

인공지능, 무엇이고 어디로 가고 있는가?

2025. 6. 12

정 송

KAIST AI연구원장 및 AI대학원장

정송 교수

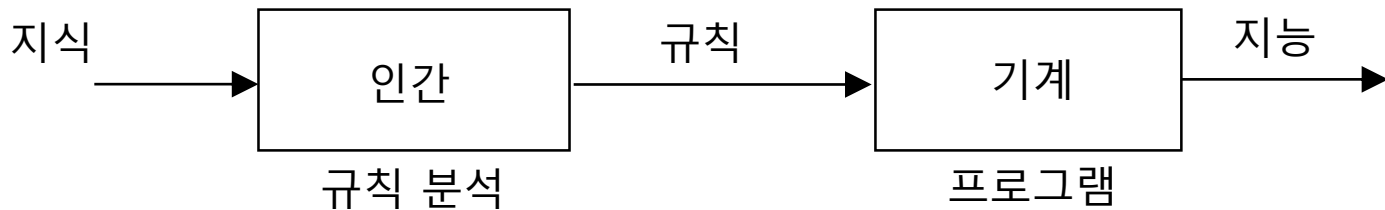
- 연구분야: 인공지능, 최적화, 통신네트워크
- 학력
 - ✓ 서울대학교 제어계측공학과 학사(1983-88), 석사(1988-90)
 - ✓ 텍사스 오스틴 전기컴퓨터공학과 박사 (1991-94)
- 경력
 - ✓ 2018 - 현재: KAIST ICT 석좌교수
 - ✓ 2019 - 현재: KAIST AI대학원 원장/교수
 - ✓ 2025 - 현재: KAIST AI연구원 원장
 - ✓ 2024 - 현재: KAIST Edge AI 연구센터 센터장
 - ✓ 2013-2015: KAIST-LG전자 5G 공동연구센터 센터장
 - ✓ 2000-2019: KAIST 전기전자공학부 교수
 - ✓ 1996-2000: 서강대학교 전자공학과 교수
 - ✓ 1994-1996: 미국 AT&T Bell Labs, Holmdel 연구원
- 논문 및 수상
 - ✓ 국제저널 71편, 국제학회 100편 (인용 11,238회, h-index 43)
 - ✓ IEEE William R. Bennett Prize Paper Award (2013, 2016) - 아시아 최초
 - ✓ ACM MobiSys 학회 Best Paper Award (2021) - 아시아 최초
 - ✓ IEEE SECON 학회 Best Paper Award (2013) - 아시아 최초
 - ✓ 한국통신학회 해동학술대상 (2016)
 - ✓ KAIST 기술혁신대상 (2016), KAIST 전략리더십상 (2025)
 - ✓ 삼성 휴먼테크 논문대상 (금 3, 은 6, 동 4)
 - ✓ AT&T 개척자상 (1995)
- 제자
 - ✓ 박사 21명, 석사 45명
 - ✓ 대학교수(국내 9명, 해외 1명), 기업/연구소(해외 7명, 국내 49명)



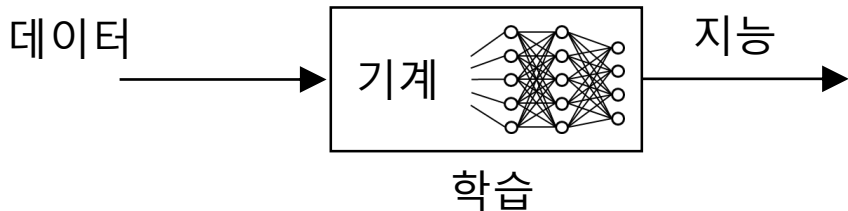
- 국제활동
 - ✓ IEEE Alexander Graham Bell Medal 선정위원
 - ✓ IEEE/ACM Trans. on Networking 편집위원
 - ✓ IEEE Trans. on Mobile Computing 편집위원
 - ✓ IEEE Trans. on Wireless Communications 편집위원
 - ✓ Computer Networks 편집위원
 - ✓ IEEE/IFIP WiOpt 조정위원장
 - ✓ ACM MobiHoc 2022 조직위원장
 - ✓ IEEE WCNC 2020 학술위원장
 - ✓ IEEE SECON 2015 학술위원장
 - ✓ IEEE/IFIP WiOpt 2009 조직위원장
- 국내활동
 - ✓ 前 인공지능대학원협의회 회장
 - ✓ 現 국가AI연구거점 자문위원회 위원
 - ✓ 現 문화예술교육진흥원 AI문화예술교육정책자문단 단장
 - ✓ 現 외교부 과학기술외교자문위원회 위원
 - ✓ 現 공군 정책발전자문위원회 위원
 - ✓ 現 6G포럼 자문위원회 부위원장
 - ✓ 現 한국지능정보사회진흥원 국가AI백서 편찬위원

규칙 기반 vs. 학습 기반

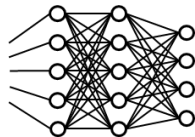
- 규칙 기반 논리적 추론 (기호주의 AI)



- 데이터 기반 통계적 학습 (연결주의 AI)



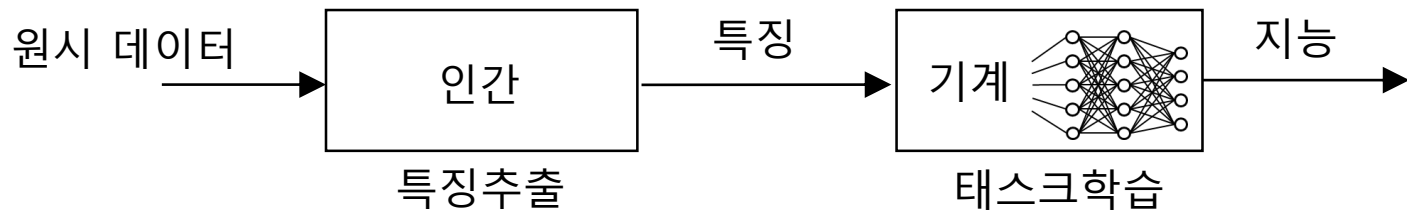
1957년 인공신경망
(Artificial Neural Network)



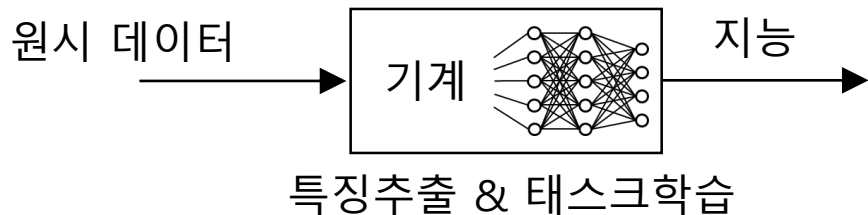
1986년 오류역전파법
(Backpropagation)

표현학습: 특징 자동추출 & 추상화

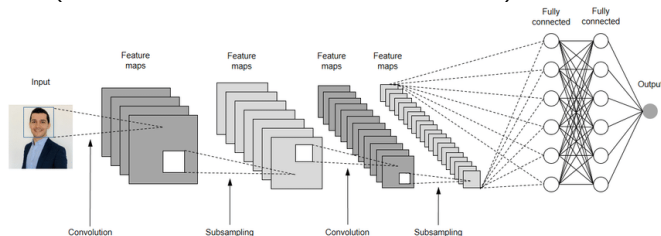
- 기존 머신러닝



- 딥러닝



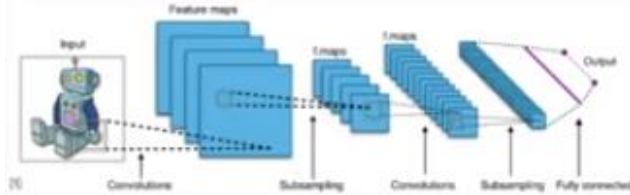
1989년/2012 합성곱신경망
(Convolutional Neural Network)



모델 중심 기계학습

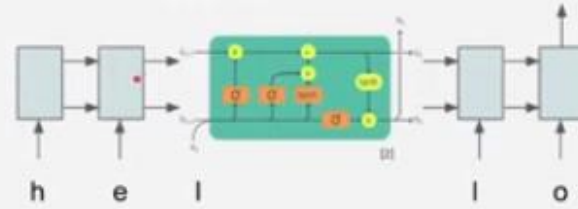
Computer Vision

Convolutional NNs (+ResNets)



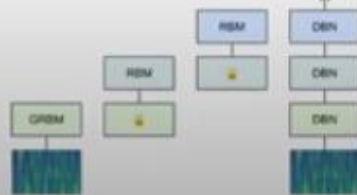
Natural Lang. Proc.

