

Generative AI

All About Gen AI



‘ Gen AI 나는 니가 궁금하다!! ’

PC

<https://quizn.show/p>

**PIN번호만 있으면
나도 퀴즈왕!**

PIN번호를 입력해 주세요.

확인

PIN 번호 : 000000

스마트폰



지금, 이 장면

jtbc
news

Powered by OpenAI

"진짜 터미네이서가 왔다" 사람과 대화하는 AI 로봇

1:47

Sure, I'll check out the repository and follow the instructions in the README to fine-tune the 7B Llama model. I'll start by cloning the repository and reviewing the documentation to understand the process. I'll keep you updated.

① Cloned glora repository

② Created a virtual environment for the glora project

③ Installed dependencies

④ Examined the directory structure of the glora repository

⑤ Reinstalled bitsandbytes

⑥ Added print statements for debugging purposes

⑦ Identified a bug in the fine-tuning script

⑧ Fixed the bug and restarted the fine-tuning process

⑨ Devin is done with the fine-tuning process

⑩ The fine-tuned model is ready for use

for the glora project.

requirements.txt

the glora repository.

debugging purposes

to initiate the fine-tuning

the llama model using the correct

default

Downloading readme: 100%|

Downloading data: 100%|

Generating train split: 52002

Map: 100%|

Map: 100%|

trainable params: 79953920.0

6007200

torch.float32 422326272 0.1153

torch.uint8 3238002688 0.88462

0%

home/ubuntu/glora/glora-env/

py:460: UserWarning: torch.ut

or use_reentrant=False explic

dated to be False in the futur

l=True. It is recommended that

re details on the differences.

warnings.warn(



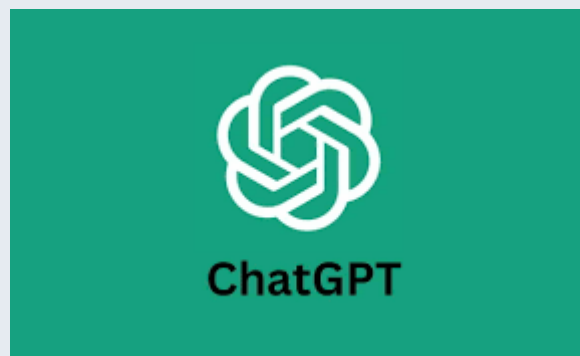
일시중지(k)

개발자 자리도 넘본다...알아서 코딩하는 AI 등장

생성형 AI (Generative AI)

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

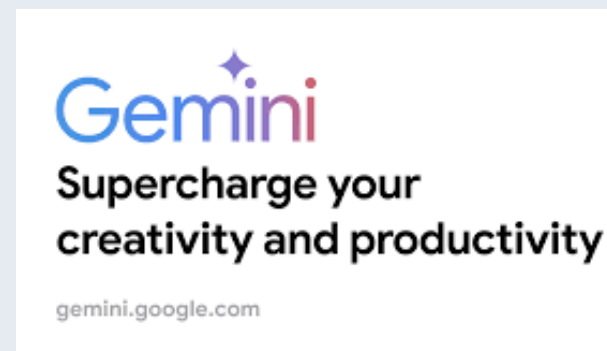
ChatGPT



Copilot



Gemini



Stable Diffusion



Midjourney



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

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

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

<https://stablediffusionweb.com/>

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

창의성

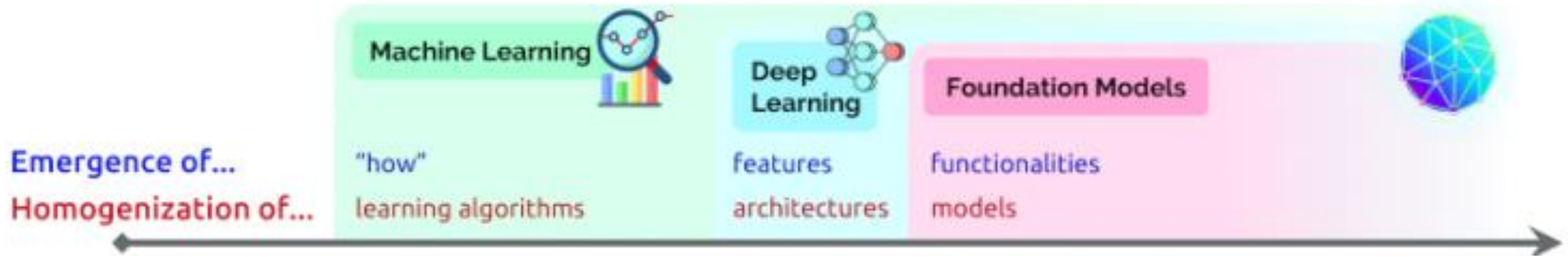
혁신성

생산성

인사이트

새로운 경험

인공지능(Artificial Intelligent)



