

Generative AI

1. 생성형 AI 개요





인공지능(Artificial Intelligent)



인공지능(AI)

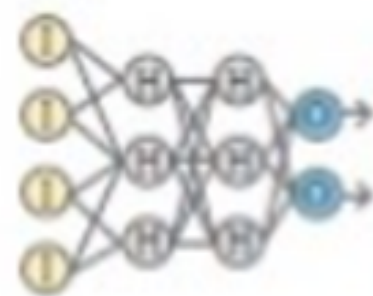
사람의 지적능력(추론, 인지)을 구현하고 모방하는 모든 기술



머신러닝

명시적인 프로그래밍 없이 학습할 수 있는 능력

선형회귀
로지스틱회귀
K-최근접 이웃
결정트리
랜덤포레스트
서포트 벡터 머신
클러스터링
차원축소



딥러닝

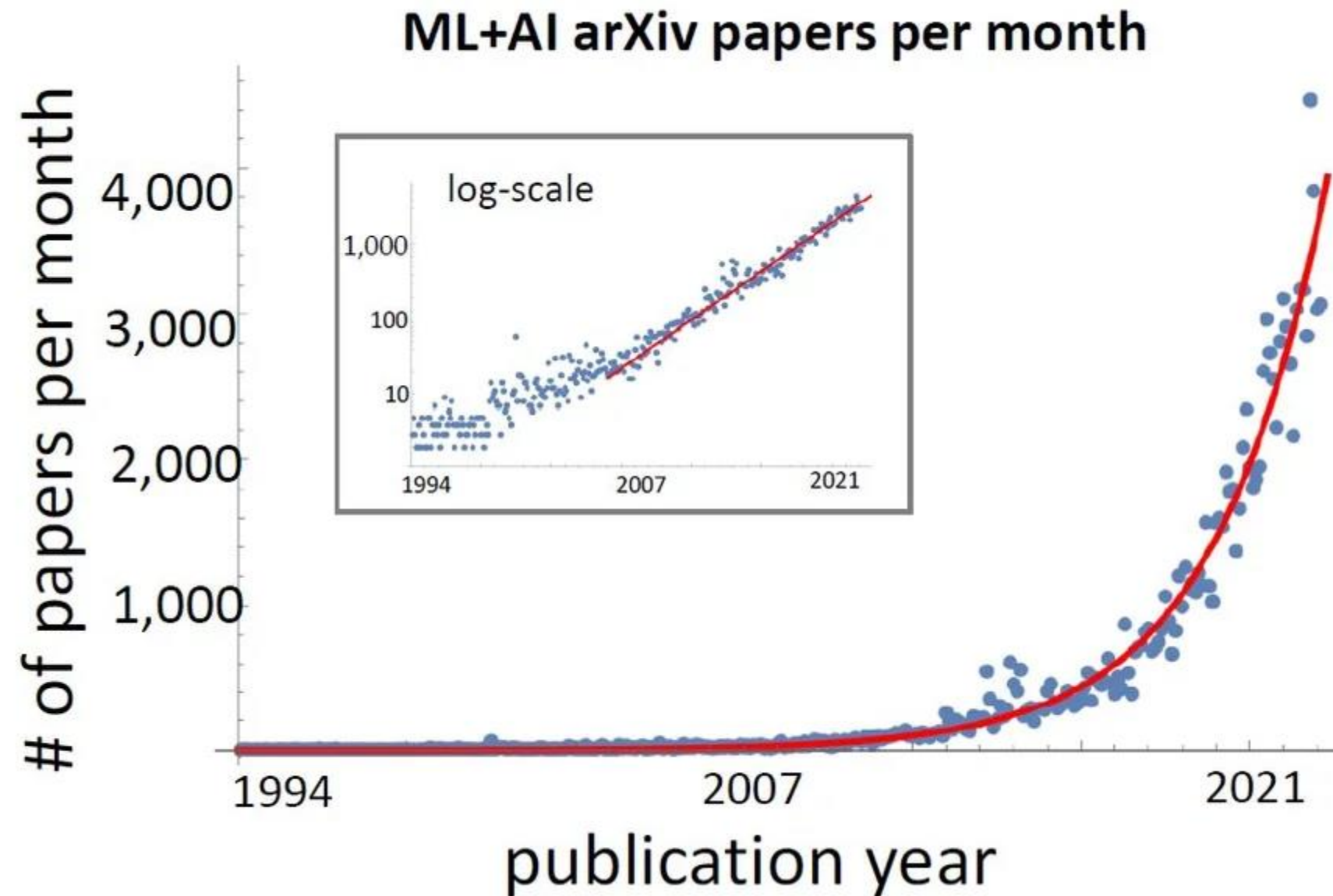
인공신경망 이용해 데이터에서 패턴을 찾아내는 기술

심층신경망(DNN)
합성곱신경망(CNN)
순환 신경망(RNN)
생성적 적대 신경망(GAN)
강화학습(RL)

트랜스포머

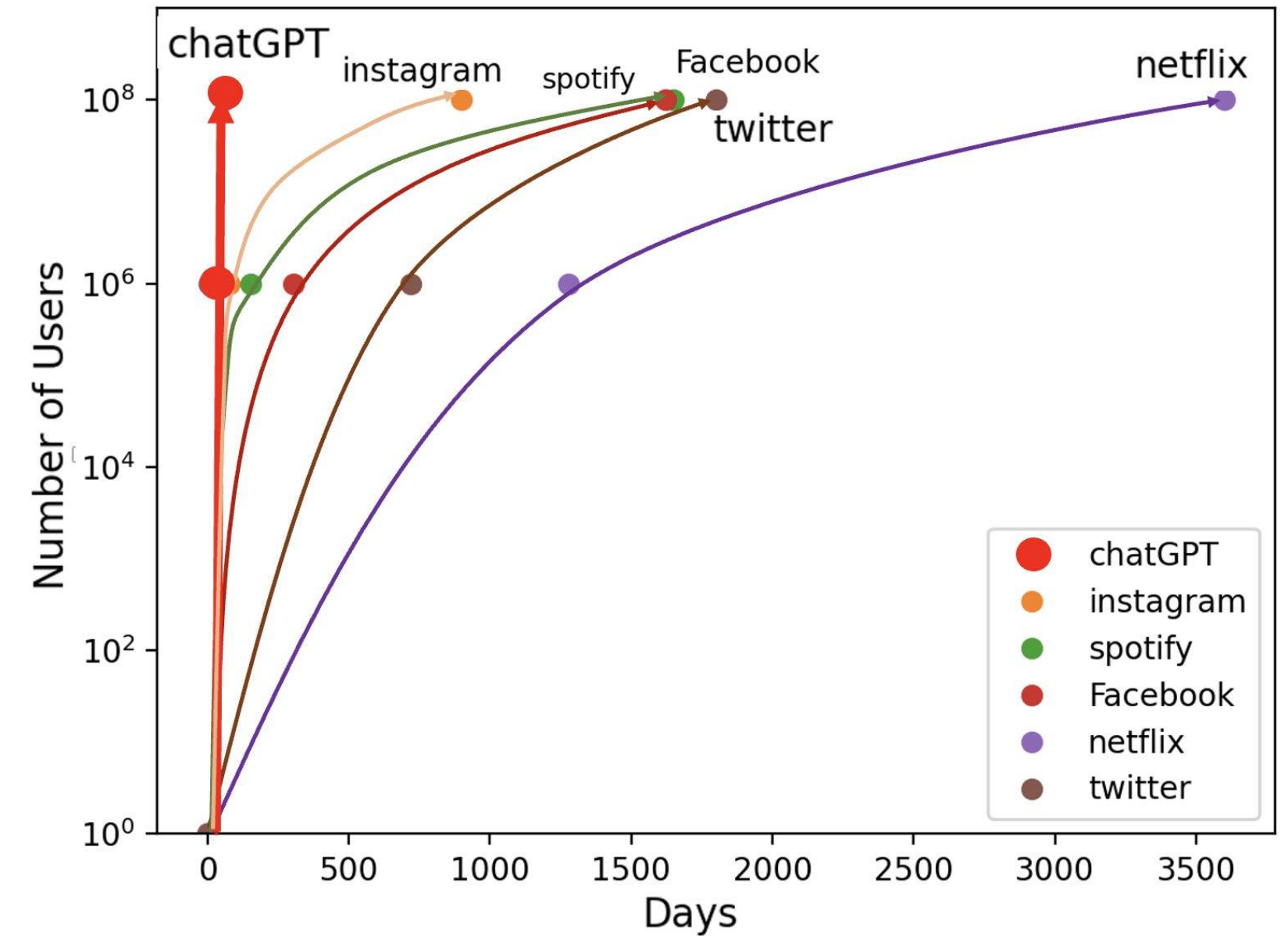
BERT
GPT

AI 기술의 발전속도



https://www.reddit.com/r/singularity/comments/xwdzr5/the_number_of_ai_papers_on_arxiv_per_month_grows/

Something different is happening.

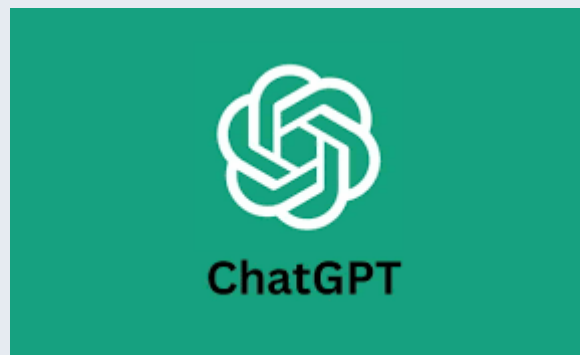


<https://twitter.com/kylelf/status/1623679176246185985?t=g9wnm52DZEfe42CJAjooRA&s=03>

Generative AI

생성형 AI는 인공지능망을 이용하여 새로운 데이터를 생성해내는 기술로 명령어(Prompt)를 통해 사용자의 의도를 스스로 이해하고, 주어진 데이터로 학습, 활용하여 텍스트, 이미지, 오디오, 비디오 등 새로운 콘텐츠를 생성해내는 인공지능입니다.

