데이터 사이언스 전망

2024. 12.

Industrial Data Science Lab

- **Scaling Laws (Open AI, 2020)**
  - 컴퓨팅 리소스, 데이터, 모형 크기를 늘릴 수록 성능 개선
  - 새로운 능력이 생겨남 (Emergent Abilities)

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.

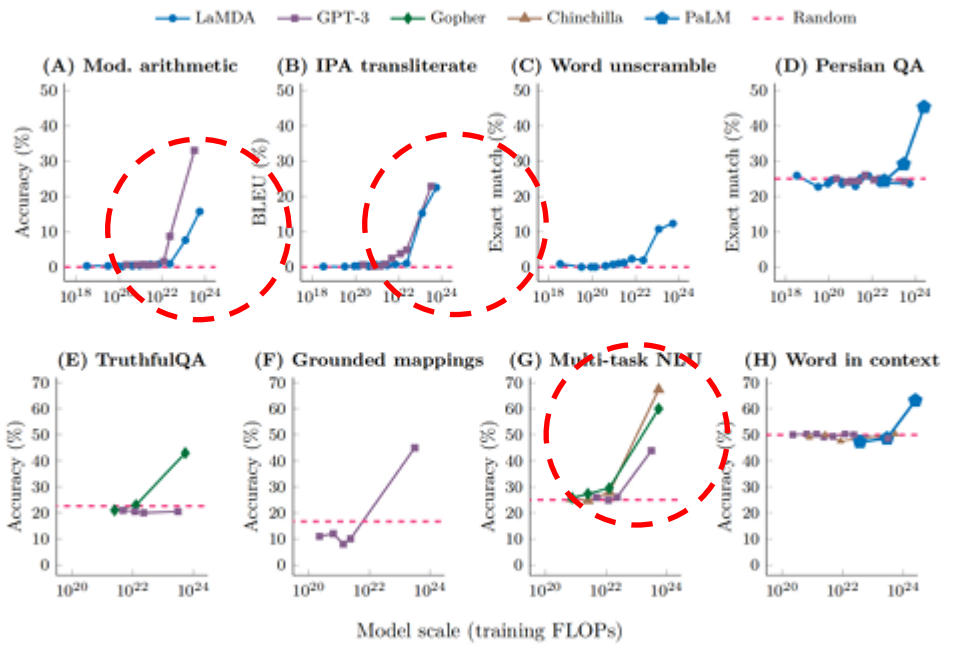
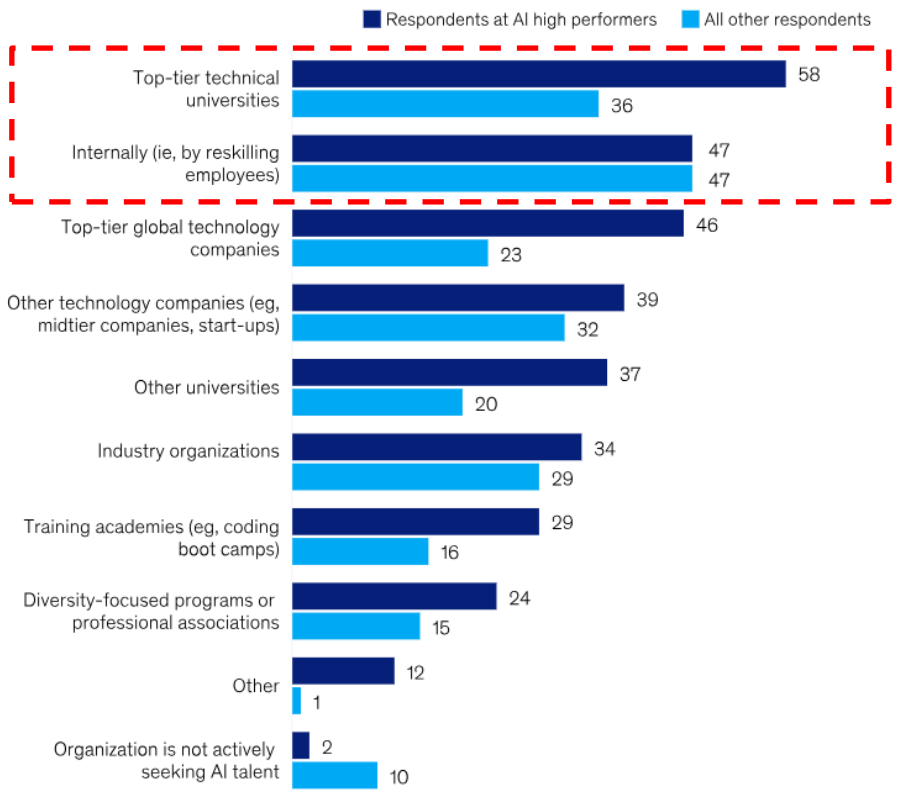


Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [33].

- **Brain drain의 문제**
  - "The state of AI in 2022—and a half decade in review, McKinsey, 2022"

Respondents from AI high performers report sourcing AI-related talent in a broader variety of ways than other respondents.

Sources that respondents' organizations are using for AI-related talent, % of respondents[1]



[1]Only asked of respondents whose organizations have adopted AI in at least one function. For respondents at AI high performers, n = 51. For all other respondents, n = 413.

McKinsey & Company

Stanford's 2021 Artificial Intelligence Index Report

"… the number of new AI PhD graduates in North America entering the AI industry post-graduation grew from 44.4% in 2010 to around 48% in 2019. By contrast, the share of new AI PhDs entering academia dropped from 42.1% in 2010 to 23.7% in 2019…."

- AI 확산과 함께, 산업 분야에서의 AI 인력 급증
- 연구기관/대학에서의 Brain drain

- **생성형 AI**
  - 대형 언어모델이나 이미지 생성모델 활용, User가 원하는 것을 생성하는 AI 분야
  - LLM: 테라바이트 단위의 대용량 코퍼스로 모형을 학습, 문장을 생성하는 모형
  - IGM(Image Generation Model): 텍스트를 입력하면 그에 대한 이미지를 생성, DALL-E2, 미드저니 등

II. AI 활용 이슈

- AI 활용과 이슈

Radiology Case Reports | Open access

Articles & Issues | About | Publish | Search in this journal | Guide

Article publishing charge

$550

Article publishing charge for open access

This journal offers authors the option to publish their research via open access. To publish open access, a publication fee (APC) author or research funder.

Publishing timeline

Acceptance rate

Abstracting and indexing

Fig. 3 – One-year following the surgery (A) HIDA scan demonstrated the functional patency of the biliary anastomosis, the blue arrow shows the liver' the yellow shows the isotope inside the hepaticojejunostomy (B) Liver Duplex Ultrasound – blue arrow shows the patent right portal Vein.

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Conclusion

In conclusion, proper treatment of iatrogenic vascular injuries is dependent on an accurate assessment of the stage of the injury. The injury should be recognized quickly. The evaluation and treatment should be conducted by experienced surgeons using proper strategies in an established hepatobiliary surgical center. Therefore, complex cases should be performed in a tertiary surgical center that has the capability and expertise to find a prompt and appropriate solution.

Radiology Case Reports 19 (2024) 2106–2111

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/radcr

RCR RADIOLOGY CASE REPORTS

Case Report

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review ☆,☆☆

Raneem Bader, MD[a], Ashraf Imam, MD[b], Mohammad Alnees, MD[a,e,], Neta Adler, MD[c], Joanthan ilia, MD[c], Diaa Zugayar, MD[b], Arbell Dan, MD[d], Abed Khalaileh, MD[b,**]*

[a] Hadassah Medical Center, Jerusalem, Israel
[b] Department of General Surgery, Hadassah Medical Center and Hebrew University, Jerusalem, Israel
gery, Hadassah medical center and Hebrew university, Jerusalem,
edical center and Hebrew university, Jerusalem, Israel
l Education, Global Clinical Scholars Research Training program,