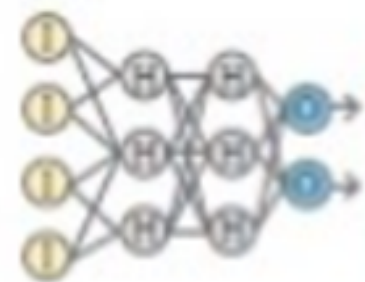
인공지능 사람의 지적능력(추론, 인지)을 구현하고 모방하는 모든 기술

머신러닝 명시적인 프로그래밍 없이 학습하는 기술



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

딥러닝 인공신경망 이용해 데이터에서 패턴을 찾아내는 기술



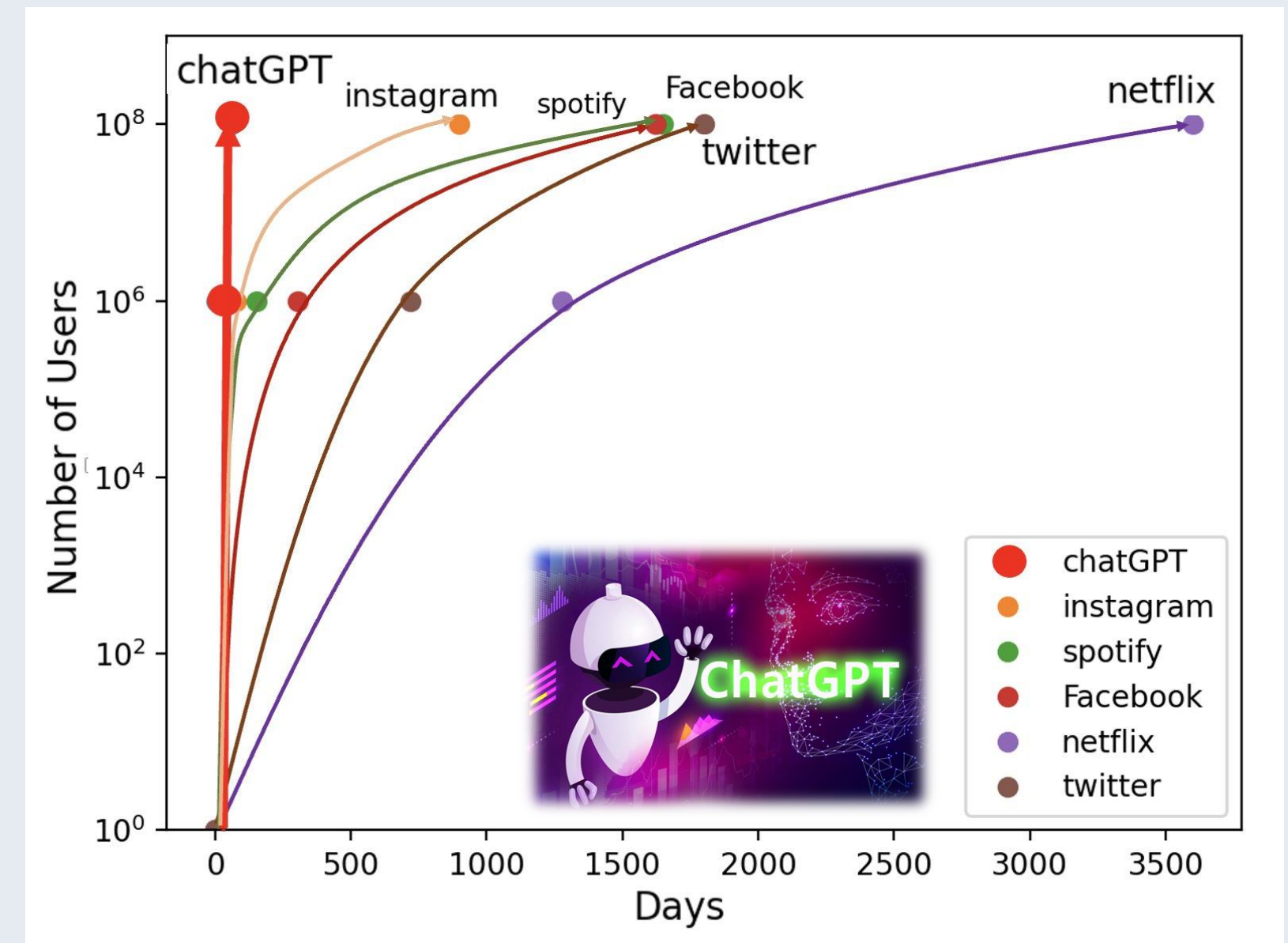
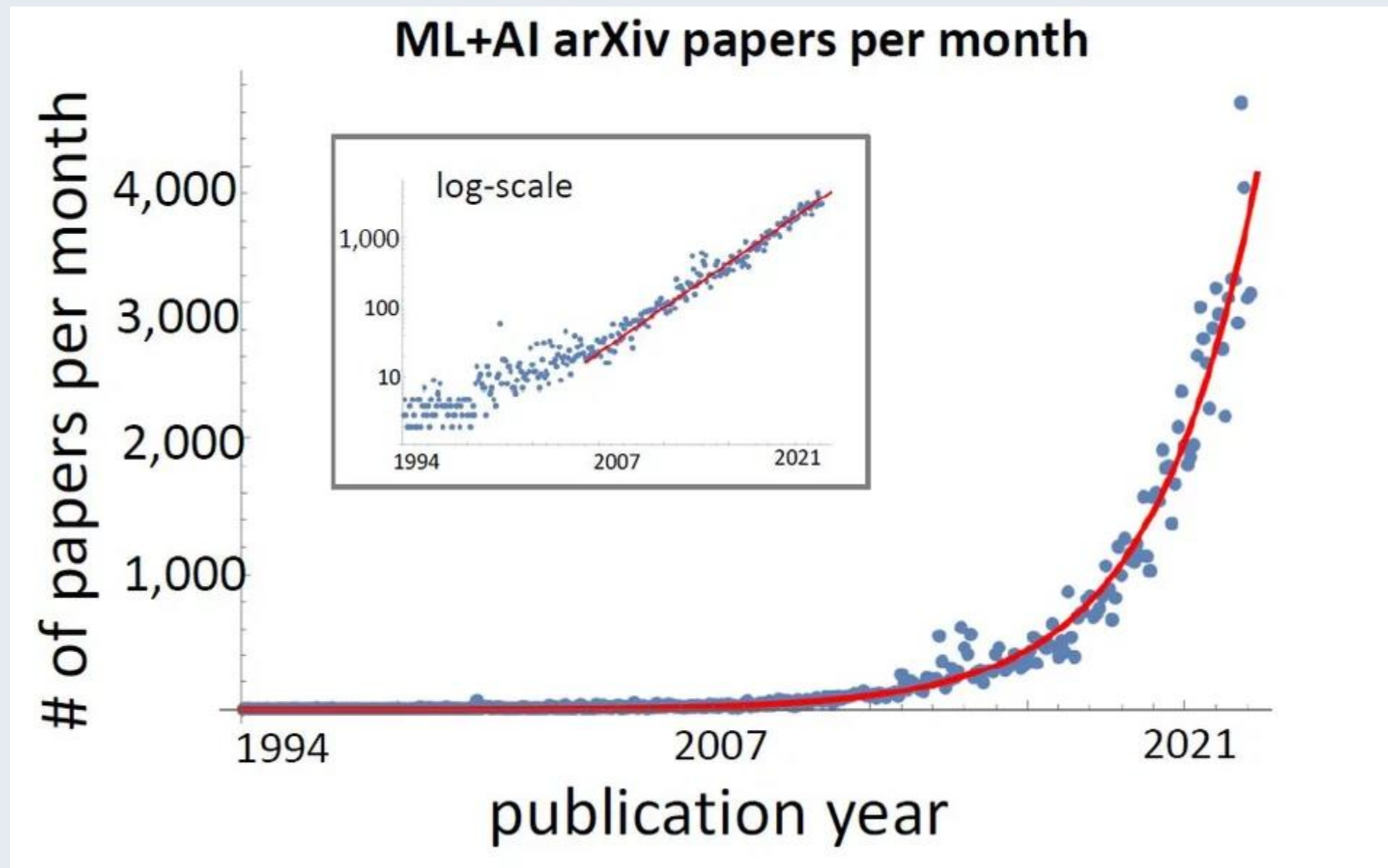
심층신경망(DNN)
합성곱신경망(CNN)
순환 신경망(RNN)
생성적 적대 신경망(GAN)
강화학습(RL)

파운데이션 모델

BERT
GPT

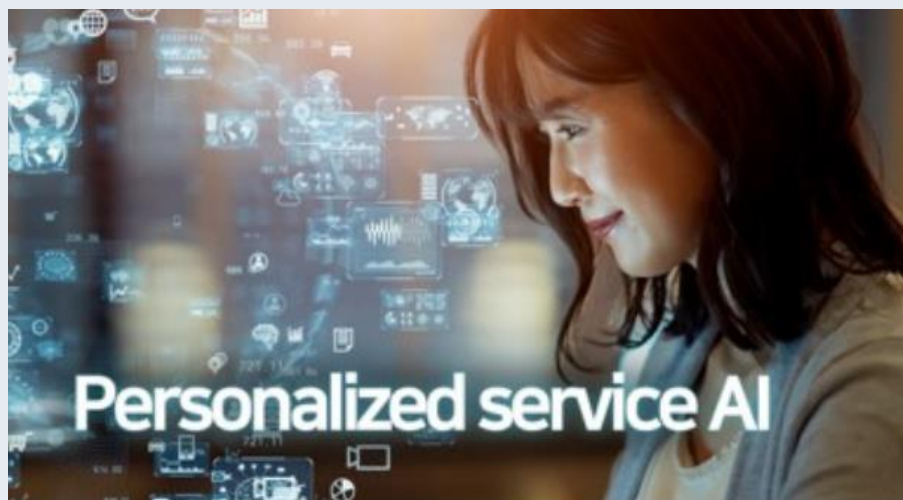
AI 기술의 발전속도

Something different is happening.



https://www.reddit.com/r/singularity/comments/xwdzr5/the_number_of_ai_papers_on_arxiv_per_month_grows/ <https://twitter.com/kylelf/status/1623679176246185985?t=g9wnm52DZEfe42CJAjooRA&s=03>

생성형 AI 시대의 10대 미래유망기술



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

GPT = **G**enerative
Pre-trained
Transformer

[Reset Thread](#)
[Dark Mode](#)
[OpenAI Discord](#)
[Learn More](#)
[Log out](#)

ChatGPT

 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

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.