Recurrent NNs (+LSTMs)



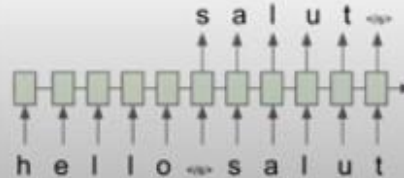
Speech

Deep Belief Nets (+non-DL)



Translation

Seq2Seq



RL

BC/GAIL

Algorithm 1 Generative adversarial imitation learning

1. **Input:** Expert trajectories $\tau_0 \sim \pi_0$, initial policy and discriminator parameters θ_0, w_0
2. **for** $i = 0, 1, 2, \dots$ **do**
3. Sample trajectories $\tau_i \sim \pi_i$
4. Update the discriminator parameters from π_i to π_{i+1} with the gradient

$$\nabla_{w_i} \mathbb{E}_{\tau_i} \log(D_i(x, \pi_i)) + \mathbb{E}_{\tau_0} \nabla_{w_i} \log(1 - D_i(x, \pi_i)) \quad (17)$$
5. Take a policy step from θ_i to θ_{i+1} using the TRPO rule with cost function $\log(D_{i+1}(x, \pi_i))$. Specifically, take a KL-constrained natural gradient step with

$$\hat{\nabla}_{\theta_i} \mathbb{E}_{\tau_i} \log \pi_i(x) \nabla^2 \pi_i(x) \leftarrow \nabla^2 \pi_i(x)$$

$$\text{where } \nabla^2 \pi_i(x) = \mathbb{E}_{\tau_i} \nabla^2 \log \pi_i(x, \pi_i) \mid \pi_i \sim \pi_i, \pi_0 = \pi_0 \quad (18)$$
6. **end for**

데이터 중심 기계학습

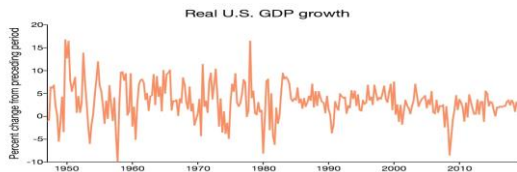
- 2017년 트랜스포머 (Transformer) => 시퀀스 이해 문제로 통일

나는 너를 사랑해

Language Transformer



Vision Transformer



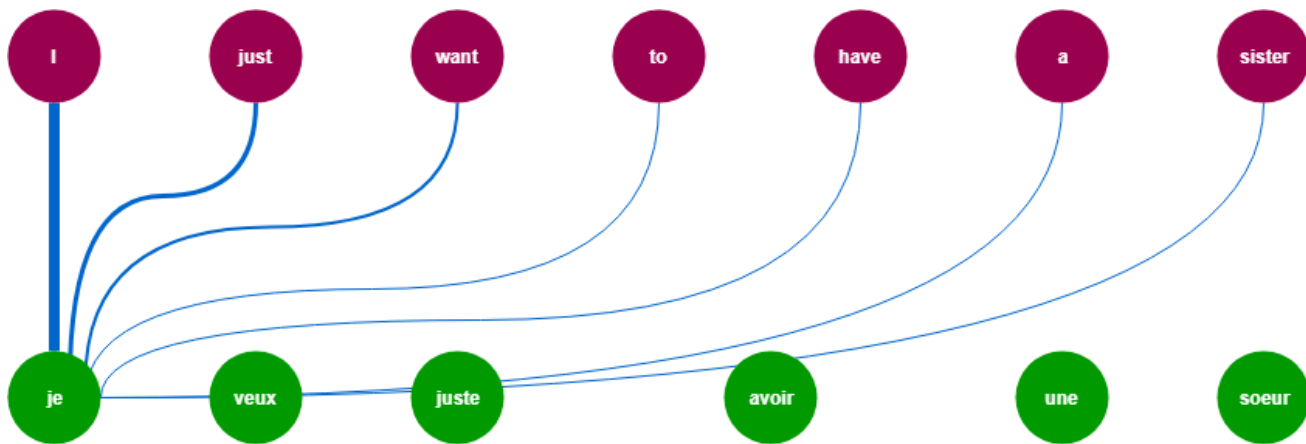
Time Series Transformer



Decision Transformer

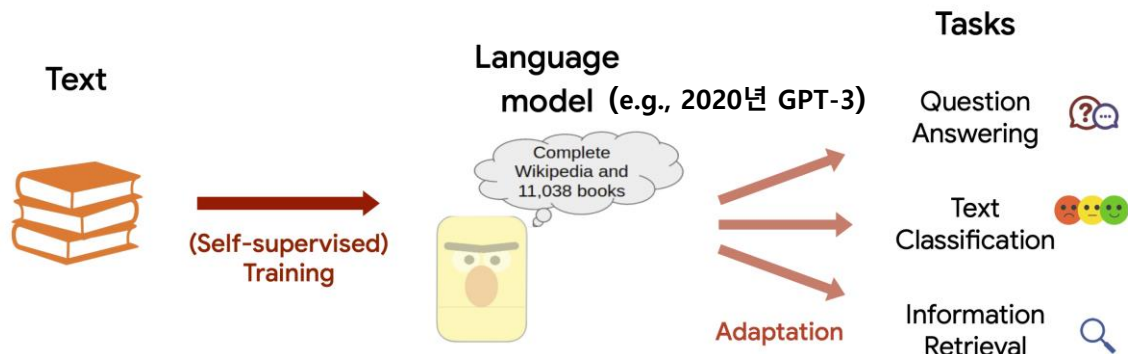
인간의 주의집중을 모방

- 2015년 어텐션 메커니즘 (Attention Mechanism)

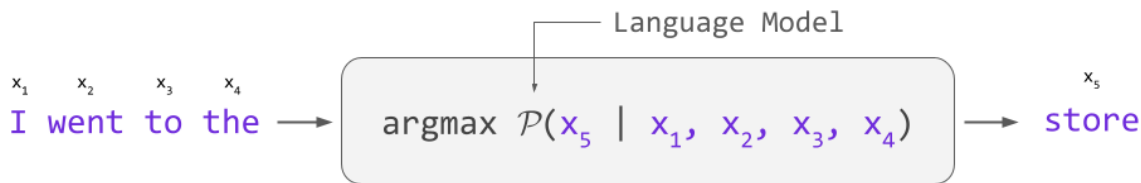


사전학습을 통한 대형언어모델(LLM)

- 기반모델 (Foundation Model)