ABSTRACT

Serious iatrogenic vascular injuries are considered uncommon; however, they are underreported. There are limited studies discussing the proper management of these injuries; therefore, the management is often anecdotal. A 4-month-old female patient presented with elevated liver enzymes and cholecystitis with sludge. Her HIDA scan suggested biliary atresia. During the surgery, there was a bilateral iatrogenic injury to the hepato-duodenal ligament, including the portal vein, hepatic artery, and bile ducts. The patient underwent splenectomy and cholecystectomy, and the hepatic artery transection was successfully managed with a splenic artery jump graft and a portal vein bypass initiated with the SMV using a Gore-Tex® vascular graft. The management of iatrogenic vascular injury depends primarily on the assessment of the stage of the injury, which should be conducted by experienced surgeons using proper strategies in an established hepato-biliary surgical center. Additionally, there is little data provided in the literature, mostly case reports. Therefore, no preferred or specific approach can be found.

- **DX to AX: AI Transformation**

  - AI도입과 관련한 기업 의견 및 기술 활용도 인식

(단위: %; N=1,000)

구분	비중	구분	비중
기업 수요에 맞는 AI 기술 및 솔루션 부족	35.8	데이터 활용(개인정보 및 데이터 접근)	15.6
		성과 창출의 불확실성	11.2
AI 기술 및 솔루션 개발 비용	20.6	기타	1.0
전문인력 및 역량 부족	15.7	모름/무응답	0.1

* 자료: KDI(2020)

(단위: %; N=738)

ID	세부 기술(토픽)	기술 활용도		수용도	기술 유용성		개발 시급성 정도	
		현재 활용도	미래 전망	수용 의사	성과 도움	경쟁력 제고	R&D 시급	국고 지원
1	자연어 이해 및 인식 처리 기술	54.7	64.8	51.9	60.3	64.6	58.1	56.5
2	인간 감정 분석 기술	49.2	58.8	45.5	54.2	57.9	51.4	50.1
3	지식 추론 기술	56.9	67.2	54.3	60.8	60.6	55.6	58.7
4	생성형 인공지능 기술	70.1	76.0	62.2	74.9	75.6	70.5	73.8
5	인공지능 신뢰성 기술	64.2	76.2	62.1	70.7	76.3	67.8	71.5
6	경로 탐색 및 모델 최적화	68.6	73.2	64.9	71.5	73.6	72.8	72.2
7	객체 감지 및 추적을 위한 비전 딥러닝 기술	66.8	77.5	65.9	70.7	71.7	73.4	74.1
8	그래프 분석 기반 진단 및 예측 기술	66.0	74.3	62.7	63.7	68.6	67.3	69.8
9	강화학습 기술	66.0	74.7	65.9	68.0	71.3	68.6	70.6
10	머신러닝 기반 데이터 보안 및 보호 기술	68.7	76.8	68.4	75.1	76.3	72.9	72.6
11	딥러닝 기반 이미지 분석 및 처리 기술	73.3	78.3	70.1	74.4	76.8	75.2	73.2
12	딥러닝 모델 알고리즘 및 성능 최적화	71.1	82.8	66.0	76.4	78.0	73.3	75.9

* 주: 각 토픽별 최고값과 최저값의 항목을 각각 파란색, 주황색으로 표시함

Source: 우리나라 및 주요국 인공지능(AI) 기술수준의 최근 변화 추이 (SPRI,2023)

II. AI 활용 이슈

- **LLM 성능 개선: 최신 정보 습득, Hallucination, 수학 계산 등**
- **LMM: 멀티모달, 컴퓨터 비전과 언어모형의 결합**
  - Diffusion Model: 이미지 합성과 변환
- **강화학습: Human Preference의 효과적 반영, Diffusion에의 적용**

