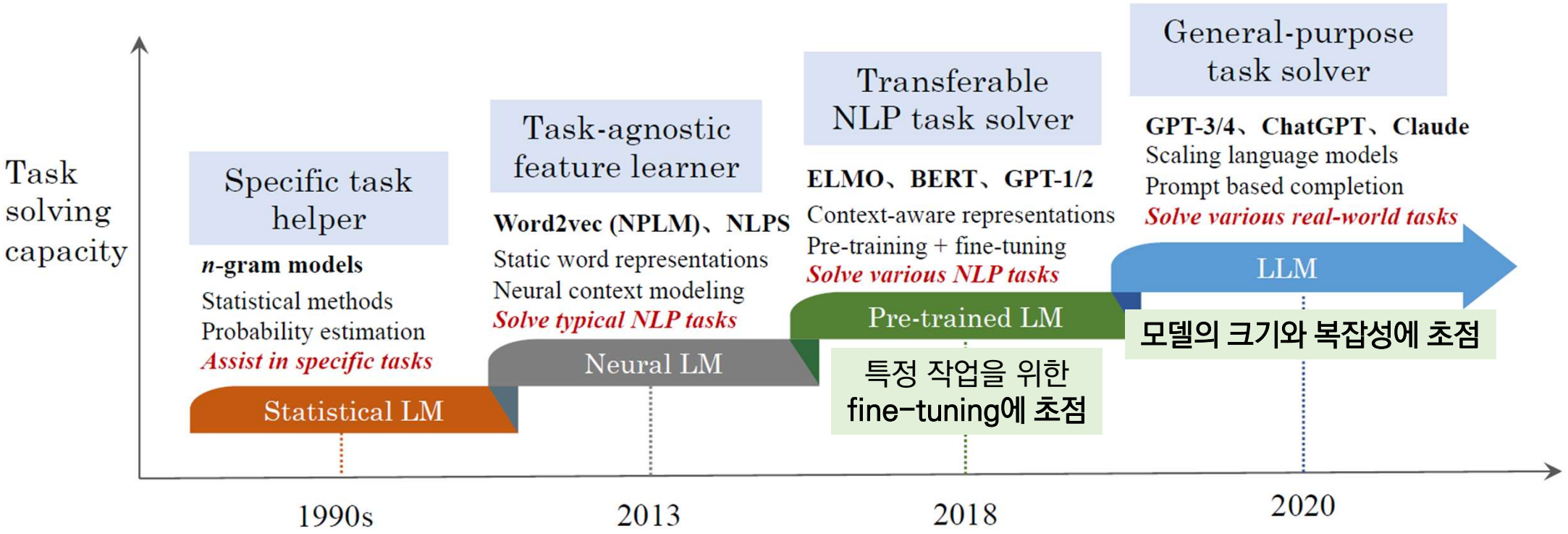
LLM 개념 및
최신 기술 동향 소개

2020305039 유하영

2024.05.30

Language Model