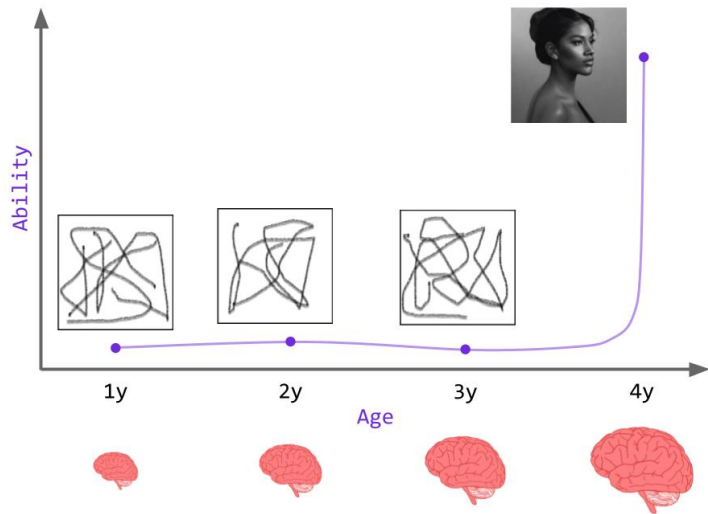
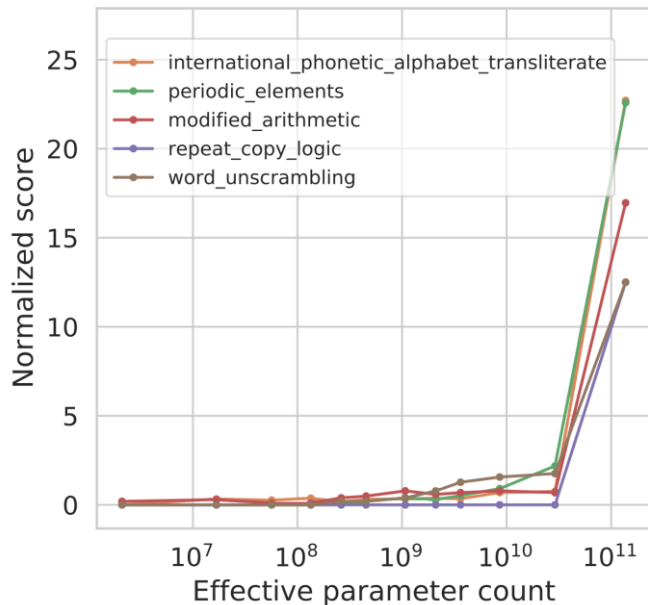


- 생성형 인공지능 모델 - 확률적으로 다음 단어 생성



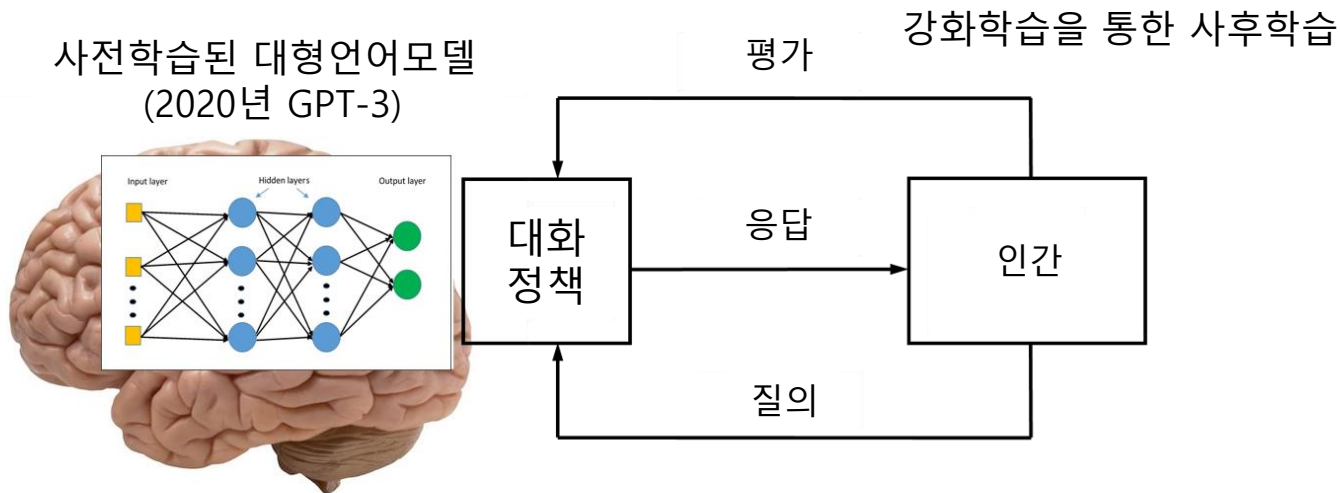
대형언어모델의 창발성

- 트랜스포머 크기가 커짐에 따라 질의응답, 문장요약, 추론, 수학문제 풀기, 코딩 등 다음 단어 생성 능력을 넘어서는 새로운 능력 창발

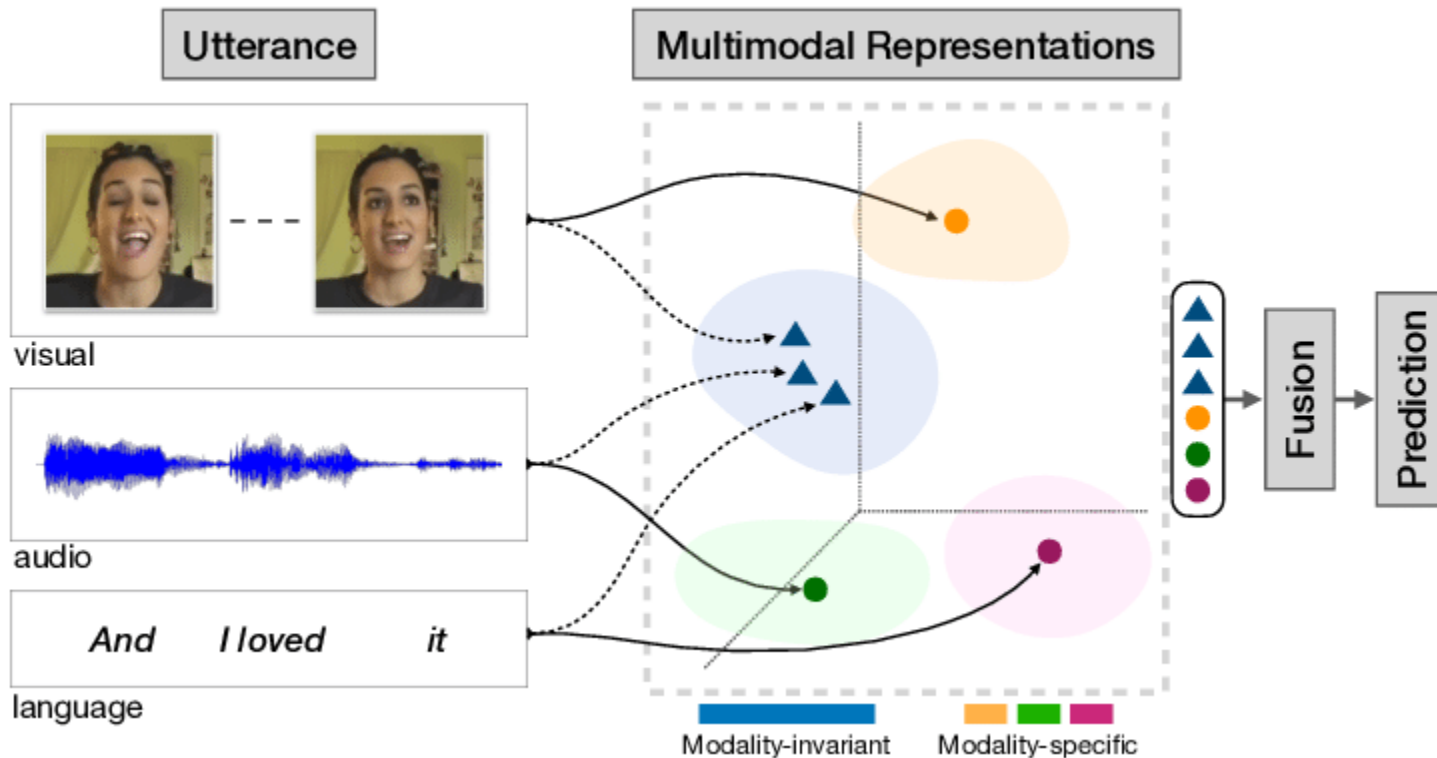


인간 피드백을 통한 AI의 인간 가치 정렬

- ChatGPT (2022년)