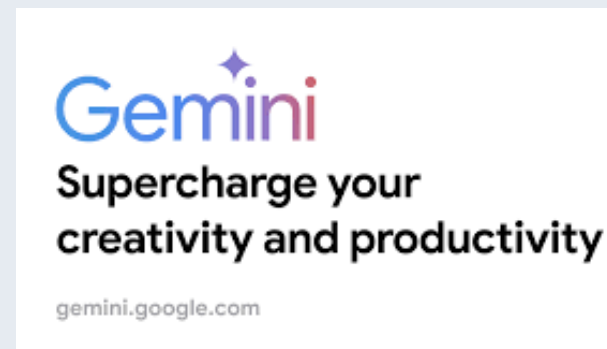
ChatGPT



Copilot



Gemini



Stable Diffusion



Midjourney



<https://chat.openai.com/>

<https://www.midjourney.com/>

<https://copilot.microsoft.com/>

<https://stablediffusionweb.com/>

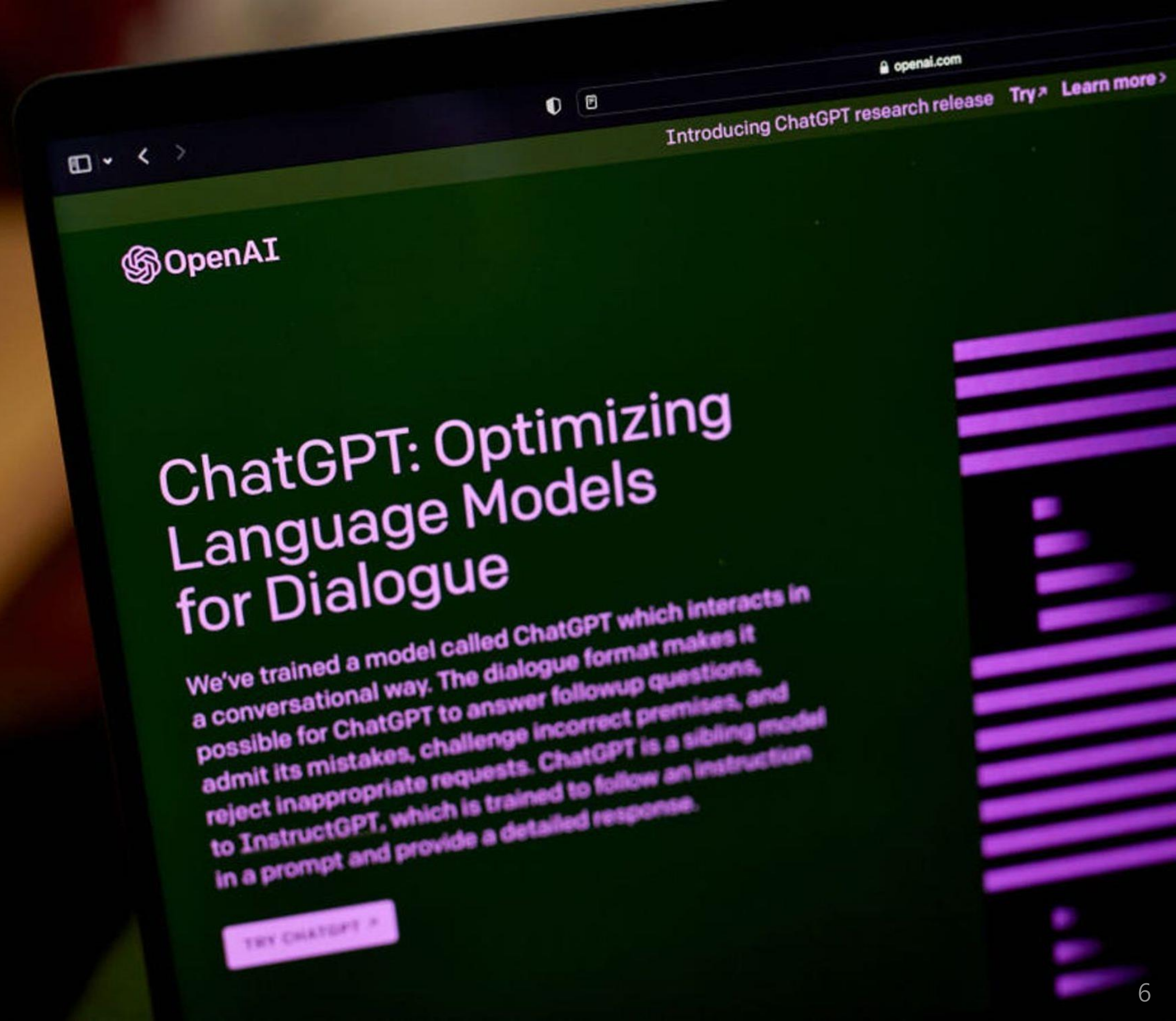
<https://gemini.google.com/>

ChatGPT

<https://chat.openai.com/>



GPT = Generative
Pre-trained
Transformer



ChatGPT

<https://chat.openai.com/>

텍스트 이해
질문에 대한 답변
콘텐츠 생성
리스트 생성
코드 작성/디버깅
장단점 비교 제시
교육 지원
창의적인 글쓰기
아이디어 기획
번역
단계별 지침 제공
가상 비서 역할

The screenshot shows the ChatGPT web interface. On the left is a dark sidebar with the following links: 'Reset Thread', 'Dark Mode', 'OpenAI Discord', 'Learn More', and 'Log out'. The main content area has a white background with the 'ChatGPT' title at the top. Below the title are three columns: 'Examples' (with a sun icon), 'Capabilities' (with a lightning bolt icon), and 'Limitations' (with a warning triangle icon). Each column contains three text boxes. At the bottom of the main area is a large text input field with a cursor and a send button. Below the input field is a small disclaimer: 'Free Research Preview: ChatGPT is optimized for dialogue. Our goal is to make AI systems more natural to interact with, and your feedback will help us improve our systems and make them safer.'

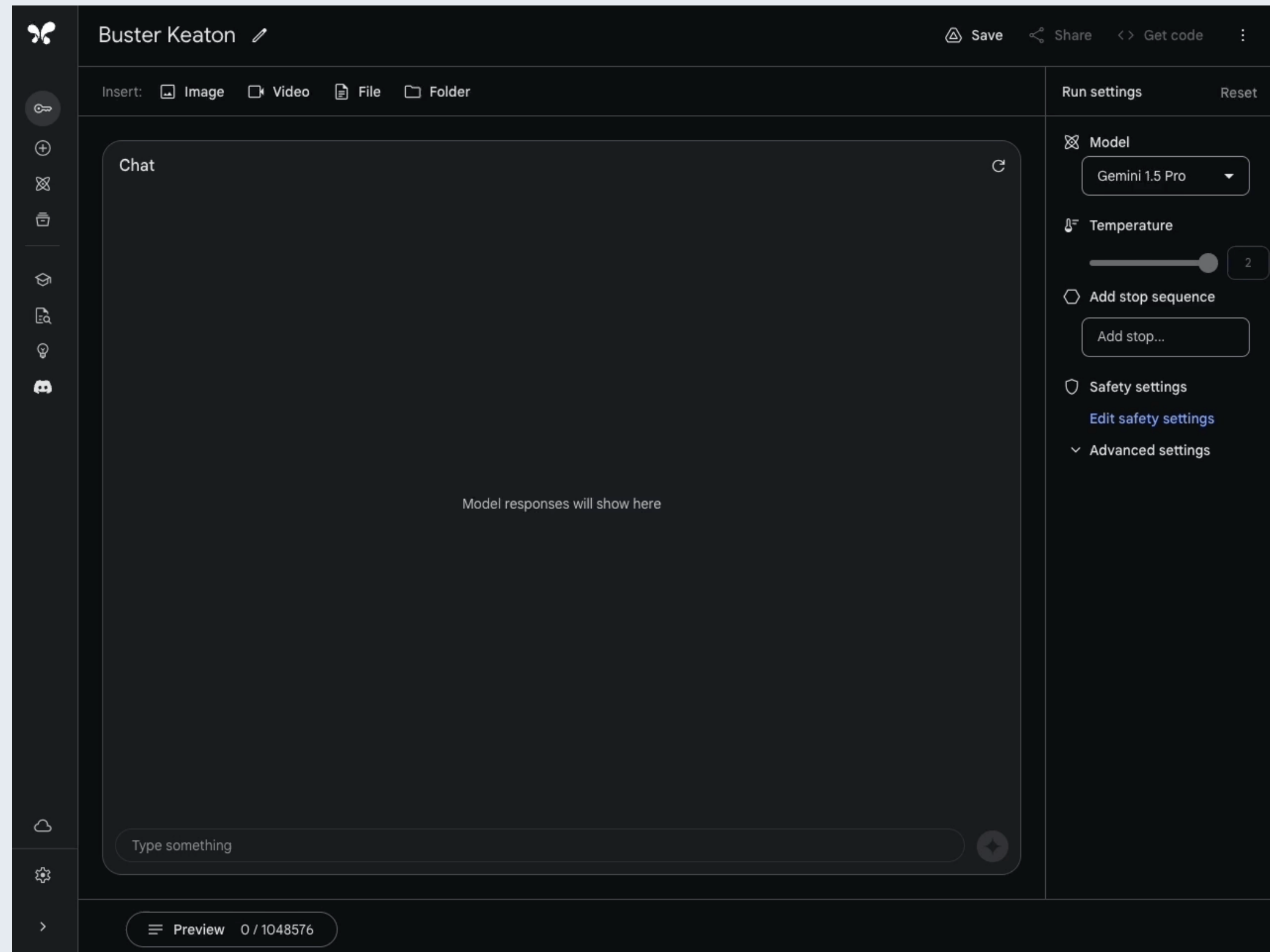
Examples	Capabilities	Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Gemini

<https://gemini.google.com/>

- 멀티모달 추론 기능을 갖춘 대화형 생성 인공지능
- 전문가 혼합(MoE, Mixture-of-Experts) 아키텍처를 통해 효율을 높임
- 70만 개 이상의 단어, 3만 줄의 코드, 1시간 분량의 동영상, 11시간 분량의 음성에 해당하는 방대한 양의 정보를 한 번에 처리

구글코리아 블로그



제미나이(Gemini) : <https://blog.google/intl/ko-kr/products/explore-get-answers/google-gemini-next-generation-model-february-2024-kr/>
젬마(Gemma) : <https://blog.google/intl/ko-kr/products/explore-get-answers/-gemma-open-models-kr/>

Gemini



Sundar Pichai

CEO, Google and Alphabet

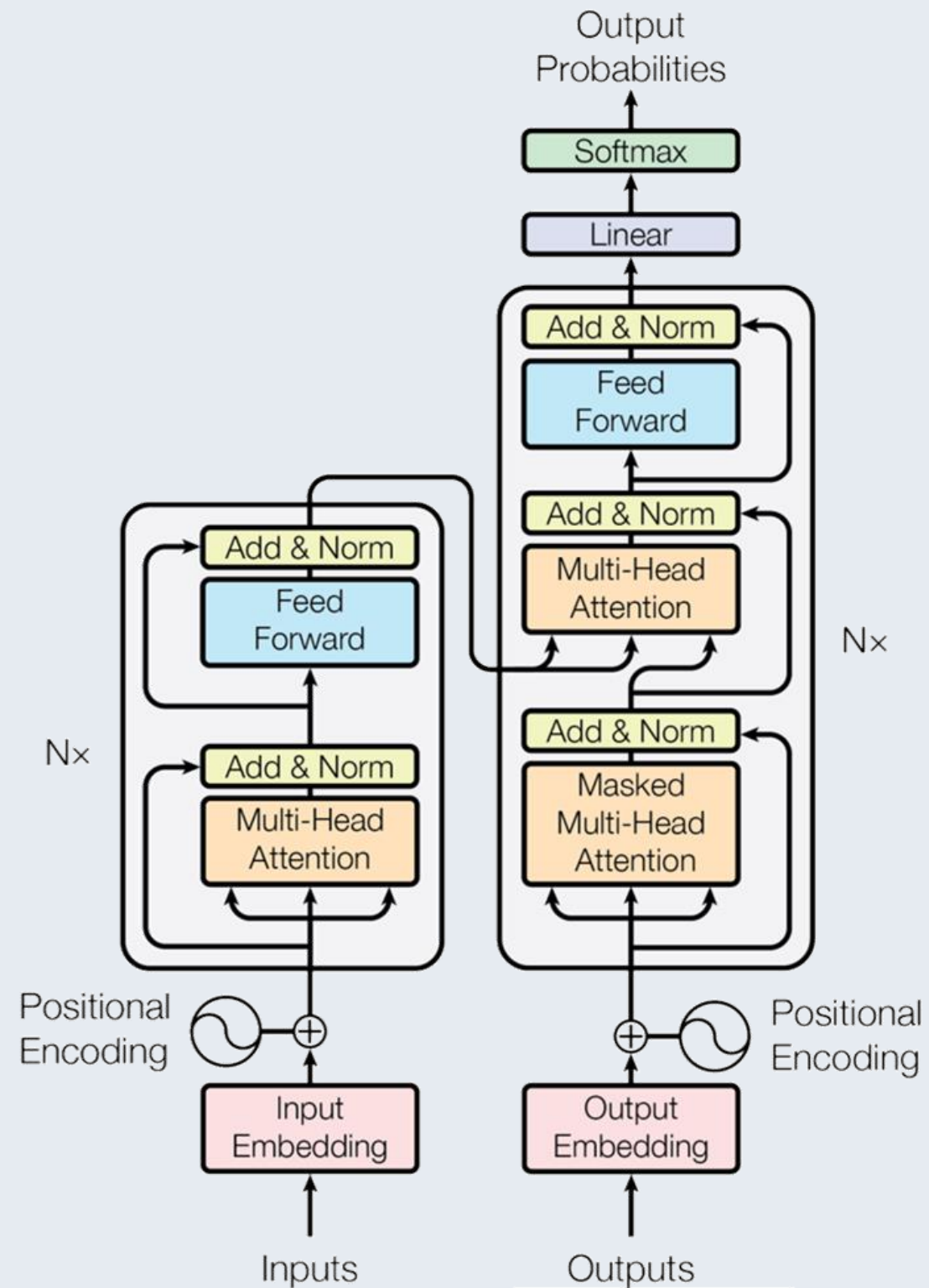
처음부터 AI에 관심을 갖게 된
이유 중 하나는

LLM (Large Language Model)