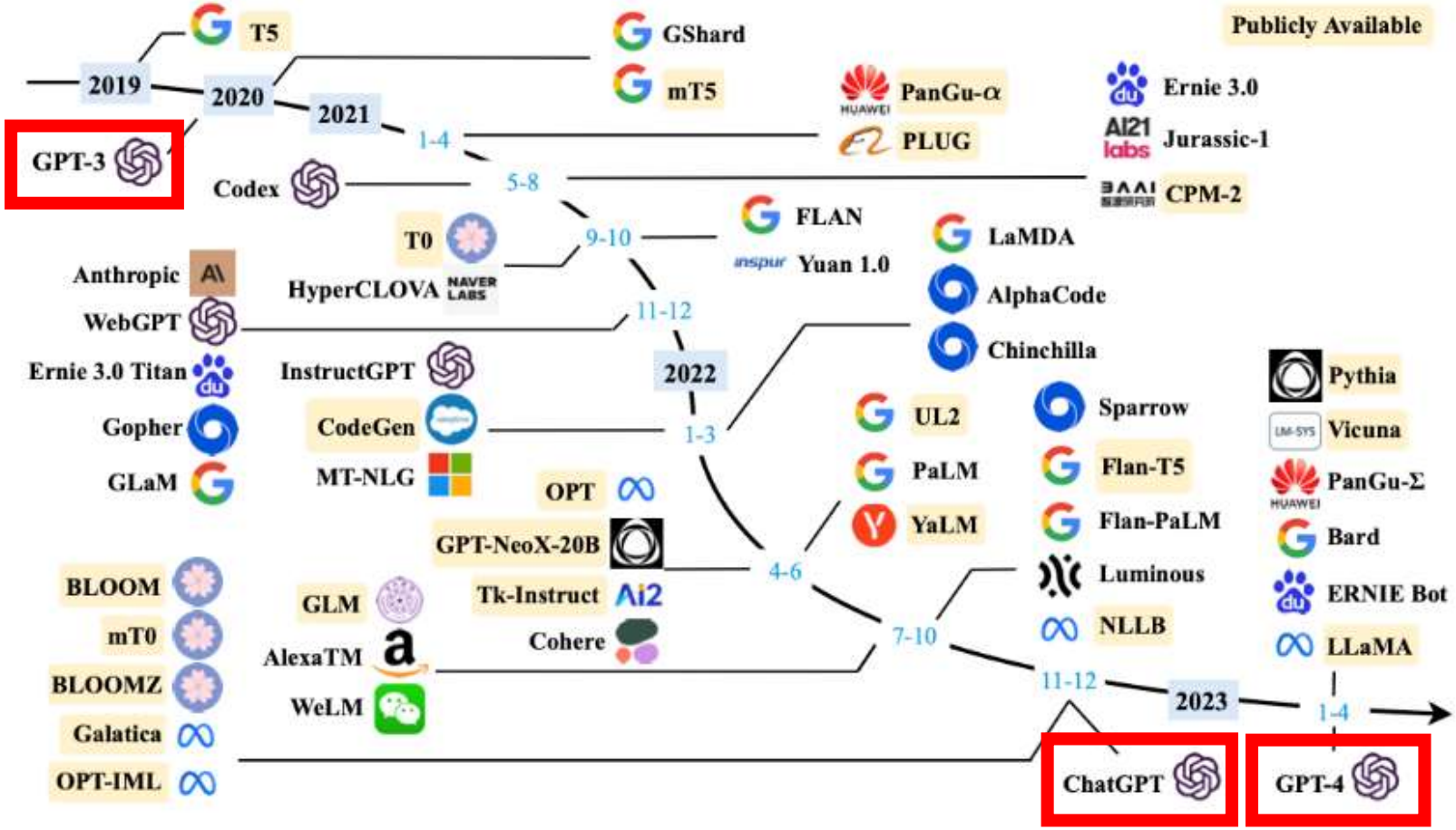
: 주어진 텍스트 데이터로부터 언어의 패턴을 학습하여,
문맥에 맞는 단어나 문장을 예측하거나 생성하는 인공지능 시스템