- **Leveraged by AI: AI 활용 전략의 중요성**

서울경제 구독

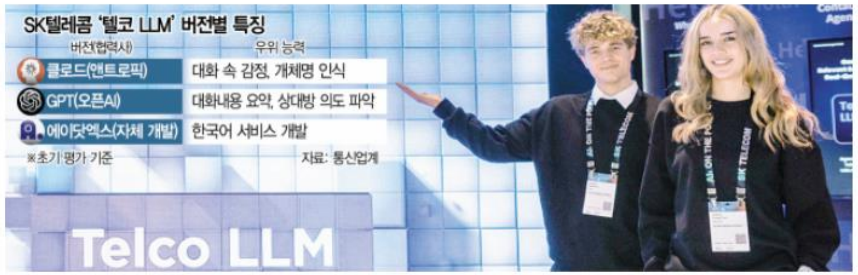
공감은 클로드, 요약은 GPT, 한국어 서비스는 에이닷… SKT '통신 언어모델' 세가지 버전 만든다

입력 2024.03.18. 오후 5:36 · 수정 2024.03.18. 오후 6:37 기사원문

김윤수 기자

1 댓글 가가

[텔코LLM 멀티전략 구체화]
버전별 장점 살리는 파인튜닝 돌입
클로드 버전 상담봇 구현에 제격
GPT, 트렌드 분석·사후관리 활용
골라 쓰게 지원…연내 출시 목표



[서울경제]

SK텔레콤이 이동통신사에 특화한 대형언어모델(LLM)을 서로 다른 세 가지 버전으로 출시한다. SK텔레콤은 앤트로픽의 클로드 버전, 오픈AI의 GPT 버전, 자체 개발한 에이닷엑스(A.X)로 구성