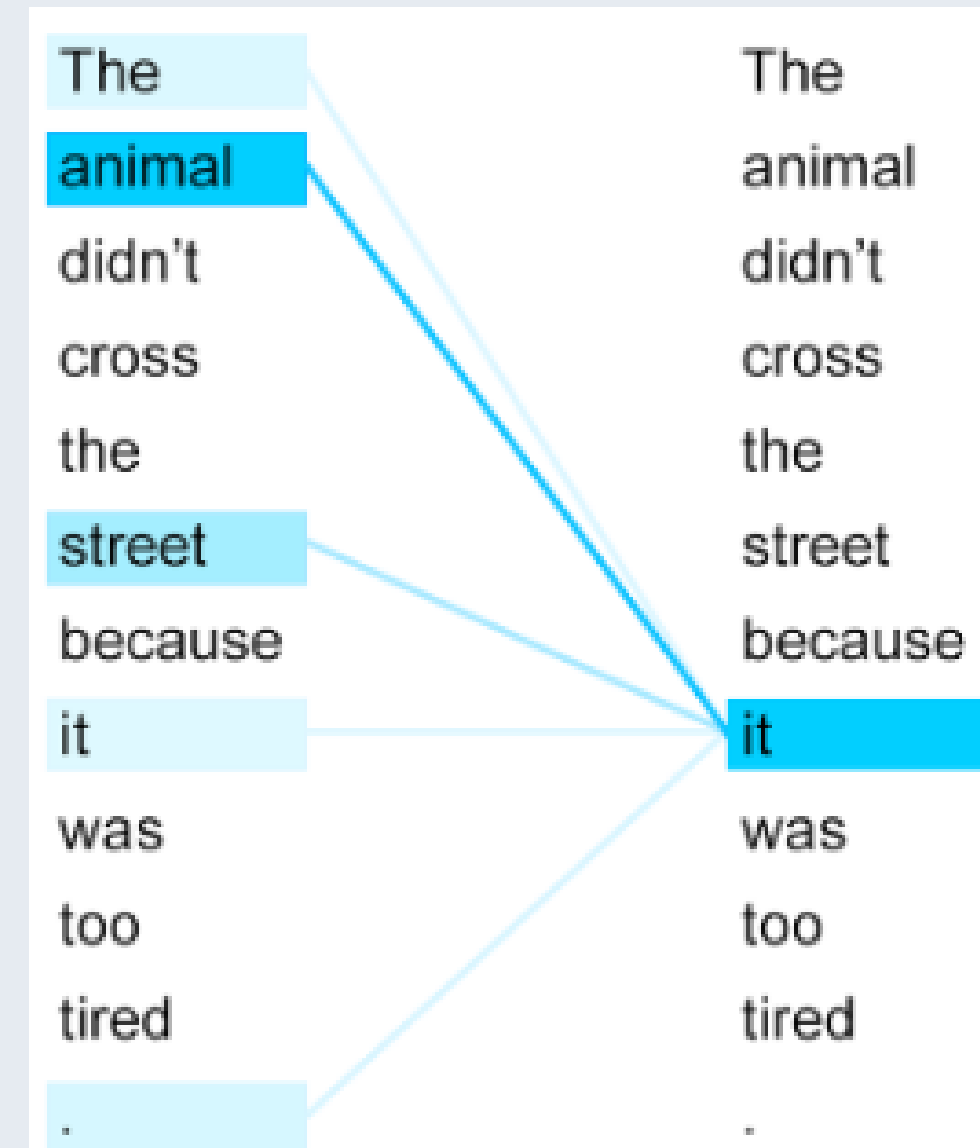


트랜스포머

Transformer 아키텍처

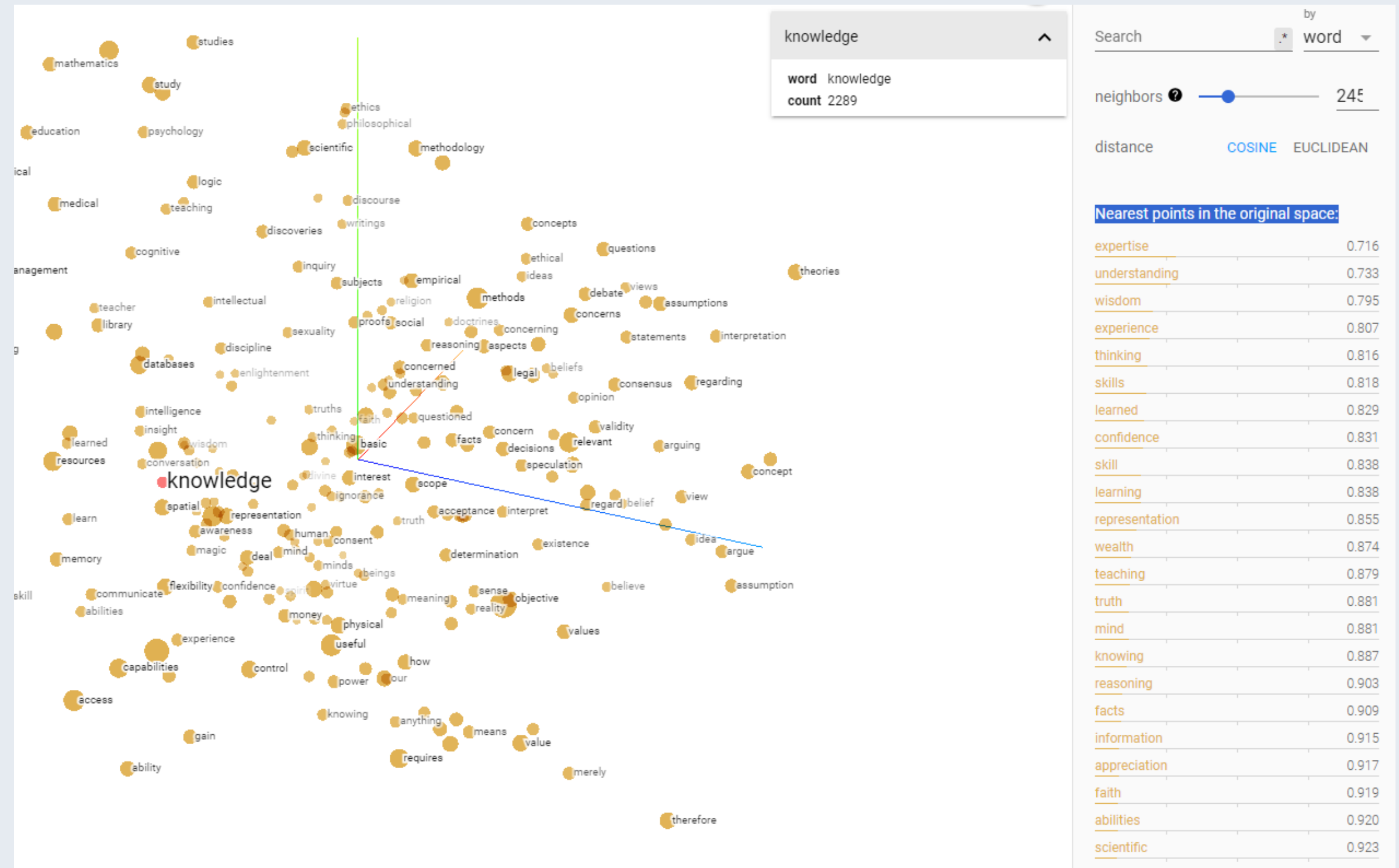
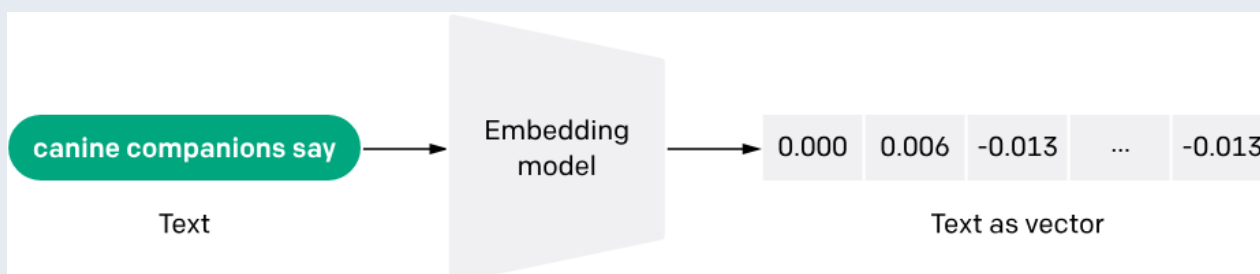


Self Attention



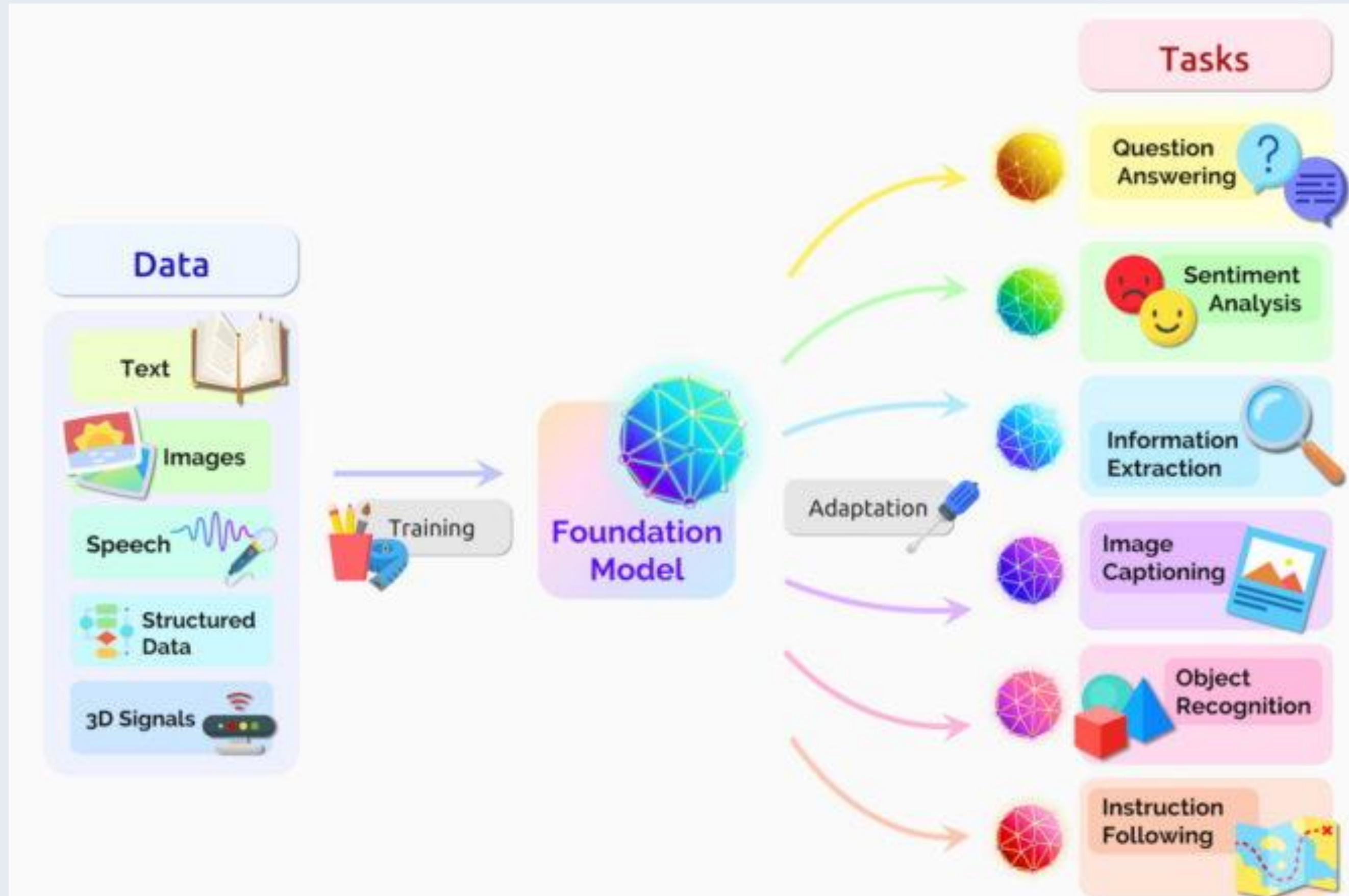
임베딩 (Embedding)

$$\text{banking} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$



<https://projector.tensorflow.org/>

파운데이션 모델 (FM, Foundation Model)