〈4세대 언어 모델(LM)의 진화 과정〉

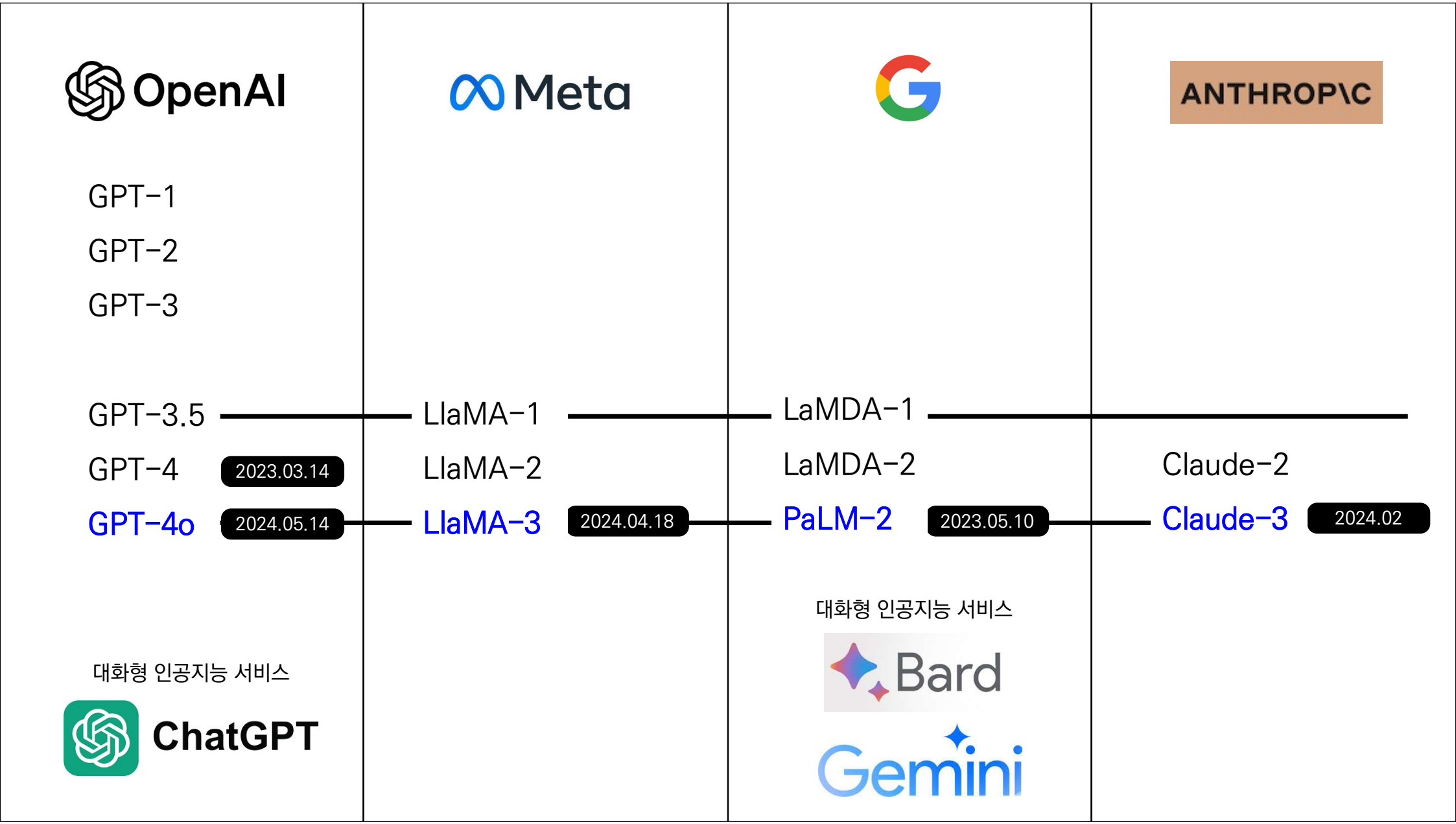
https://wikidocs.net/237619

주요 LLM Timeline



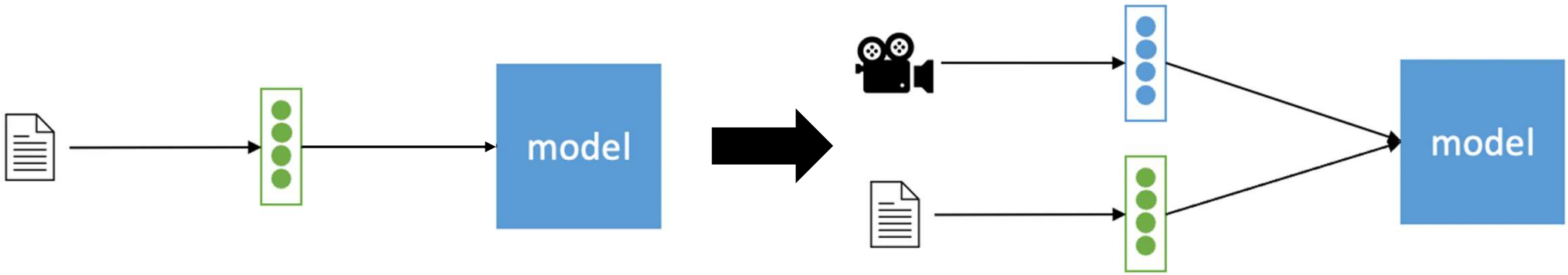
〈최근 몇 년 동안의 LLM 타임라인〉

https://brunch.co.kr/@brunchgpjz/49


OpenAI
Meta
ANTHROPIC
GPT-1
GPT-2
GPT-3
GPT-3.5
GPT-4
2023.03.14
GPT-4o
2024.05.14
LlaMA-1
LlaMA-2
LlaMA-3
2024.04.18
LaMDA-1