멀티모달 이해와 생성



대형 멀티모달 모델 (2023년)

- GPT-4, Gemini

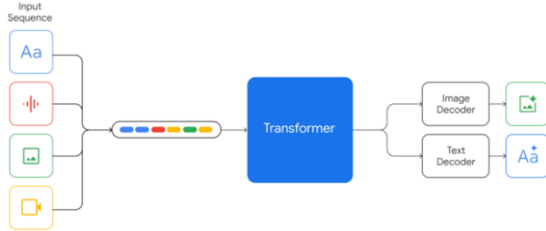
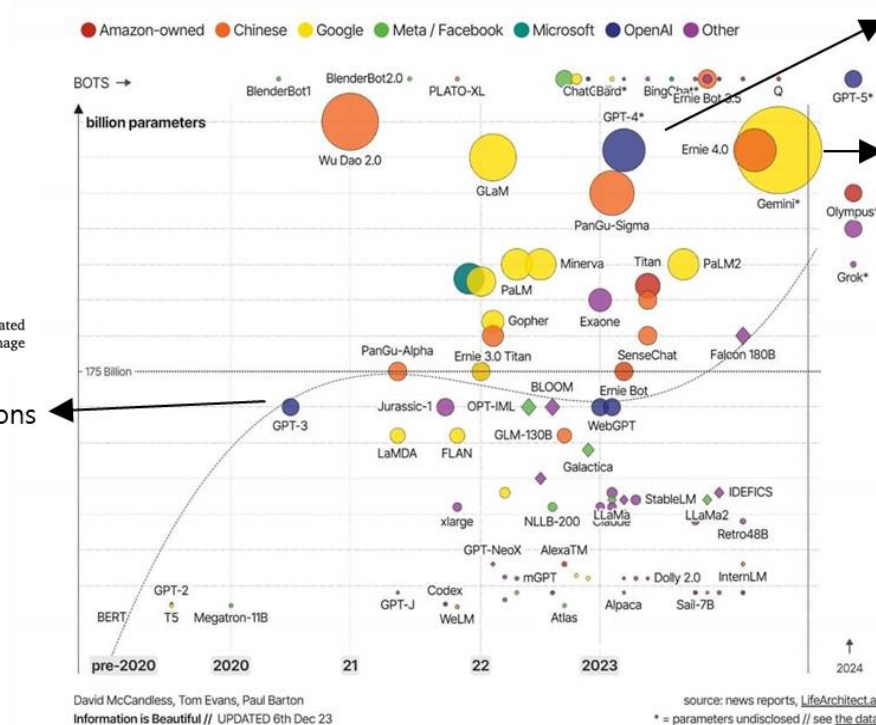


Figure 2 | Gemini supports interleaved sequences of text, image, audio, and video as inputs (illustrated by tokens of different colors in the input sequence). It can output responses with interleaved image and text.

Parameters: 175 billions
Date: May 2020
Owner: OpenAI



Parameters: 1,000 billions
Date: March 2023
Owner: OpenAI

Parameters: 8,000 billions
Date: December 2023
Owner: Google

플랫폼의 플랫폼

플러그인/스킬/툴



Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



OpenTable

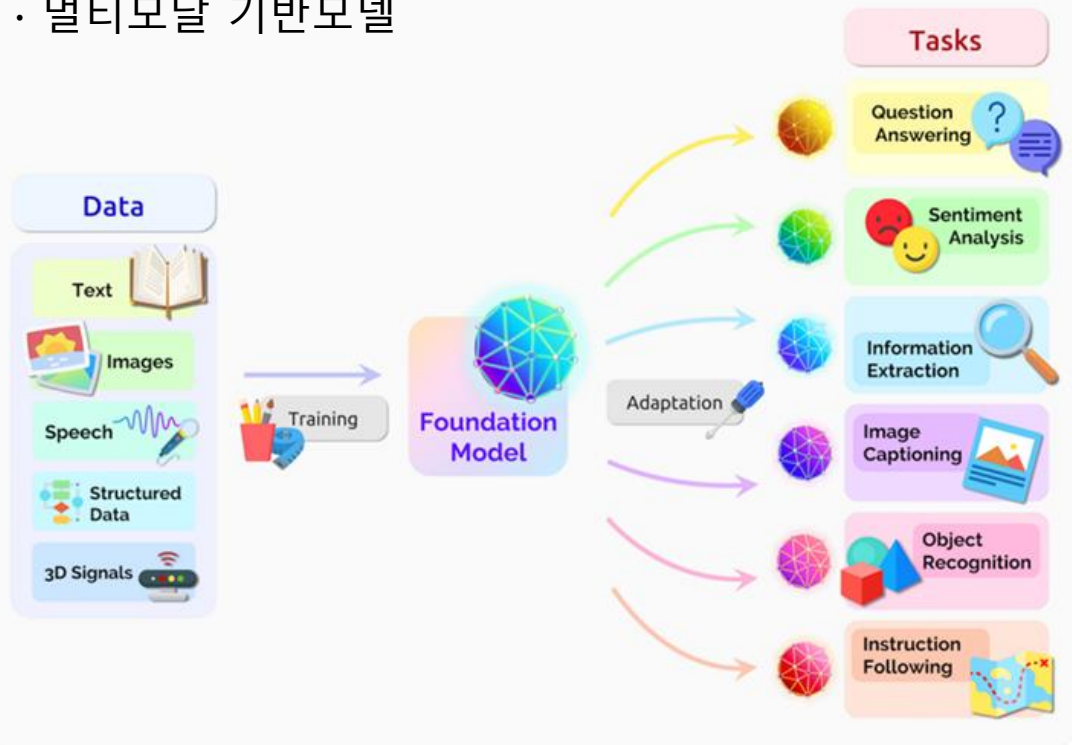
Provides restaurant recommendations, with a direct link to book.



Wolfram

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and

· 멀티모달 기반모델



앱스



대형언어모델과 검색의 결합

Microsoft Bing SEARCH CHAT

What's the high and low for weather in Toronto, Ontario today?

✓ Searching for: Toronto, Ontario weather today
✓ Generating answers for you...

Your personal and company data are protected in this chat

Today in Toronto, Ontario, the high temperature is **15 degrees Celsius** and the low temperature is **8 degrees Celsius**. The weather is predicted to be **cloudy** with a **30% chance of showers early this afternoon**. The wind speed is expected to be **20 km/h gusting to 40 km/h from the northeast**.

I hope this helps!

Learn more: [1. a.msn.com](#) [2. cbc.ca](#) [3. theweathernetwork.com](#) [+3 more](#) 1 of 30

Toronto, Ontario
Updated a few minutes ago

14°C 15° 8°
Precipitation: 8% Wind: 22 KM/H Humidity: 67%

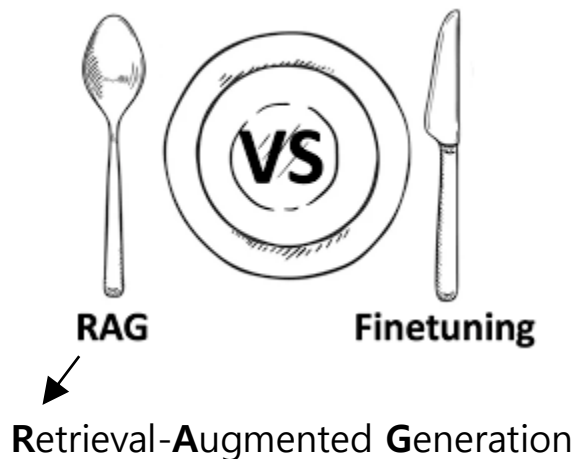
Cloudy - Mon 16, 1:42 p.m.

