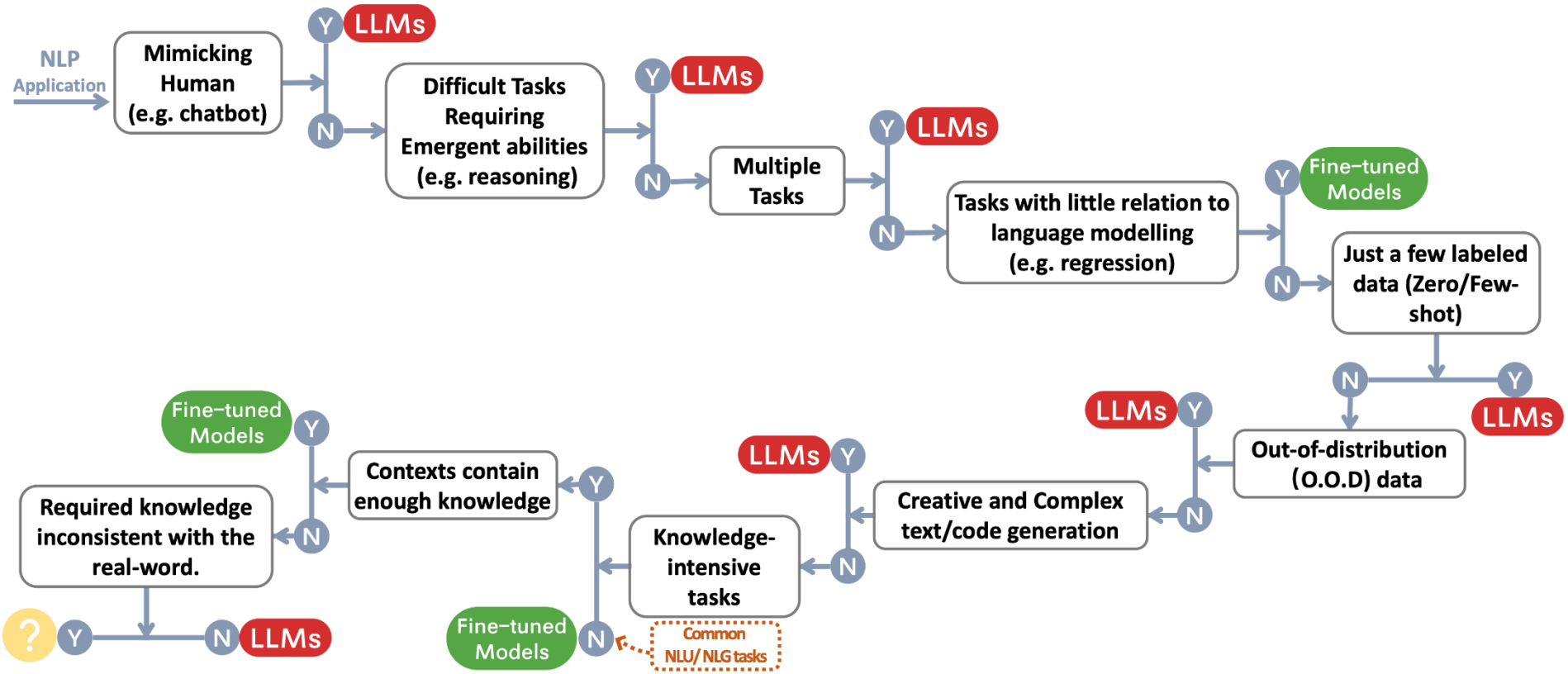
- **장점**
  - 저렴한 운영비용(월 28000원)
  - 좋은 성능

- **단점**
  - 입력 정보가 OpenAI 서버로 전송: **보안**
  - 보안
  - Customizing 불가
    - Fine-tuning을 제공하나 부족함

Amang Kim, Facebook

LLM for everything

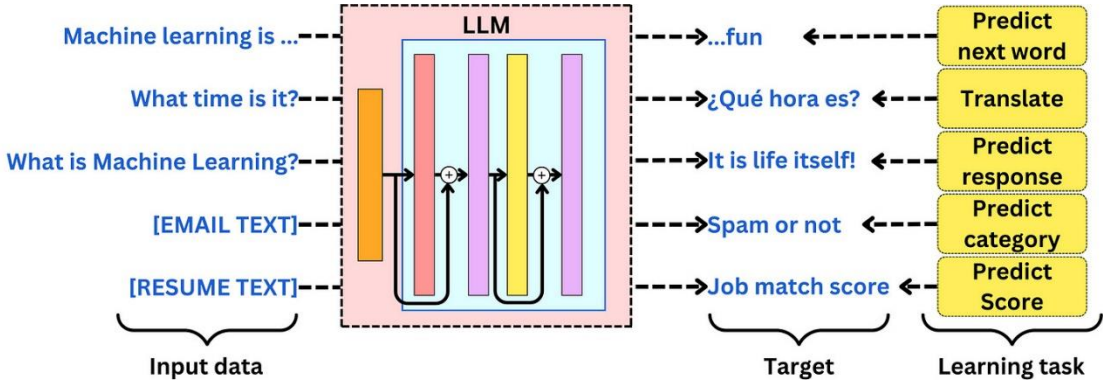
- 거의 모든 task는 LLM으로 해결 가능



https://github.com/Mooler0410/LLMsPracticalGuide/blob/main/imgs/decision.png

ChatGPT 고려사항

- **지식 단절**
  - ~2021년 9월 데이터로 학습
- **답변**
  - 일반적인 질의응답 가능
- **분야**
  - 범용 분야 답변 가능
- **컨트롤**
  - fine-tuning 가능
- **데이터 보안**
  - API를 이용한 질의 응답



https://newsletter.theaiedge.io?utm_source=navbar&utm_medium=web

ChatGPT 고려사항

- **지식 단절**
  - ~2021년 9월 데이터로 학습
  - **최신/기업 데이터 답변 불가**
- **답변**
  - 일반적인 질의응답 가능
  - **환각(hallucination)**
- **분야**
  - 범용 분야 답변 가능
  - **특정 도메인/테스크 전문성 부족**
- **컨트롤**
  - fine-tuning 가능
  - **특화 모델/파이프라인, 이슈 대응 불가**
- **데이터 보안**
  - API를 이용한 질의 응답
  - **고객/기업 정보 유출 가능**

30년 경력 美변호사, 챗GPT 믿었다 '망신'

이채완 기자

입력 2023-05-30 03:00 | 업데이트 2023-05-30 03:00

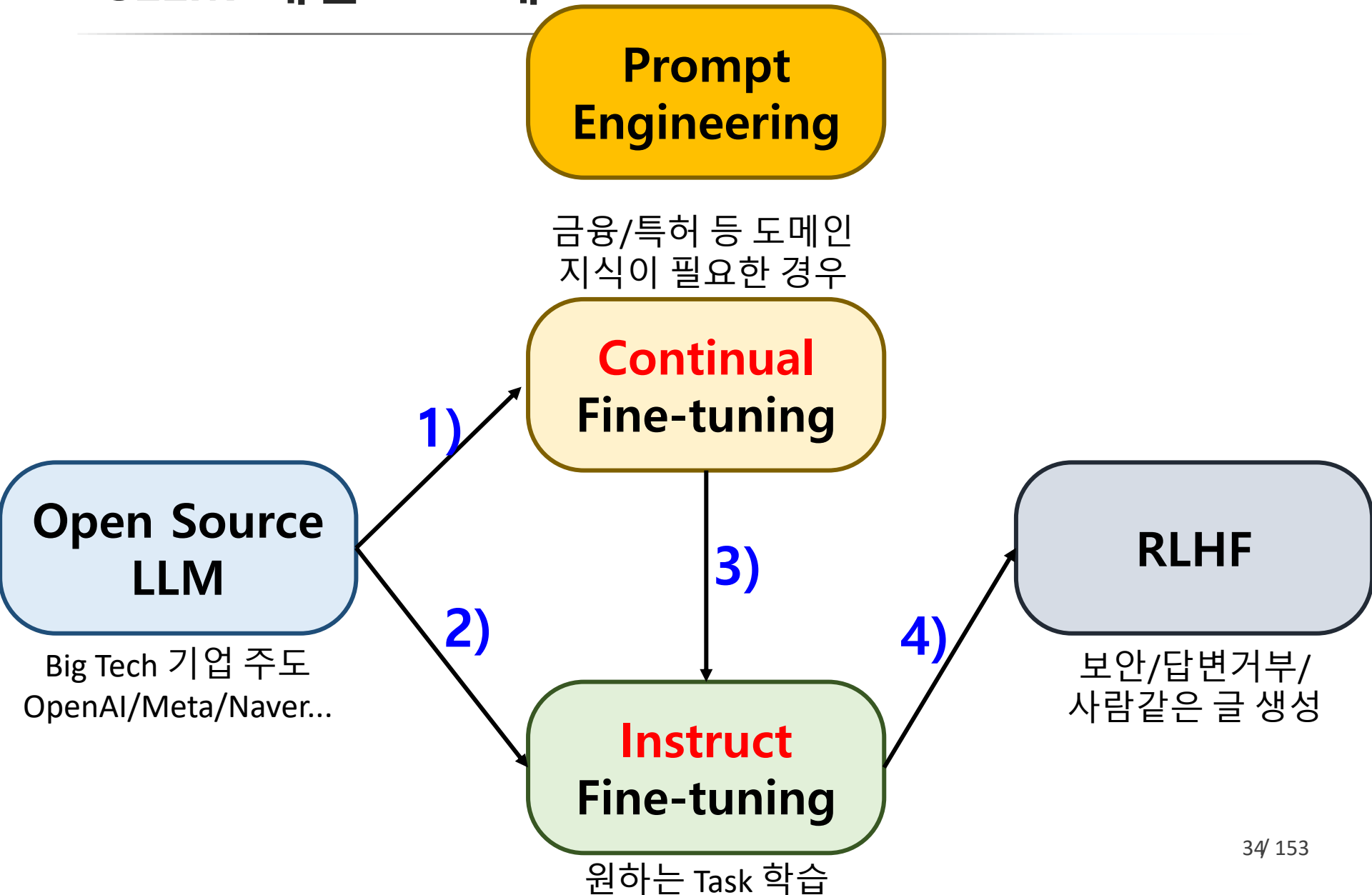
허위-엉뚱한 판례 담은 의견서 제출
판사 "거짓 내용 가득" 청문회 회부
챗GPT는 끝까지 "실제 사례" 주장

30년 경력의 미국 베테랑 변호사가 법원에 서류를 제출하면서 생성형 인공지능(AI) '챗GPT'를 사용해 판례를 인용했다가 제재를 받을 처지에 놓였다. 챗GPT를 통해 인용한 판례가 실제로 존재하지 않는 '거짓'임이 밝혀졌기 때문이다. AI가 만들어낸 각종 거짓 정보에 따른 부작용이 속출하는 가운데 전문직 종사자 또한 AI의 윤리적 사용에 상당한 주의를 기울여야 함을 보여준다는 지적이 나온다.

https://www.donga.com/news/Inter/article/all/20230530/119530095/1

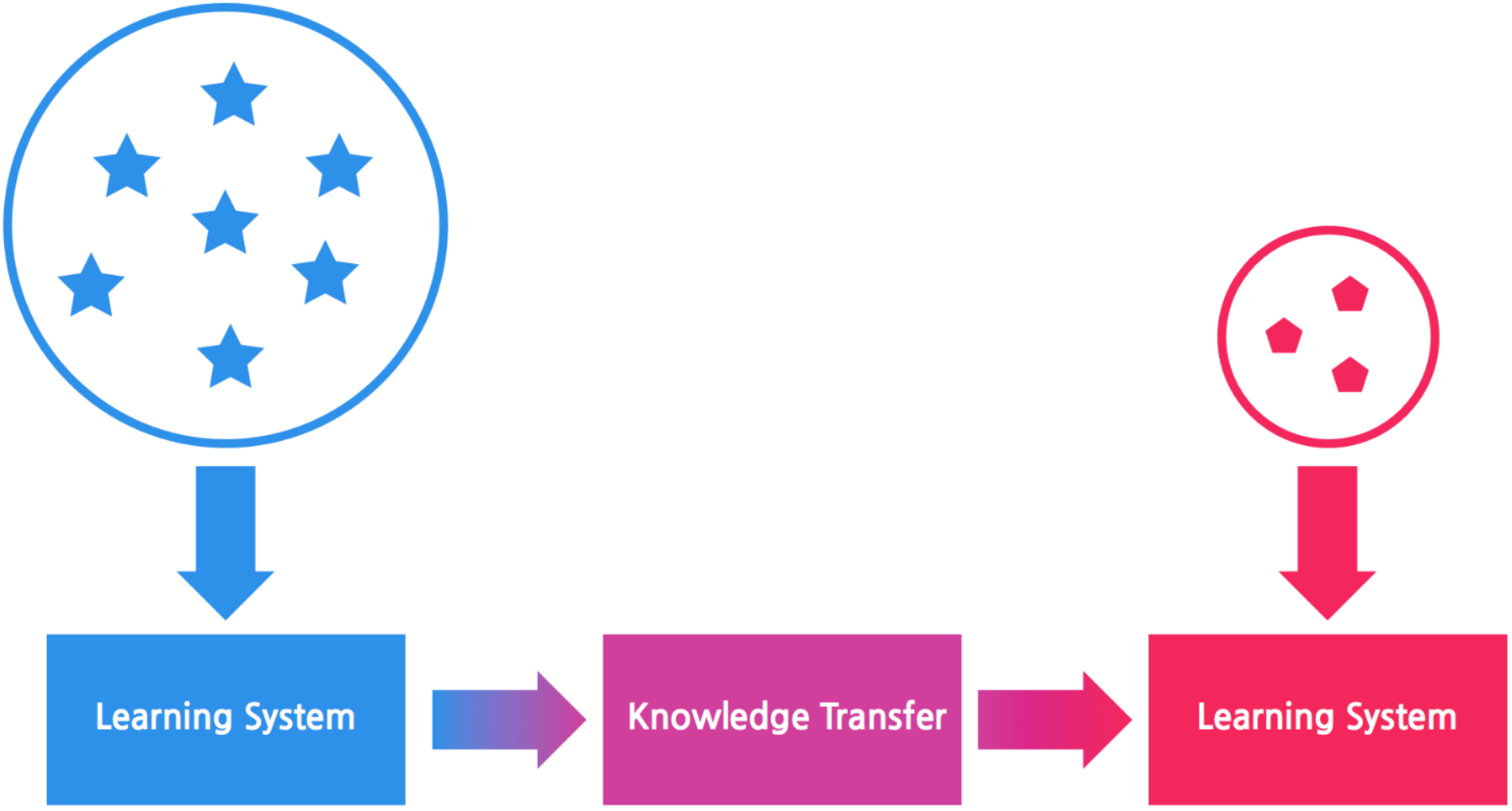
LLM Model Types

sLLM 개발 프로세스

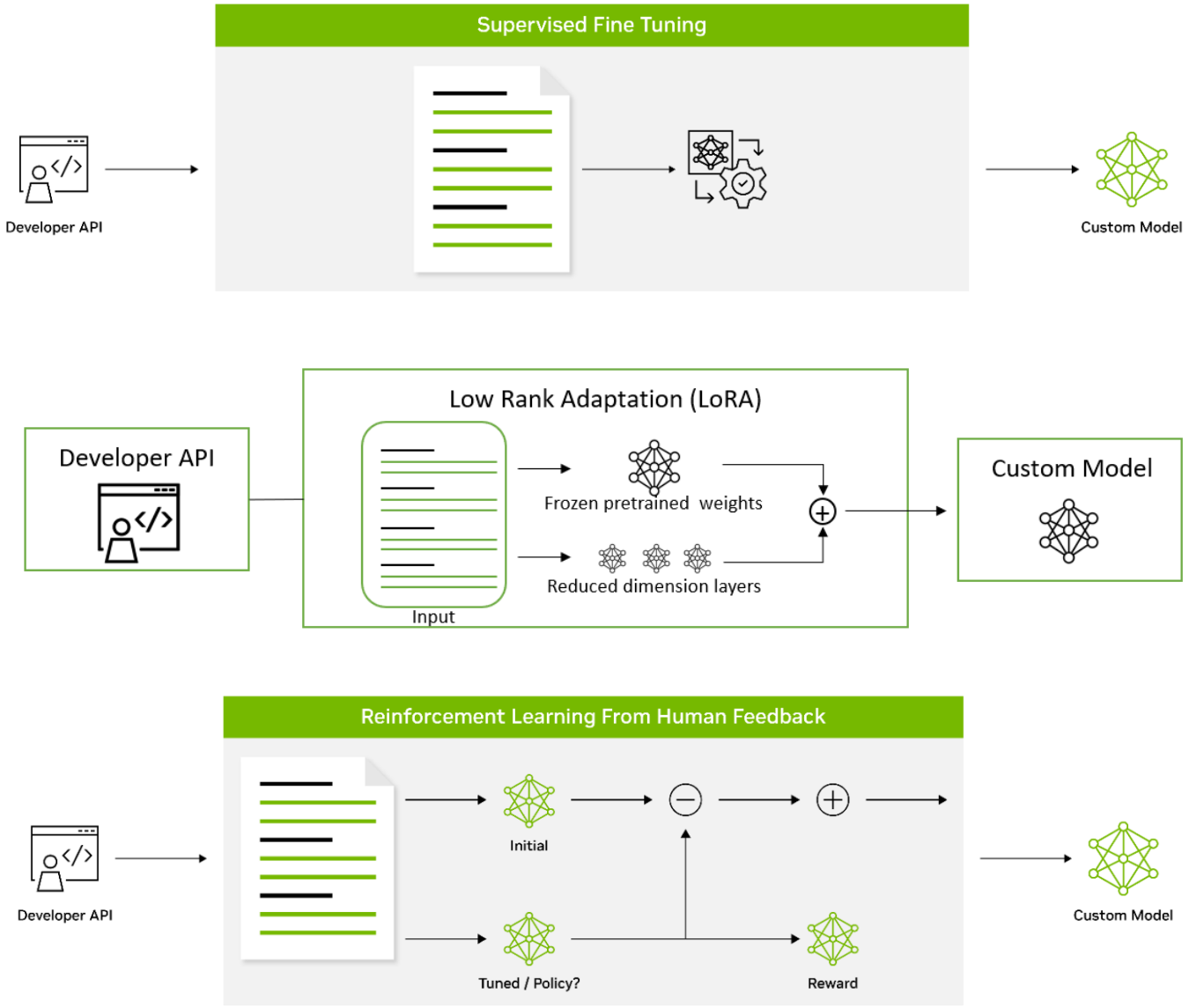

Prompt Engineering
금융/특허 등 도메인 지식이 필요한 경우
Continual Fine-tuning
1)
Open Source LLM
Big Tech 기업 주도
OpenAI/Meta/Naver...
2)
3)
4)
RLHF
보안/답변거부/ 사람같은 글 생성
Instruct Fine-tuning
원하는 Task 학습