LaMDA-2
PaLM-2
2023.05.10
Claude-2
Claude-3
2024.02
대화형 인공지능 서비스
ChatGPT
대화형 인공지능 서비스
Bard
Gemini

Multi Modal AI

: 시각, 청각을 비롯한 여러 인터페이스를 통해
다양한 채널의 모달리티(양상)를 동시에 받아들여 학습하고 사고하는 AI
(인간이 사물을 받아들이는 방식과 동일하게 학습하는 AI)



GPT-4o

2024.05.14

omni(모든)

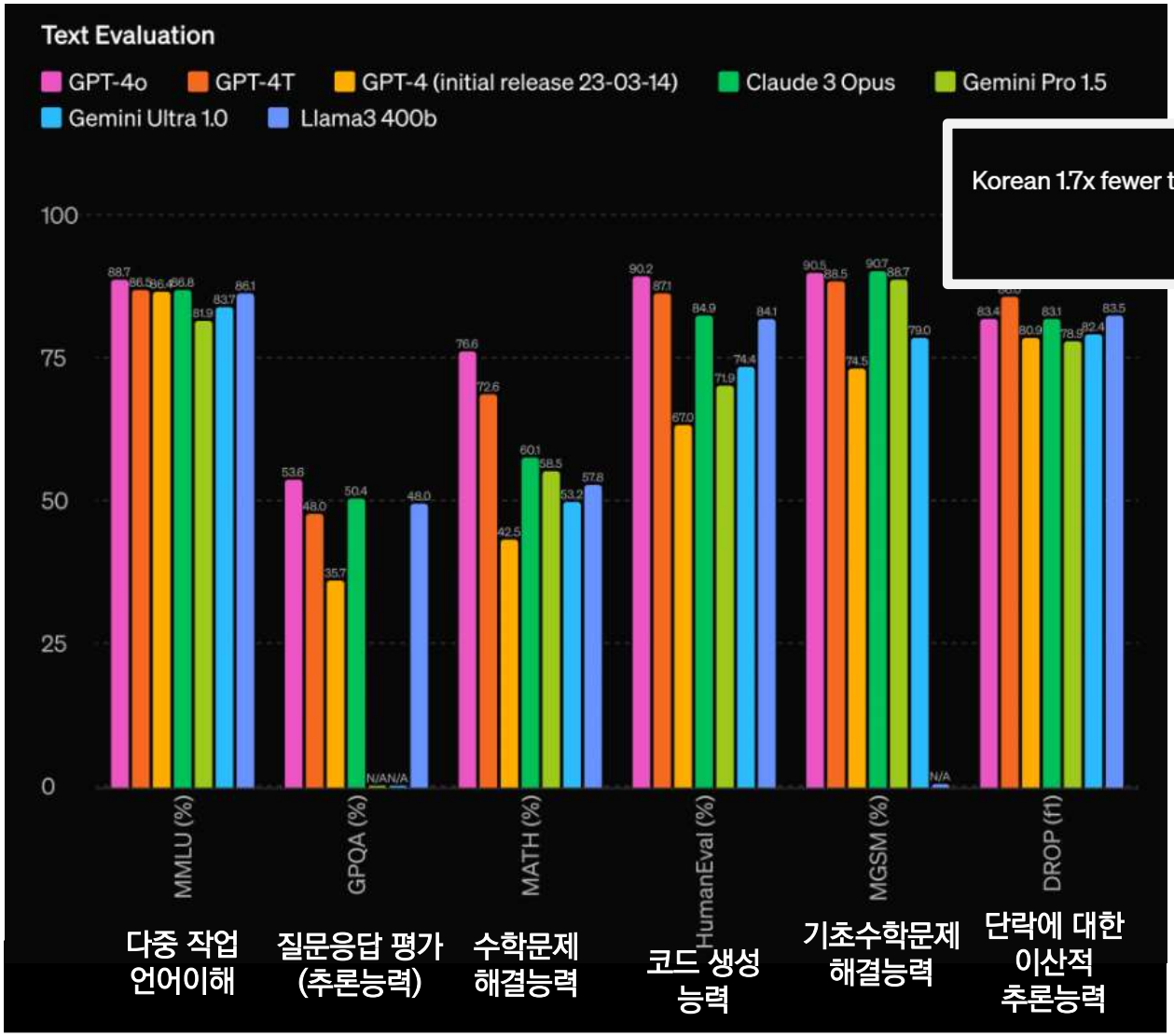
텍스트 · 오디오 · 이미지 · 비디오 등
다양한 형태의 데이터를 이해할 수 있는
HCI를 강화한 모델