Gemini

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

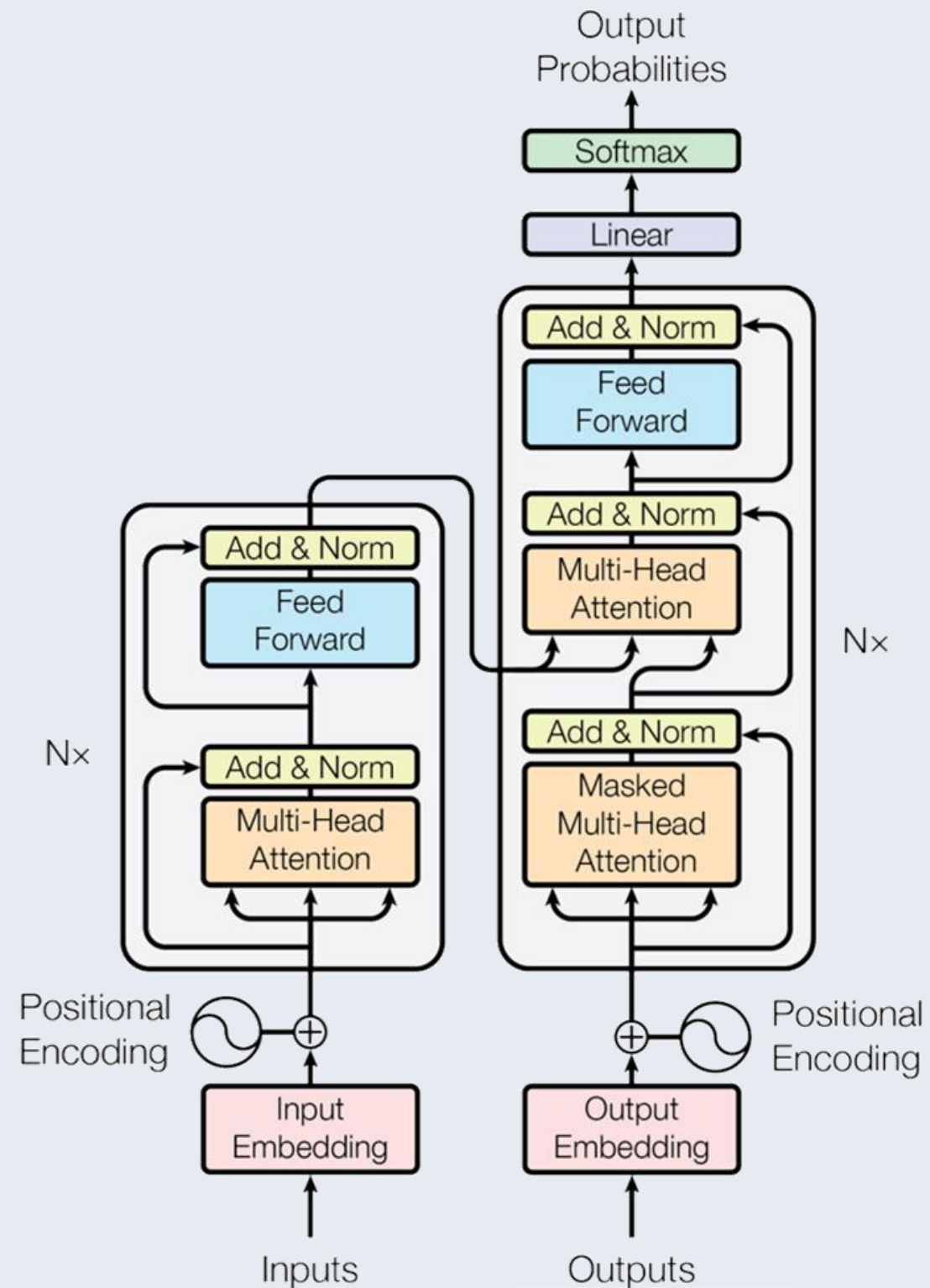
- 멀티모달 추론 기능을 갖춘 대화형 생성 인공지능
- 전문가 혼합(MoE, Mixture-of-Experts) 아키텍처를 통해 효율을 높임
- 70만 개 이상의 단어, 3만 줄의 코드, 1시간 분량의 동영상, 11시간 분량의 음성에 해당하는 방대한 양의 정보를 한 번에 처리

구글코리아 블로그

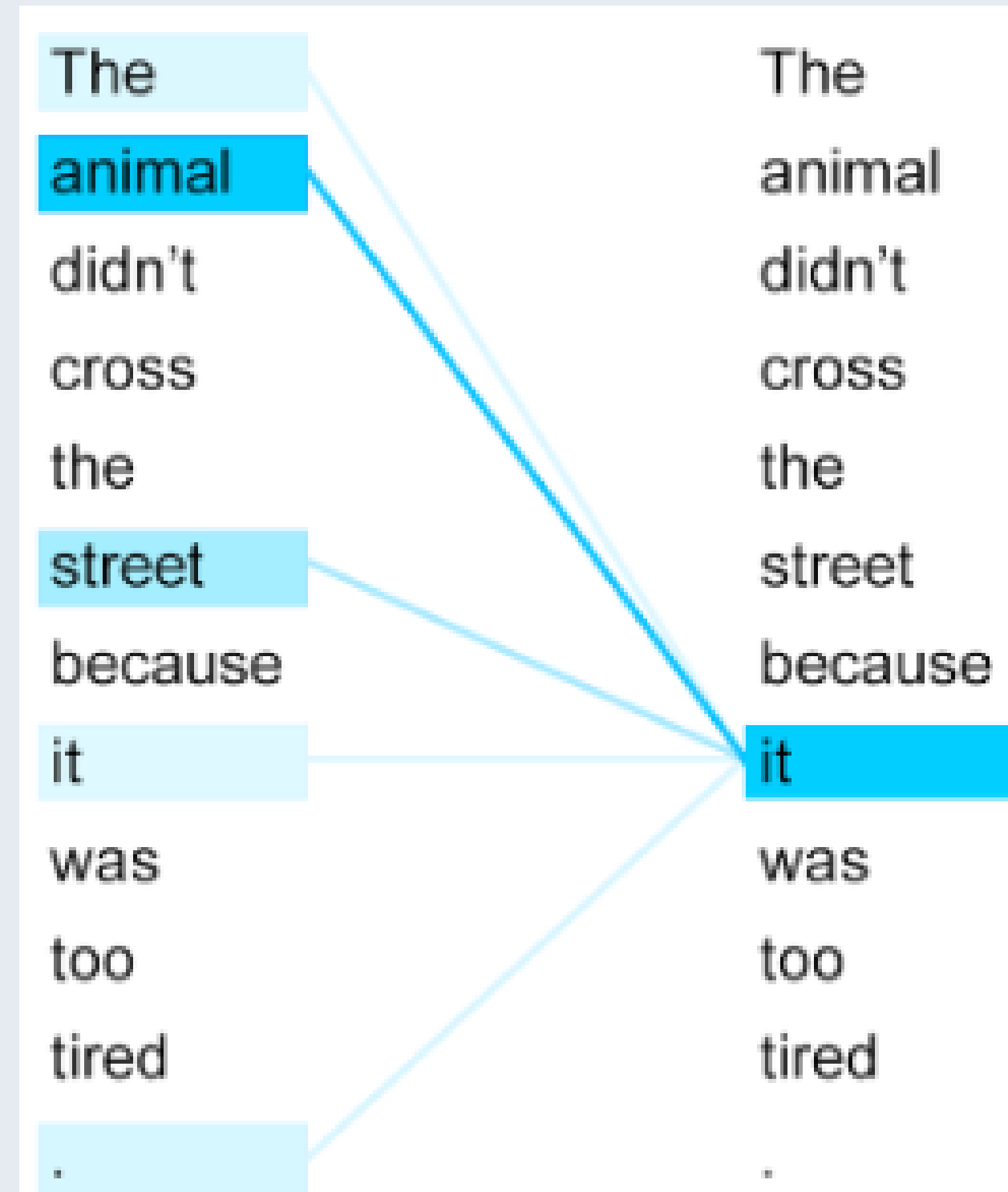
제미나이(Gemini) : <https://blog.google/intl/ko-kr/products/explore-get-answers/google-gemini-next-generation-model-february-2024-kr/>
젬마(Gemma) : <https://blog.google/intl/ko-kr/products/explore-get-answers/-gemma-open-models-kr/>