매일경제 구독 PICK

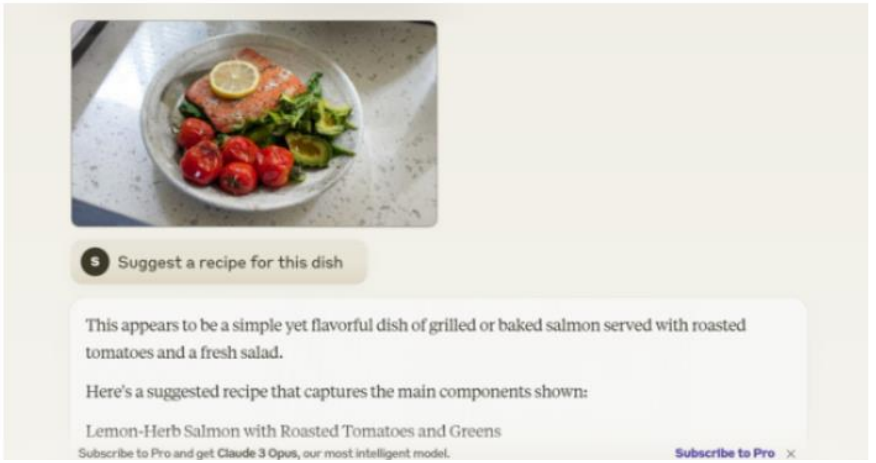
8개월만에 대학원생 지적수준 갖췄다고?…초거대 AI 초고속 업그레이드

입력 2024.03.06. 오전 6:39 기사원문

이상덕 기자 · 이덕주 기자

7 3 가가

앤스로픽 멀티모달 클로드 공개
오픈AI GPT 1년보다 개선 빨라
"AI 산업 2030년 1천조 육박"
경쟁 밀리면 도태 우려에 속도전



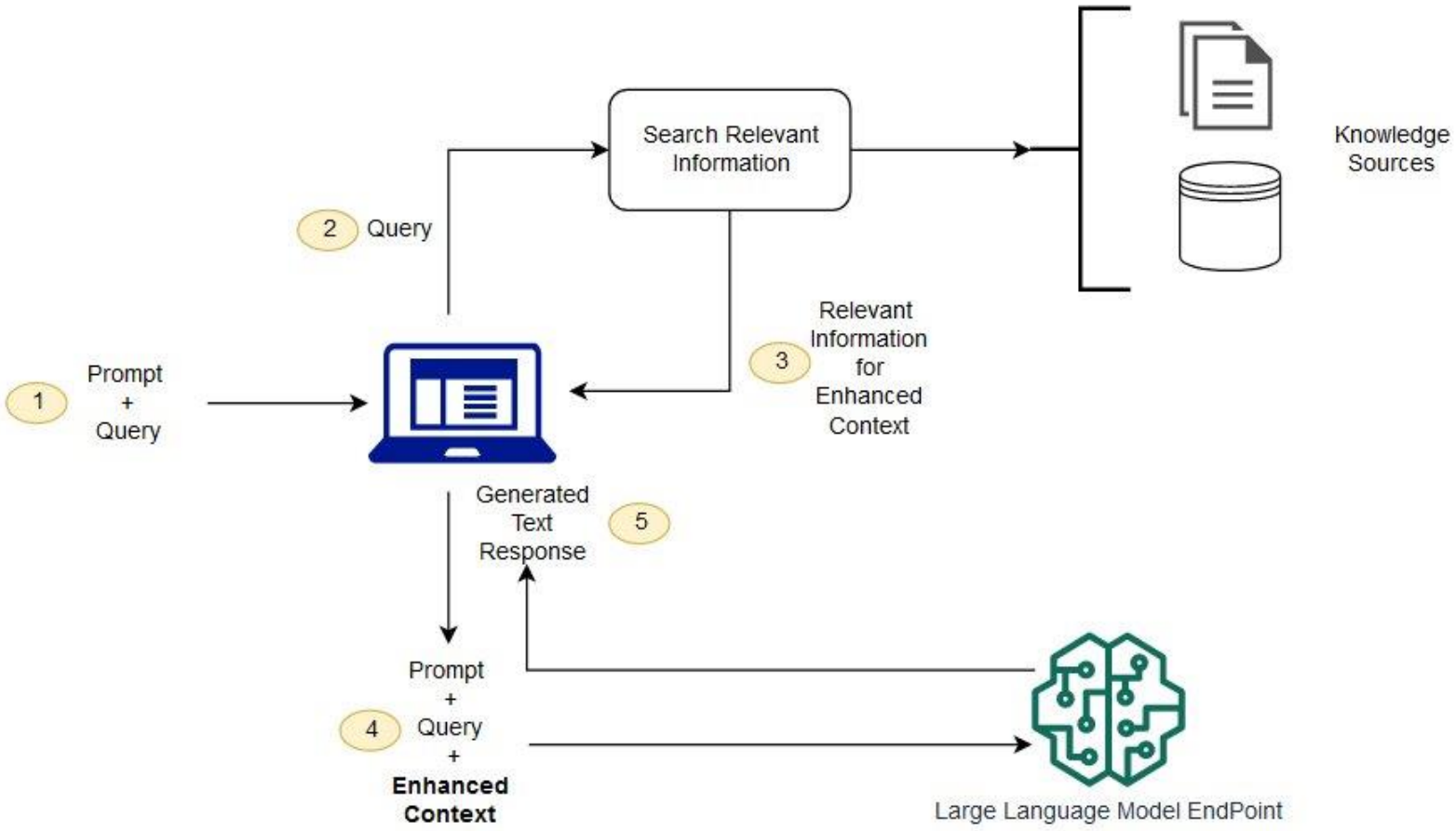
- **효율적인 LLM을 위한 트렌드: sLLM**
  - Small Large Language Model
  - LLM 대비 파라미터 수를 수십억~수백억 개 수준으로 유지
  - Meta의 LLaMA부터 관심
  - 효율적 / On device 활용 / 높은 유연성 / Fine tuning을 적용