이미지 인식기능

토큰 효율성 개선

처리속도 향상

https://openai.com/index/hello-gpt-4o/



Korean 1.7x fewer tokens (from 45 to 27)

안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!

* 한국어 토큰 처리 방식 변경
- 사용비용 감소
- 속도 개선

https://openai.com/index/hello-gpt-4o/

OpenAI's **GPT** model

GPT-3.5

GPT-4 2021.09까지의 데이터 기반
멀티모달 기능을 갖춘 고성능 LM

GPT-4 Turbo 2023.04까지의 데이터 기반

GPT-4o GPT-4의 최적화된 버전

- 언어나 단어가 아닌 것(호흡소리) 이해
- 4o가 기존 모델에 비해 코드 생성의 정확도가 높아짐
- 실시간 번역(리얼타임 보이스로)
- 리얼타임 비전으로 수학문제를 풀 수 있음
- 노래 부르기 가능(노래 느낌의 보이스)
- GPT-4o : 풍부한 답변, 정확도는 떨어짐..

GPT-4o

step 1. 직접 질의
반반차가 무슨 뜻을 가진 단어인지 아니?

step 2. 간접적인 질의
팀장님이 반반차 쓰고 은행 다녀오래.
위 문장을 보고 반반차의 뜻을 추론할 수 있겠니?

step 3. 연관단어를 포함한 질의
아… 반차쓰긴 아까우니 반반차 쓰고 갔다와야겠다 ㅋㅋ
위 문장을 보고 반반차의 뜻을 추론할 수 있겠니?

GPT-3.5	전혀 추론 X
GPT-4	step3에서 간신히 추론
GPT-4o	step1에서 바로 알 때도 있고, step3에서 추론할 경우도 있음

LLM의 경량화 트렌드

LLM

Large Language Model

경량화

SLLM

Small Large Language Model

적은 파라미터와 훈련 데이터를 가지고도
최대한의 성능을 낼 수 있는 언어모델

필요에 따라 fine-tuning을 통해 정확도를 높이는
맞춤형 LLM

- 고품질의 데이터 사용
- 효율적인 모델 아키텍처
- 메모리 사용량과 계산 복잡도를 줄이기 위한
  양자화 및 저순위 근사화 기법 사용
- 학습 과정에서의 여러 최적화 기법 사용

https://blog-ko.superb-ai.com/what-is-the-alternative-to-llm-in-chatgpt/

LlaMA

Large Language Model Meta AI

GPT-3 : 175B parameters

LlaMA	Model hyper parameters					
Number of parameters	dimension	n heads	n layers	Learn rate	Batch size	n tokens
7B	4096	32	32	3.0E-04	4M	1T
13B	5120	40	40	3.0E-04	4M	1T
33B	6656	52	60	1.5.E-04	4M	1.4T
65B	8192	64	80	1.5.E-04	4M	1.4T

model

당시, GPT-3.5와 비교했을 때
훨씬 적은 파라미터로 동일한 성능을 나타냄

https://blog-ko.superb-ai.com/what-is-the-alternative-to-llm-in-chatgpt/

LlaMA 2 is 30X cheaper

Llama2가 요약에 있어
GPT-4만큼 정확하며 30배 더 저렴
상업적으로도 무료 !