GPT3 & 전이학습

- 다음 단어만 '잘' 맞추는 모델 → 전이학습

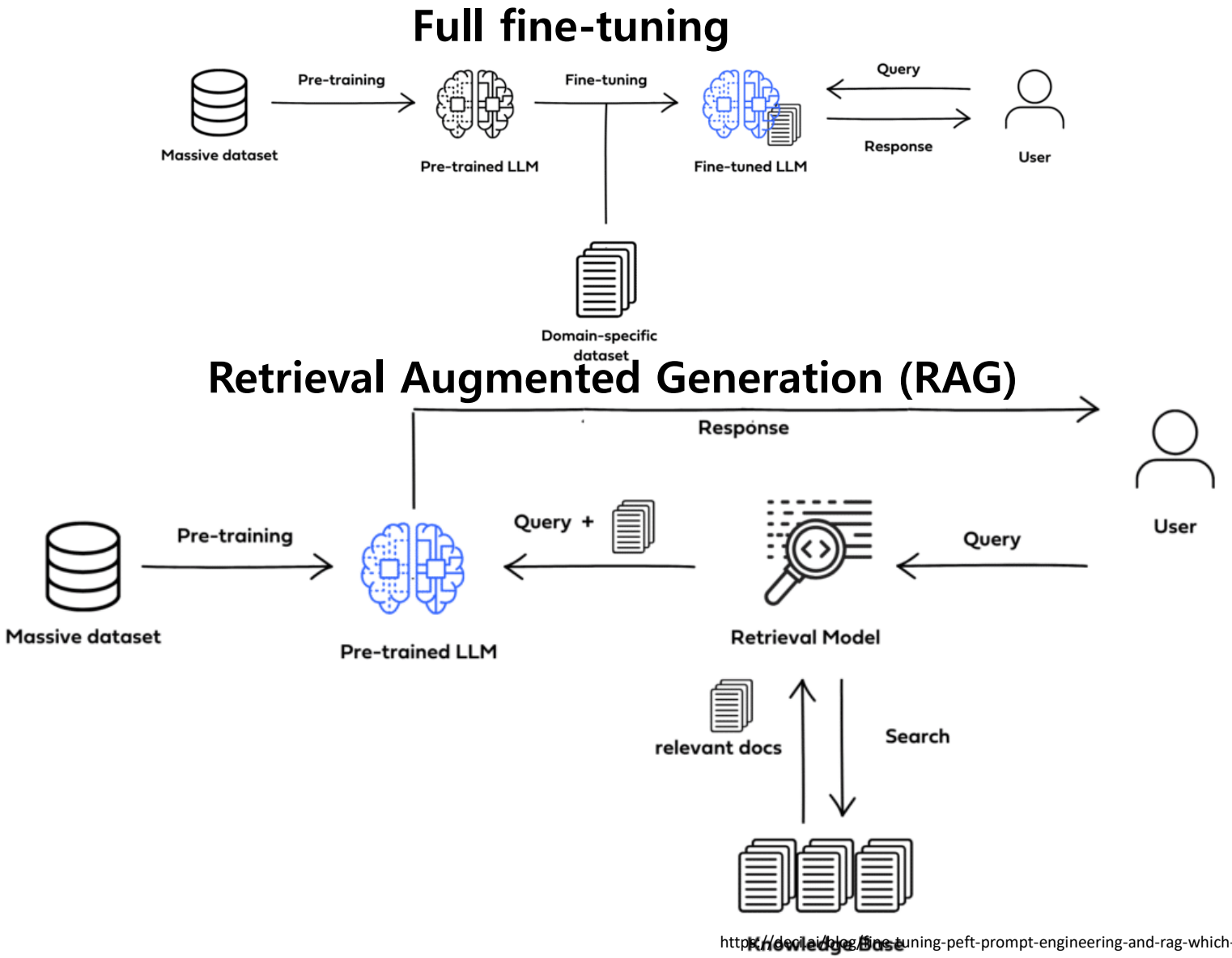


LoRA for parameter-efficient fine-tuning



https://developer.nvidia.com/blog/selecting-large-language-model-customization-techniques/

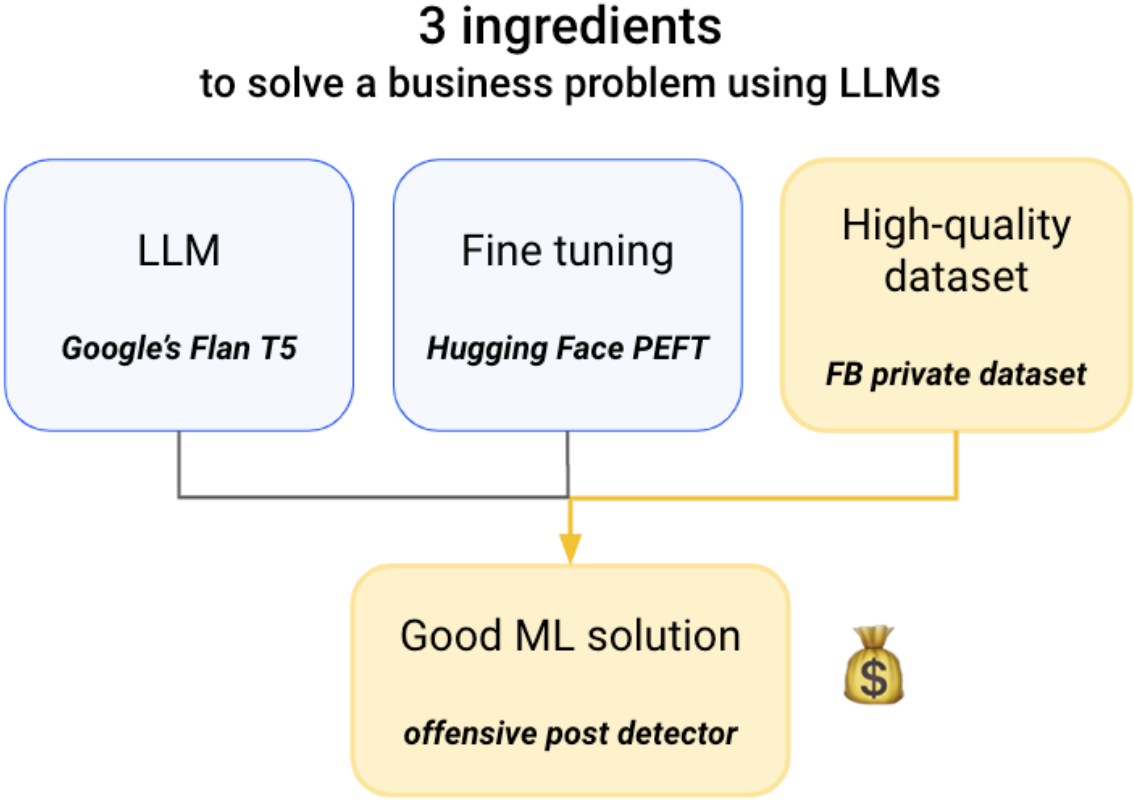
LoRA for parameter-efficient fine-tuning



https://deci.ai/blog/fine-tuning-peft-prompt-engineering-and-rag-which-one-is-right-for-you

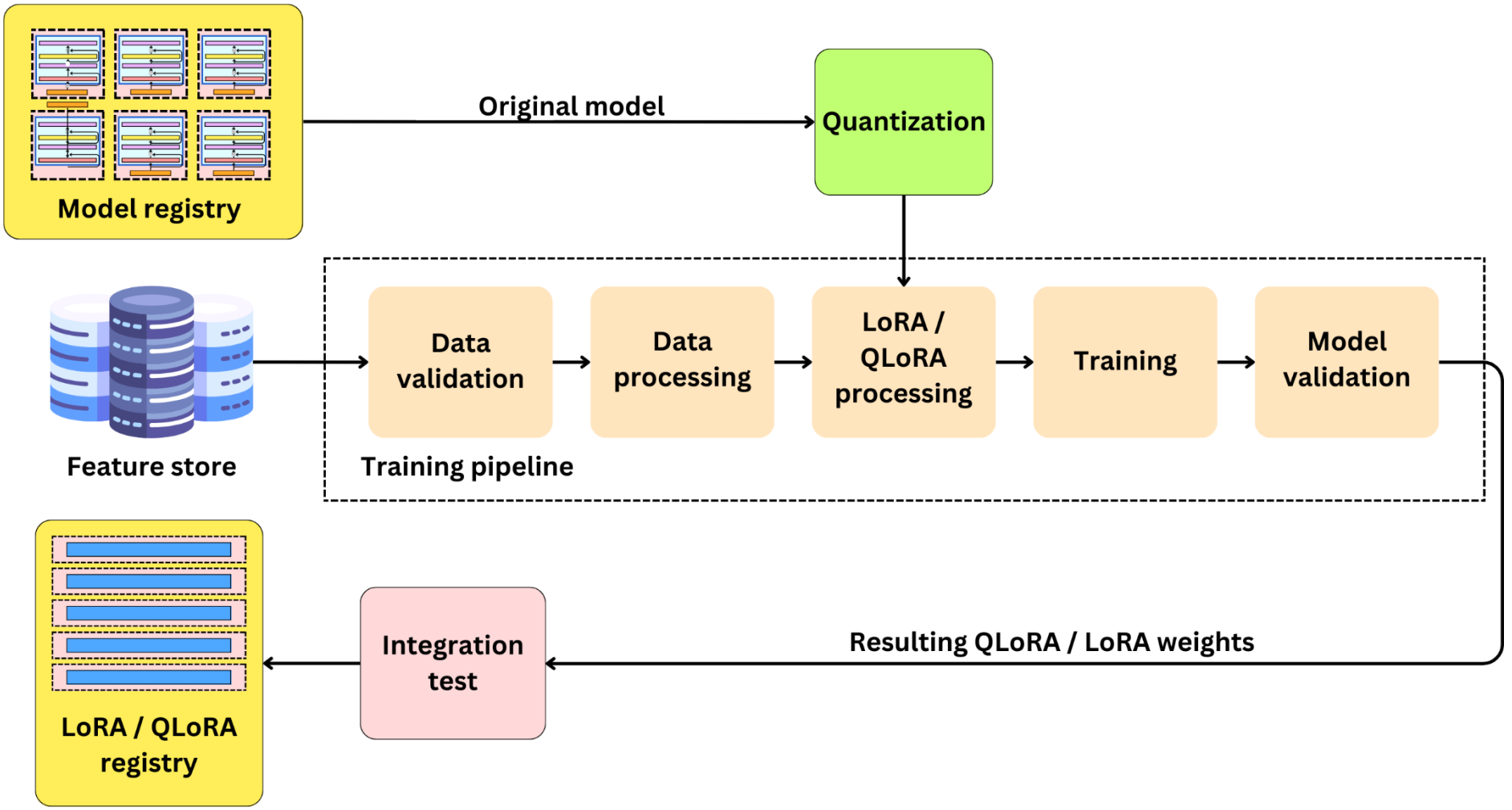
Good ML solution= LLM + Fine tuning + High-quality dataset

- **데이터 중심 접근법(Data-centric approach)**
  - 모델의 품질을 향상시키기 위해 데이터셋의 품질을 개선하는 기술
  - LLMs(대규모 언어 모델)의 예측 오류를 최대 37%까지 줄일 수 있음



https://www.realworldml.xyz/blog/llms-in-real-world-projects

Fine-tuning



https://newsletter.theaiedge.io/p/augmenting-llms-fine-tuning-or-rag?utm_campaign=post&utm_medium=web