  - 예:
    - 알파카7B: 스탠포드대학교, LLaMA의 가장 작은 버전 활용한 소형 언어모형, GPT-3.5에 버금가는 성능이면서도 100만원 내외의 비용과 8대 PC로 3시간 훈련
    - Dolly: 데이터브릭스, 60억개 파라미터, 1대의 서버로 3시간 훈련

- **RAG: 검색 증강 생성**
  - Retrieval-Augmented Generation
  - Hallucination을 줄이는 비용 효율적인 접근 방식
    - LLM의 기능을 특정 도메인이나 특정 기관의 내부 지식 기반으로 확장
    - LLM 외부의 데이터를 활용해서 보다 정확한 답변을 찾는 방안
    - 모형 학습을 다시 할 필요는 없음

- **RAG: 검색 증강 생성**

- LLM 학습 데이터 외의 데이터인 외부 데이터 생성 후 임베딩을 통해 벡터DB저장
- 질문에 대한 관련 정보 검색
- 외부 데이터 업데이트: 실시간 및 배치 방식



- **생성형 AI 평가**
  - 다양한 도메인에서의 지식 평가, 정확성, 강건함, 효율성 등의 입체적 평가 추세
  - 다양한 지표와 다양한 태스크의 사용, 언어 생성 능력, 지식 활용 추론 능력 등 평가

  - 평가 방식
    - Human Evaluation : 관련 전문가 평가
    - AI Evaluation : 자동화된 Metric의 사용
    - 두 평가의 Correlation을 통한 검증

- **생성형 AI 평가**
  - 언어 생성 능력: 문장을 완성 또는 요약
  - 지식 활용 능력: 상식 또는 지식에 근거한 답변 생성 능력 평가
    - closed-book QA : 사전에 학습한 지식만을으로 답변
    - open-book QA: 외부 지식을 활용해 답변
    - 평가 데이터: Natural Questions, ARC, TruthfulQA 등의 데이터셋, Wikipedia 등 외부 소스 활용
  - 추론 능력: 주어진 정보의 이해나 학습된 지식 활용으로 논리적 추론을 통한 답변 평가
    - 일반 상식 추론:PiQA, HellaSwag, WinoGrande 등 활용
    - 수학적 논리 테스트: GSM8K 등 활용