Model	Input Words	Input Tokens Total	Summary ratio	Output Tokens Total	Price/M input ($)	Price/M output ($)	Cost to summarize 100K words
gpt-4	96522	125902	0.2	25180	30	60	$5.48
gpt-3.5-turbo	96522	125902	0.2	25180	1.5	2	$0.25
Llama-2-7b	96522	149238	0.2	29848	0.25	0.25	$0.05
Llama-2-13b	96522	149238	0.2	29848	0.5	0.5	$0.09
Llama-2-70b	96522	149238	0.2	29848	1	1	$0.19

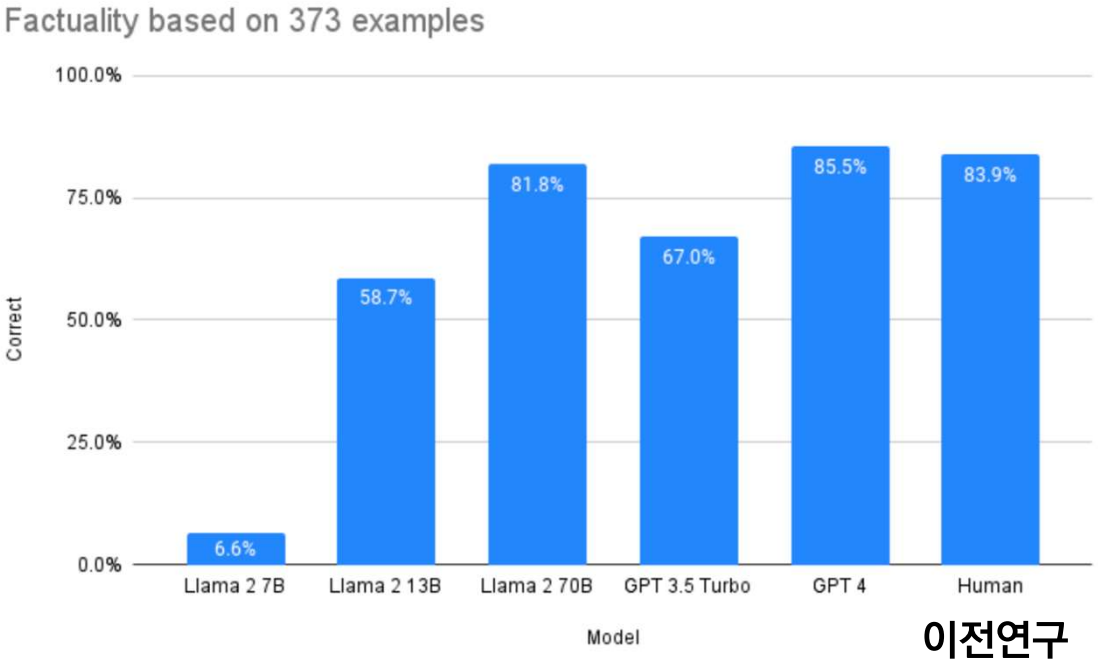
Alpaca

LlaMA 7B를 Instruction tuning한 모델

- self-Instruct 방식으로 학습데이터 생성
- 모델은 GPT-3, GPT-3.5 사용

▶ Instruction tuning ?
특정 명령(instruction)에 맞춰 적절한 답변을 생성할 수 있도록 훈련하는 것

▶ self-Instruct ?
LLM로 데이터를 생성해서
그 데이터로 다시 LLM을 학습하는 것



이전연구

https://discuss.pytorch.kr/t/gn-llama2-gpt-4-30/2376

PaLM2, Gemma

Transformer

LaMDA
Language Model for
Dialogue Applications

대화에 중점을 두는(언어이해+생성) LM / 약 1370억개 파라미터

PaLM 2
Pathways
Language Model 2

약 5400억개 파라미터
(최고점수 기록)