14° 14° 12° 11° 10° 9° 8° 12°
2 PM 5 PM 8 PM 11 PM 2 AM 5 AM 8 AM 11 AM

Mon 16 Tue 17 Wed 18 Thu 19 Fri 20 Sat 21 Sun 22 Mon 23 Tue
15° 8° 15° 7° 16° 10° 18° 13° 16° 9° 14° 9° 13° 6° 10° 3° 13°

Learn more from Microsoft Start Weather

What is the weather forecast for tomorrow? How do I prepare for a rainy day in Toronto? What are some indoor activities to do on a cloudy day in Toronto?



대형 행동모델 (LAM)

- 행동 가능 LLM

- 플러그인을 통해 외부 도구(프로그램, 서비스, 데이터, 행동체 등)와 연결되어 특정 작업 수행

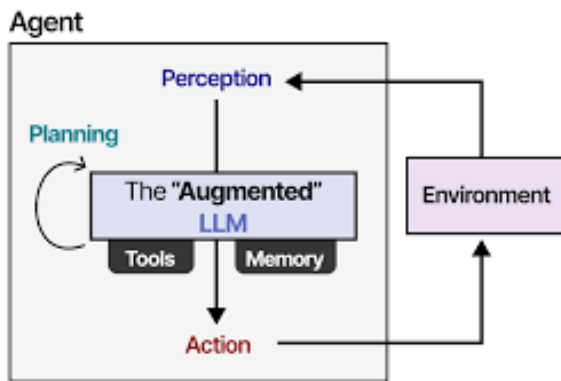
- AI 에이전트로의 진화

- 자율적인 상황 인식
 - 자율적인 의사 결정
 - 실시간 행동 수행
 - 지속적인 학습

- 피지컬 AI로의 진화

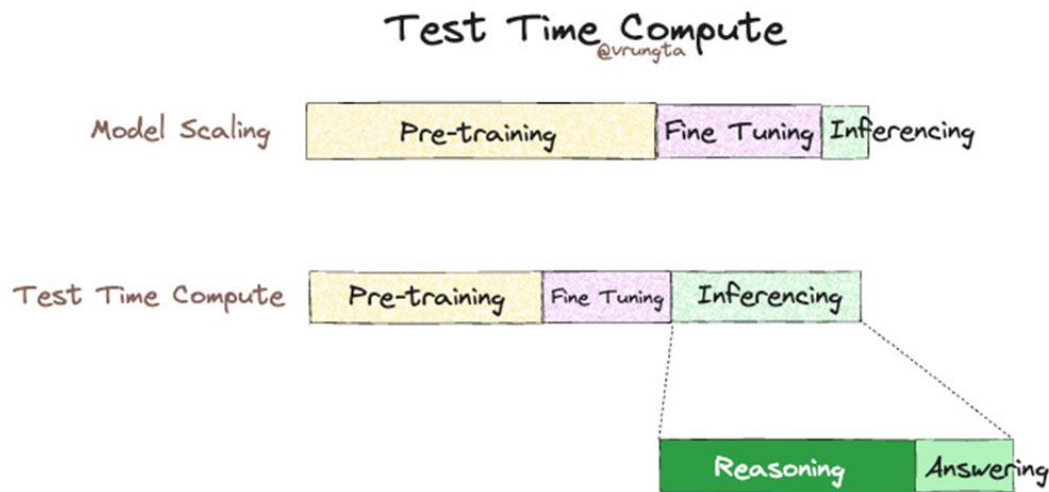
- 물리적인 세계를 이해하고 물리적인 행동을 수행
 - 로봇, 자율주행

- 인간-컴퓨터/머신/실제세계 상호작용의 혁명



추론모델: LLM 스케일링의 패러다임 전환

- 규모를 더 크게 vs 생각을 더 오래



- 사고의 연쇄 (Chain of Thought), 사고의 나무 (Tree of Thoughts), 탐색 (Exploration), 강화학습(Reinforcement Learning)

AGI?

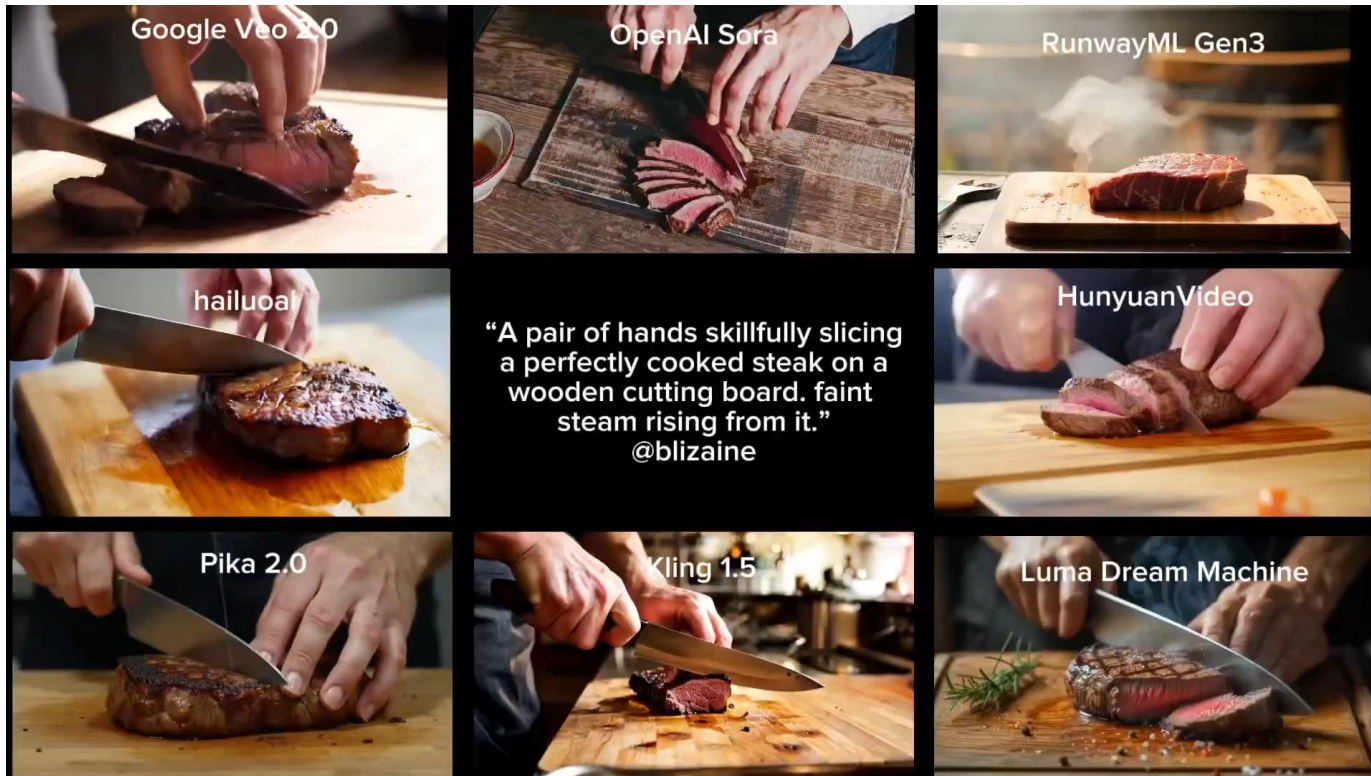
- AGI의 요구사항?
 - 추론 (Reasoning)
 - 사고의 연쇄, 탐색 및 평가
 - 행위자적 추론 (Agentic Reasoning)
 - 능동적인 목표 설정, 계획 수립 및 의사결정
 - 환경과 상호작용하며 환경 변화에 적응
 - 자기 진화 (Self-evolving)
 - 자기 평가
 - 지속적인 학습 및 성능 개선
- 인간의 지능이 궁극적 목표인가?
 - 비행기 vs. 새
 - AGI vs. 인간의 뇌
- AGI의 위험성?
 - 상태/행동 공간의 제약을 통해 관리 가능

AI 스택: 어디에 가장 큰 기회가 있나?