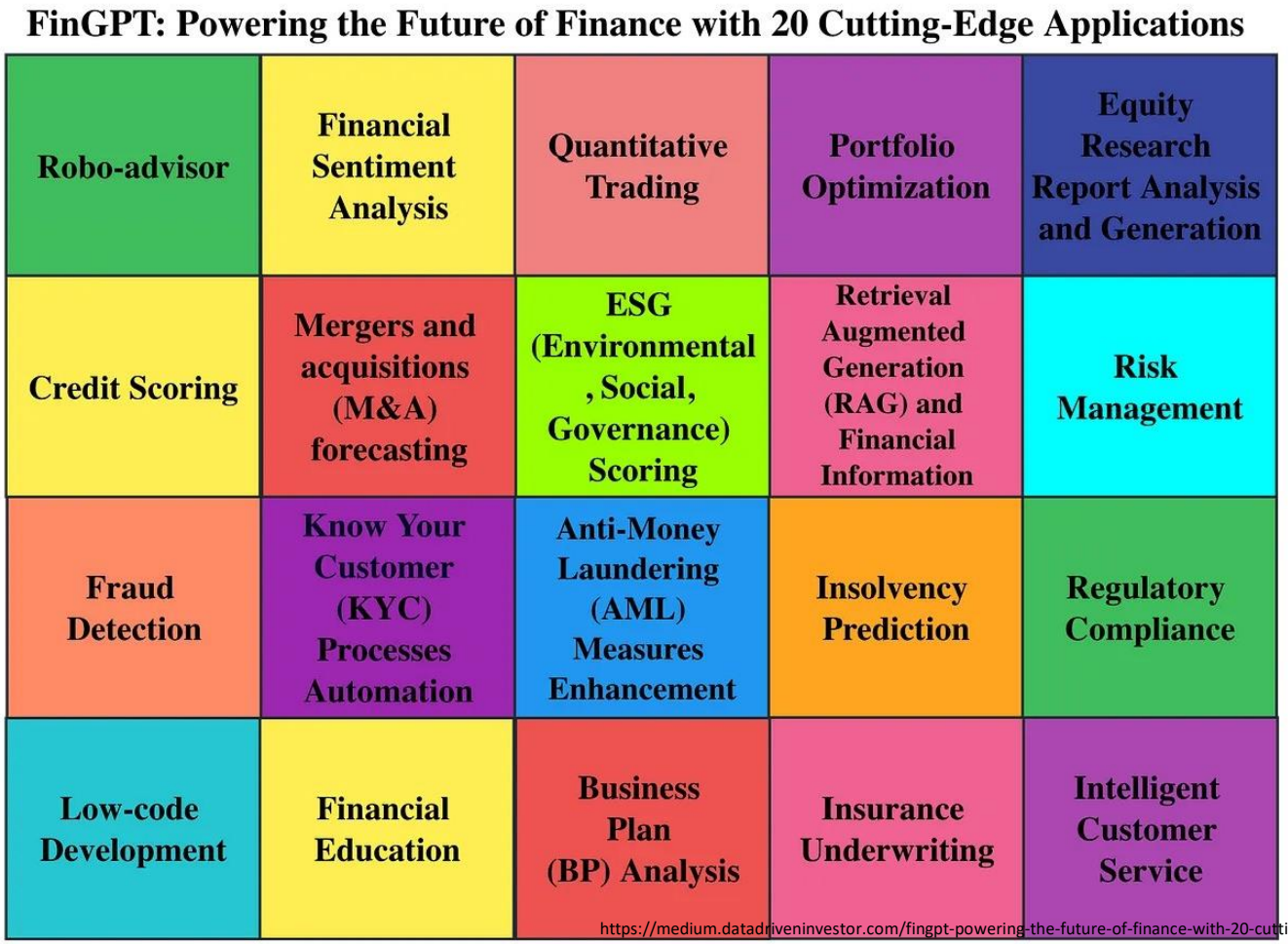
Cost to deploy Llama 2

VS

Cost to deploy ChatGPT

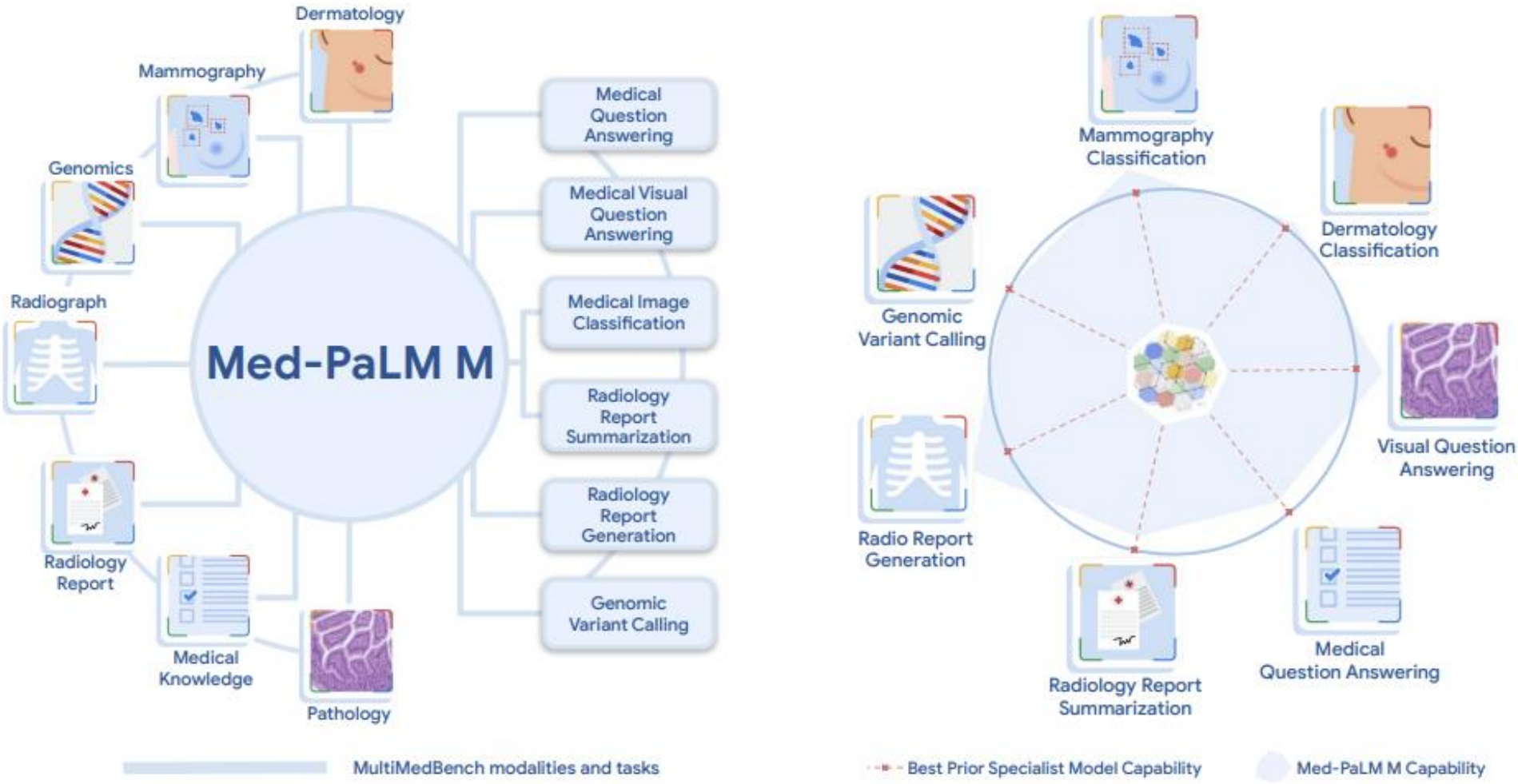
Finance

- 금융분야 20가지의 활용과 LLM 의 사용예



https://medium.datadriveninvestor.com/fingpt-powering-the-future-of-finance-with-20-cutting-edge-applications-7c4d082ad3d8

Biomedicine



with textual context including view orientation and reason for the study in addition to the question. (bottom) shows the task prompt for the dermatology classification task. We formulate the skin lesion classification task as a multiple choice question answering task with all the class labels provided as individual answer options. Similar to the chest X-ray report generation task skin lesion image tokens are interleaved with the patient clinical history as additional context to the question. The blue <img> denotes the position in the prompt where the image tokens are embedded.

Medicine

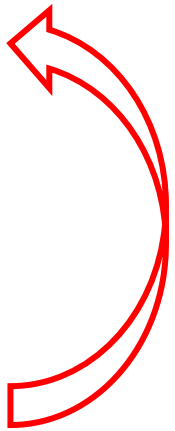
- **A Survey of Large Language Models in Medicine: Progress, Application, and Challenge**

Domains	Model Development	Models	# Params	Pre-train/Fine-tune Data Scale	Data Source
Medical-domain LLMs (Sec. 2.3)	Pre-training (Sec. 2.3.1)	BioBERT [61, 62]	110M	18B tokens	PubMed [63]
		PubMedBERT [64]	110M/340M	3.2B tokens	PubMed [63]
		SciBERT [65]	110M	3.17B tokens	Literature [66]
		ClinicalBERT [67]	110M	112k clinical notes	MIMIC-III [68]
		BlueBERT [69, 70, 71]	110M/340M	>4.5B tokens	PubMed [63] MIMIC-III [68]
		BioCPT [72]	330M	255M articles	PubMed [63]
		BioGPT [73]	1.5B	15M articles	PubMed [63]
		BioMedLM [74]	2.7B	110GB	PubMed [75]
		OphGLM[76]	6.2B	20k dialogues	MedDialog [77]
		GatorTron [78, 23]	8.9B	>82B tokens 6B tokens 2.5B tokens 0.5B tokens	EHRs [23] PubMed[63] Wiki MIMIC-III [68]
		GatorTronGPT[79]	5B/20B	277B tokens	EHRs[79]
	Fine-tuning (Sec. 2.3.2)	DoctorGLM [20]	6.2B	323MB dialogues	CMD. [80]
		BianQue[81]	6.2B	2.4M dialogues	BianQueCorpus [81]
		ClinicalGPT [82]	7B	96k EHRs 192 medical Q&A 100k dialogues	MD-EHR [82] VariousMedQA [83, 22] MedDialog [77]
		Qilin-Med [84]	7B	3GB	ChiMed [84]
		Qilin-Med-VL [85]	13B	>1M medical image-text	ChiMed-VL [85]
		ChatDoctor[19]	7B	110k dialogues	HealthCareMagic [86] iCliniq [87]
		BenTsao [17]	7B	8k instructions	CMeKG-8K [88]
		HuatuoGPT [89]	7B	226k instructions&dialogues	Hybrid SFT[89]
		LLaVA-Med [90]	7B	600k medical image-text	PMC-15M [91])
		Baize-healthcare	7B	101K dialogues	Quora+MedQuAD[92]
		Visual Med-Alpaca [93]	7B	54k medical Q&A	VariousMedQA[93]
		Med-Flamingo [94]	8.3B	0.8M images&584M tokens 1.3M medical image-text	MTB [94] PMC-OA [95]
		MedAlpaca [16]	7B/13B	160k medical Q&A	Medical Meadow [16]
		PMC-LLaMA [21]	13B	79.2B tokens	Books+Literature[96] MedC-I [21]
		Clinical Camel [18]	13B/70B	70k dialogues 100k articles 4k medical Q&A	ShareGPT [97] PubMed [63] MedQA [22]
		MedPaLM 2 [15]	340B	193k medical Q&A	MultiMedQA [15]
		MedPaLM M [98]	12B/84B/562B	>1M medical image-text	MultiMedBench [98]
	Prompting (Sec. 2.3.3)	DeID-GPT [99]	ChatGPT/GPT-4	Chain-of-Thought [100]	-
		ChatCAD [101]	ChatGPT	Zero-shot Prompting	-
		Dr. Knows [102]	ChatGPT	Zero-shot Prompting	UMLS [103, 104]
		MedPaLM [14]	PaLM (540B)	40 instructions	MultiMedQA [15]

https://arxiv.org/abs/2311.05112

도메인 특화 LLM 구축

- 도메인 데이터 수집
- Instruct 정의/학습데이터 구축
- LLM 학습
- 추론 테스트



sLLM

https://modulabs.co.kr/blog/llama-2-intro/

sLLM

- **Small Large Language Model: <20B**
- **LLM은 성능이 좋지만 비용/속도 이슈**
- **특정 task에 특화된 작지만 똑똑한 LM을 만들자**
- **장점:** 비용은 적고 **특정 task** 성능 우수해

대규모언어모델과 소형 대규모언어모델 비교

구분	대규모언어모델(LLM)	소형 대규모언어모델(sLLM)
크기(파라미터)	약 수천억 개	약 수십억 개
예시	오픈AI GPT-4, 구글 팜2, 네이버 하이퍼클로바X, LG 엑사원	메타 라마, 스탠퍼드대 알파카, 스캐터랩 핑퐁
특징	정확하고 복잡한 작업, 방대한 컴퓨팅 자원 필요	적은 컴퓨팅 활용, 특정 영역 언어에 특화, 신속한 파인튜닝

스캐터랩이 자체 개발한 언어모델이 적용된 AI 챗봇 이루다

https://www.mk.co.kr/news/it/10791394?fbclid=IwAR0AEP2UvL8zli7G2KH0Yeo2JzE-4Z4hStsG_2_-Mx07dL1aGZH-6GkLGhA

Use Cases

Get comparable performance to full finetuning by adapting LLMs to downstream tasks using consumer hardware

GPU memory required for adapting LLMs on the few-shot dataset `ought/raft/twitter_complaints`. Here, settings considered are full finetuning, PEFT-LoRA using plain PyTorch and PEFT-LoRA using DeepSpeed with CPU Offloading.

Hardware: Single A100 80GB GPU with CPU RAM above 64GB

Model	Full Finetuning	PEFT-LoRA PyTorch	PEFT-LoRA DeepSpeed with CPU Offloading
bigscience/T0_3B (3B params)	47.14GB GPU / 2.96GB CPU	14.4GB GPU / 2.96GB CPU	9.8GB GPU / 17.8GB CPU
bigscience/mt0-xxl (12B params)	OOM GPU	56GB GPU / 3GB CPU	22GB GPU / 52GB CPU
bigscience/bloomz-7b1 (7B params)	OOM GPU	32GB GPU / 3.8GB CPU	18.1GB GPU / 35GB CPU

Performance of PEFT-LoRA tuned `bigscience/T0_3B` on `ought/raft/twitter_complaints` leaderboard. A point to note is that we didn't try to squeeze performance by playing around with input instruction templates, LoRA hyperparams and other training related hyperparams. Also, we didn't use the larger 13B mt0-xxl model. So, we are already seeing comparable performance to SoTA with parameter efficient tuning. Also, the final checkpoint size is just `19MB` in comparison to `11GB` size of the backbone `bigscience/T0_3B` model.

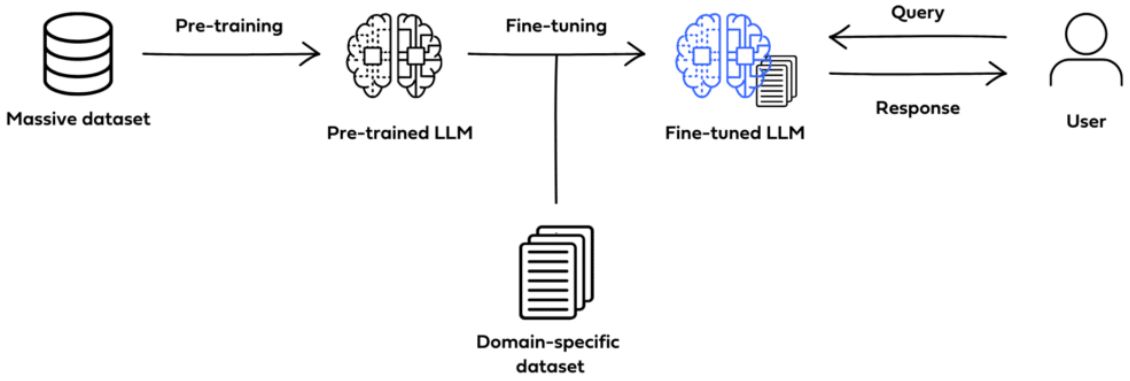
https://arxiv.org/abs/2106.09685
magazine.sebastianraschka.com/p/finetuning-llms-with-adapters?utm_campaign=post&utm_medium=web

PEFT: LoRA

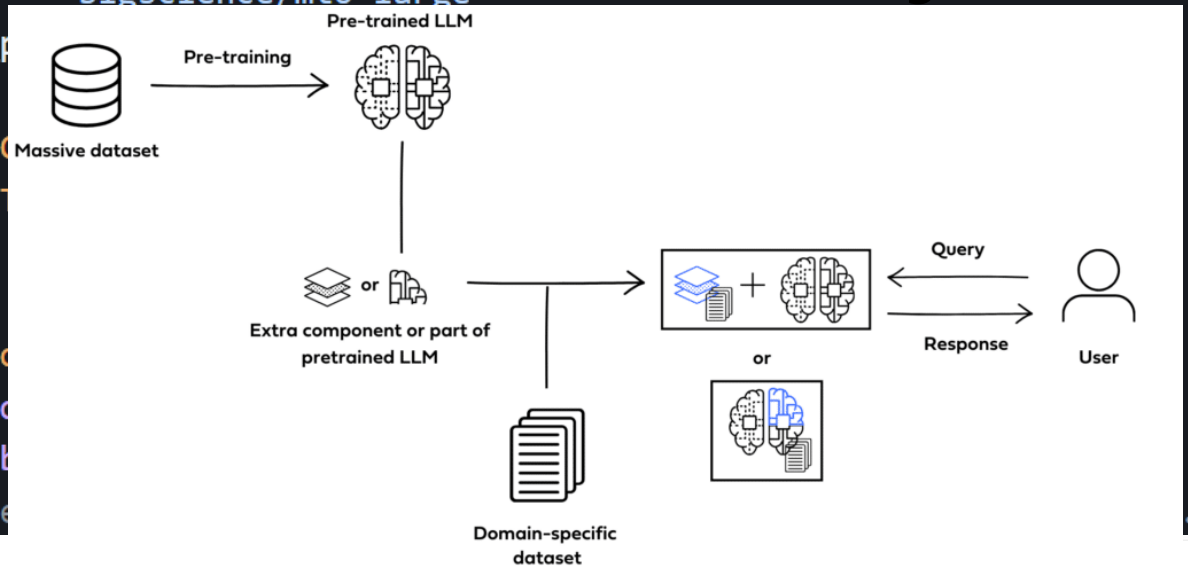
- LLM full f
- PEFT: LLM
- LLM의 아
- 연산과 비

on 시키는 방법

Full fine-tuning



Parameter-efficient fine-tuning




```
from transformers import AutoModelForSeq2SeqLM
from peft import get_peft_config, get_peft_model, LoraConfig, TaskType
model_name_or_path = "bigscience/mt0-large"
tokenizer_name_or_p

peft_config = LoraC
    task_type=TaskT                                    lora_dropout=0.1
)

model = AutoModelFo
model = get_peft_mo
model.print_trainab
# output: trainable                                    .19151053100118282
```


https://arxiv.org/abs/2106.09685

magazine.sebastianraschka.com/p/finetuning-llms-with-adapters?utm_campaign=post&utm_medium=web