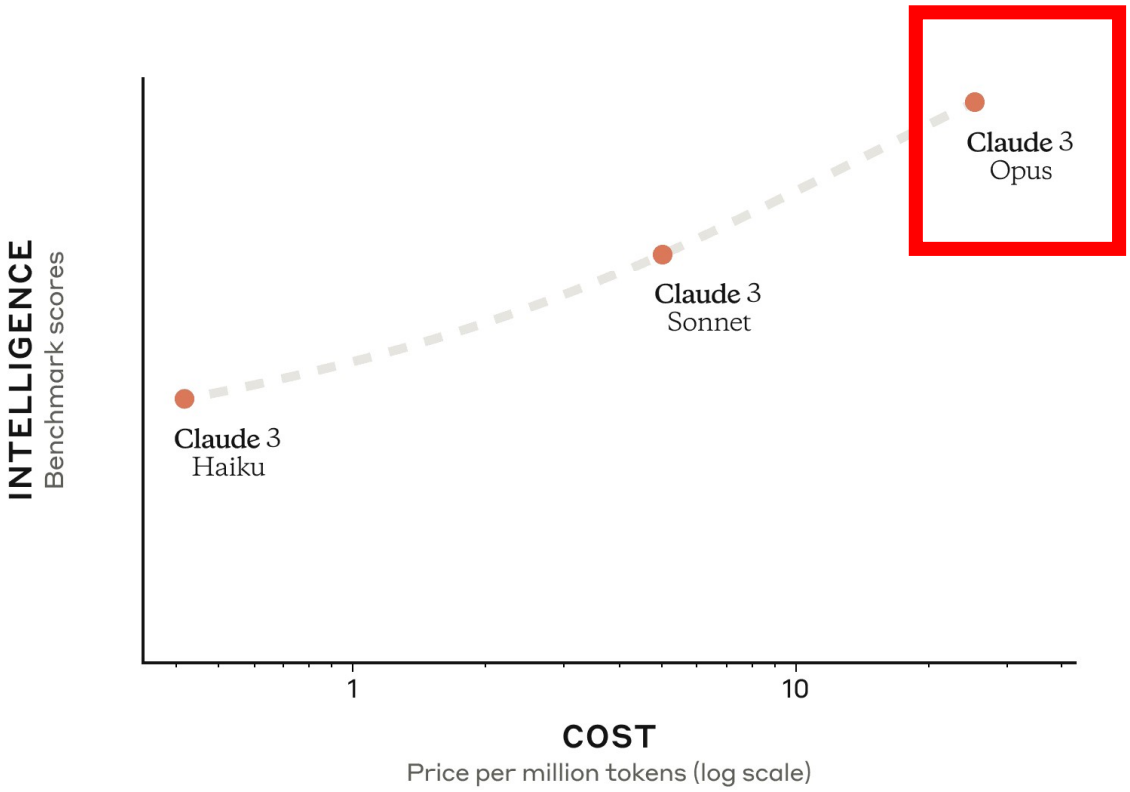
Gemma 경량화된 오픈소스 모델(SLLM)

모델의 규모가 클수록
성능이 향상된다는 가설을 다시한번 입증

https://blog-ko.superb-ai.com/what-is-the-alternative-to-llm-in-chatgpt/

Claude

ANTHROP\C



〈비용, 성능 차이에 따른 모델〉

GPT-4, Gemini Ultra보다 강력하다고?!

(Haiku – 논문 한 편을 3초도 안 되는 시간에 읽을 수 있음!)

https://modulabs.co.kr/blog/llama-3-intro/

Claude

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

https://modulabs.co.kr/blog/anthropic-claude-3-0/

Kensho에서 제공하는 AI 모델의 벤치마크 테스트 결과

Leaderboard

특정 도메인에 대한 지식 정도

양적정보 추출능력

프로그램 생성능력

종합성능

Rank	Model Name	Organization	Model Size (in billions)	Domain Knowledge (%)	Quantity Extraction (%)	Program Synthesis (%)	Overall (%)
1	**GPT-4o**	OpenAI	unknown	83.97	93.72	86.18	87.96
2	**GPT-4 Turbo**	OpenAI	unknown	81.68	95.52	85.77	87.66
3	**GPT-4**	OpenAI	unknown	80.15	96.41	79.67	85.41
4	Claude 3 Opus	Anthropic	unknown	74.05	92.83	82.52	83.13
5	Claude 3 Sonnet	Anthropic	unknown	71.76	95.52	71.14	79.47
6	Llama 3 70B	Meta	70	77.1	93.27	67.89	79.42
7	Mistral Large	Mistral AI	unknown	61.83	92.83	69.92	74.86
8	Claude 2	Anthropic	unknown	68.7	87	66.67	74.12

https://benchmarks.kensho.com/

트랜스포머를 대체할 차세대 아키텍처 ?