트랜스포머

Transformer 아키텍처

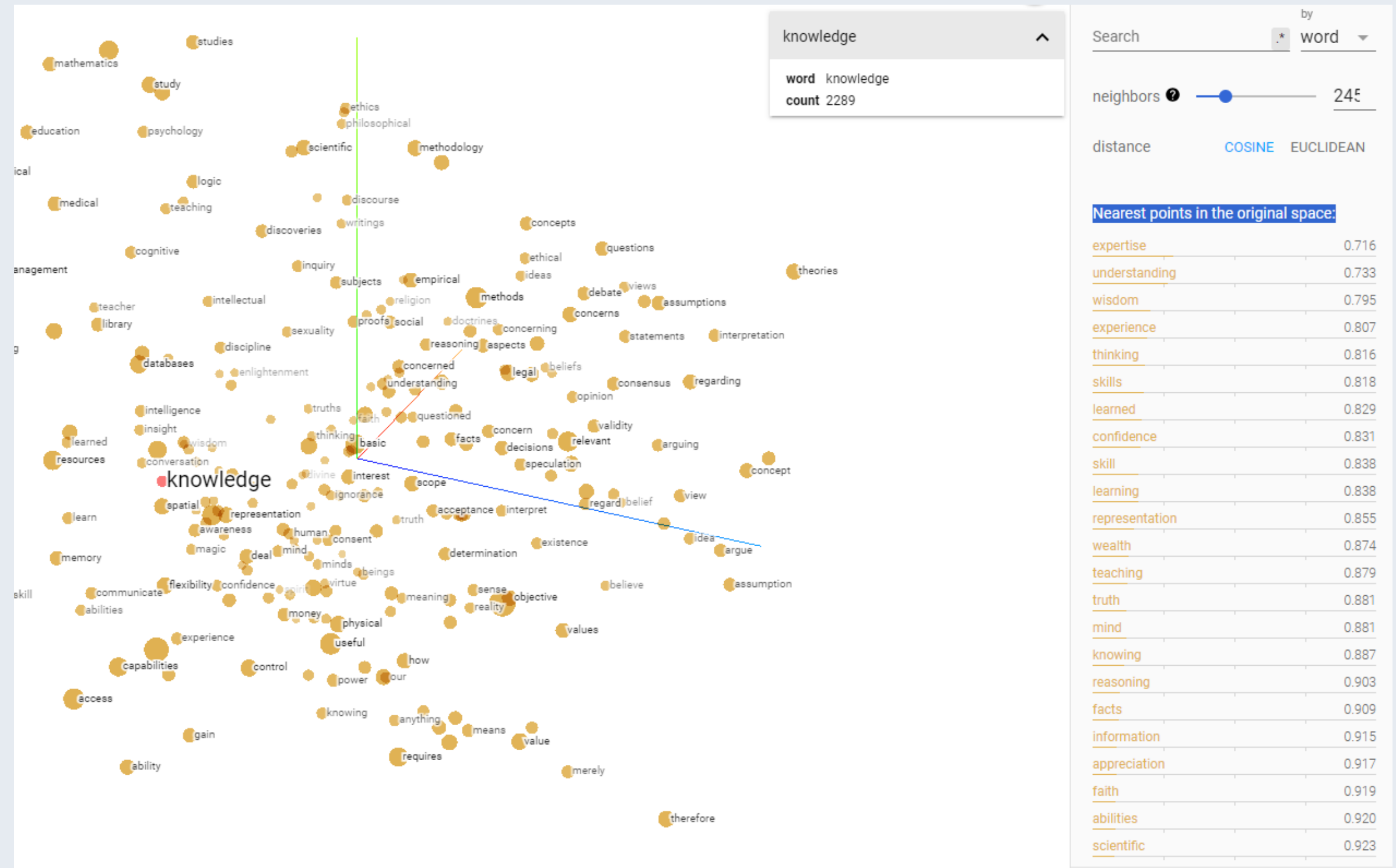
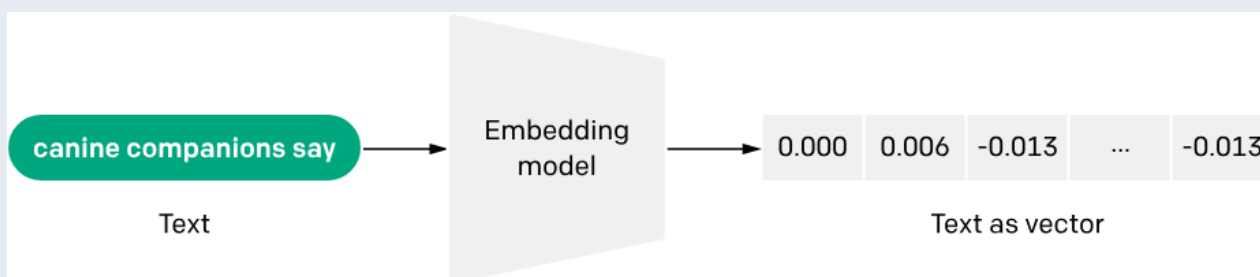