PEFT: LoRA

Model & Method	# Trainable Parameters	E2E NLG Challenge BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M ($\text{Adapter}^{\text{L}}$)*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M ($\text{Adapter}^{\text{L}}$)*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M ($\text{Adapter}^{\text{H}}$)	11.09M	$67.3_{\pm.6}$	$8.50_{\pm.07}$	$46.0_{\pm.2}$	$70.7_{\pm.2}$	$2.44_{\pm.01}$
GPT-2 M (FT^{Top2})*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.85}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{71.8}_{\pm.1}$	$\mathbf{2.53}_{\pm.02}$
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L ($\text{Adapter}^{\text{L}}$)	0.88M	$69.1_{\pm.1}$	$8.68_{\pm.03}$	$46.3_{\pm.0}$	$71.4_{\pm.2}$	$\mathbf{2.49}_{\pm.0}$
GPT-2 L ($\text{Adapter}^{\text{L}}$)	23.00M	$68.9_{\pm.3}$	$8.70_{\pm.04}$	$46.1_{\pm.1}$	$71.3_{\pm.2}$	$2.45_{\pm.02}$
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.89}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{72.0}_{\pm.2}$	$2.47_{\pm.02}$

https://arxiv.org/abs/2106.09685

magazine.sebastianraschka.com/p/finetuning-llms-with-adapters?utm_campaign=post&utm_medium=web

sLLM: LLaMA2

- **LLaMA2: Large Language Model Meta AI2**
  - META에서 공개한 거대언어모델. 70억~650억 파라미터 크기
  - Open source / free access / can be used commercially
  - 무료로 상업적 이용이 가능
  - 월간 활성 사용자(MAU)가 7억 명의 회사가 활용할 경우 메타와 별도의 라이센스 계약이 필요
  - **MS**와의 파트너십
  - "소프트웨어가 개방돼 있으면 더 많은 사람이 빠르게 문제를 찾아내고 식별하고 해결할 수 있어 안전과 보안을 향상시킬 수 있다." by 마크 저커버그(메타 CEO)

https://huggingface.co/blog/llama2

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
Llama 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
Llama 1	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
Llama 1	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
Llama 1	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
Llama 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
Llama 2	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
Llama 2	70B	**37.5**	**71.9**	**63.6**	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

https://huggingface.co/meta-llama
https://huggingface.co/blog/llama2

Llama 2

Model	Llama2	Llama2-hf	Llama2-chat	Llama2-chat-hf
7B	Link	Link	Link	Link
13B	Link	Link	Link	Link
70B	Link	Link	Link	Link

- 사용법

```python
from transformers import AutoTokenizer
import transformers
import torch

model = "meta-llama/Llama-2-7b-chat-hf"

tokenizer = AutoTokenizer.from_pretrained(model)
pipeline = transformers.pipeline(
    "text-generation",
    model=model,
    torch_dtype=torch.float16,
    device_map="auto",
)

input_message = 'I liked "Breaking Bad" and "Band of Brothers".

sequences = pipeline(
    input_message,
    do_sample=True,
    top_k=10,
    num_return_sequences=1,
    eos_token_id=tokenizer.eos_token_id,
    max_length=200,
)
for seq in sequences:
    print(f"Result: {seq['generated_text']}")
```

https://huggingface.co/blog/llama2
https://gist.github.com/younesbelkada/9f7f75c94bdc1981c8ca5cc937d4a4da#file-finetune_llama_v2-py

Llama 2 fine-tuning

```
## 1) install
>> pip install trl

## 2) fine-tuning
python trl/examples/scripts/sft_trainer.py \
    --model_name meta-llama/Llama-2-7b-hf \
    --dataset_name timdettmers/openassistant-guanaco \
    --load_in_4bit \
    --use_peft \
    --batch_size 4 \
    --gradient_accumulation_steps 2
```

cuadráticamente separables?### Assistant: El método del Perceptrón biclásico es un algoritmo de aprendizaje automático que se utiliza para clasificar patrones en dos…

https://huggingface.co/blog/llama2
https://huggingface.co/datasets/timdettmers/openassistant-guanaco

Llama 2 추론

- **TGI(Text Generation Inference) 기준**
  - 7B: Nvidia A10
  - 13B: Nvidia A100
  - 70B: Nvidia A100 x 4

https://huggingface.co/blog/llama2

Midjourney by Bruce Yang

RAG

- **Question**
  - 이번에 유럽연합에서 발의한 인공지능법안에 대해 설명해주세요

유럽연합 인공지능법안의 개요 및 대응방안

1. AI법안의 개요
   - 가. 적용 범위와 체계
   - 나. 수인불가 리스크(Unacceptable Risk)를 가진 AI시스템
   - 다. 높은 리스크(High Risk)를 가진 AI시스템
   - 라. 제한적 리스크(Limited Risk)를 가진 AI시스템
     - — 투명성 의무 대상
   - 마. 최저의 리스크(Minimal Risk)를 가진 AI시스템
   - 바. 기타
2. AI법안에 대한 평가와 시사점
   - 가. 리스크 기반 접근의 수단으로서의 인증(적합성평가)
     - — 무분별한 수용의 위험
   - 나. AI 규제 정책에 대한 조율 메커니즘의 정립
   - 다. EU와의 상호인정협정(Mutual Recognition Agreement)의 선제적 준비
   - 라. 미국과의 조율 및 공조

고학수
서울대학교
법학전문대학원
교수

임용
서울대학교
법학전문대학원
부교수

박상철
서울대학교
법학전문대학원
조교수

1) European Commission, COM/2021/206 final, Brussels, 21.4.2021.

2) European Commission, "White Paper on Artificial Intelligence - A Europe-an approach to excellence and trust," COM(2020) 65 final, 2020.

3) European Parliament resolution of 20 October 2020 on a framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020/2012(INL). European Parliament resolution of 20 October 2020 on a civil liability regime for artificial intelligence, 2020/2014(INL). European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI). European Parliament Draft Report, Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters, 2020/2016(INI). European Parliament Draft Report, Artificial intelligence in education, culture and the audiovisual sector, 2020/2017(INI).

4) 우리나라의 경우에도 2021. 7. 31. 현재 7건의 인공지능 관련 법안이 국회에 발의된 상태이다. 그 중 5건('인공지능교육진흥법안', '인공지능 기술 기본법안', '인공지능 집적단지의 육성에 관한 특별법안', '인공지능산업 육성에 관한 법률안', '인공지능 연구개발 및 산업 진흥, 윤리적 책임 등에 관한 법률안')은 소관 상임위원회에서 심사 중이고, 최근 발의된 '인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안'과 '인공지능에 관한 법률안'은 소관 상임위원회에 접수된 상태이다.

5) AI법안의 설명메모(Explanatory Memorandum) 1.1.항.

6) 단, 군용으로 개발되었거나 군용으로만 활용되는 AI시스템(Art. 2(3)), 제3국의 관할 또는 국제기구가 EU 또는 회원국과의 사법공조를 위한 국제협약의 틀 내에서 AI시스템을 활용할 경우(Art. 2(4))에는 적용이 배제된다.

7) AI법안상의 "risk"는 단순한 위험(danger)뿐만 아니라 불확실성(uncertainty)에 기반한 우려(도덕적 내용)하는 개념이어서 "위험" 이라는 일반적인 역어를 사용하지 않고 "리스크" 라고 표기한다. AI법안은 리스크 기반 접근을 취한다고 설명하고 있지만, 개별 AI시스템의 리스크를 어떻게 평가하고 분류할 것인지에 대한 내용을 담고 있지는 않아서, 그에 대한 비판적인 시각도 있다.

유럽집행위원회(European Commission, 이하 'EC')는 2021. 4. 21. 유럽의회(European Parliament)에 「인공지능에 관한 통일규범(인공지능법)의 제정 및 일부 연합제정법들의 개정을 위한 법안(Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts)」(이하 'AI법안')을 발의하였다.[1] AI법안은 EC의 2020. 2. 19.자 "인공지능 백서 – 수월성과 신뢰성에 대한 유럽식 접근방식" 보고서[2] 및 후속한 유럽의회의 2020~2021년 간의 AI 관련 윤리, 책임, 저작권, 형사, 교육·문화·시청각에 관한 각종 결의들[3]을 계승하고 있으며, 주요국 최초의 AI에 관한 일괄(omnibus) 규제 법안으로 이해된다.[4] AI법안은 공식적으로는 기본권과 유럽연합(이하 'EU')의 가치의 보호, 투자와 혁신 촉진, 기존 법령의 집행권한 및 집행력 강화, EU 단일시장의 발전을 지향한다.[5] EU가 AI의 법적 규율에 있어 미국식 자유시장과 중국식 권위주의 사이의 "제3의 길"을 모색하는 계기가 될 것이라는 기대도 있다. 그러나, 개별 조항들을 살펴보면, 2020. 12. 15. 입안된 디지털서비스법 패키지(The Digital Services Act package)와 마찬가지로 미·중에 AI 기술 내지 플랫폼 산업의 주도권을 빼앗겼다는 위기의식 및 강도 높은 보호 무역적 규제 체계를 통해 이를 돌파하고자 하는 의지도 엿보인다. 주요 내용은 이하와 같다.

1. AI법안의 개요

가. 적용 범위와 체계

AI법안은 통과될 경우 AI시스템을 EU 내에서 출시(placing on the market) 또는 서비스(putting into service)하는 제공자들(providers)이나 EU 내에 위치한 AI시스템의 활용자들(users)에게 적용되고(Art. 2(1)(a), (b)), 제3국에 위치한 AI시스템의 경우에도 그 시스템의 산출물이 EU에서 활용될 경우 그러한 시스템의 제공자들과 활용자들에게도 적용된다(Art. 2(1)(c)).[6] 여기서 활용자란 사적·비전문적 활동 과정에서 AI시스템을 활용하는 경우를 제외하므로(Art. 3(4)), 소비자가 아닌 영업 목적의 활용자를 의미한다는 점에 특히 유의해야 한다.

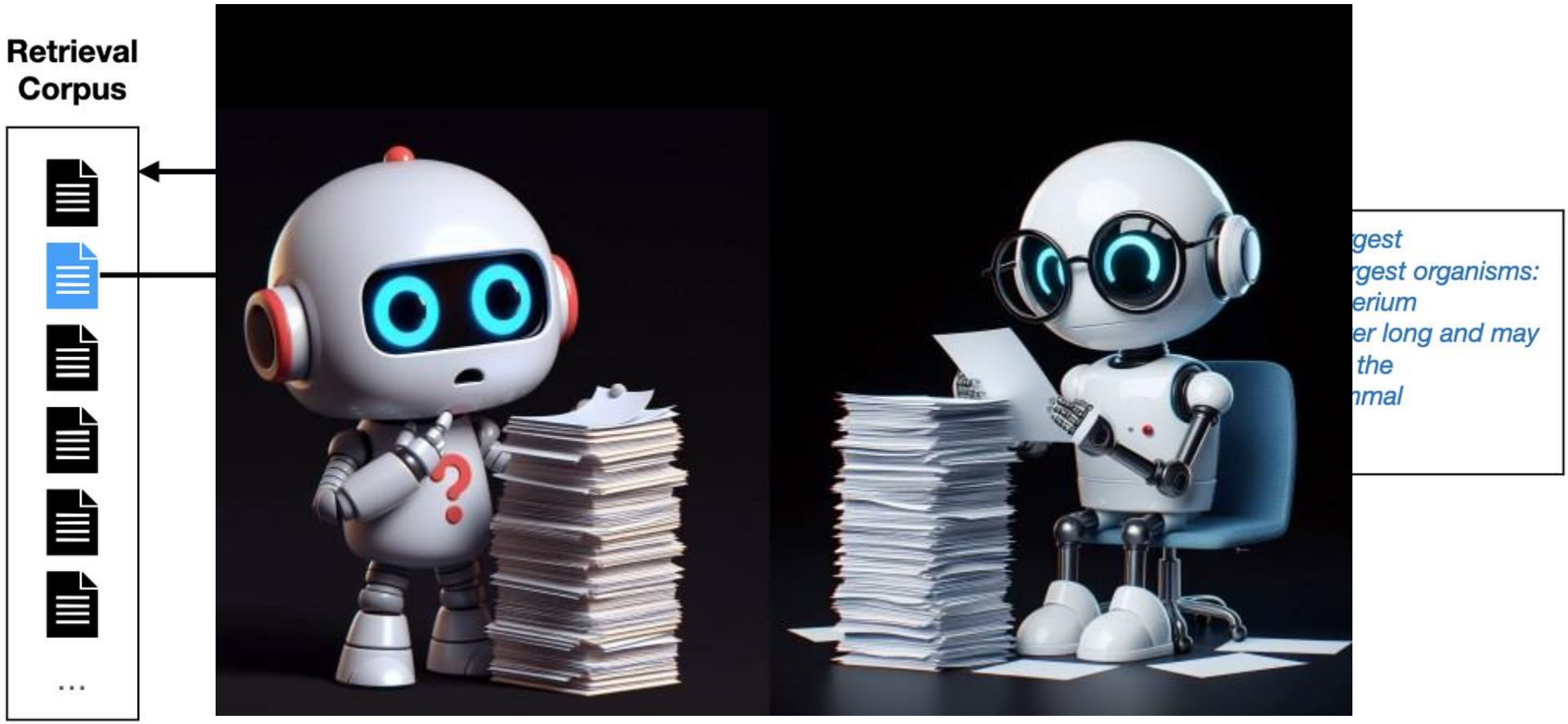
동 법안에서 AI시스템(artificial intelligence system)은 (i) 인간이 정의한 일련의 목적을 위해, (ii) 기계학습(machine learning), 논리/지식 기반 접근(logic- and knowledge-based approaches), 통계적 접근(statistical approaches), 베이즈 추정법(Bayes estimation), 검색 및 최적화 방법(search and optimization methods) 중 하나 또는 복수의 기술 내지 기법을 활용하여 개발되고, (iii) 해당 시스템이 상호작용하는 환경에 영향을 미치는 콘텐츠·예측·추천·결정 등의 출력을 생성하는 소프트웨어로 정의되고 있다(Art. 3(1), Annex I).

이처럼 폭넓게 정의된 AI시스템에 대해 AI법안은 이를 관련 리스크에 따라 구분하여 달리 취급하는 **리스크 기반 접근(risk-based approach)**을 취하고 있다(Preamble (14)).[7] 즉, AI시스템의 리스크를 수인불가 리스크(unaccept-

http://sapi.co.kr/wp-content/uploads/2021/09/%EC%9C%A0%EB%9F%BD%EC%97%B0%ED%95%A9-%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%EB%B2%95%EC%95%88%EC%9D%98-%EA%B0%9C%EC%9A%94-%EB%B0%8F-%EB%8C%80%EC%9D%91%EB%B0%A9%EC%95%88_0923_2.pdf
https://bbs.ruliweb.com/news/board/1005/read/2788207
https://www.ziksir.com/news/articleView.html?idxno=5722

1) DB 기반 챗봇

- 1) 질문과 유사한 DB 내 정보 검색
- 2) 질문+유사정보를 입력으로 LLM에게 답변 생성



Facebook by Mark Kim
https://www.together.xyz/blog/openchatkit?fbclid=IwAR3sB9IPx0WdPHubEo2dPqP1-rfgVDF650RnlT6oWznKuh5FzRi8GhSmjp0