Mamba

: 시퀀스 모델링에 특화된 AI 아키텍처로,
긴 데이터 시퀀스를 효과적으로 처리할 수 있도록 설계된 모델

-> Attention 사용 X

Transformer

다음단어를 읽을 때,
이전의 문맥을 모두 기억해야 함.

Mamba

하지만, 인간은 대략적인 줄거리는 기억
하지만
신데렐라가 어떤 종류의 집안일을 했는
지는 기억하지 못함.

모든 정보는 기억하되,
스토리에 대한 압축된 표현을 만들어서
관련있는 것 -> 유지
관련없는 것 -> 삭제

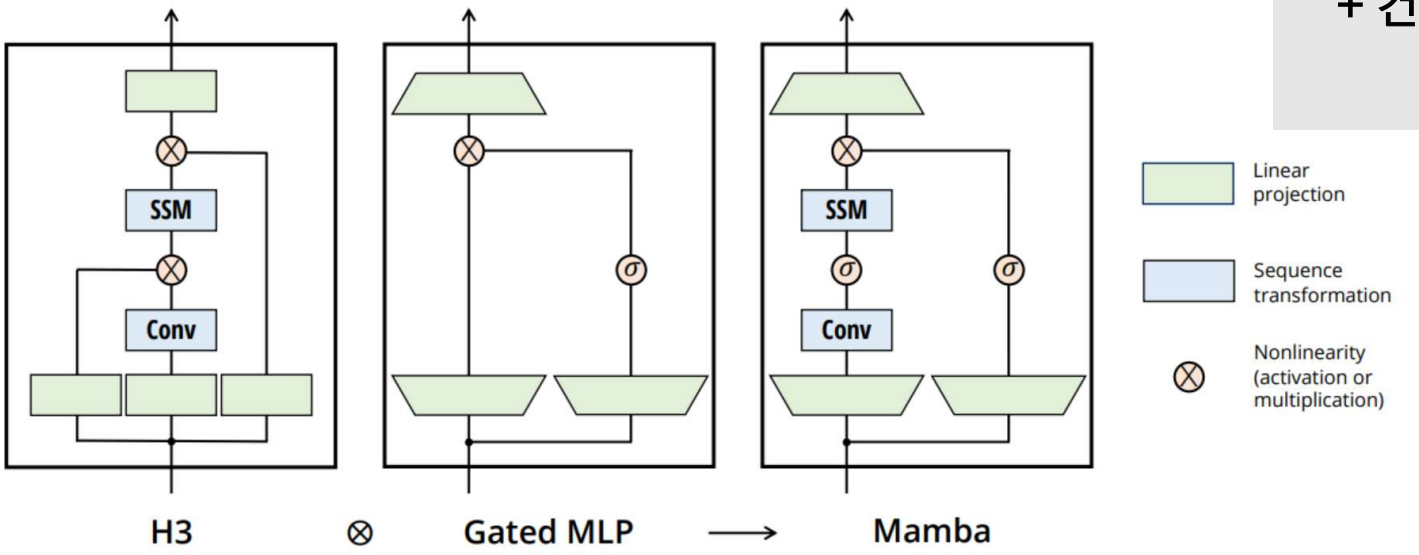
https://modulabs.co.kr/blog/introducing-mamba/

Mamba

: 시퀀스 모델링에 특화된 AI 아키텍처로,
긴 데이터 시퀀스를 효과적으로 처리할 수 있도록 설계된 모델

SSMs
(Selective State Space Models)
: 선택적 상태공간을 활용한 선형시간 시퀀스 모델링
(제어 이론에서 주로 이용되는 상태공간 모델)

Transformer	Mamba
훈련데이터 + 컨텍스트 데이터	훈련데이터 + 컨텍스트 데이터 압축/필터링



https://modulabs.co.kr/blog/introducing-mamba/