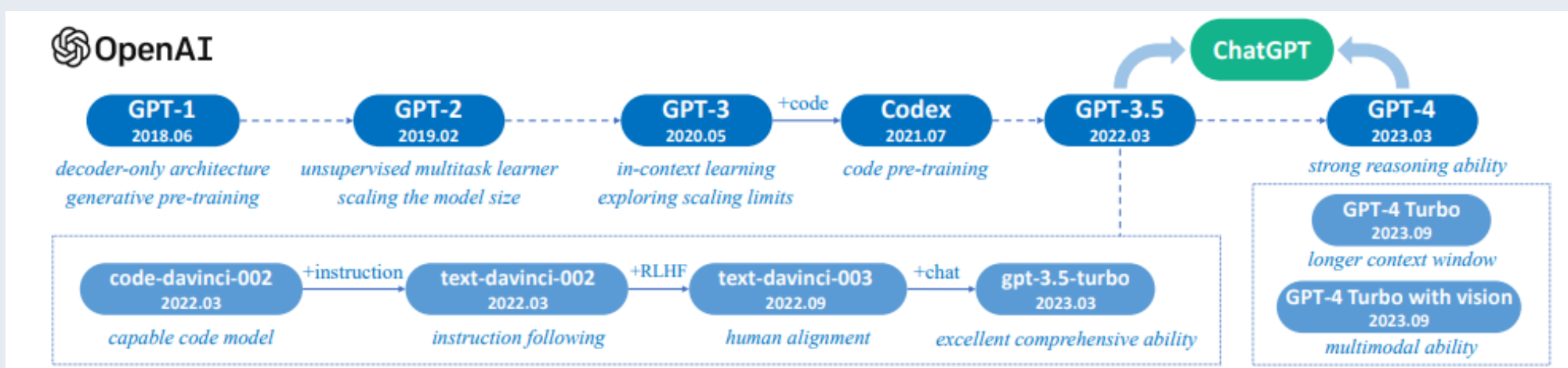
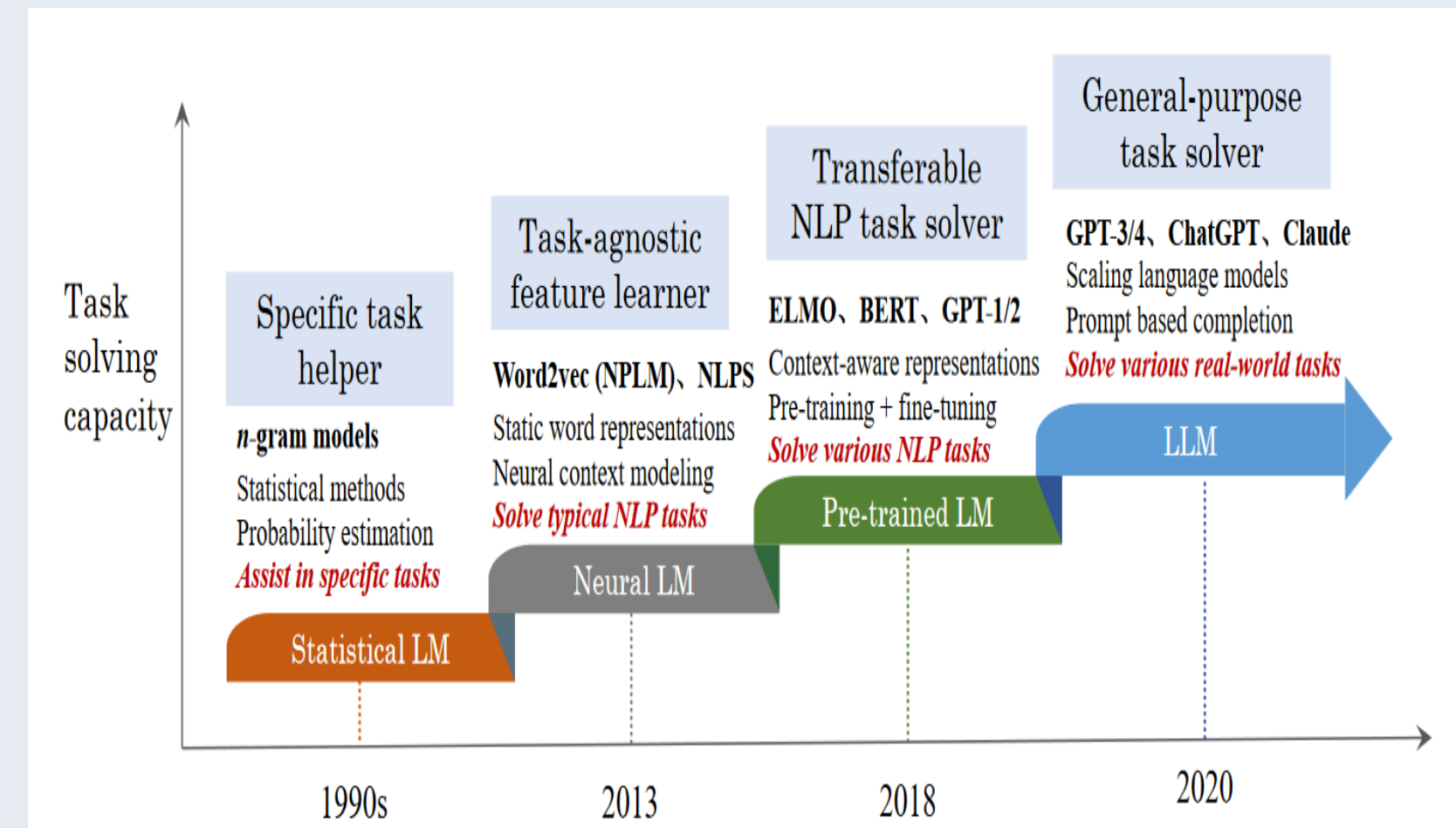
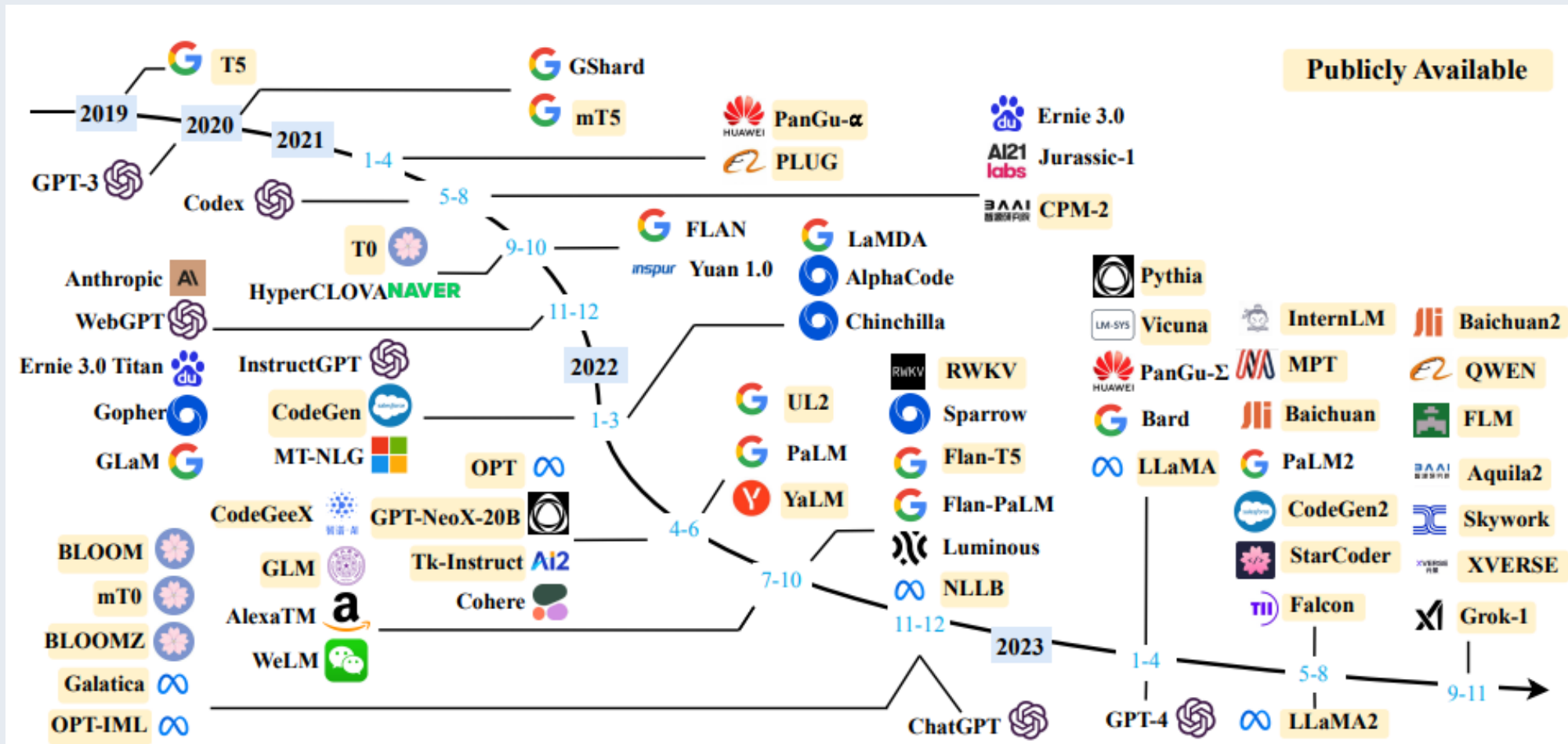
LLM Model

Model	Release Time	Size (B)	Base Model	Adaptation IT RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL CoT
T5 [82]	Oct-2019	11	-	- -	1T tokens	Apr-2019	1024 TPU v3	-	✓ -
mT5 [83]	Oct-2020	13	-	- -	1T tokens	-	-	-	✓ -
PanGu- α [84]	Apr-2021	13*	-	- -	1.1TB	-	2048 Ascend 910	-	✓ -
CPM-2 [85]	Jun-2021	198	-	- -	2.6TB	-	-	-	- -
T0 [28]	Oct-2021	11	T5	✓ -	-	-	512 TPU v3	27 h	✓ -
CodeGen [86]	Mar-2022	16	-	- -	577B tokens	-	-	-	✓ -
GPT-NeoX-20B [87]	Apr-2022	20	-	- -	825GB	-	96 40G A100	-	✓ -
Tk-Instruct [88]	Apr-2022	11	T5	✓ -	-	-	256 TPU v3	4 h	✓ -
UL2 [89]	May-2022	20	-	- -	1T tokens	Apr-2019	512 TPU v4	-	✓ ✓
OPT [90]	May-2022	175	-	- -	180B tokens	-	992 80G A100	-	✓ -
NLLB [91]	Jul-2022	54.5	-	- -	-	-	-	-	✓ -
CodeGeeX [92]	Sep-2022	13	-	- -	850B tokens	-	1536 Ascend 910	60 d	✓ -
GLM [93]	Oct-2022	130	-	- -	400B tokens	-	768 40G A100	60 d	✓ -
Flan-T5 [69]	Oct-2022	11	T5	✓ -	-	-	-	-	✓ ✓
BLOOM [78]	Nov-2022	176	-	- -	366B tokens	-	384 80G A100	105 d	✓ -
mT0 [94]	Nov-2022	13	mT5	✓ -	-	-	-	-	✓ -
Galactica [35]	Nov-2022	120	-	- -	106B tokens	-	-	-	✓ ✓
BLOOMZ [94]	Nov-2022	176	BLOOM	✓ -	-	-	-	-	✓ -
OPT-IML [95]	Dec-2022	175	OPT	✓ -	-	-	128 40G A100	-	✓ ✓
LLaMA [57]	Feb-2023	65	-	- -	1.4T tokens	-	2048 80G A100	21 d	✓ -
Pythia [96]	Apr-2023	12	-	- -	300B tokens	-	256 40G A100	-	✓ -
CodeGen2 [97]	May-2023	16	-	- -	400B tokens	-	-	-	✓ -
StarCoder [98]	May-2023	15.5	-	- -	1T tokens	-	512 40G A100	-	✓ ✓
LLaMA2 [99]	Jul-2023	70	-	✓ ✓	2T tokens	-	2000 80G A100	-	✓ -
Baichuan2 [100]	Sep-2023	13	-	✓ ✓	2.6T tokens	-	1024 A800	-	✓ -
QWEN [101]	Sep-2023	14	-	✓ ✓	3T tokens	-	-	-	✓ -
FLM [102]	Sep-2023	101	-	✓ -	311B tokens	-	192 A800	22 d	✓ -
Skywork [103]	Oct-2023	13	-	- -	3.2T tokens	-	512 80G A800	-	✓ -