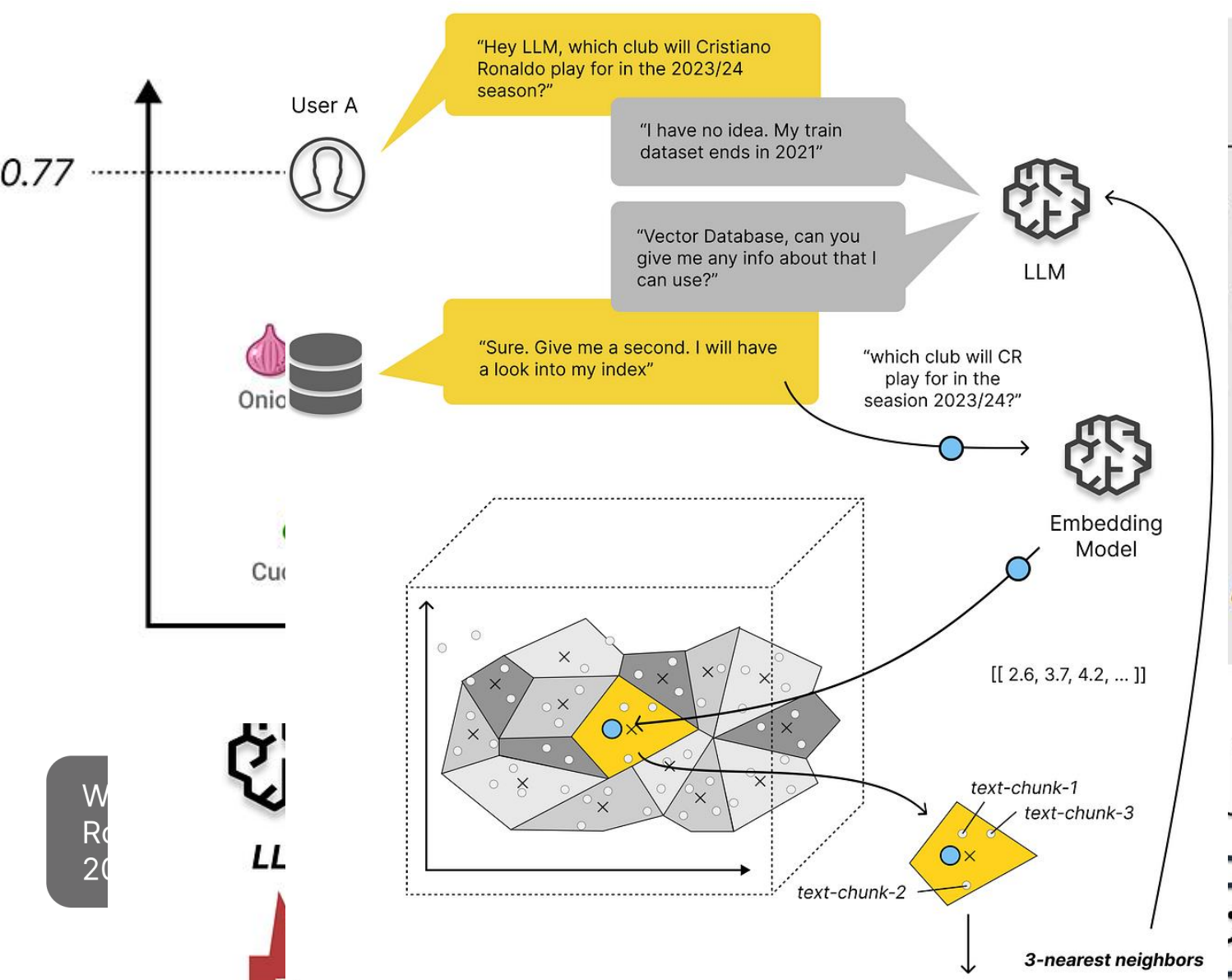
검색 증강 생성: RAG(Retrieval-Augmented Generation)

- **LLM을 애플리케이션에 직접 도입할 때 문제**
  - LLM의 정보 부족(long-tail), 제한된 답변 능력
  - 2021년 10월 이후 데이터가 없어 최신 답변 제공 불가
  - 업데이트(재학습) 어려움
  - 결과물은 해석/검증 어려움

- **해결방안**
  - 최신/도메인 데이터로 fine-tuning: 매일??
  - 원하는 정보가 담긴 문서를 프롬프트에 추가해서 질의: 비용, 길이 제한
  - **RAG**: 주요 정보를 DB에 저장, 사용자의 질의가 들어올 때 관련정보를 검색해서 프롬프트에 추가하여 LLM에 질의

https://eugeneyan.com/writing/llm-patterns

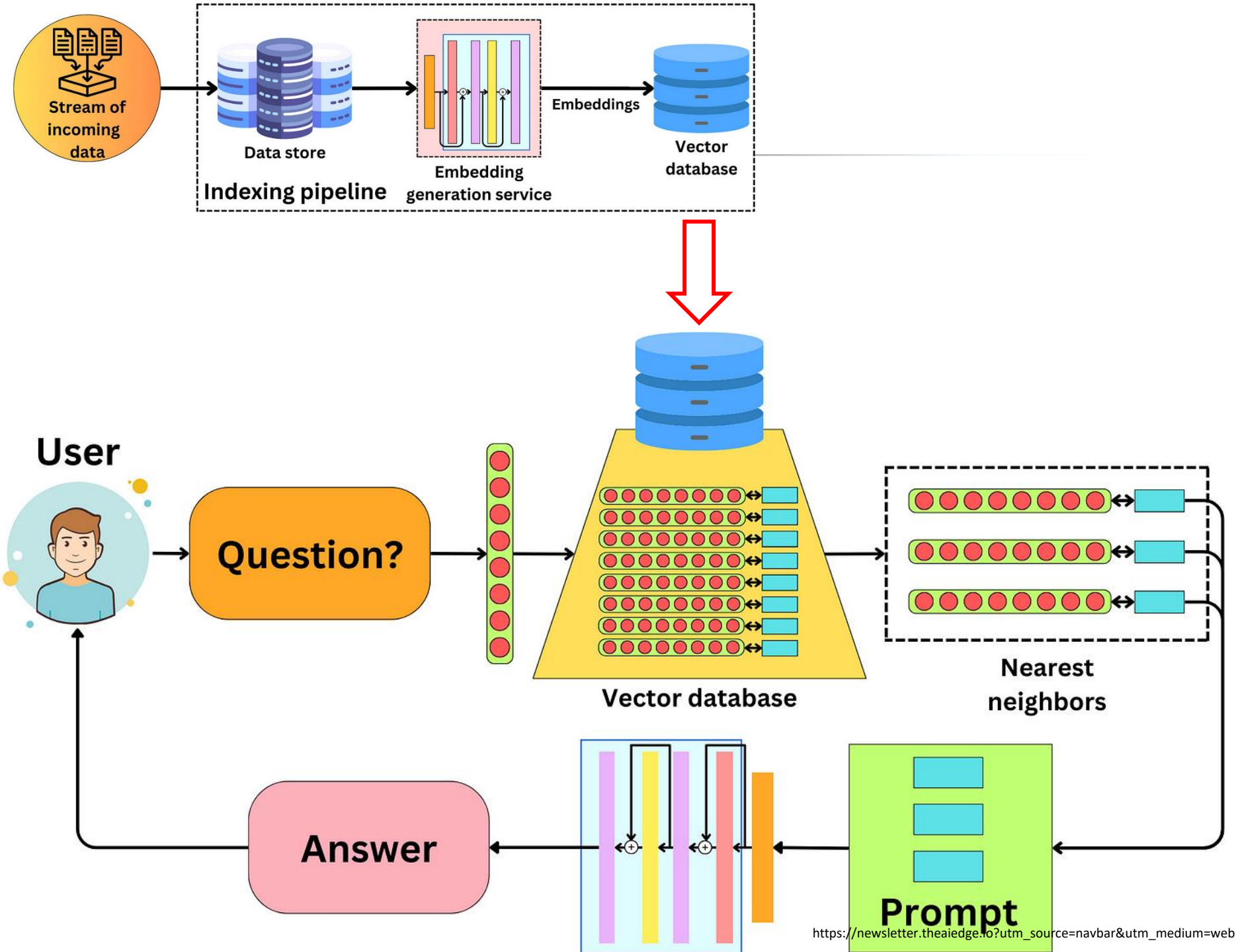
https://towardsdatascience.com/survival-of-the-fittest-compact-generative-ai-models-are-the-future-for-cost-effective-ai-at-scale-6bbdc138f618

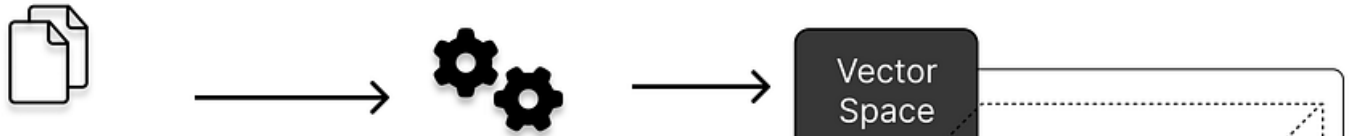


	similarity score
	1.1
	1.2
	3.6
	4.2
	5.1

```
[[
  1, "... Christiano Ronaldo is set to move to Saudi Arabia from the start of 2023 ...",
  2, "... next match for Al Nassr as 2023/24 Saudi Pro League season begins ...",
  3, "... the full list of Al Nassr's fixtures ..."
]]
```

https://towardsdatascience.com/all-you-need-to-know-about-vector-databases-and-how-to-use-them-to-augment-your-llm-apps-596f39adfed


Stream of incoming data
Data store
Indexing pipeline
Embedding generation service
Embeddings
Vector database
User
Question?
Vector database
Nearest neighbors
Answer
Prompt
https://newsletter.theaiedge.io?utm_source=navbar&utm_medium=web



elasticsearch

OpenSearch

milvus

Pinecone

Weaviate
Open Source

Text Search
Databases

Pure Vector
Databases

Vector Libraries

Vector
Databases

Vector-Capable NoSQL
Databases

redis

Annoy

FAISS

Vector-Capable SQL
Databases

mongoDB.

*Pgvector for
Postgres*

OpenAI

SBERT

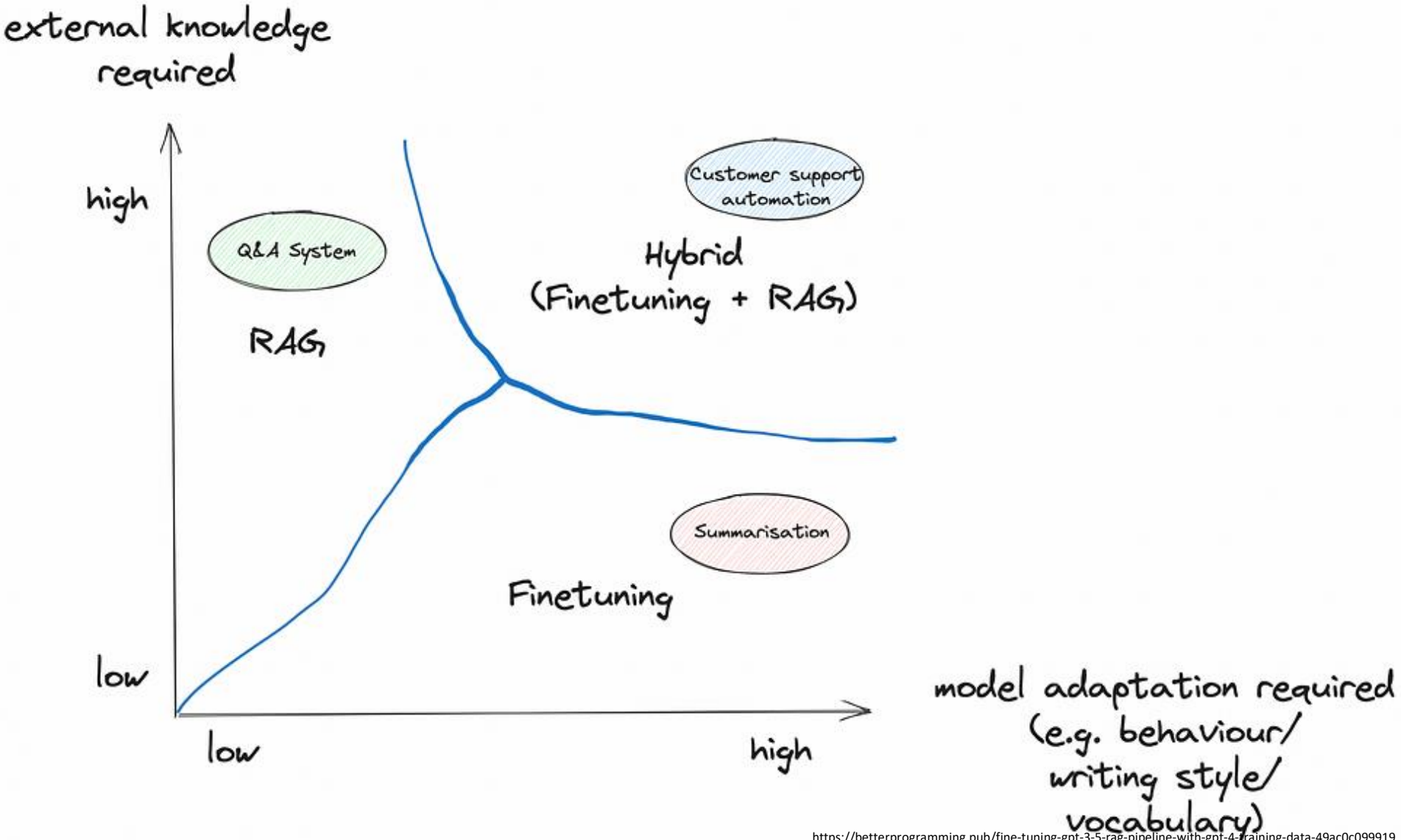
Facebook by Damien Benveniste

https://betterprogramming.pub/fine-tuning-gpt-3-5-rag-pipeline-with-gpt-4-training-data-49ac0c099919

https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7

https://towardsdatascience.com/all-you-need-to-know-about-vector-databases-and-how-to-use-them-to-augment-your-llm-apps-596f39adfedb

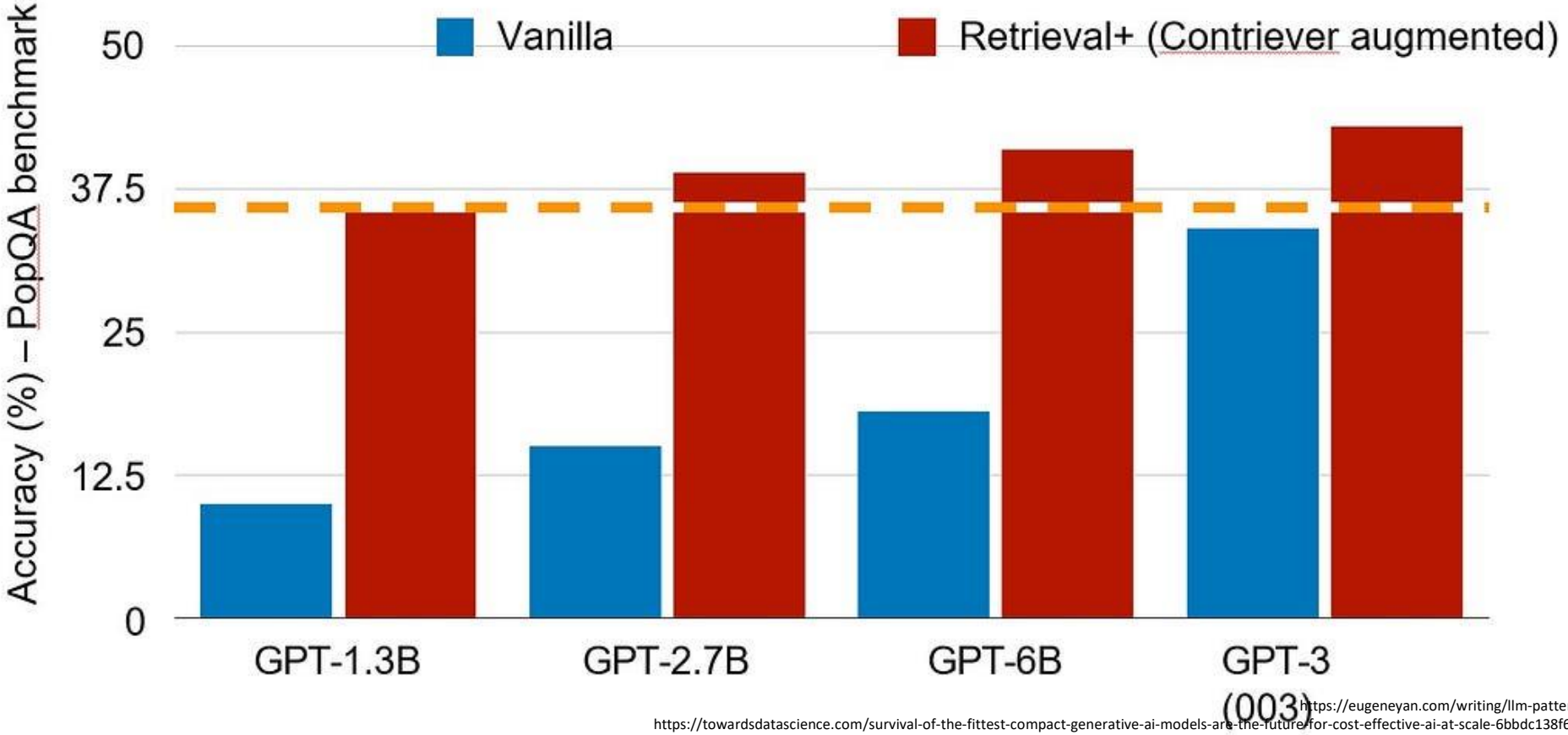
Finetuning vs RAG



https://betterprogramming.pub/fine-tuning-gpt-3-5-rag-pipeline-with-gpt-4-training-data-49ac0c099919
https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7

검색 증강 생성: RAG(Retrieval-Augmented Generation)