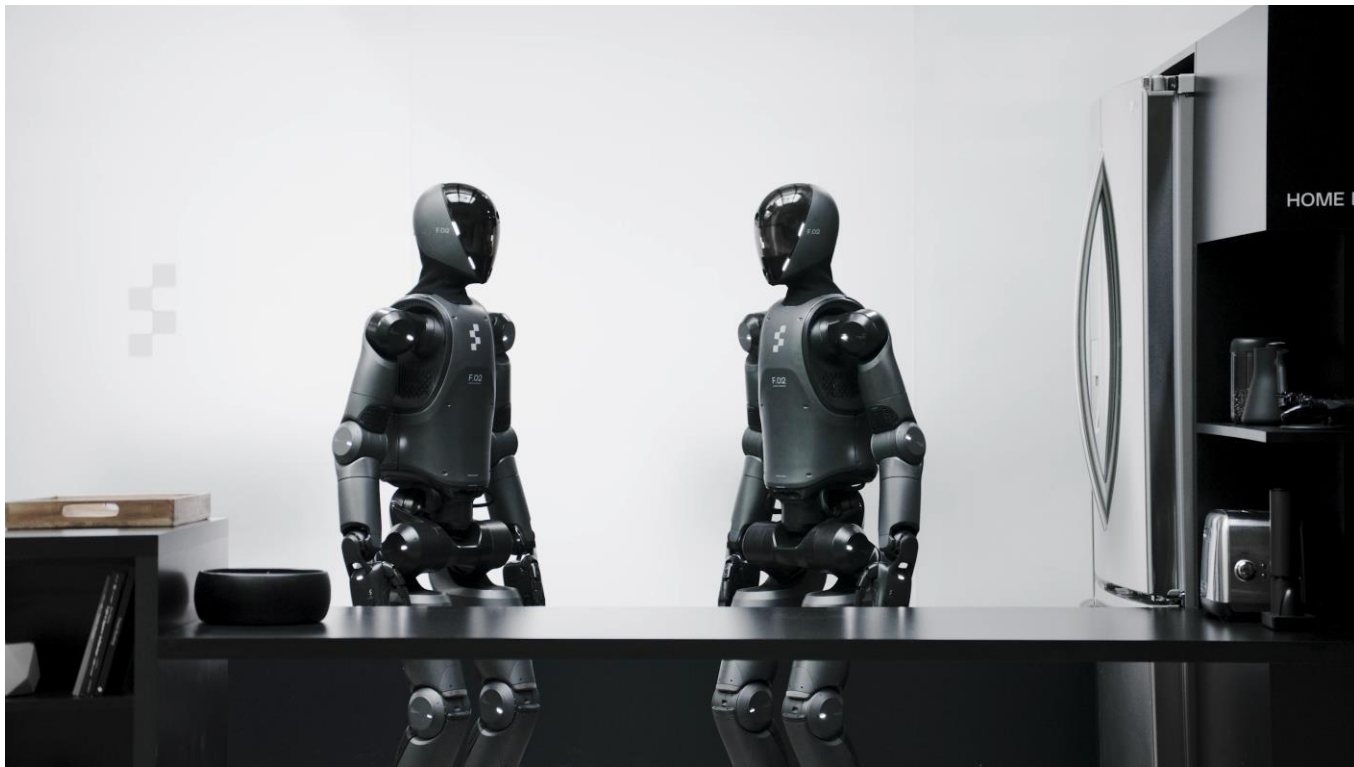
- 트랜스포머 대안·로봇 기반모델 등 차세대 기반모델이 글로벌 AI 기술 경쟁의 핵심으로 부상
- 다양한 버티컬 AI 개발을 통해 기반모델의 산업 적용과 혁신 잠재력 실현
- 동일선상에서 출발하는 지금이 AI G3 도약의 절호의 기회



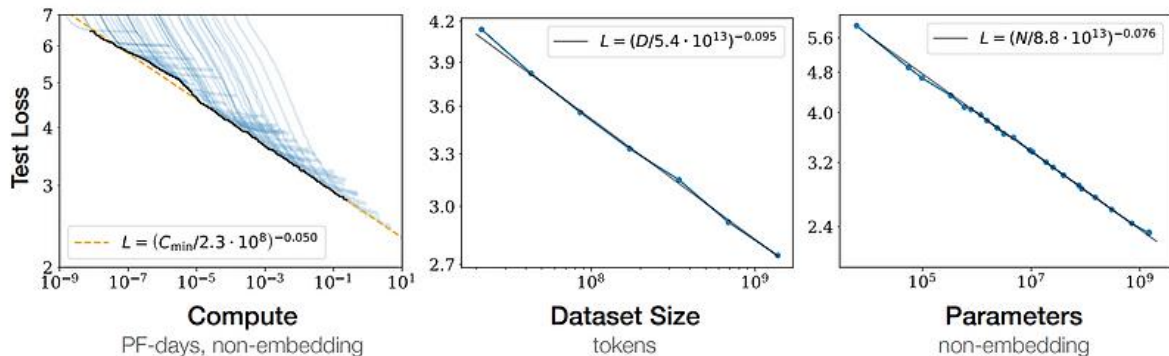
비디오생성 기반모델 ⇒ 물리법칙 이해 월드모델



월드모델 \Rightarrow 로봇 기반모델



뉴럴 비용곡선 법칙 (Neural Scaling Law)



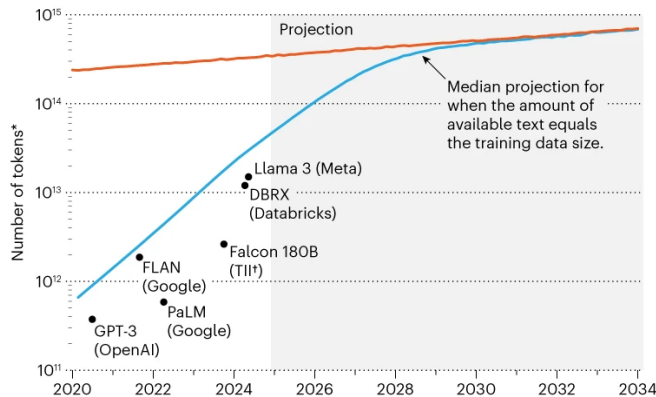
The Stargate Project: Trump Touts \$500 Billion Bid For AI Dominance

OpenAI reportedly wants to build 5-gigawatt data centers, and nobody knows who could supply that much power

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly approaching a crisis point. An estimate suggests that, by 2028, developers will be using data sets that match the amount of text that is available on the Internet.