Model	Release Time	Size (B)	Base Model	Adaptation IT RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL CoT
GPT-3 [55]	May-2020	175	-	- -	300B tokens	-	-	-	✓ -
GShard [104]	Jun-2020	600	-	- -	1T tokens	-	2048 TPU v3	4 d	- -
Codex [105]	Jul-2021	12	GPT-3	- -	100B tokens	May-2020	-	-	✓ -
ERNIE 3.0 [106]	Jul-2021	10	-	- -	375B tokens	-	384 V100	-	✓ -
Jurassic-1 [107]	Aug-2021	178	-	- -	300B tokens	-	800 GPU	-	✓ -
HyperCLOVA [108]	Sep-2021	82	-	- -	300B tokens	-	1024 A100	13.4 d	✓ -
FLAN [67]	Sep-2021	137	LaMDA-PT	✓ -	-	-	128 TPU v3	60 h	✓ -
Yuan 1.0 [109]	Oct-2021	245	-	- -	180B tokens	-	2128 GPU	-	✓ -
Anthropic [110]	Dec-2021	52	-	- -	400B tokens	-	-	-	✓ -
WebGPT [81]	Dec-2021	175	GPT-3	- ✓	-	-	-	-	✓ -
Gopher [64]	Dec-2021	280	-	- -	300B tokens	-	4096 TPU v3	920 h	✓ -
ERNIE 3.0 Titan [111]	Dec-2021	260	-	- -	-	-	-	-	✓ -
GLaM [112]	Dec-2021	1200	-	- -	280B tokens	-	1024 TPU v4	574 h	✓ -
LaMDA [68]	Jan-2022	137	-	- -	768B tokens	-	1024 TPU v3	57.7 d	- -
MT-NLG [113]	Jan-2022	530	-	- -	270B tokens	-	4480 80G A100	-	✓ -
AlphaCode [114]	Feb-2022	41	-	- -	967B tokens	Jul-2021	-	-	- -
InstructGPT [66]	Mar-2022	175	GPT-3	✓ ✓	-	-	-	-	✓ -
Chinchilla [34]	Mar-2022	70	-	- -	1.4T tokens	-	-	-	✓ -
PaLM [56]	Apr-2022	540	-	- -	780B tokens	-	6144 TPU v4	-	✓ ✓
AlexaTM [115]	Aug-2022	20	-	- -	1.3T tokens	-	128 A100	120 d	✓ ✓
Sparrow [116]	Sep-2022	70	-	- ✓	-	-	64 TPU v3	-	✓ -
WeLM [117]	Sep-2022	10	-	- -	300B tokens	-	128 A100 40G	24 d	✓ -
U-PaLM [118]	Oct-2022	540	PaLM	- -	-	-	512 TPU v4	5 d	✓ ✓
Flan-PaLM [69]	Oct-2022	540	PaLM	✓ -	-	-	512 TPU v4	37 h	✓ ✓
Flan-U-PaLM [69]	Oct-2022	540	U-PaLM	✓ -	-	-	-	-	✓ ✓
GPT-4 [46]	Mar-2023	-	-	✓ ✓	-	-	-	-	✓ ✓
PanGu- Σ [119]	Mar-2023	1085	PanGu- α	- -	329B tokens	-	512 Ascend 910	100 d	✓ -
PaLM2 [120]	May-2023	16	-	✓ -	100B tokens	-	-	-	✓ ✓

LLM Model

A Survey of Large Language Models : <https://arxiv.org/pdf/2303.18223.pdf> , 번역자료 : <https://wikidocs.net/222912>



생성형 AI 유스케이스

Code Generation, Documentation, and Quality Assurance

- [Code Snippets AI](#)
- [ChatGPT](#)
- [Google Gemini](#)
- [Tabnine](#)

Product Development and Management

- [Viable Generative Analysis](#)
- [Stability AI](#)
- [AI21 Labs](#)
- [GPT-4](#)

Blog and Social Media Content Writing

- [Jasper](#)
- [Notion AI](#)
- [Phrasee](#)
- [HubSpot Content Assistant](#)

Inbound and Outbound Marketing and Sales

- [Twain](#)
- [Salesforce Einstein GPT](#)
- [HubSpot AI](#)

Project Management and Operations

- [Wrike](#)
- [ClickUp](#)
- [monday.com](#)
- [Notion](#)

Graphic Design and Video Marketing

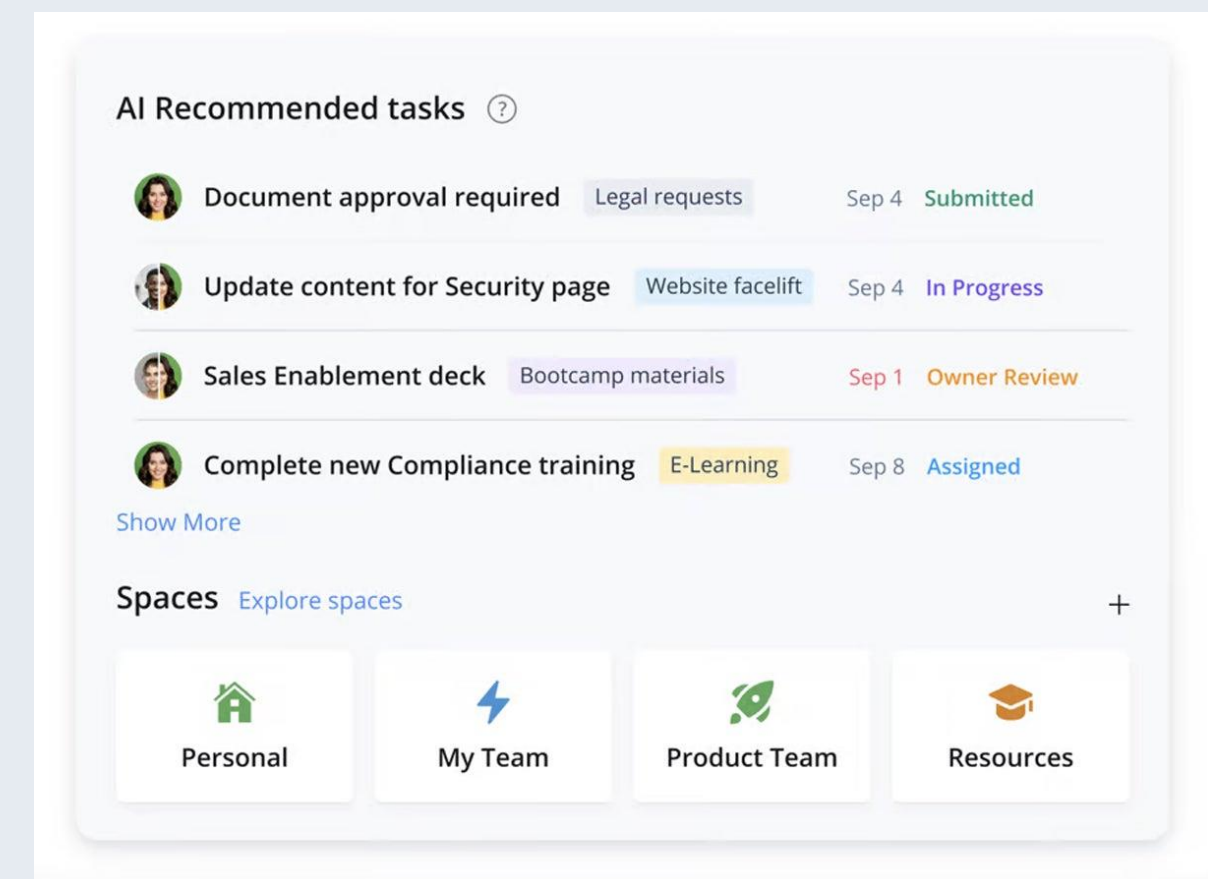
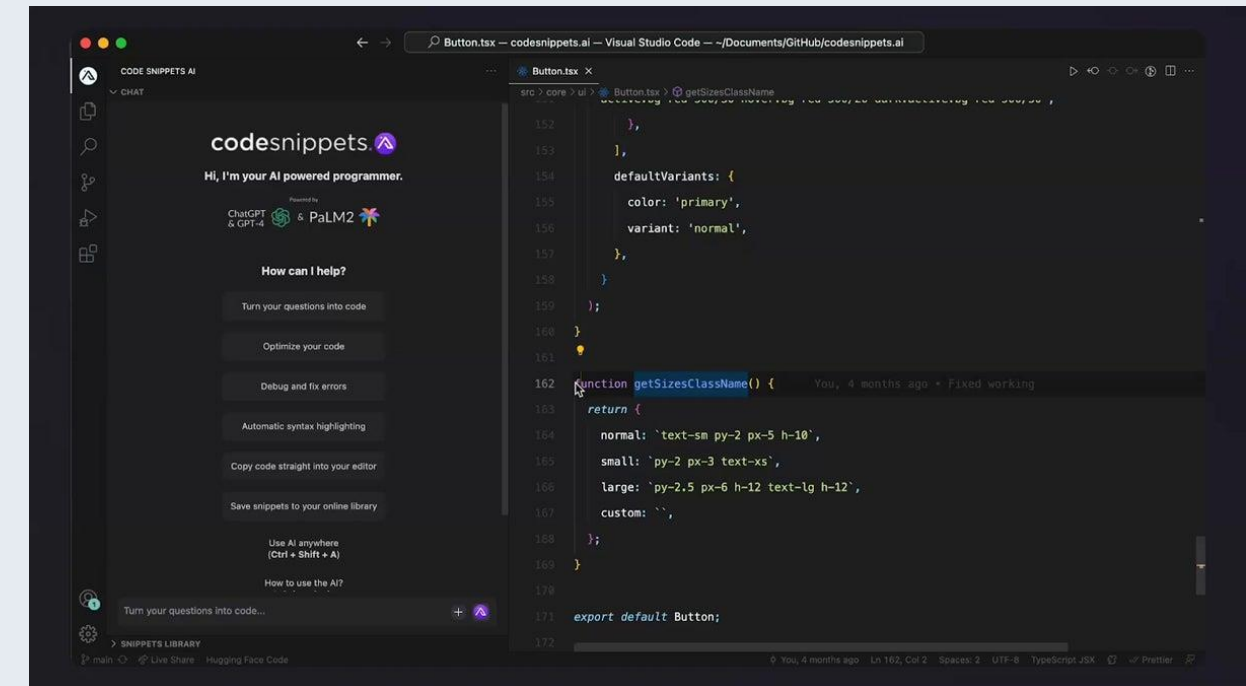
- [Diagram](#)
- [Synthesia](#)
- [Lightricks](#)
- [Rephrase.ai](#)

Entertainment Media Generation

- [Stability AI's Stable Diffusion](#)
- [Plask](#)
- [Charisma](#)
- [Latitude Unchained](#)

Performance Management and Coaching

- [Anthropic Claude](#)
- [Gong](#)
- [CoachHub AIMY](#)



생성형 AI 유스케이스

Business Performance Reporting and Data Analytics

- [SparkBeyond Discovery](#)
- [Dremio](#)
- [Narrative BI](#)
- [Copilot for Power BI](#)

Customer Support and Customer Service

- [Gridspace](#)
- [IBM watsonx Assistant](#)
- [UltimateGPT](#)
- [Zendesk Advanced AI](#)
- [Forethought SupportGPT](#)

Medical Diagnostics and Pharmaceutical Drug Discovery

- [Paige.ai](#)
- [Google Med-PaLM 2](#)
- [Insilico Medicine](#)
- [Iambic Therapeutics](#)

Consumer-Friendly Synthetic Data Generation

- [Syntho Engine](#)
- [Synthesis AI](#)
- [MOSTLY AI](#)
- [Infinity AI](#)

Smart Manufacturing and Predictive Maintenance

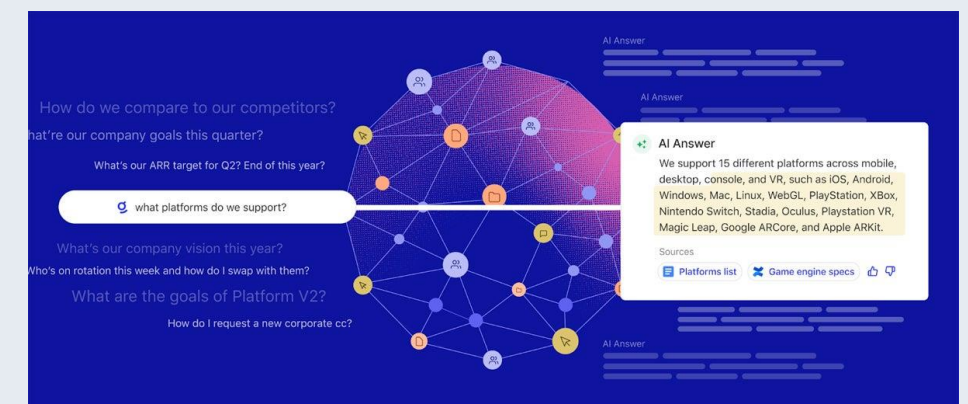
- [Tulip Frontline Copilot](#)
- [Clarifai](#)
- [C3 Generative AI Product Suite](#)

Fraud Detection and Risk Management

- [Simplifai InsuranceGPT](#)
- [Docugami](#)
- [ChatGPT](#)