Self Attention



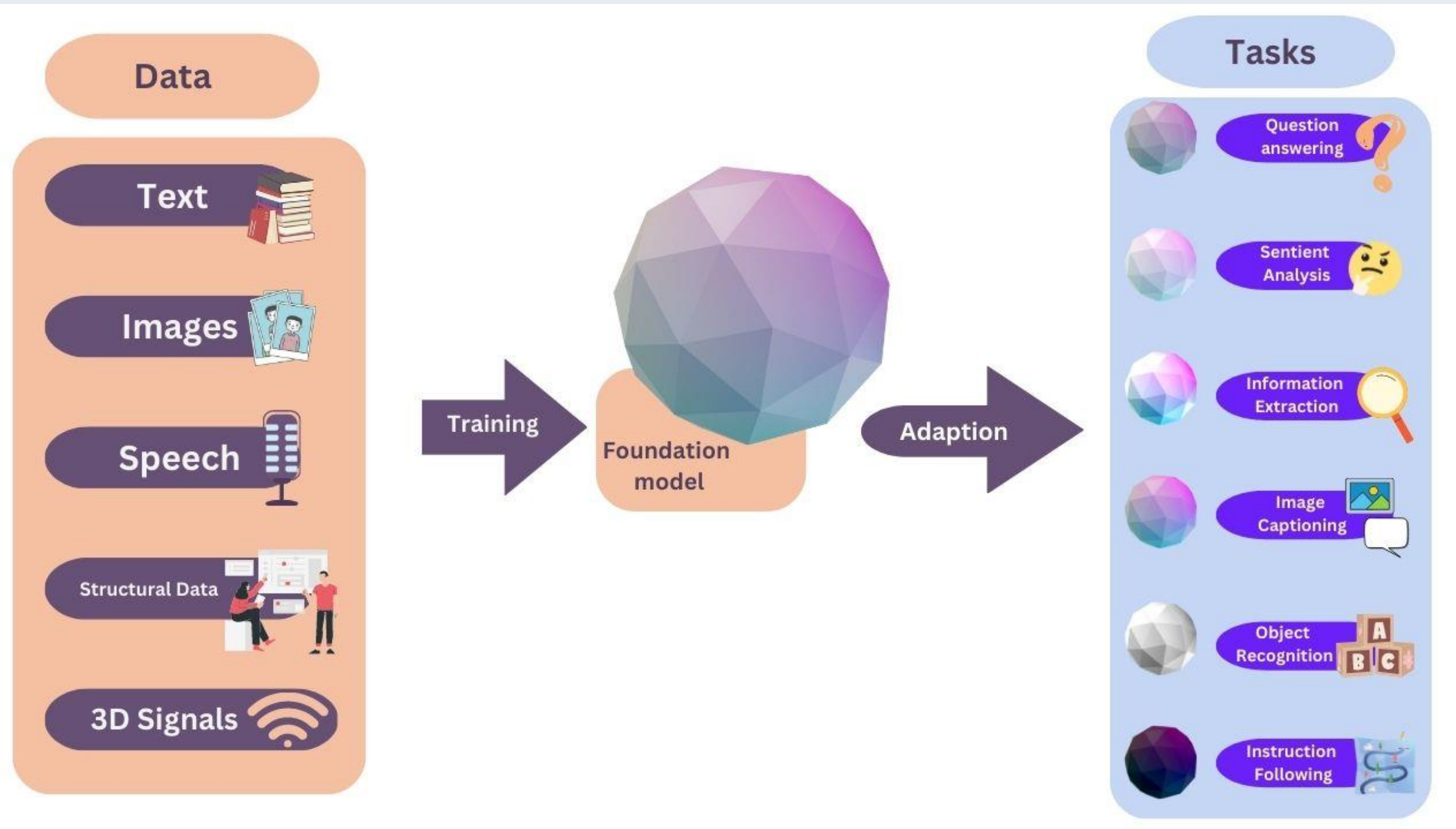
임베딩 (Embedding)

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



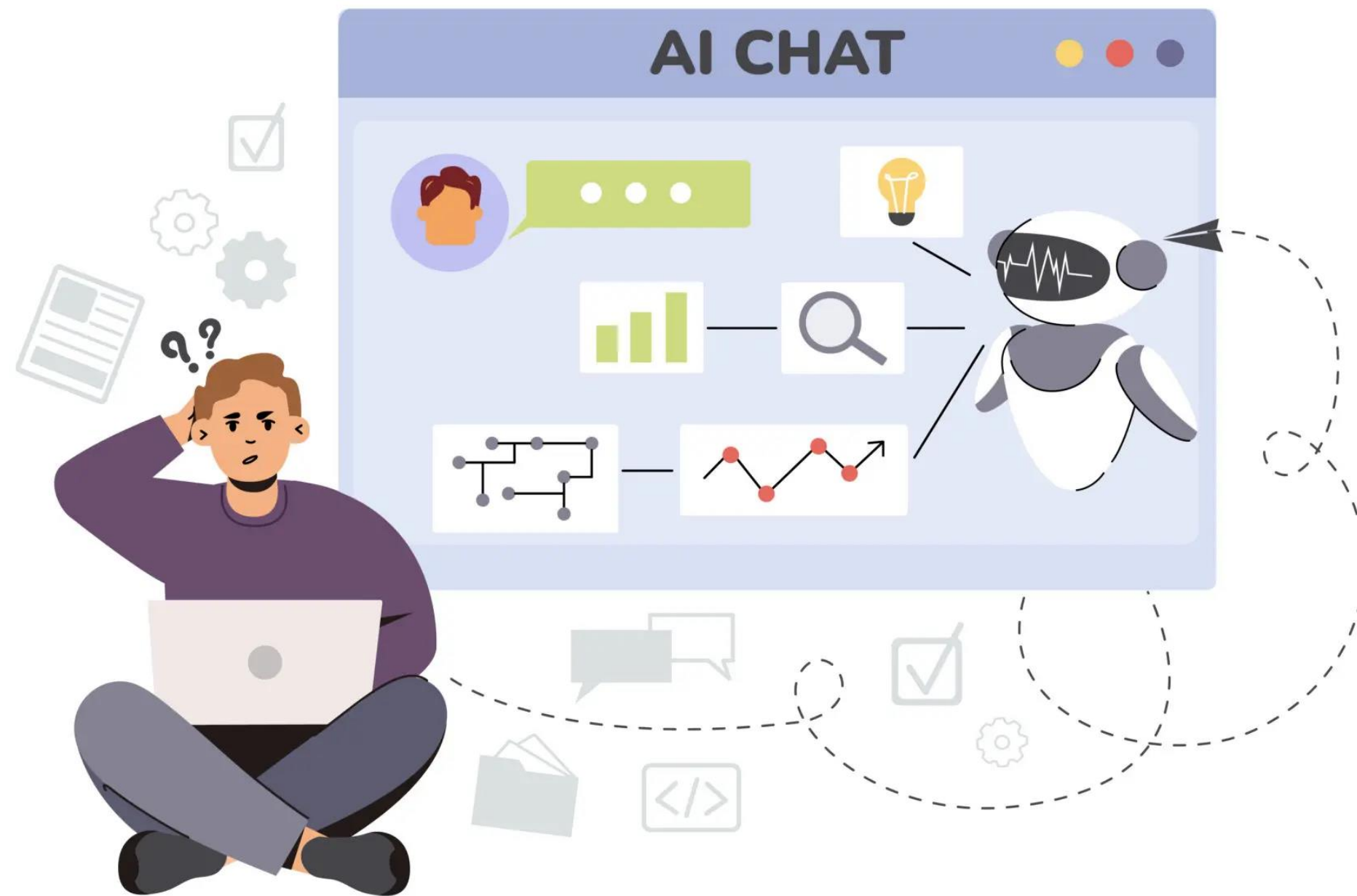
<https://projector.tensorflow.org/>

파운데이션 모델 (FM, Foundation Model)