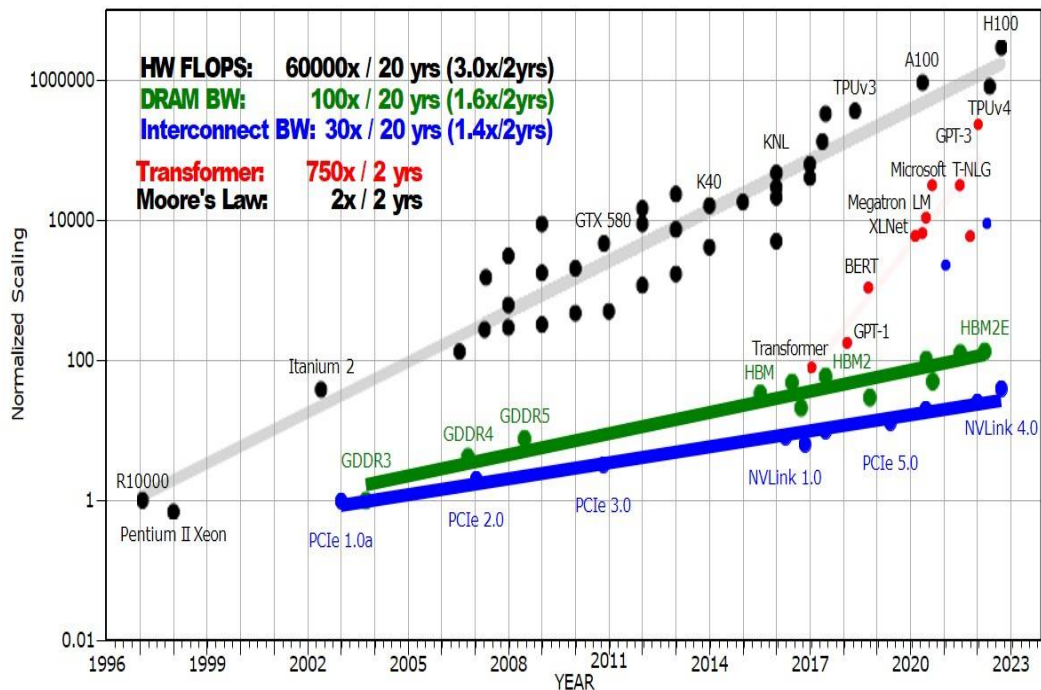
— Amount of available text on the Internet — Size of training data sets for LLMs
• Individual LLMs



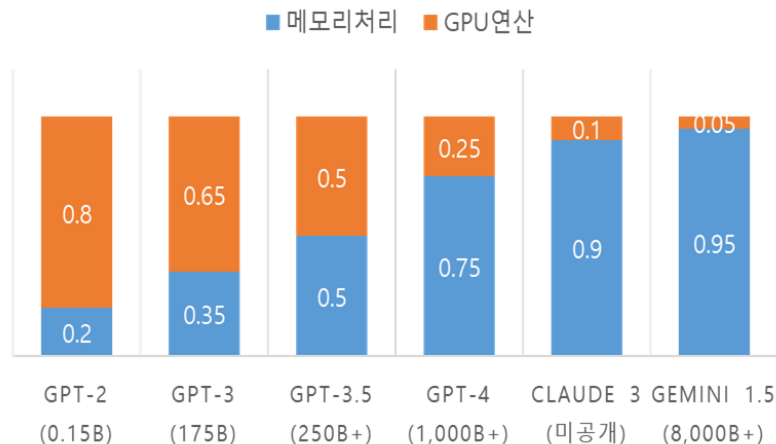
©nature

*One token is about 0.8 words. *Technology Innovation Institute, Abu Dhabi.

LLM과 메모리 장벽



GPU연산 대비 메모리처리 비중



뉴럴 비용곡선 법칙의 한계 돌파

	상태공간 모델	뉴로모픽 모델
기반 철학	수학적 시스템 해석 (제어이론 기반)	생물학적 뉴런 모사 (뇌 신경망 구조)
연산 구조	시간 상태 천이 + 커널 기반 처리	이벤트 기반 스파이킹 연산 (비동기)
Attention 사용	사용하지 않음	사용하지 않음
계산 복잡도	선형 복잡도 $O(N)$	이벤트 수에 비례 $O(\text{활성 뉴런 수})$
학습	- Backpropagation 기반 e2e 학습 - 구조적 커널로 안정적 학습	- STDP, Hebbian 등 - 학습 방법 미성숙
병렬성	GPU 친화적, 병렬 연산 가능	이벤트 기반, 병렬성 낮음
응용	언어 모델, 시계열, 긴 문맥에 강함	실시간 반응, 로봇, 센서 입력 특화
대형모델 확장성	GPT 대체 가능성 있음 (Mamba, S4 등)	대형 LLM 급 확장 없음

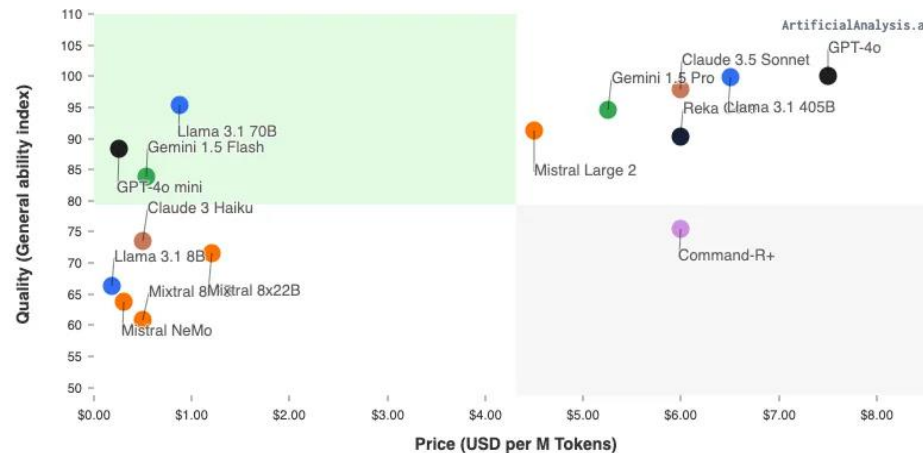
오픈소스 모델 vs. 상용 모델

Quality vs. Price

Quality: General reasoning index, Price: USD per 1M Tokens

Most attractive quadrant

■ GPT-4o ■ GPT-4o mini ■ Gemini 1.5 Pro ■ Gemini 1.5 Flash ■ Llama 3.1 405B ■ Llama 3.1 70B
■ Llama 3.1 8B ■ Mistral Large 2 ■ Mixtral 8x22B ■ Mistral NeMo ■ Mistral 8x7B
■ Claude 3.5 Sonnet ■ Claude 3 Haiku ■ Command-R+ ■ Reka Core

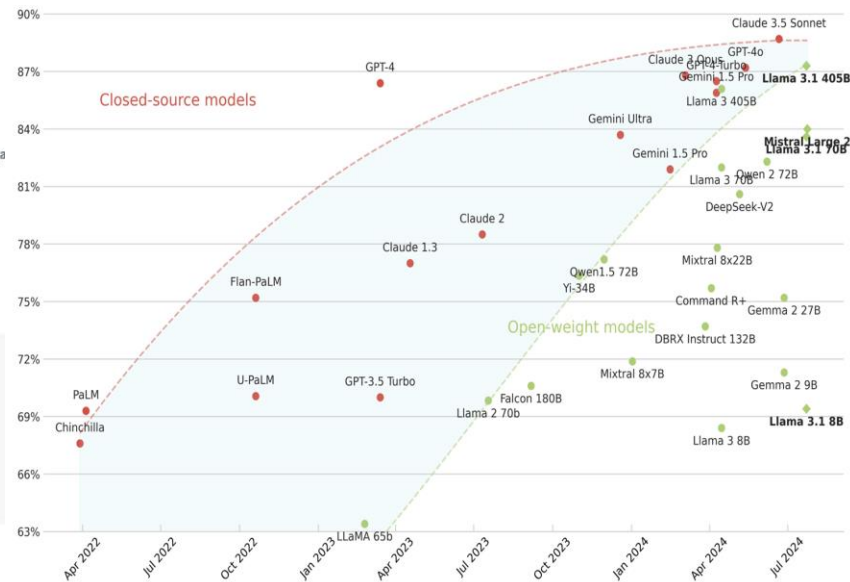


Closed-source vs. open-weight models

Llama 3.1 405B closes the gap with closed-source models for the first time in history.