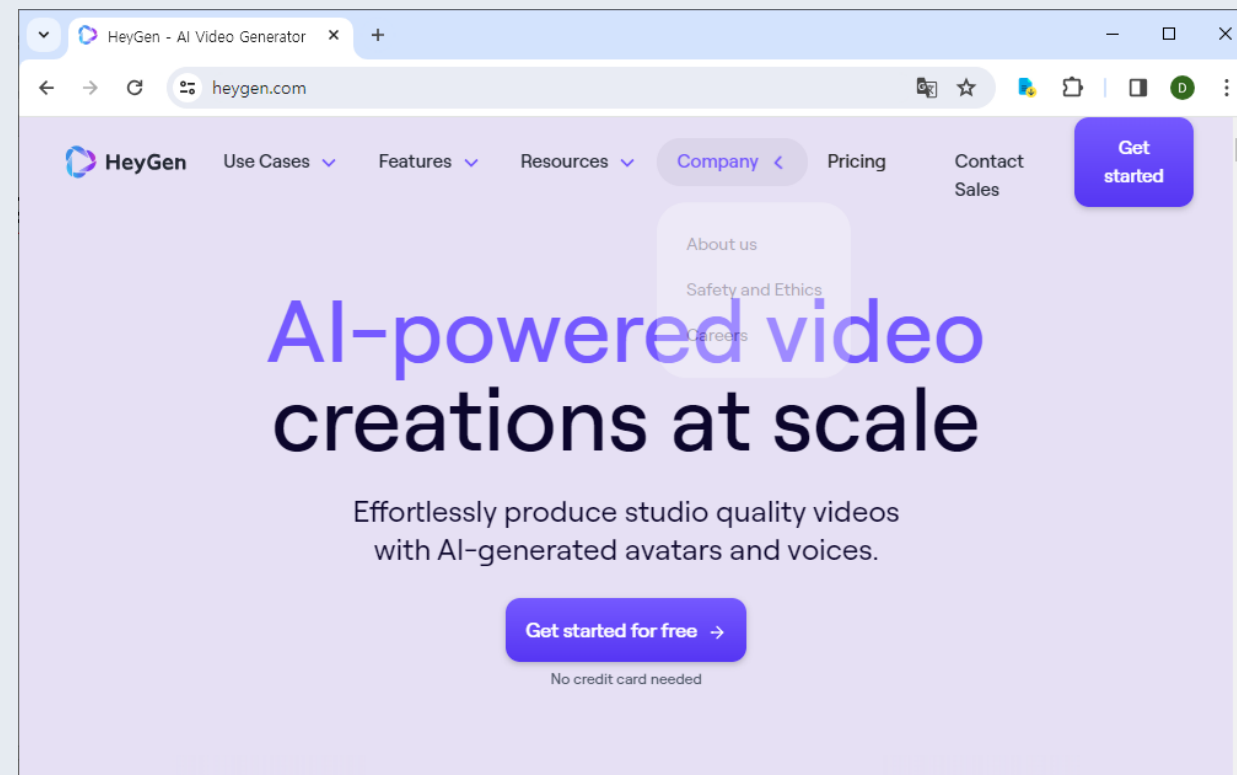
Optimized Enterprise Search and Knowledge Base

- [Glean](#)
- [Coveo Relevance Generative Answering](#)
- [Elasticsearch Relevance Engine](#)

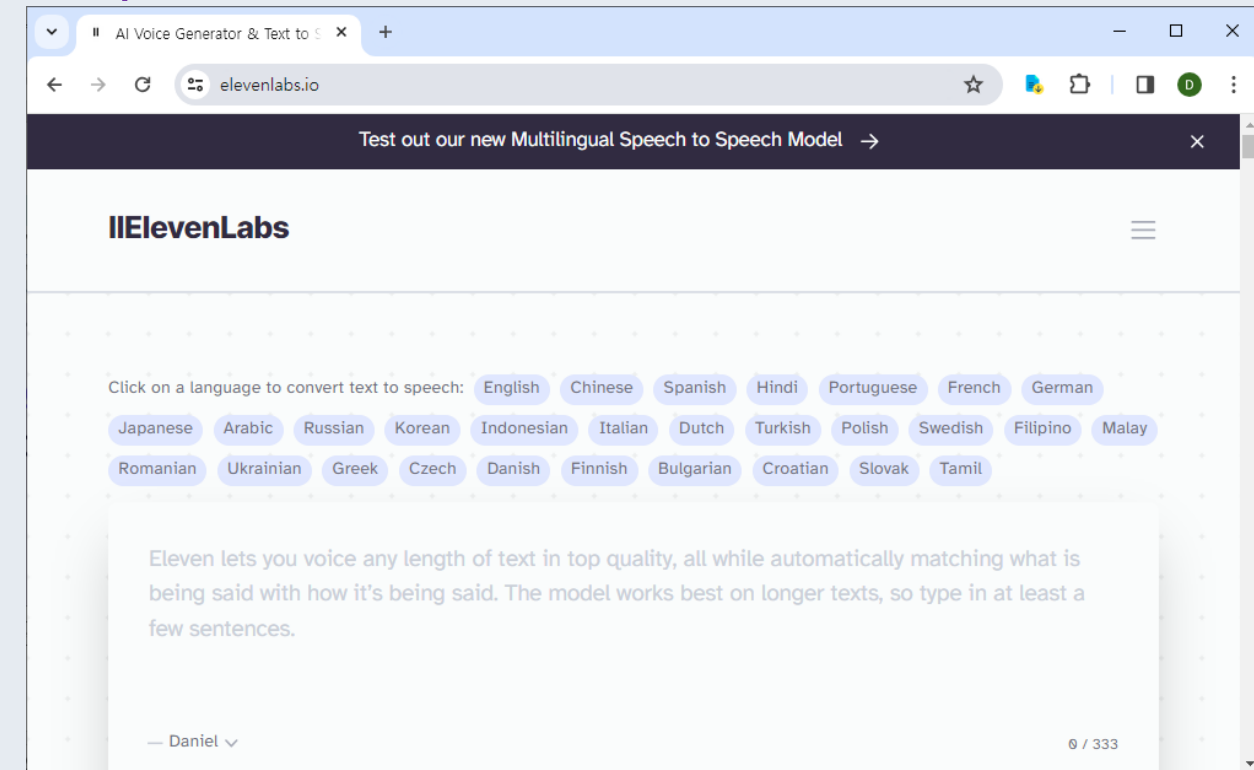


생성형 AI 유스케이스

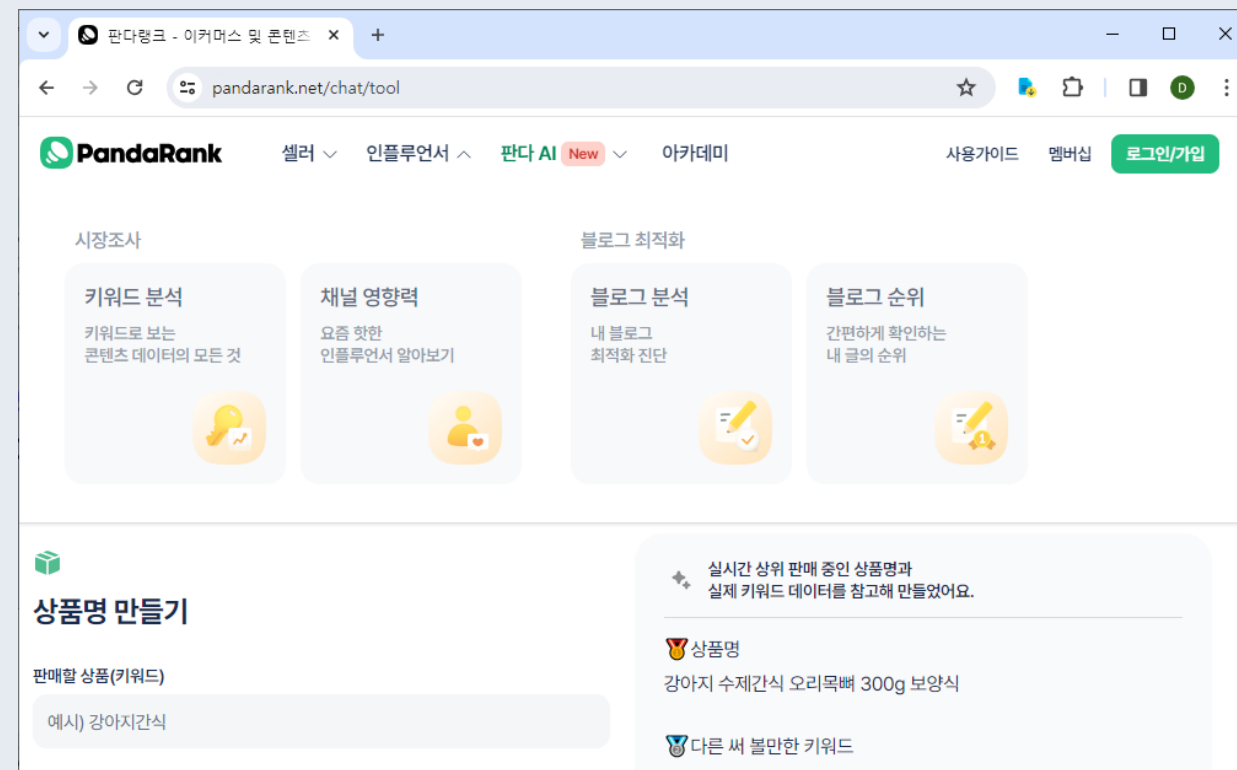
<https://www.heygen.com/>



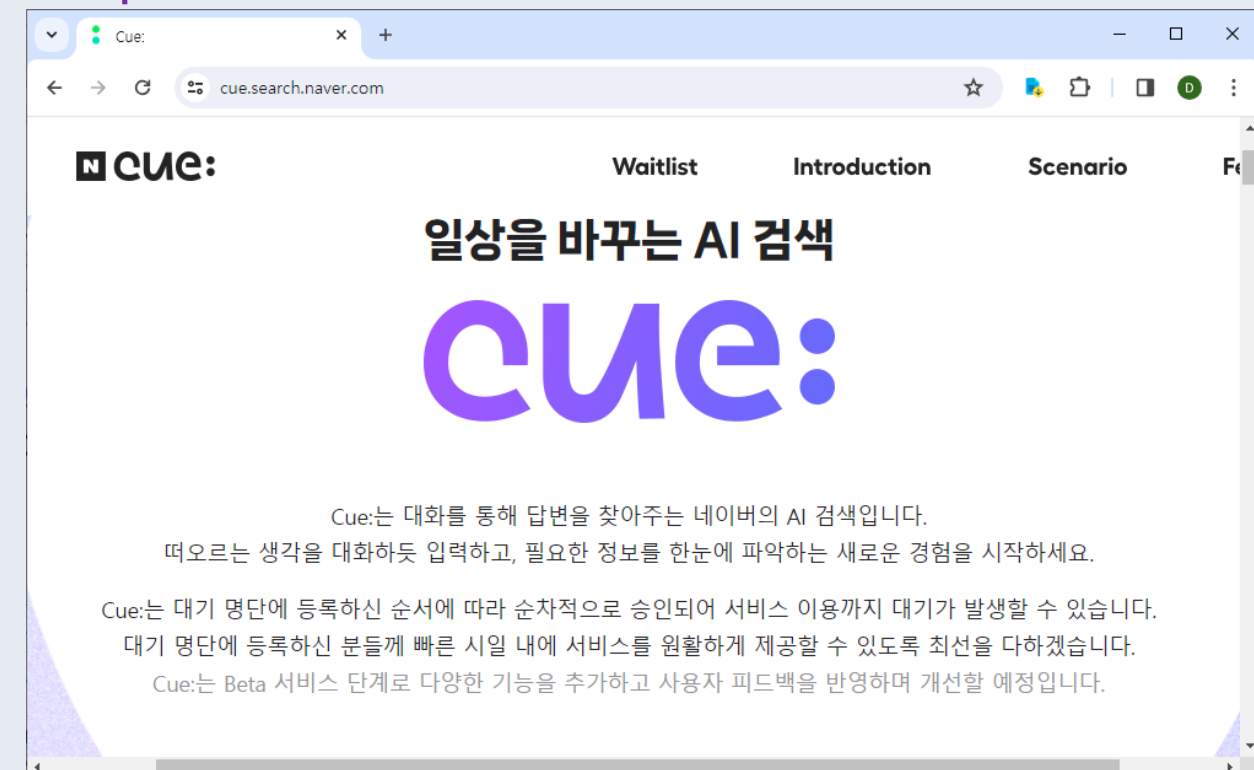
<https://elevenlabs.io/>



<https://pandarank.net/chat/tool>



<https://cue.search.naver.com/>



실습 - ChatGPT

<https://openai.com/>

The image shows a browser window with the OpenAI website and a separate window showing the 'Get started' modal. Red boxes and numbers indicate the steps to access ChatGPT.

Step 1: On the OpenAI website (<https://openai.com/>), click the **Log in** button in the top right navigation bar.

Step 2: On the OpenAI Platform (<https://platform.openai.com/apps>), click the **ChatGPT** button in the bottom navigation bar.

Step 3: In the 'Get started' modal, click the **Sign up** button.

실습 - ChatGPT

ChatGPT

chat.openai.com/#pricing

당신의 요금제

무료 플랜

무료

당신의 현재 플랜

ChatGPT를 처음 시작하는 사람들을 위한

- ✓ 무제한 메시지, 상호 작용, 그리고 이력
- ✓ 저희의 GPT-3.5 모델에 접근
- ✓ 웹, iOS, Android에서 접근 가능

기존에 사용하던 플랜이 있나요? [결제 도움말을](#) 확인해보세요.