- **RAG 목적**

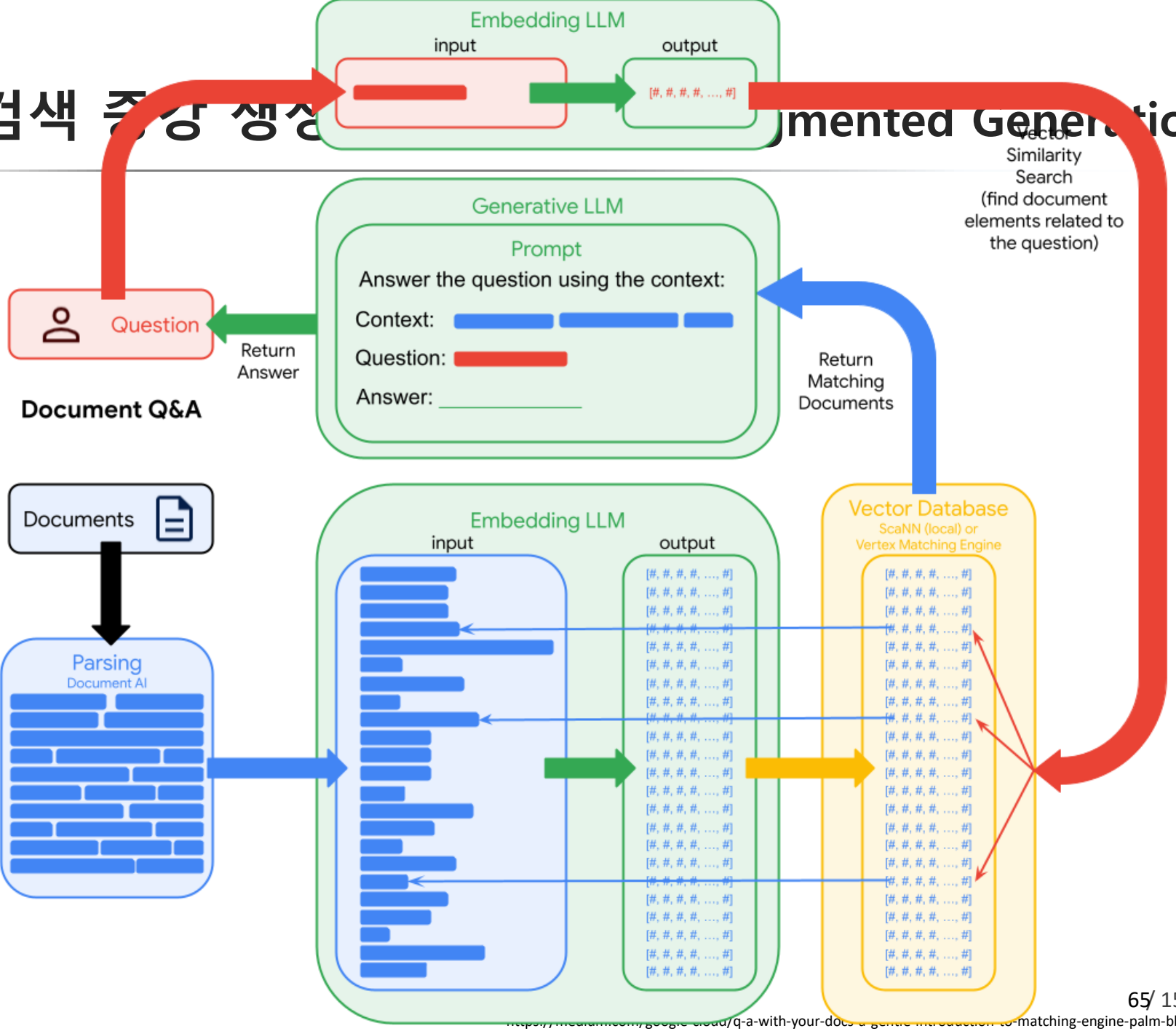


https://eugeneyan.com/writing/llm-patterns

https://towardsdatascience.com/survival-of-the-fittest-compact-generative-ai-models-are-the-future-for-cost-effective-ai-at-scale-6bbdc138f618

https://medium.com/google-cloud/q-a-with-your-docs-a-gentle-introduction-to-matching-engine-palm-bbbb6b0cff7b

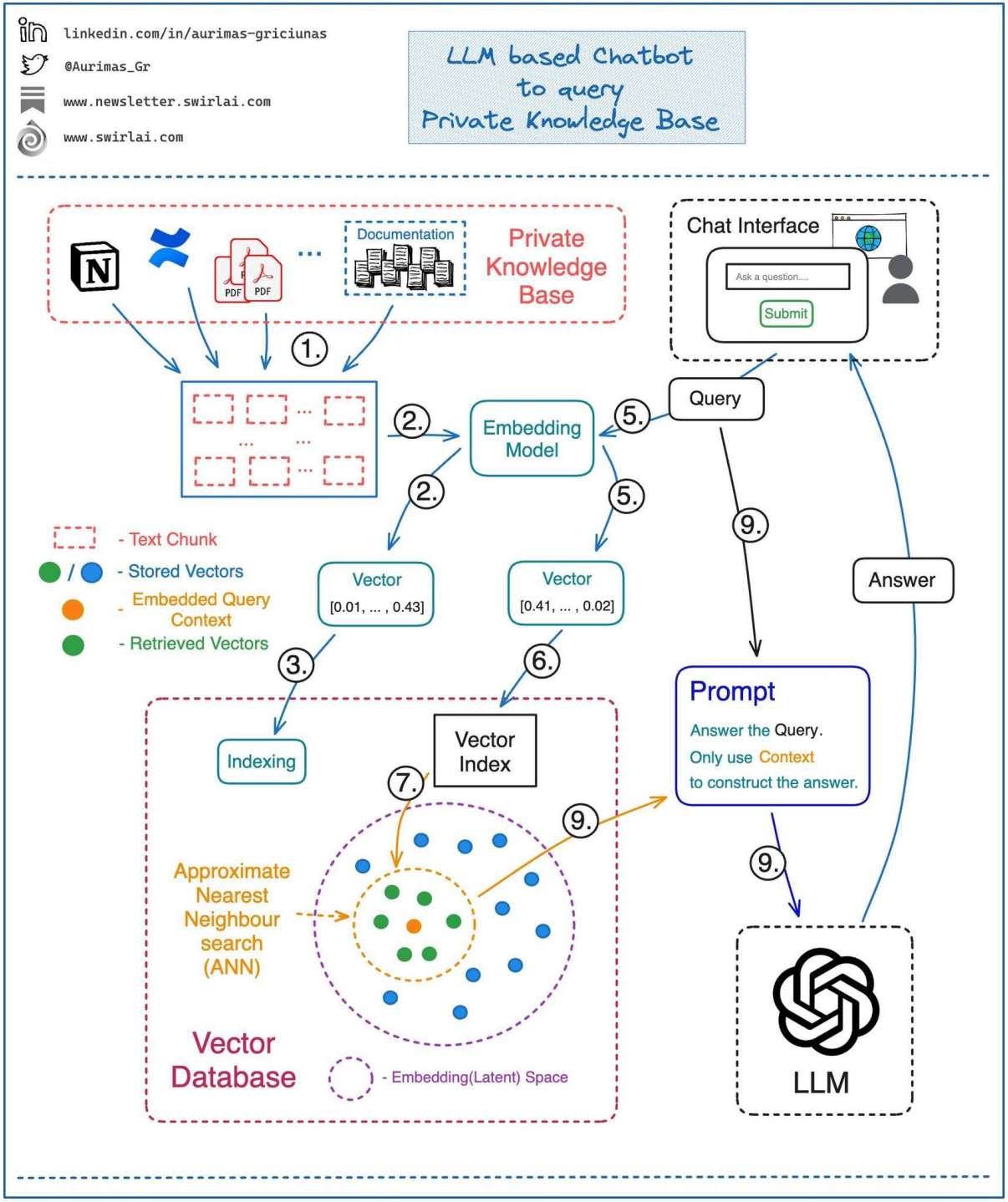
검색 증강 생성 (Retrieval Augmented Generation)



https://medium.com/google-cloud/q-a-with-your-docs-a-gentle-introduction-to-matching-engine-palm-bbbb6b0cff7b

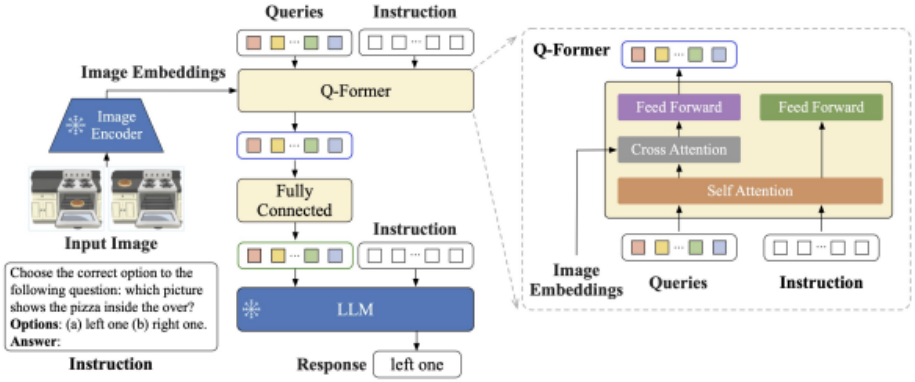
RAG S

- RAG D
  - DB
  - Int
  - Do

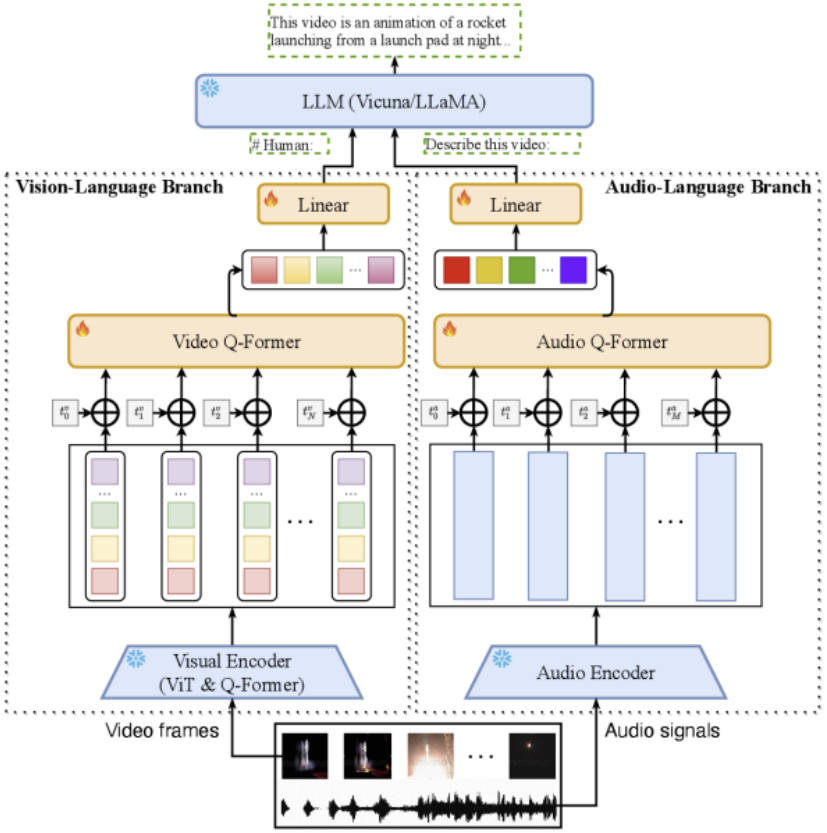


Multimodal LLM

- **An instruction-tuned audio-visual language model for video understanding**
  - ArXiv preprint arXiv:2306.02858.
- **Instructblip: Towards general-purpose visionlanguage models with instruction tuning**
  - ArXiv, abs/2305.06500.



Video-LLaMA

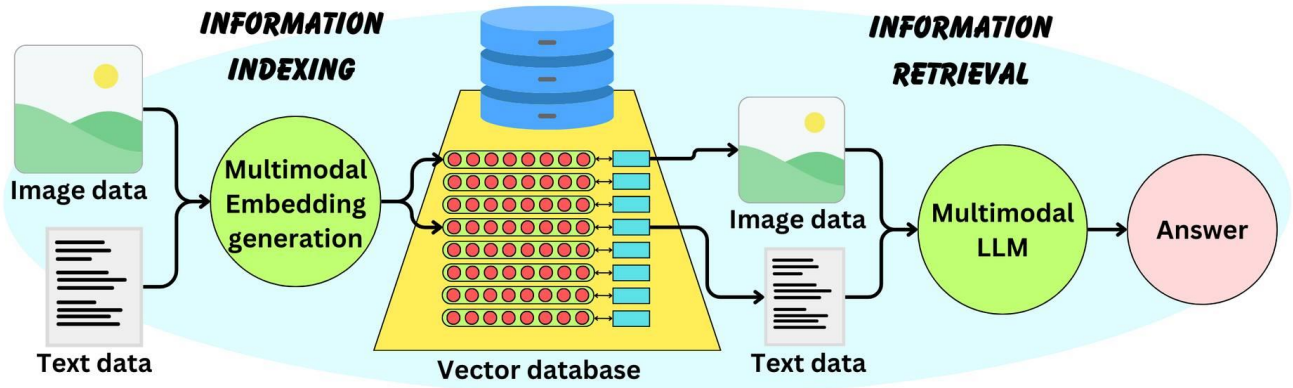


InstructBLIP

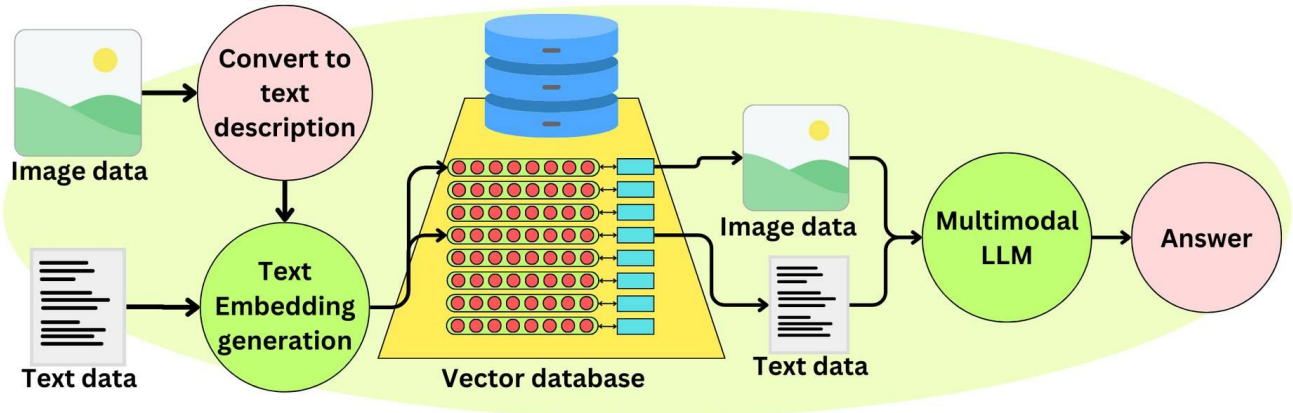
3 Ways to Build Multimodal RAG Pipelines

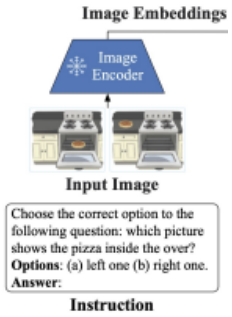
TheAiEdge.io

Multimodal Embedding generation + Multimodal Retrieval

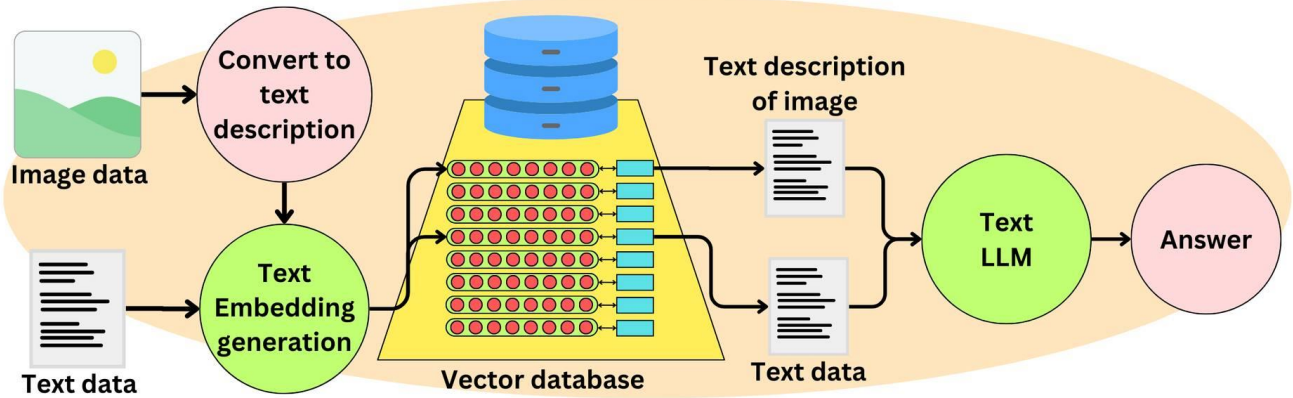


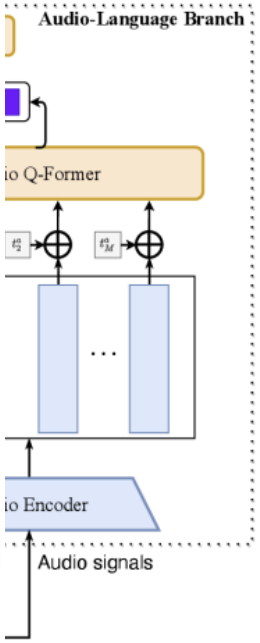
Text Embedding generation + Multimodal Retrieval