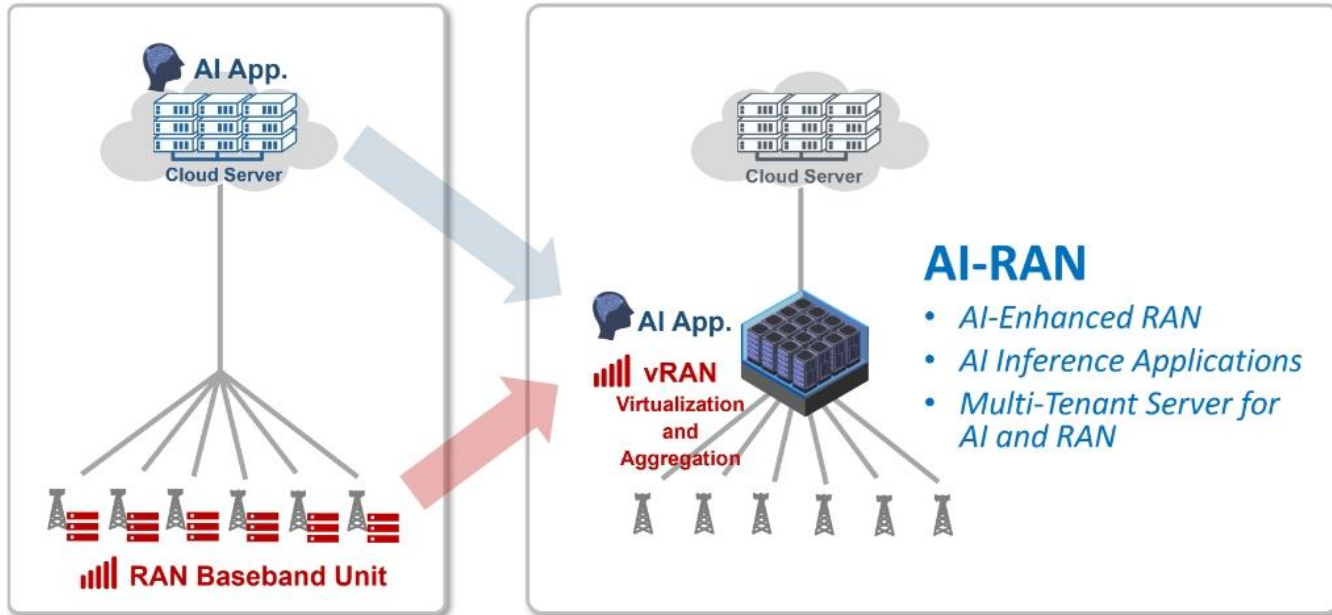
@maximelabonne

MMLU (5-shot)

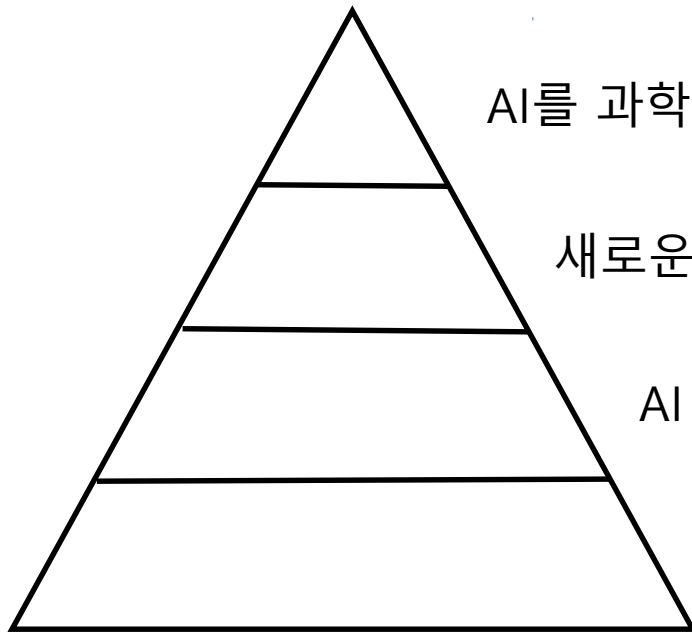


엣지 AI: 인공지능과 이동통신의 결합

AI-RAN : AI and RAN Integration



AI 인재 피라미드



AI를 과학으로 다루는 인력

새로운 AI 모델과 알고리즘을 개발하는 인력

AI 시스템과 서비스를 개발하는 IT 인력

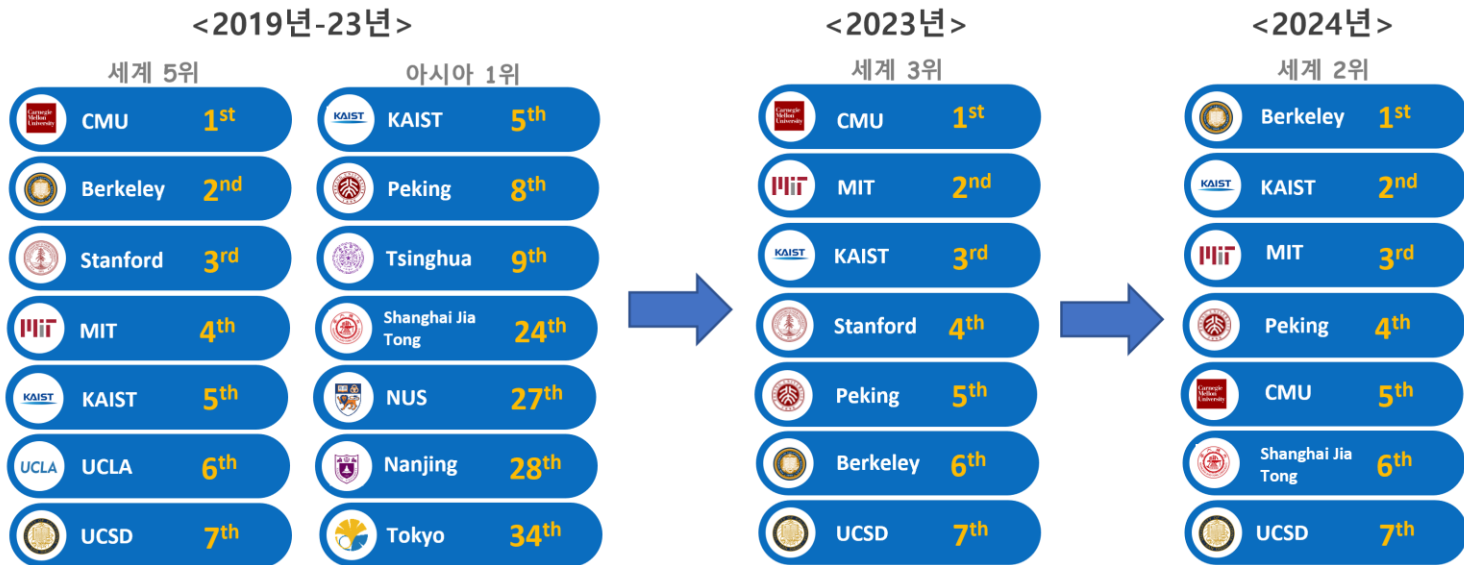
AI를 산업에 활용하는 도메인 인력

톱 클래스 AI 인재 유치 조건

- 동료 연구자 수준
- 데이터/컴퓨팅 리소스
- 연구 자율성
- 급여 수준
- 삶의 질
- 산학 겸직

KAIST, 인공지능 세계 대학랭킹 2위

기계학습 3대 Top 학회 (NeurIPS, ICML, ICLR) 총 논문 수 기준



해외 타 대학과의 비교: NYU 14위, EP리 17위, ETH Zurich 19위, Toronto 22위, Tel Aviv 23위, Technion 25위, Montreal 26위, Max Planck 37위, Oxford 53위, Cambridge 62위

국내 타 대학과의 비교: 서울대 65위 (아시아 13위), 포항공대 90위 (아시아 18위), 연세대 144위 (아시아 31위), UNIST 166위 (아시아 41위), 고려대 187위 (아시아 45위)

*Source : <https://csrankings.org>

Q&A