ChatGPT 플러스

USD \$20/월

Plus로 업그레이드

무료 플랜에 포함된 모든 것, 그리고:

- ✓ 가장 능력이 뛰어난 모델인 GPT-4에 대한 접근
- ✓ GPT를 검색, 생성, 사용하세요
- ✓ DALL·E, 브라우징, 고급 데이터 분석 등 추가 도구에 대한 접근

팀

월당 사람마다 USD \$25*

팀으로 업그레이드

Plus의 모든 것, 그리고:

- ✓ GPT-4 및 DALL·E, 브라우징, 고급 데이터 분석 등과 같은 도구에 대한 확장된 접근 권한
- ✓ 귀하의 작업 공간에서 GPT를 생성하고 공유하세요
- ✓ 워크스페이스 관리를 위한 관리자 콘솔
- ✓ 기본적으로 팀 데이터는 훈련에서 제외됩니다. [자세히 알아보기](#)

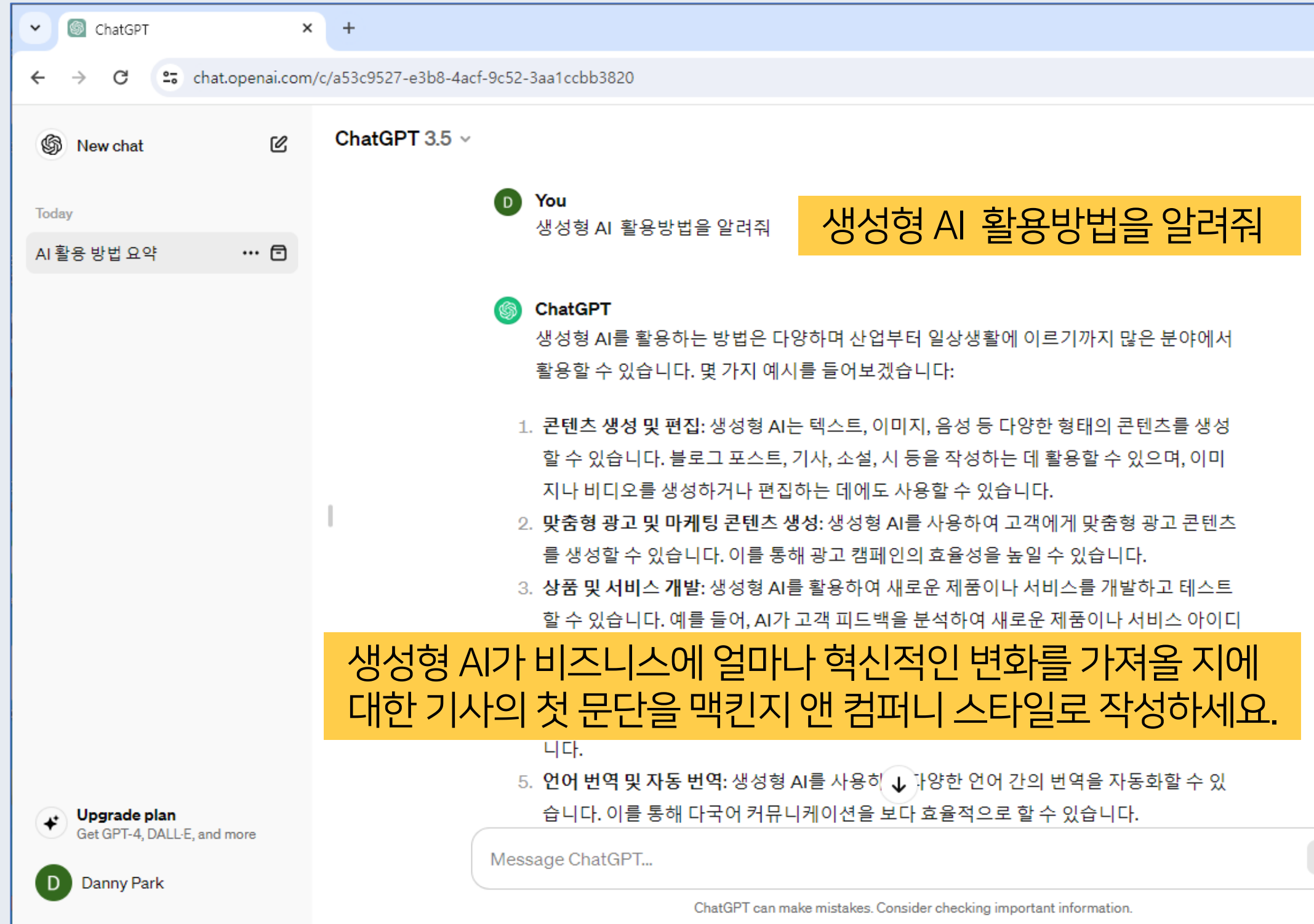
* 연간 청구, 최소 2명 사용자

플랜 업그레이드

Danny

더 많은 기능이 필요하신가요? [ChatGPT Enterprise](#) 를 확인해 보세요.

실습 - ChatGPT



The screenshot shows the ChatGPT web interface. The user has asked: "생성형 AI 활용방법을 알려줘" (Tell me how to use generative AI). The AI has responded with a list of three use cases: content generation/editing, matching ads/marketing content, and product/service development. A yellow highlight box is placed over the AI's response, containing the text: "생성형 AI가 비즈니스에 얼마나 혁신적인 변화를 가져올 지에 대한 기사의 첫 문단을 맥킨지 앤 컴퍼니 스타일로 작성하세요." (Write the first paragraph of an article about how much revolutionary change generative AI will bring to business in the style of McKinsey & Company).

ChatGPT 3.5

You: 생성형 AI 활용방법을 알려줘

ChatGPT: 생성형 AI를 활용하는 방법은 다양하며 산업부터 일상생활에 이르기까지 많은 분야에서 활용할 수 있습니다. 몇 가지 예시를 들어보겠습니다:

1. 콘텐츠 생성 및 편집: 생성형 AI는 텍스트, 이미지, 음성 등 다양한 형태의 콘텐츠를 생성할 수 있습니다. 블로그 포스트, 기사, 소설, 시 등을 작성하는 데 활용할 수 있으며, 이미지나 비디오를 생성하거나 편집하는 데에도 사용할 수 있습니다.
2. 맞춤형 광고 및 마케팅 콘텐츠 생성: 생성형 AI를 사용하여 고객에게 맞춤형 광고 콘텐츠를 생성할 수 있습니다. 이를 통해 광고 캠페인의 효율성을 높일 수 있습니다.
3. 상품 및 서비스 개발: 생성형 AI를 활용하여 새로운 제품이나 서비스를 개발하고 테스트할 수 있습니다. 예를 들어, AI가 고객 피드백을 분석하여 새로운 제품이나 서비스 아이디어를 제안할 수 있습니다.

생성형 AI가 비즈니스에 얼마나 혁신적인 변화를 가져올 지에 대한 기사의 첫 문단을 맥킨지 앤 컴퍼니 스타일로 작성하세요.

5. 언어 번역 및 자동 번역: 생성형 AI를 사용하여 다양한 언어 간의 번역을 자동화할 수 있습니다. 이를 통해 다국어 커뮤니케이션을 보다 효율적으로 할 수 있습니다.

Upgrade plan
Get GPT-4, DALL-E, and more

Danny Park

Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

파이썬 라이브러리를 활용해서 'Gen AI' 키워드로 검색한 결과를 스크래핑 하는 코드를 작성해.

```
python Copy code

import requests
from bs4 import BeautifulSoup

def scrape_gen_ai_results():
    # 검색할 키워드
    keyword = 'Gen AI'
    # 구글 검색 결과 페이지 URL (스크랩하고자 하는 검색 엔진에 맞게 변경)
    url = f'https://www.google.com/search?q={keyword}'

    # User-Agent 헤더를 설정하여 봇으로 인식되지 않도록 함
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.114 Safari/537.36'
    }

    # requests를 사용하여 URL에 GET 요청
    response = requests.get(url, headers=headers)

    # 응답 코드가 200(성공)이 아니면 에러를 발생시킴
    if response.status_code != 200:
        raise Exception(f"Failed to load page (status code: {response.status_code})")

    # BeautifulSoup을 사용하여 HTML 파싱
    soup = BeautifulSoup(response.text, 'html.parser')

    # 검색 결과 링크 가져오기
    search_results = soup.find_all('div', class_='tF2Cxc')

    # 결과 출력
    for result in search_results:
        link = result.find('a')['href']
        title = result.find('h3').text
```

실습 - Gemini

<https://gemini.google.com/>


Gemini

gemini.google.com/app/289311f9609a6f09

≡

Gemini


+



다음 문제의 답은? 한글로 설명해줘 :

A company has developed several AWS Glue jobs to validate and transform its data from Amazon S3 and load it into Amazon RDS for MySQL in batches once every day. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the AWS Glue job processes all the S3 input data on each run.

Which approach would allow the developers to solve the issue with minimal coding effort?



해당 문제는 AWS Glue 에서 S3 에서 데이터를 읽어와 매일 배치로 유효성 검사 및 변환 후 MySQL 에서 로딩하는 ETL 작업이 있는데, 현재 모든 데이터를 매번 처리하는 문제를 최소한 코딩으로 해결하는 방법을 묻고 있습니다.

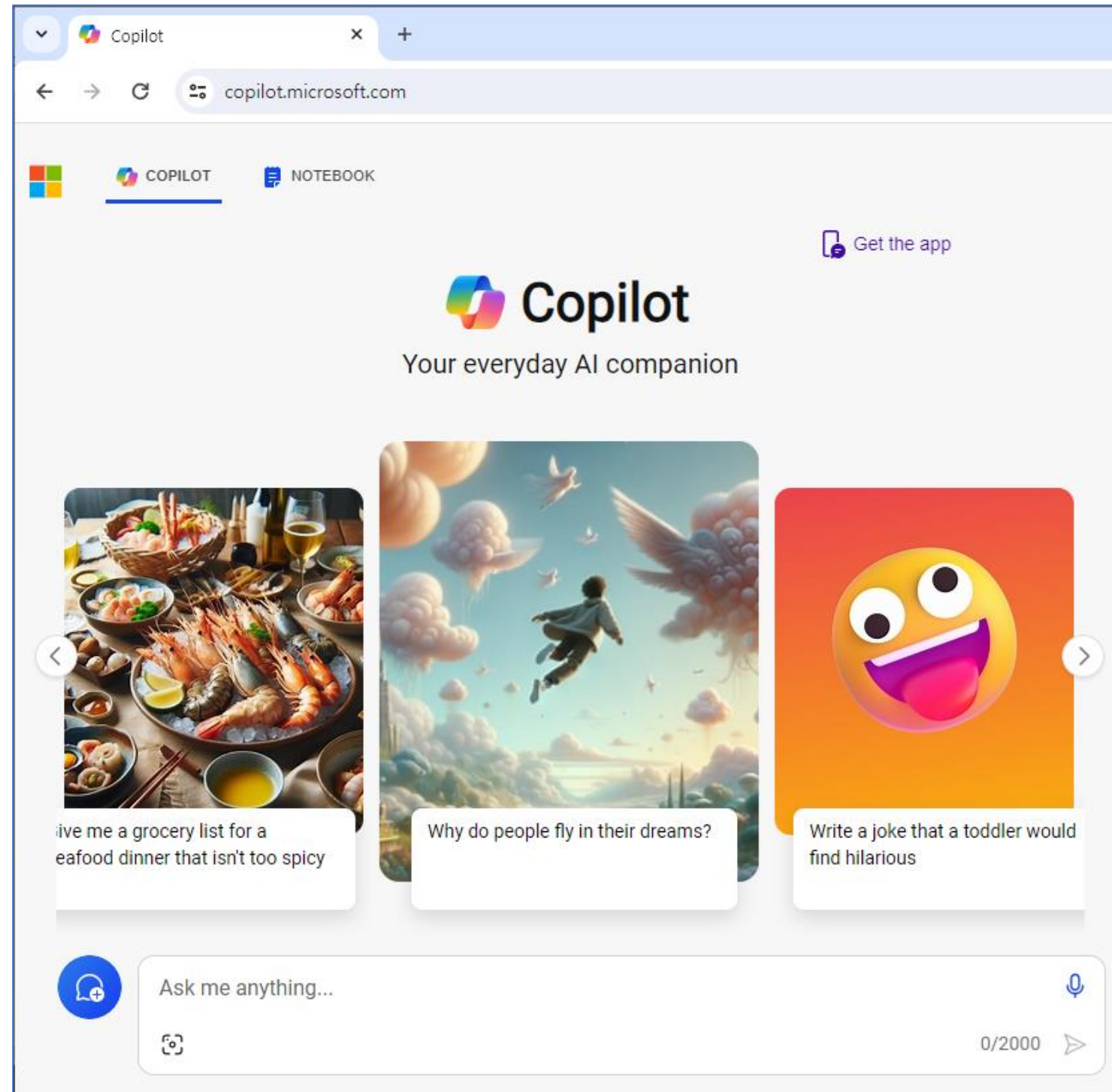
여기에는 몇 가지 옵션이 있습니다.