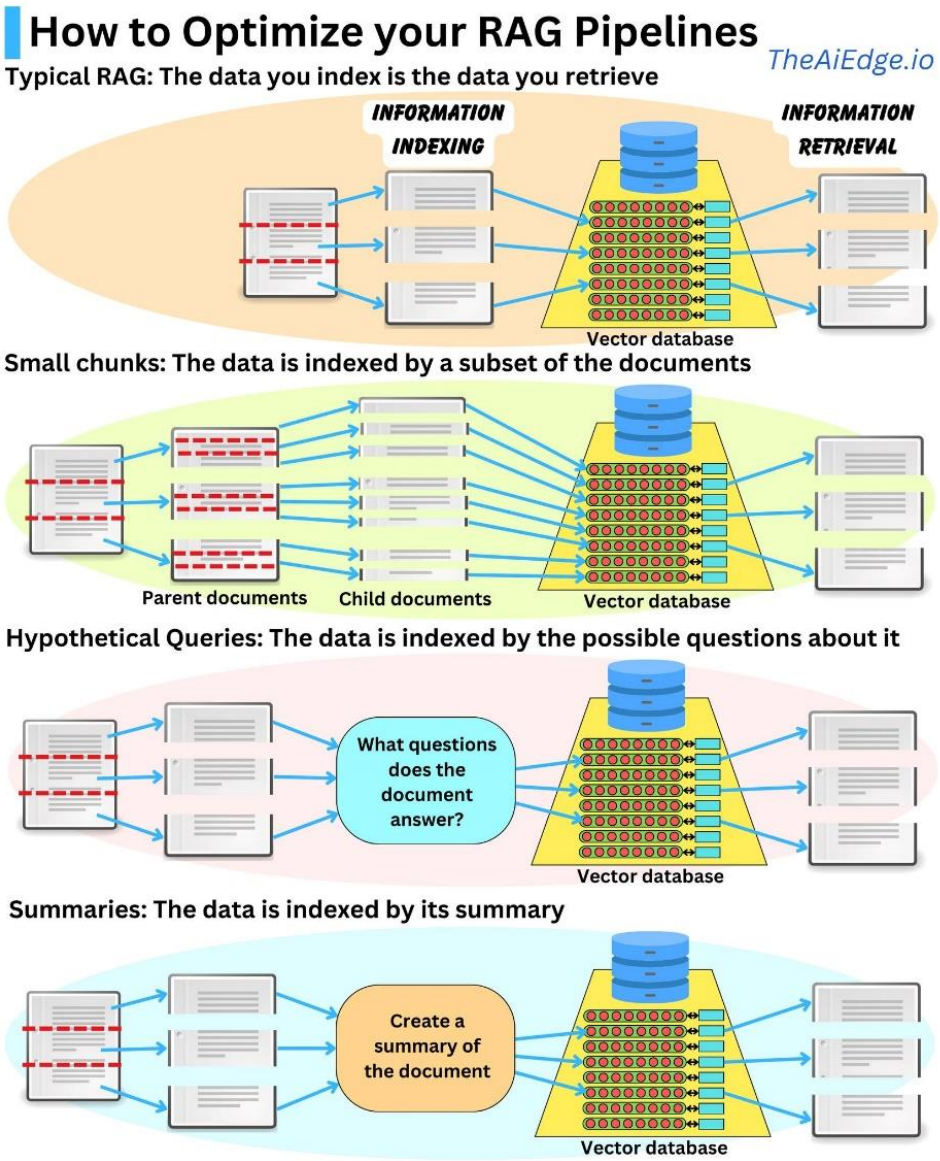


Text Embedding generation + Text Retrieval





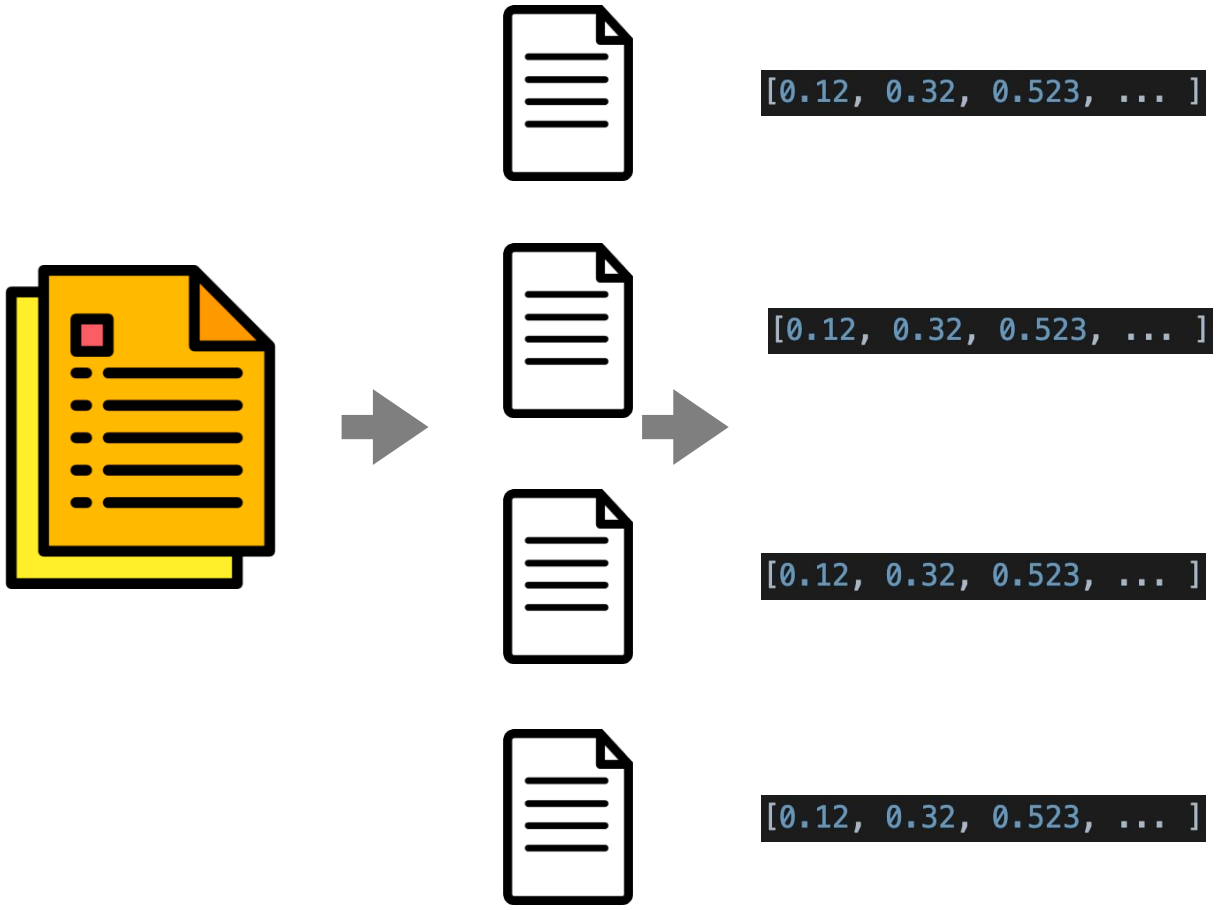
How you optimize RAG data indexing



- **Typical RAG**
  - 문서를 적당히 쪼개서 임베딩
- **Small chunks**
  - 문서를 아주 작게 쪼개서 임베딩
- **Hypothetical Queries**
  - 가능한 예상질문을 만들어서 임베딩
- **Summaries**
  - 요약문을 만들어서 임베딩

How you optimize RAG data indexing

- **Chunck: 문장 단위**

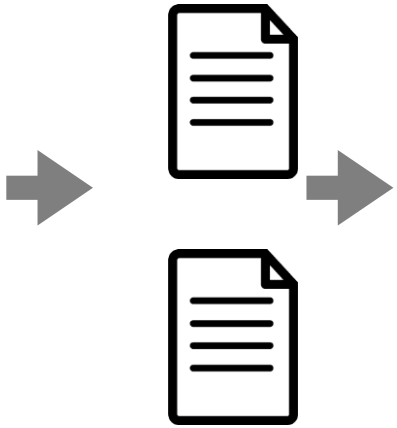


필요한 내용만 조회
임베딩 노이즈 감소
벡터 검색 정확도 상승

https://github.com/line-engineering/llm-meetup/blob/main/Seoul/1/2-rag-in-production.pdf

How you optimize RAG data indexing

- **Chunck: 페이지 단위**



임베딩 생성 비용 경감

청크 저장 비용 경감

문서 검색 비용 경감

Original

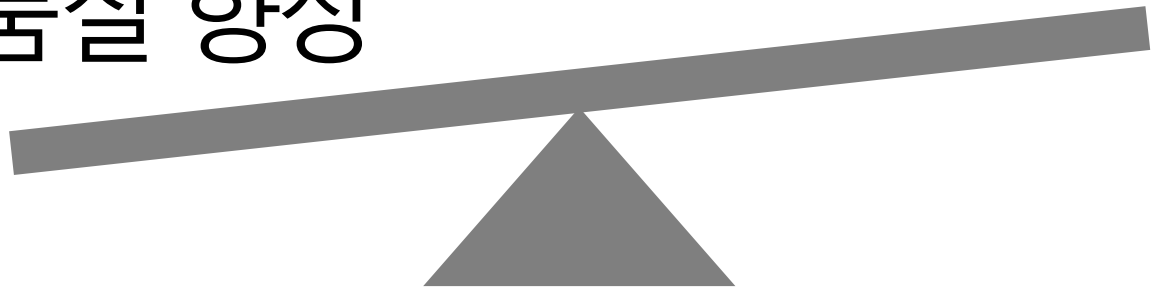
Chunks

Embedding Vectors

How you optimize RAG data indexing

답변 품질 향상

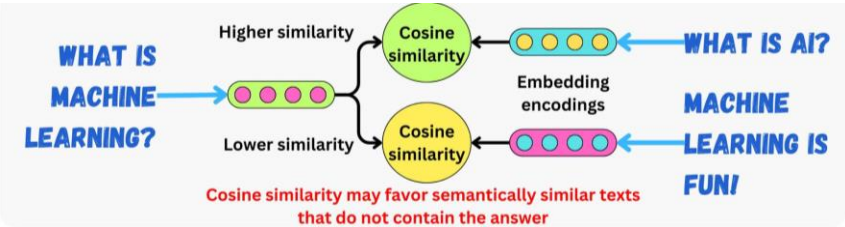
비용 증가



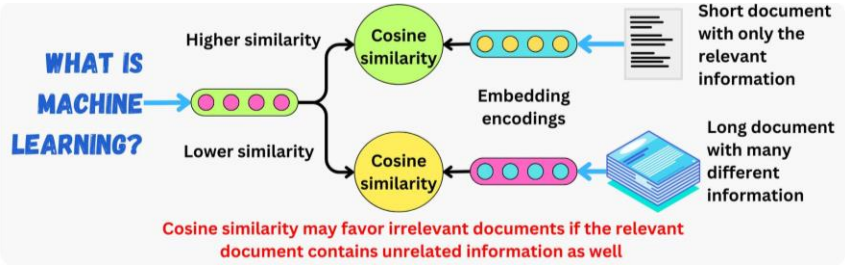
https://github.com/liner-engineering/llm-meetup/blob/main/Seoul/1/2-rag-in-production.pdf

Problems with RAG

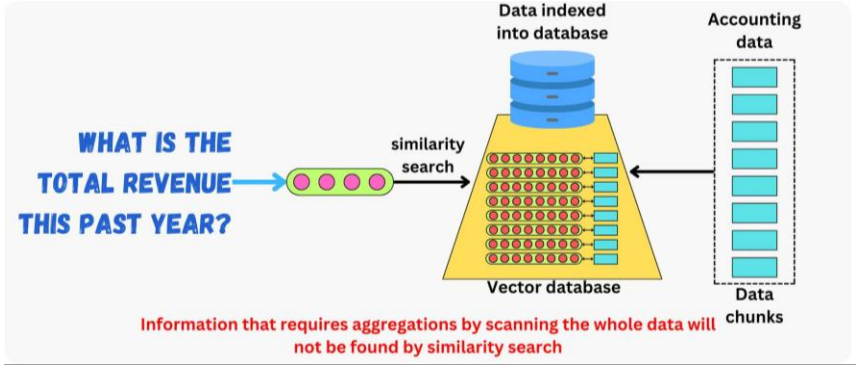
- 질문은 때론 답변과 의미적으로 유사하지 않음



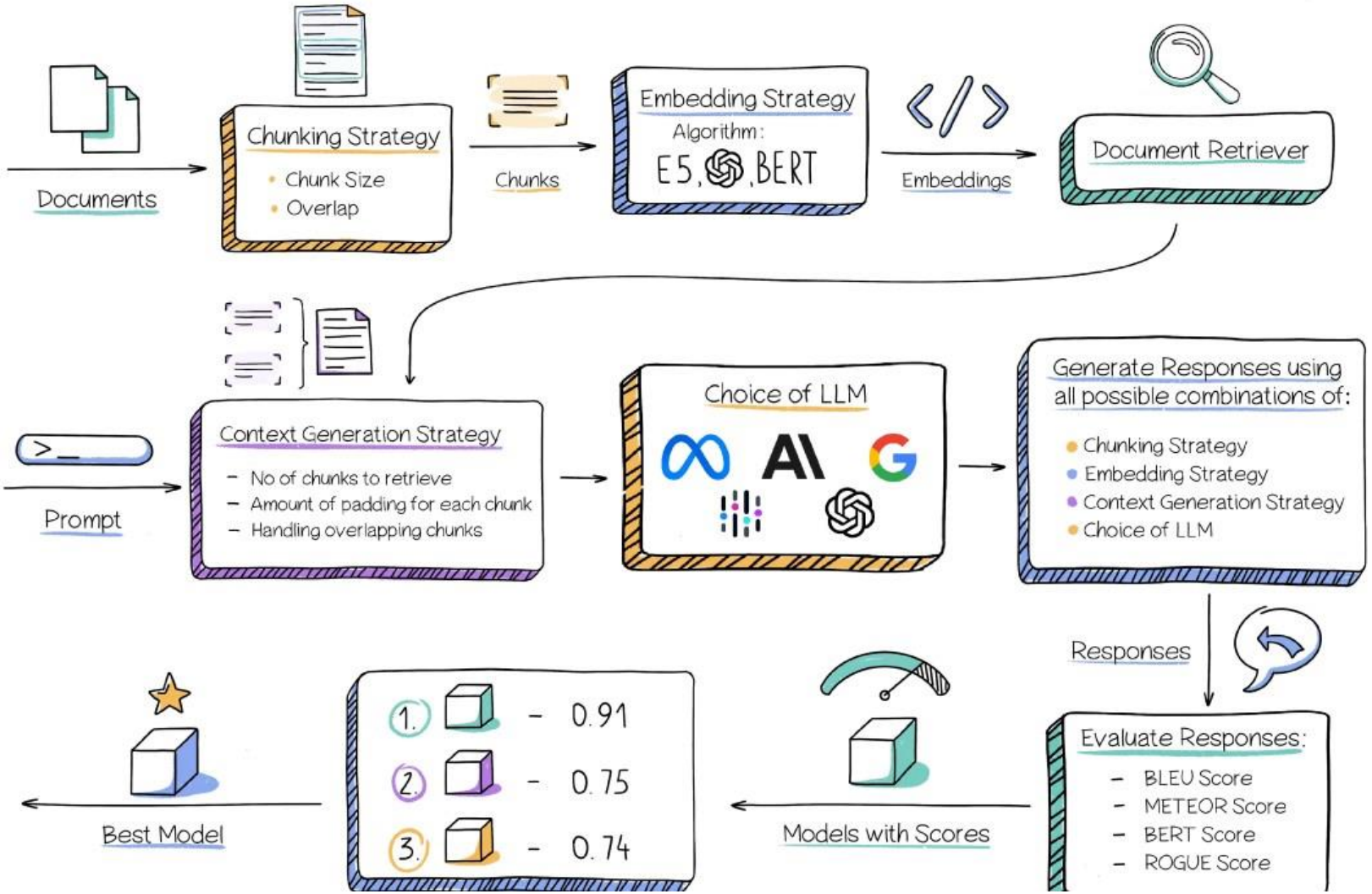
- 문서가 너무 긴 경우 의미적 유사성이 희석될 수 있음



- 정보는 하나가 아닌 여러 문서에서 조합해야 할 수 있음

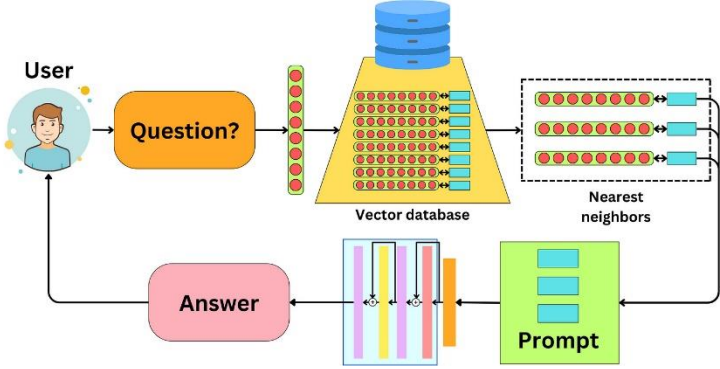


Pick the best LLM



https://blog.abacus.ai/blog/2023/08/10/create-your-custom-chatgpt-pick-the-best-llm-that-works-for-you/