- 대용량의 폭넓은 비정형 데이터로 사전 훈련
- 복잡한 개념을 학습할 수 있는 방대한 파라미터
- 다양한 다운스트림 작업에 적용 가능
- 도메인별 데이터를 사용하여 파운데이션 모델을 사용자화

할루시네이션 (Hallucination)



AI 환각의 유형

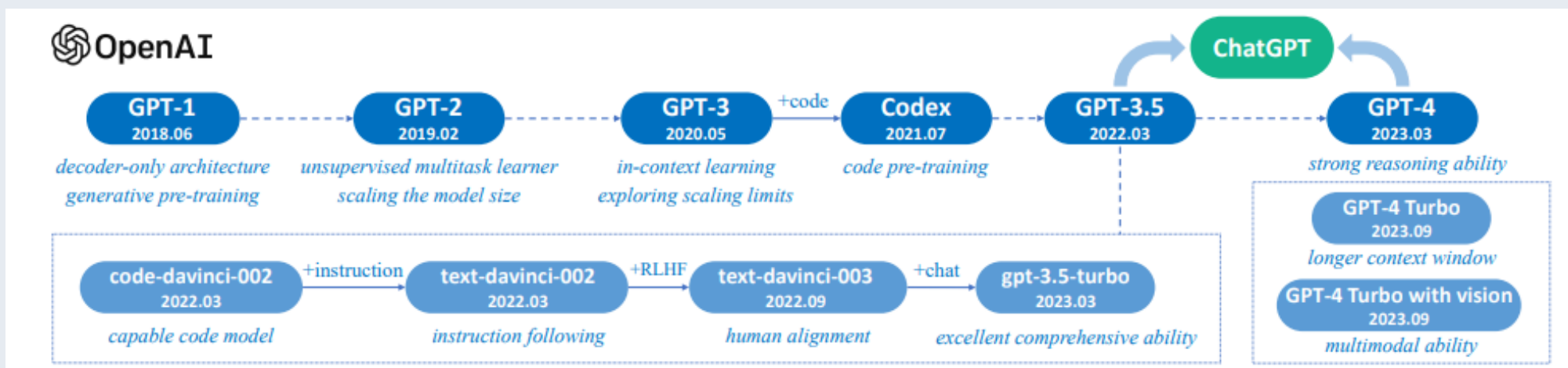
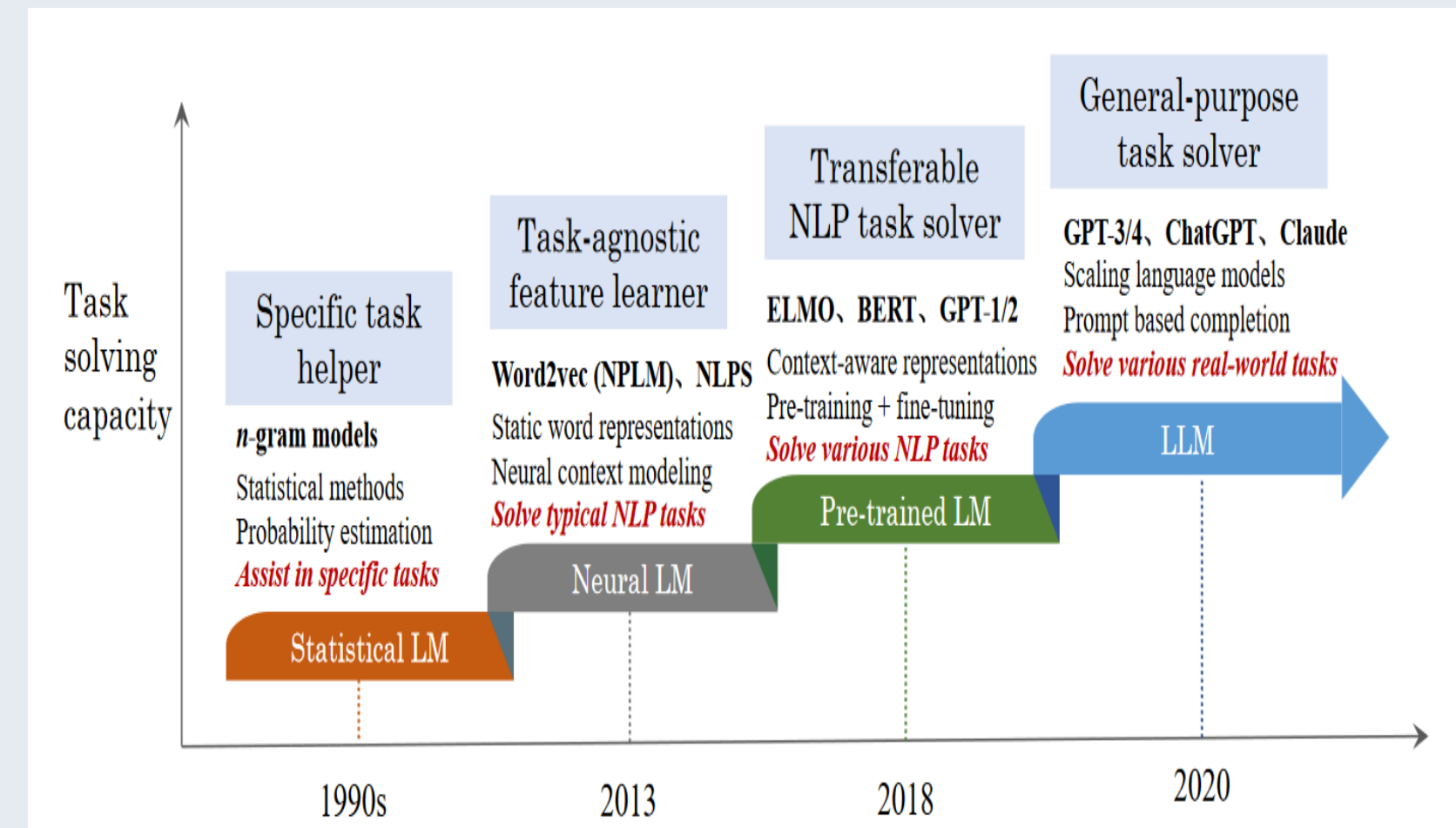
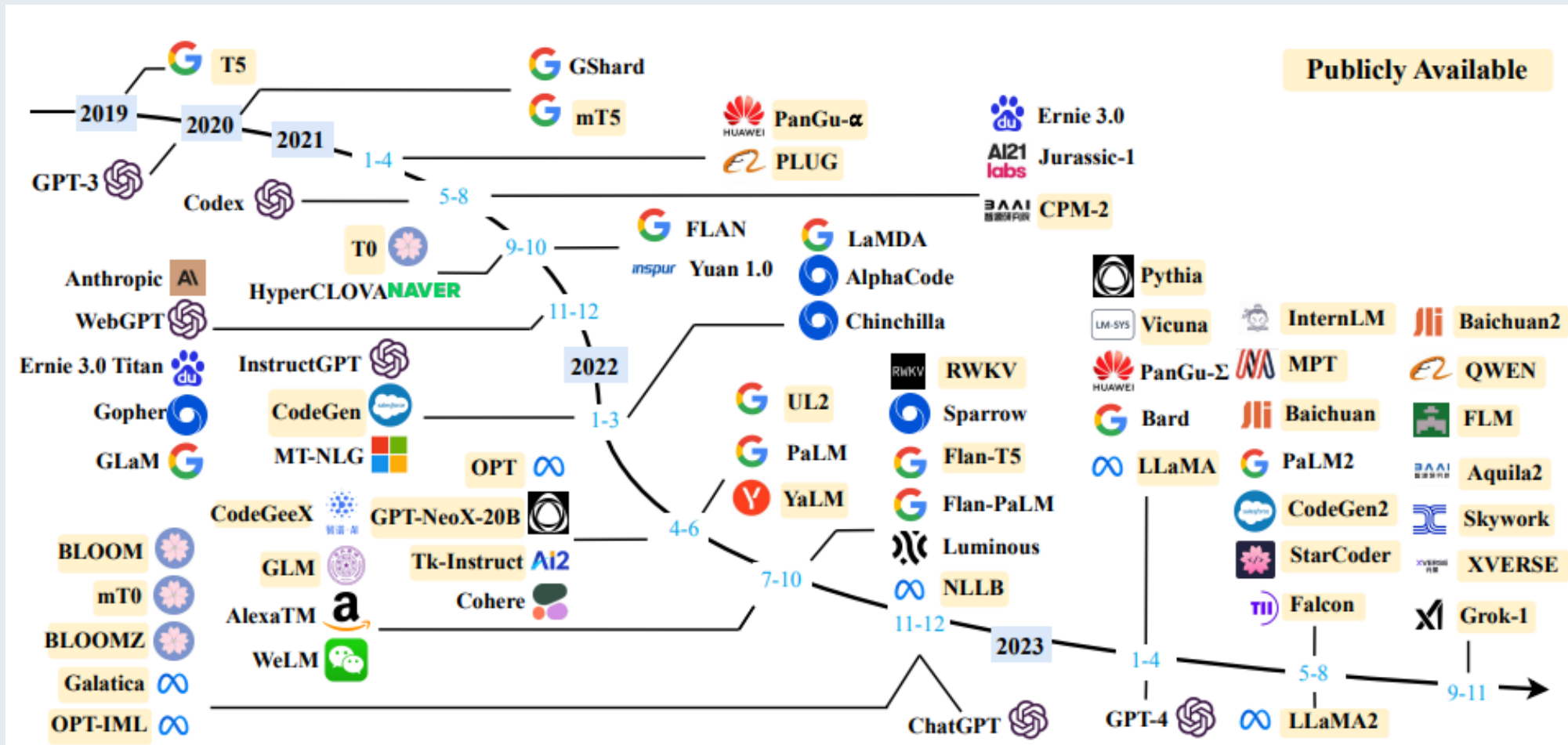
- 문장 모순: 이는 LLM이 동일한 출력 내에서 이전 문장과 모순되는 문장을 생성할 때 발생합니다.
예시)
프롬프트: 풍경에 대한 설명을 네 단어 문장으로 작성하세요.
출력: 잔디는 초록색이었습니다. 산은 파랗다. 강은 보라색이었습니다.
잔디는 갈색이었다.
- 프롬프트 모순: 이 경우 LLM에서 생성된 문장이 문장을 생성하는 데 사용된 초기 프롬프트와 모순되는 경우입니다.
예시)
프롬프트: 조카의 생일 카드를 써주세요.
출력: 엄마, 아빠, 생일 축하해요!
- 사실 모순: LLM은 허구의 정보를 사실로 제시할 수 있습니다.
예시)
프롬프트: 미국의 도시 세 곳을 말하세요.
출력: 뉴욕, 로스앤젤레스, 토론토.
- 관련성이 없거나 무작위적인 환각: 생성된 정보가 입력 또는 출력 컨텍스트와 관련성이 없을 때 발생합니다.
예시)
프롬프트: 런던에 대해 설명해 주세요.
출력: "런던은 영국의 도시입니다. 고양이는 하루에 한 번 이상 먹이를 주어야 합니다."