1. Job bookmarks 사용:

- Glue에서는 Job Bookmarks 라는 기능을 제공합니다. 이 기능은 이전 작업 실행에서 처리된 데이터를 추적하여 이후 실행에서 다시 처리하지 않도록 합니다. 이 기능을 사용하면 ETL 작업은 이전 실행 시점 이후에 추가된 데이터만 처리하도록 구성할 수 있습니다.

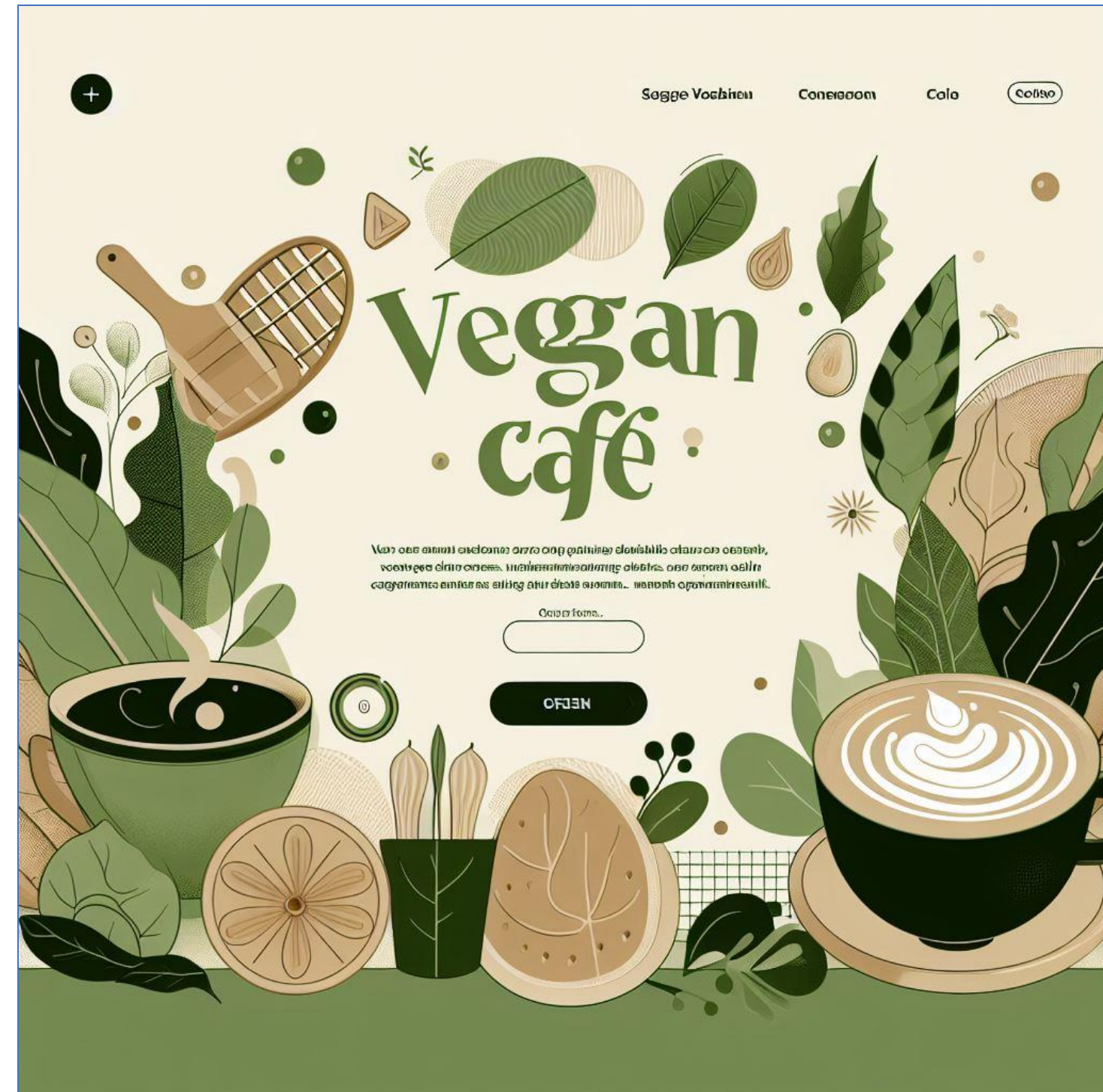
실습 - Copilot

<https://copilot.microsoft.com/>



이미지 생성 프롬프트 참고

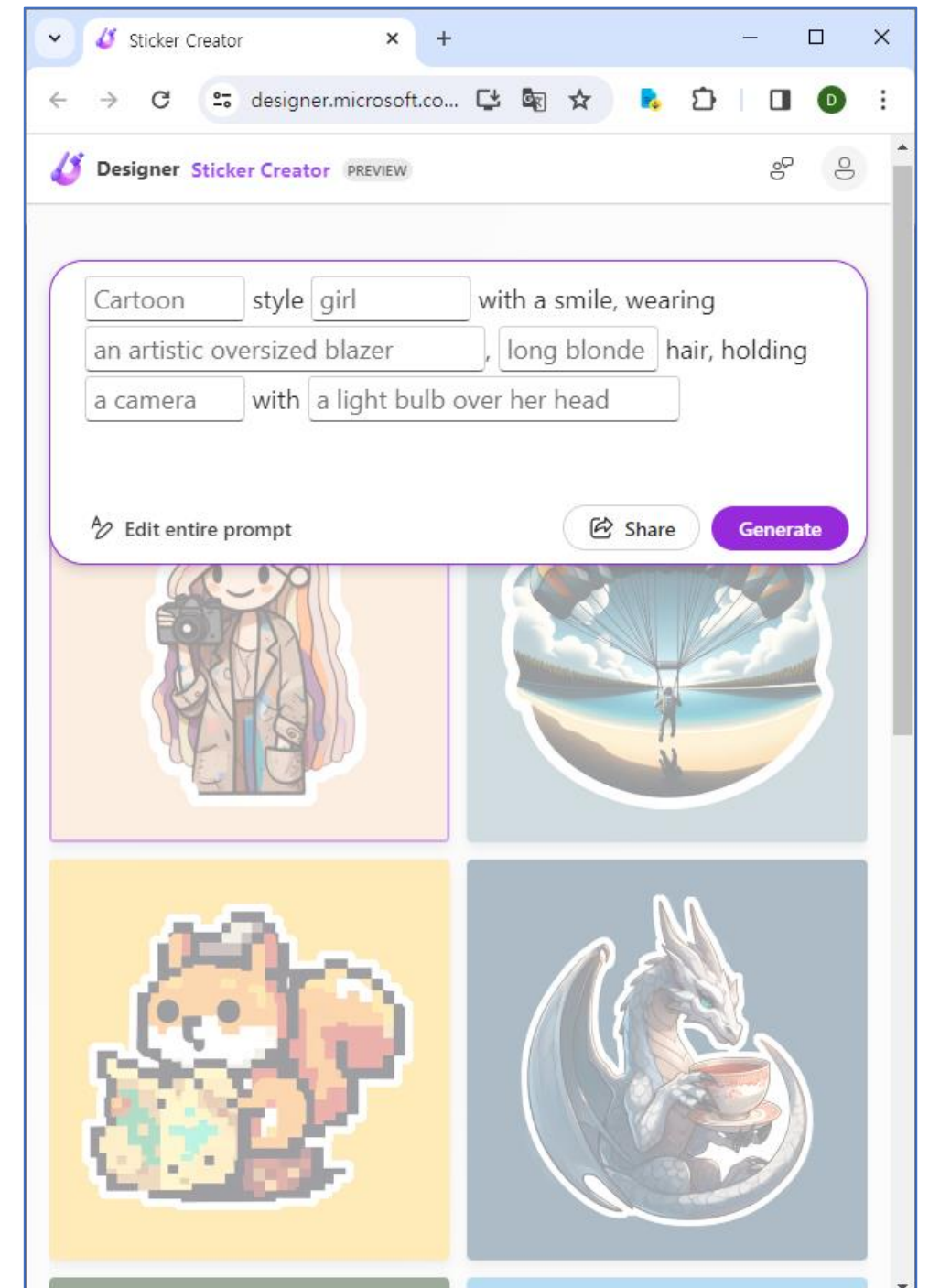
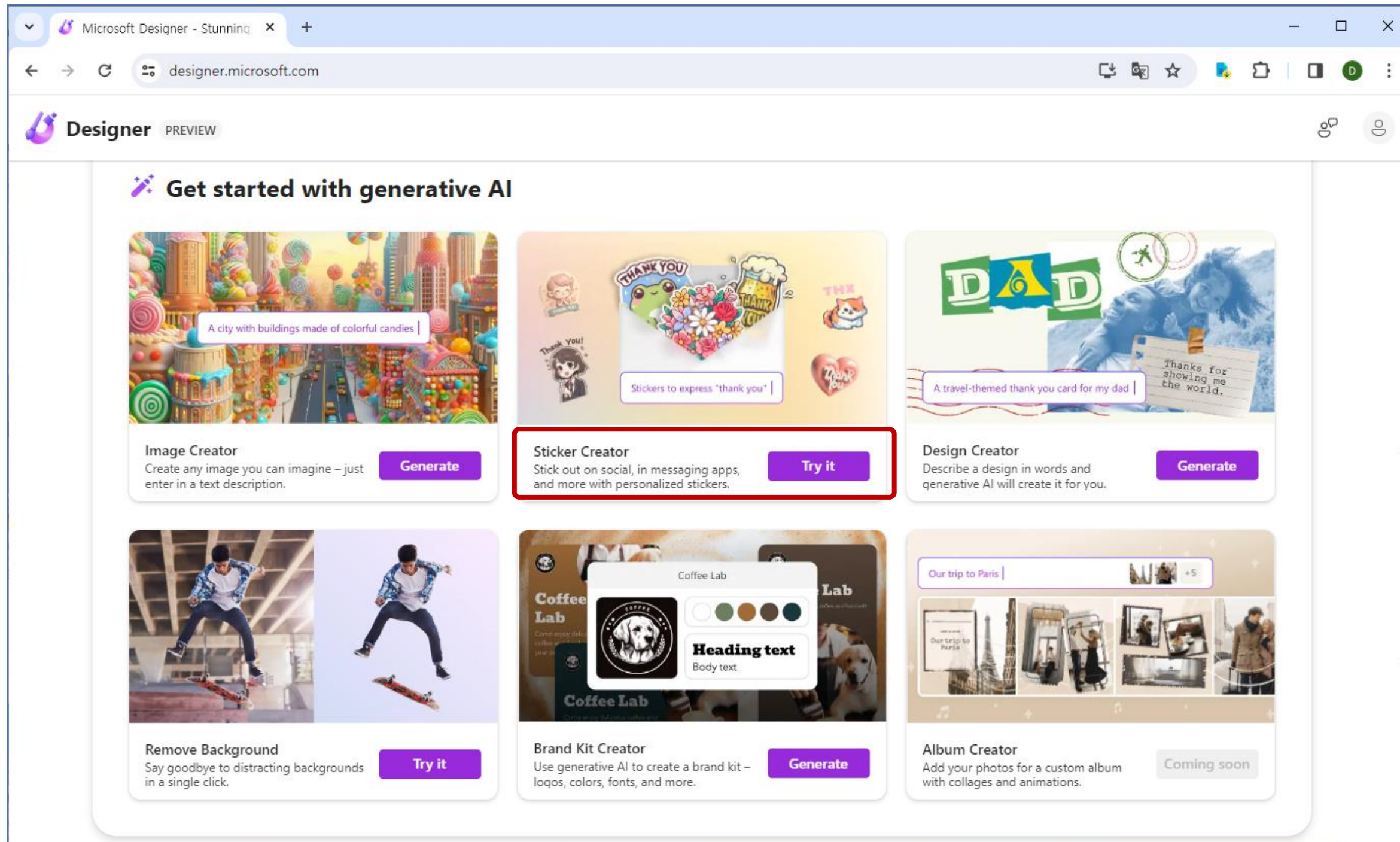
<https://eopla.net/magazines/12062>



실습 - Designer

<https://designer.microsoft.com/>

 **Get started with generative AI**



THANK YOU

kgpark88@gmail.com