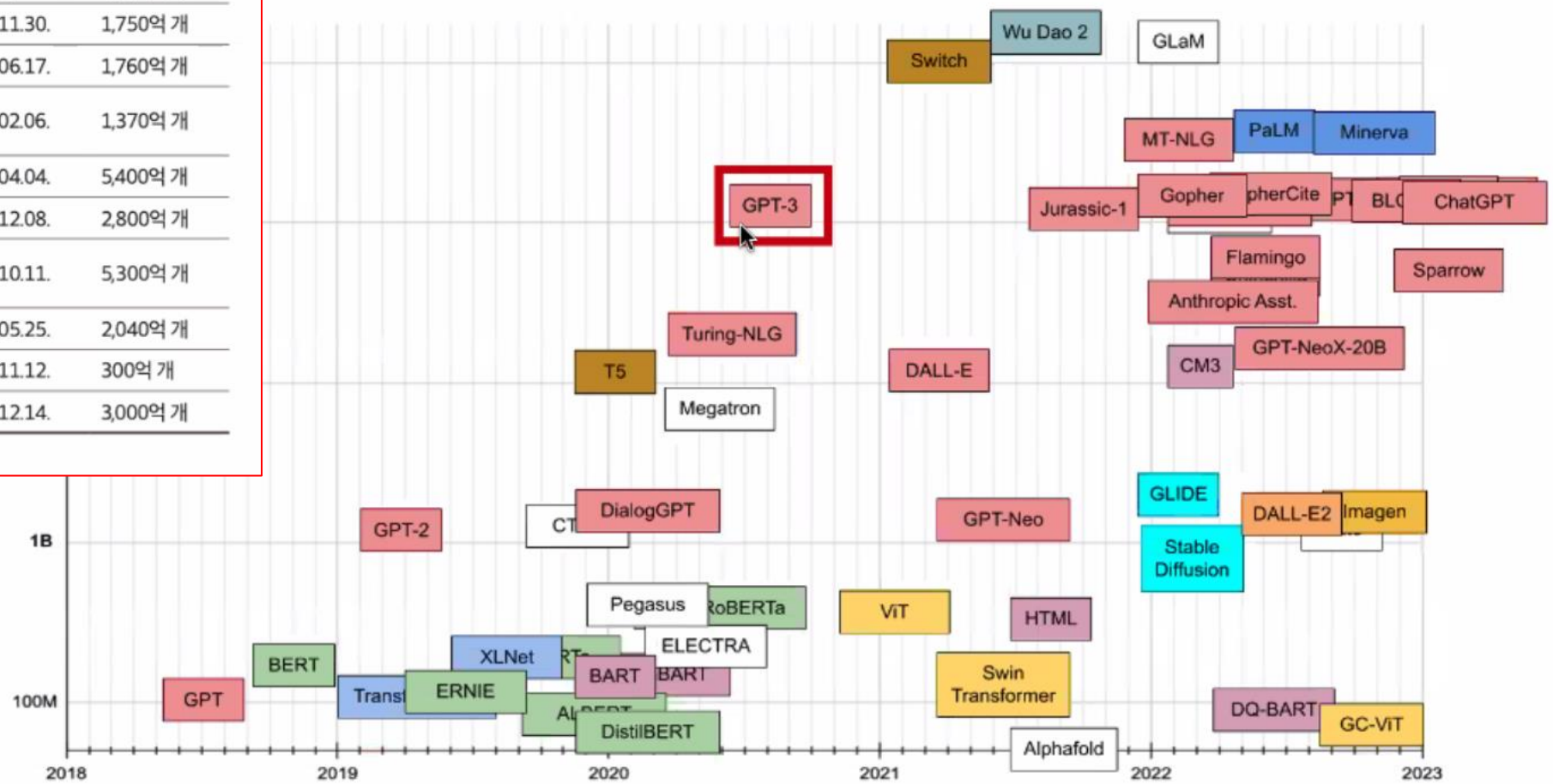
V. AI 전망

- **LLM의 명암**
  - 학습 비용
    - GPT4의 Fine tuning 약 6개월 소요
    - 범용 모형들의 Fine tuning에 2-3개월 소요
  - 학습 데이터의 최신성 유지 이슈
  - 사회적 편견과 환각(Hallucination) 현상
    - *단어 다음에 나올 높은 확률의 단어가 정답을 의미하지 않음*
  - 악의적/부정적 프롬프트에 대한 답변 / 사내 프롬프트 제한을 우회하는 문제의 가능성
  - 보안/ 저작권 등 이슈

- AI 경량화의 필요성

기업		초거대 AI 종류	출시일	파라미터 수
해외	OpenAI	GPT-3.5(챗GPT, 챗지피티)	`22.11.30.	1,750억 개
	BigScience	BLOOM(블룸, 오픈소스)	`22.06.17.	1,760억 개
	Google	Bard (바드, LaMDA(람다)기반)	`23.02.06.	1,370억 개
		PaLM(팜)	`22.04.04.	5,400억 개
		Gopher(고퍼)	`21.12.08.	2,800억 개
	MS, nVidia	Megatron (메가트론, MT NLG)	`21.10.11.	5,300억 개
국내	네이버	HyperClova(하이퍼클로바)	`21.05.25.	2,040억 개
	카카오	KoGPT(코지피티)	`21.11.12.	300억 개
	LG	Exaone(엑사원)	`21.12.14.	3,000억 개

자료 : 각 사(社) 및 언론 자료 종합(SPRi 2023)



V. AI 전망

- 지식 증류(Knowledge distillation)
- 가지치기 (Pruning)
- 양자화(Quantization)

AI is the new electricity!

Industrial Data Science Lab

Contact:
won.sang.l@gmail.com

https://sites.google.com/view/idslab