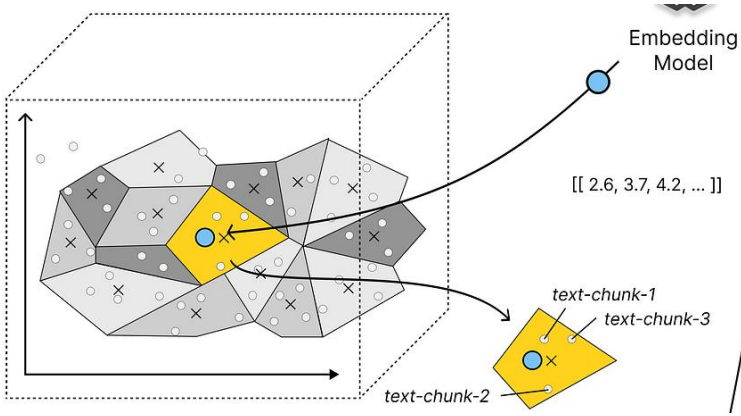
오프라인 DB기반 chatbot(on-premise)



- **0) 사전 질문&답변 리스트**
  - 자주 하는 질문에 대해 정해진 답변을 만들어 사전 질문 리스트 저장
  - 질문이 사전 질문 리스트에 있는지 먼저 검사
  - 없는 경우 벡터 검색 진행

- **1) Chunk 단위**
  - 사용자 DB에 따라 chunk 단위 고려 필수
  - 단순 글자 길이/overlap으로 정하는 건 비효율
  - 법령/규정이라면 법조항/규정 단위로, 보고서/논문이라면 문단 단위로
  - 문장 단위로 모든 내용을 품고 있고 분절이 잘되어 있다면 문장 단위로 등
  - 쓸데없이 문단을 넘나드는 것보다, 명확한 한두 문단에서 결과를 찾는게 효율적

- **2) VectorDB 선택: faiss**
  - pinecone, chroma, faiss 검토
  - offline/속도/강건성/GPU지원 면에서 faiss



오프라인 DB기반 chatbot(on-premise) 구현 시 고려 사항

- **3) embedder**
  - openai embedder: 비용 발생&online이라 패스
  - sbert: 속도/비용/offline 가능하면서 성능 좋음
  - 단점: 입력 token limit이 512token으로 chunk 길이 조정 필요. 대게 500token의 chunk면 충분히 원하는 정보가 담겨있어 무리 없음.
    - https://huggingface.co/snunlp/KR-SBERT-V40K-klueNLI-augSTS
  - 도메인 텍스트의 경우 sbert 성능 불만족 시 paraphrasing으로 NLI 추가학습 가능
- **4) 문장 유사도 기반 sorting**
  - 벡터 기반: FAISS에 임베딩 벡터를 구워 상위 N개 선택
  - 단점: exact matching 단어가 검색되지 않는 경우도 있음
  - N은 답변 생성 속도를 고려하여 2개만
  - 빈도수 기반: BM25 기반 유사도 비교. 한국어 특징인 조사 분리를 위해 mecab 사용
  - 장점: exact matching에 탁월
- **5) 최종 결과 도출 프롬프트**
  - system setting
    - 모를 땐 '모른다'라고 답변 지시
    - 한국어로 답변 지시
    - 답변 길이를 100자로 셋팅
    - reference 위치 표시

오프라인 DB기반 chatbot(on-premise) 구현 시 고려 사항

- **6) 답변 생성 모델 선택**
  - ChatGPT는 온라인이라 배제
  - 공개된 한국어 LLM 성능 비교
  - 요구 GPU 메모리 비교(납품 단가를 고려해 400만원 수준 3090/4090 24GB 기준)
  - 답변 100자 기준 응답속도 비교
    - 7B 모델은 GPU 15GB(dolly 기준)로 허용 가능. 성능은?
    - 12B 모델은 23.3GB(dolly 기준) 메모리 사용으로 3090(24GB)으로 소화하기 아슬아슬함
    - 7B 만으로 원하는 답변 성능이 나오는가?
    - 12B의 성능은? 24GB에서 돌릴 수 있는가? 응답시간은?
- **7) UI**
  - gradio: colab에서 가능하므로 강의 실습용. 안이쁘다.
  - streamlit: colab에선 불가. gradio보다 이쁘고 최적화 가능.
  - streamlit으로 제공한 후 업체의 선호에 따라 디자인 외주
- **8) Hallucination 체커 구축(환각효과를 줄이기 위한 방법)**
  - 답변이 환각인지 재질문
  - topN 문서와 생성 답변의 세세한 정보 비교 및 교정
  - 법조항, 금액, 연도, 고유명사 체커 탑재. DB에 따라 잘 못하는 부분 식별
  - 더 줄이고 싶다면 instruct-finetuning 수행(A100 4대 이상 필요)
  - role이 명확한 경우 instruct와 답변을 명확히 구축하여 추가학습. 이 때 법조항/금액/연도 등은 메타텍스트로 감싸서 보존하도록 가이드
    - Copy mechanism, sentinel tag을 적용하면 월등히 줄일 수 있음

오프라인 DB기반 chatbot(on-premise) 구현 시 고려 사항

- **9) 시스템 비용**
  - 작은 서버 1.5천만 원
  - 여러 요청을 해결하기 위해 GPU N개 스케쥴링, 400만원xN

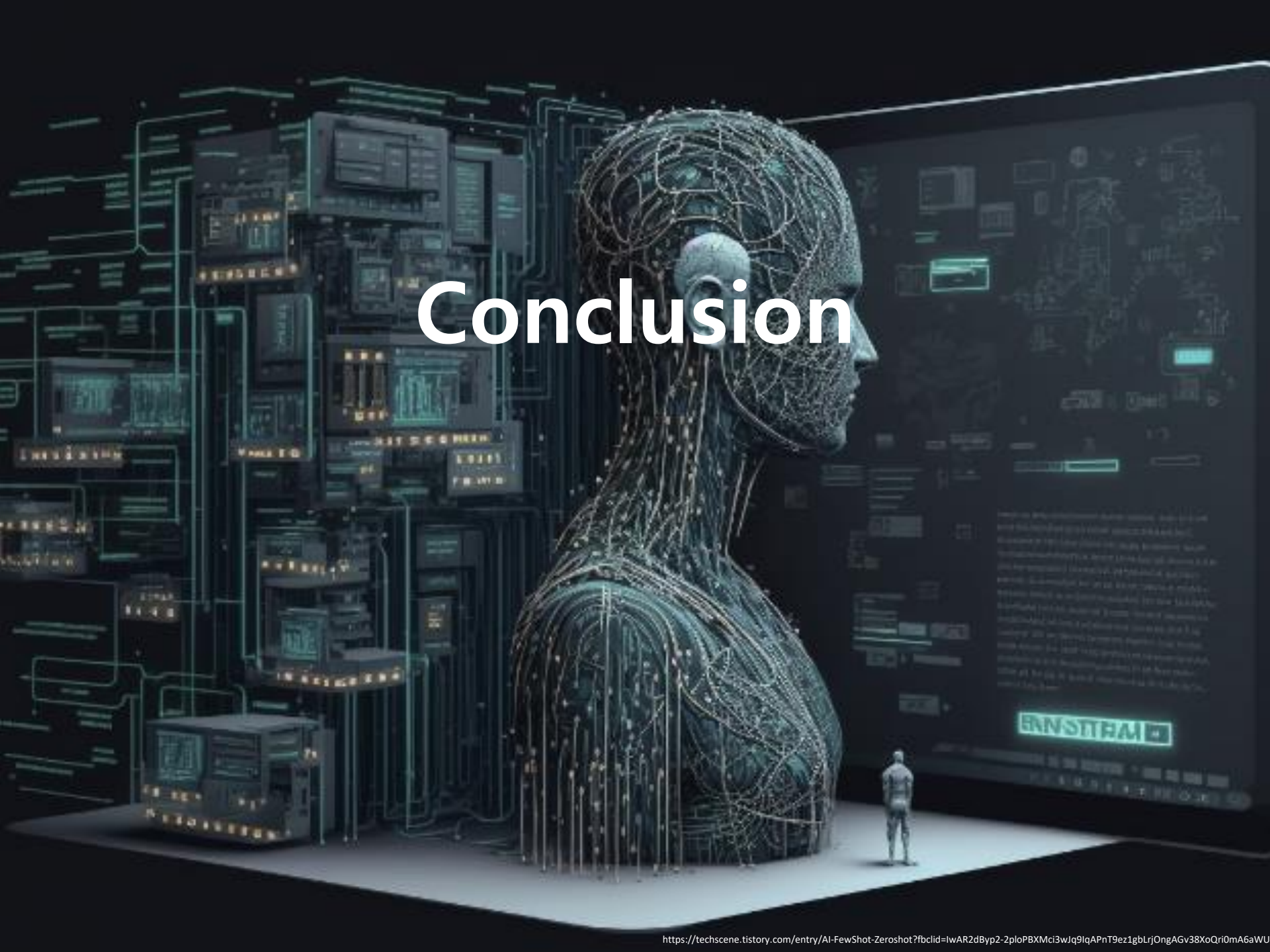
- **마무리**
  - 인공지능은 단순히 블럭을 결합해서 뚝딱 만든다고 원하는 성능이 나오지 않음
  - langchain 몇 줄 띡띡 연결한 시스템은 hallucination이 마구 터지는데 이를 만족한다면 저비용으로 쉽게 구축 가능
  - 그런데 신뢰도/정확성을 위해 추가 개선하려면 sLLM에 결국 데이터를 구축해서 instruct 학습을 해야 한다. 예산/기간을 고려해서 신중히 결정

코드실습

- URL: https://github.com/airobotlab/KoChatGPT

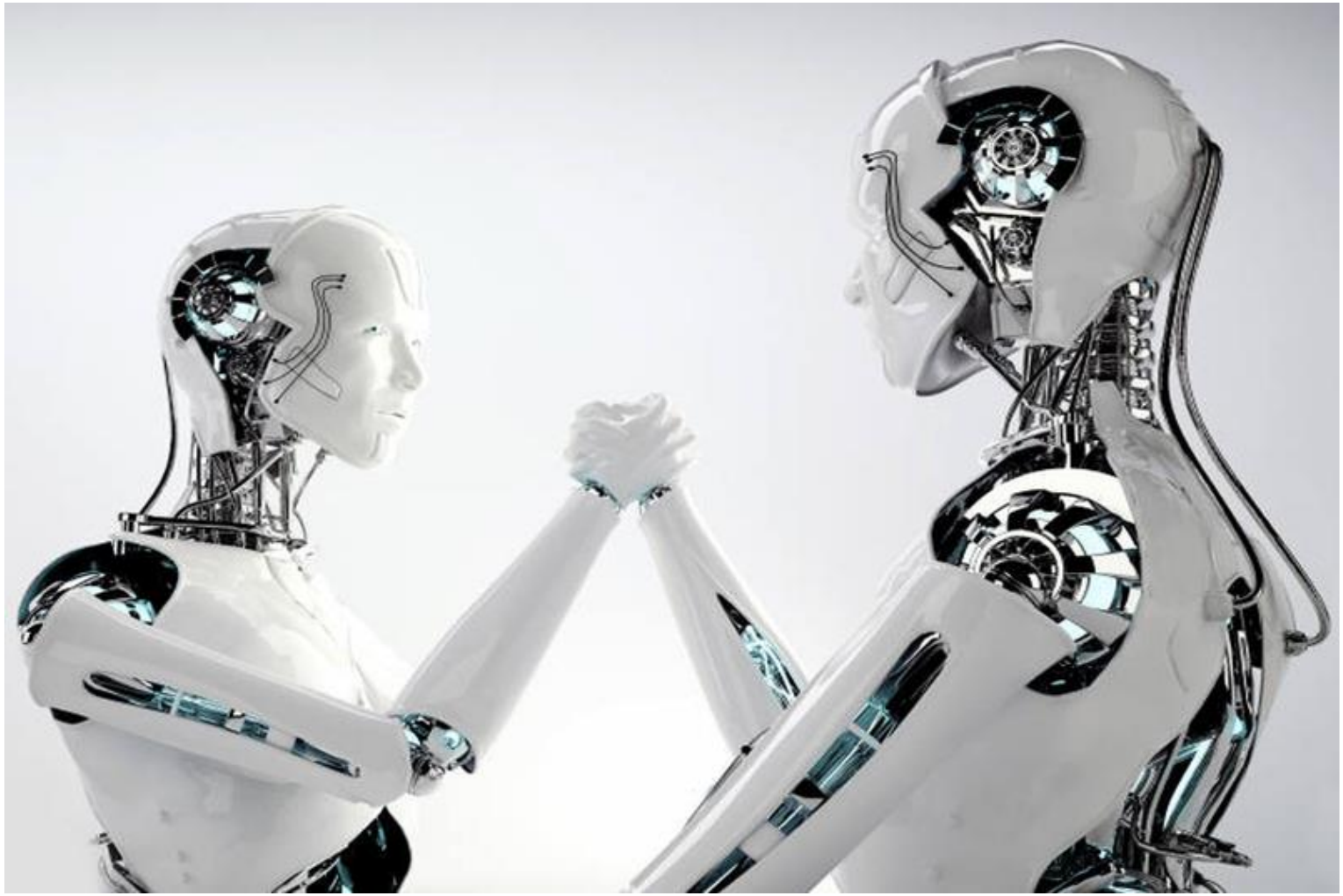
- 1_kochatgpt_code_231122.ipynb: https://colab.research.google.com/drive/1_aws1VolXkvd4xIrFExTdc3qd1hm7nNv?usp=sharing
- 2_GPT_3_5_Turbo_Fine_tuning_231122: https://colab.research.google.com/drive/1qcznV7x6awwn2kDThQ0HM6uPUuD7BeaN?usp=sharing
- 3_AutoTrain_LLM_colab_231122.ipynb: https://colab.research.google.com/drive/1QXqPVg2P7Im9A7vJsBQpzBqsRsYbzXPD?usp=sharing
- 4_instruct_fine_tuning_polyglot_ko_12B_colab_231122.ipynb: https://colab.research.google.com/drive/1SH81IOt4NYOaKVrL5Sa2tt_HeQqduU71?usp=sharing
- 5_DPO_on_llama2_실습_231122.ipynb: https://colab.research.google.com/drive/1ccQy4wrbzCWQYBvNU1Zvv9HoXM6yiUKt?usp=sharing
- 6_RAG_langchain_231122: https://colab.research.google.com/drive/1HkhOK7gec-M_QqREj_L_cC36Do3g0QpA?usp=sharing
- 7_LLM_RAG_Evaluation_231122: https://colab.research.google.com/drive/1mNwADnoDf5giPb4rh2abJFwIDAlCYlGT?usp=sharing

Conclusion



https://techscene.tistory.com/entry/AI-FewShot-Zeroshot?fbclid=IwAR2dByp2-2ploPBXMci3wJq9IqAPnT9ez1gbLrjOngAGv38XoQri0mA6aWU

What should we do?



AI will not replace you.
A person using AI will.

AI will not replace you.
A person using AI will.

Q&A

#midjourney #v4, 김성완