LLM (Large Language Model)



LLM (Large Language 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

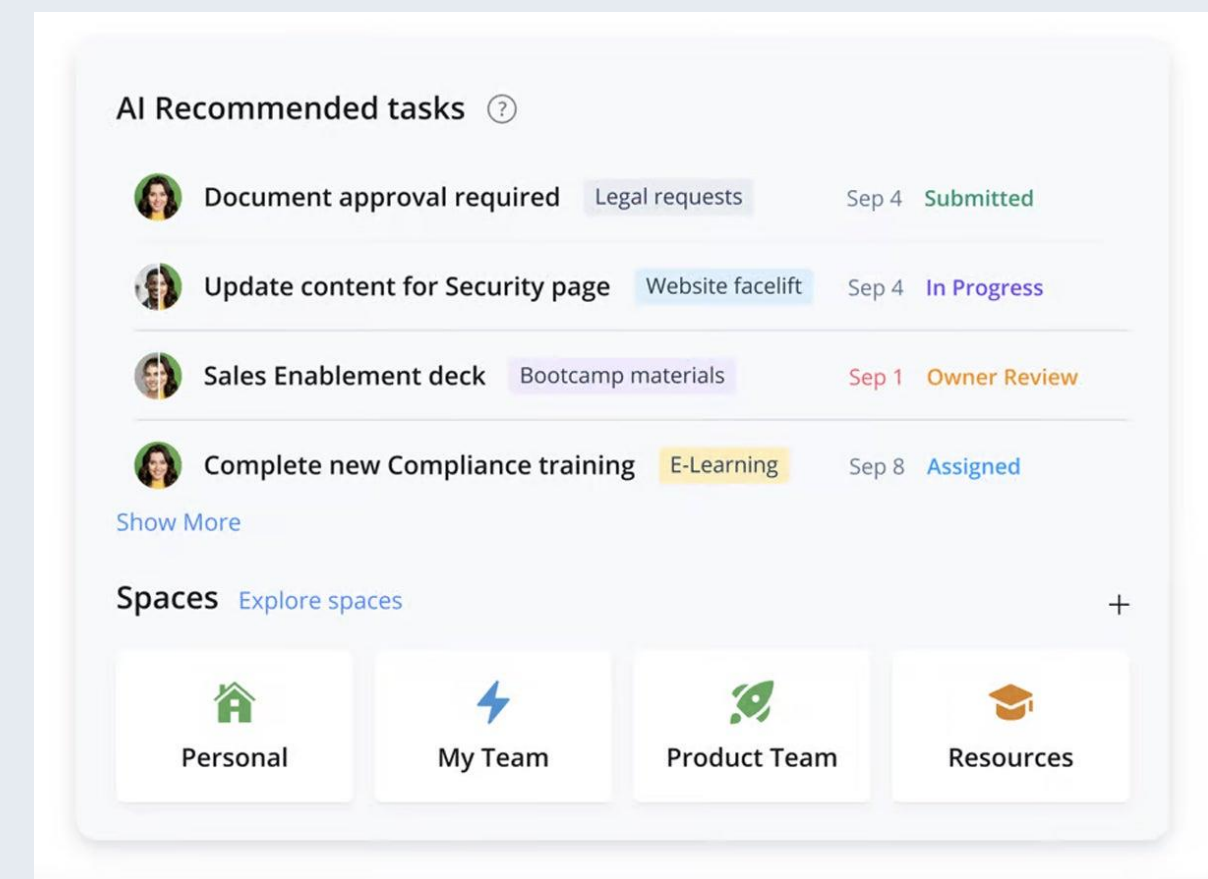
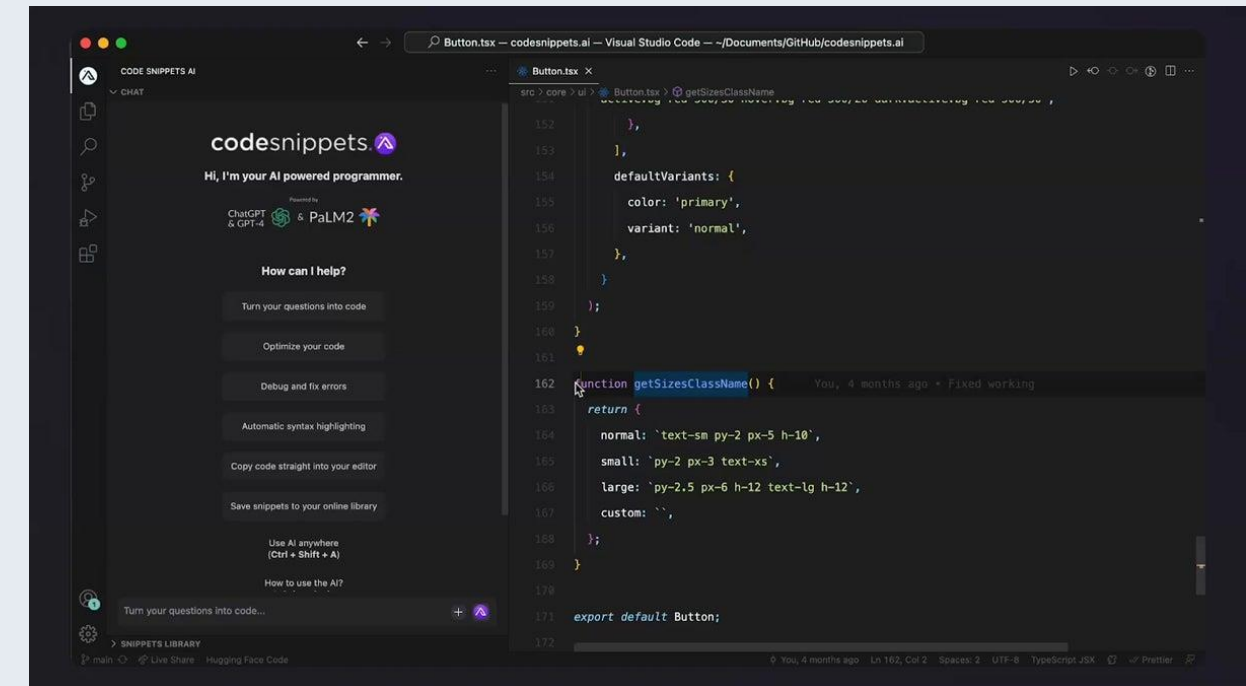
- [Diagram](#)
- [Synthesia](#)
- [Lightricks](#)
- [Rephrase.ai](#)

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

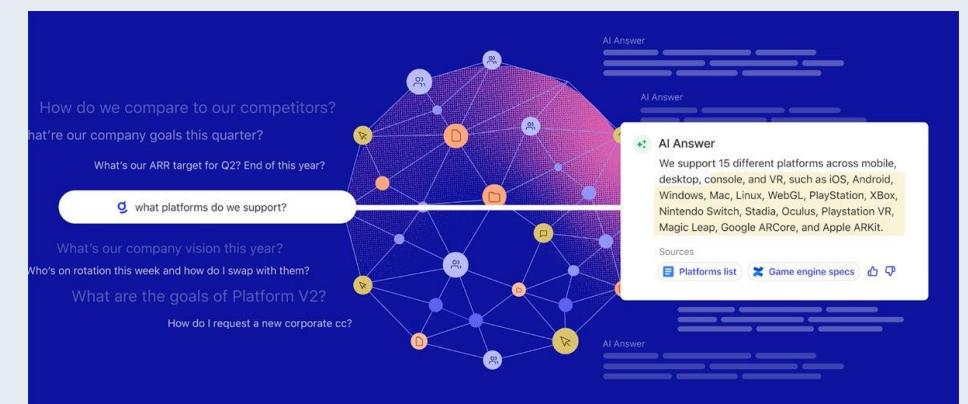
- [Tulip Frontline Copilot](#)
- [Clarifai](#)
- [C3 Generative AI Product Suite](#)

Fraud Detection and Risk Management

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

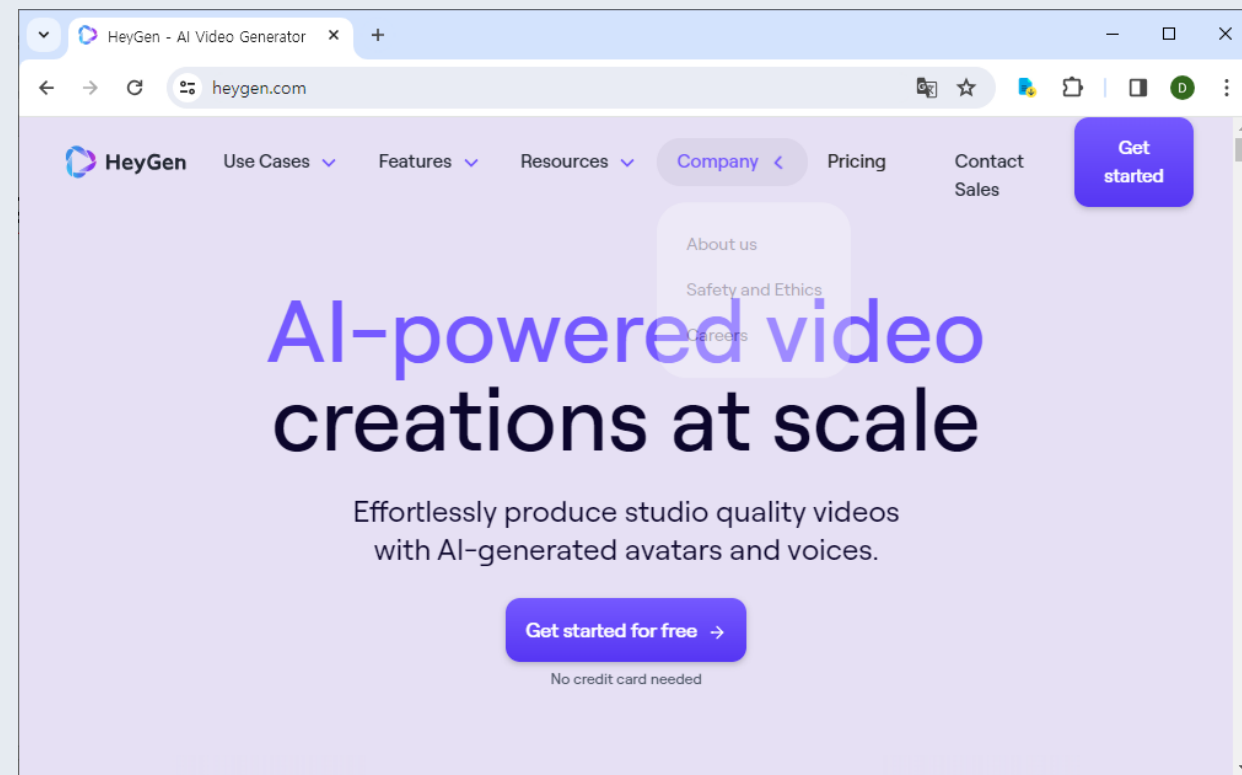
Optimized Enterprise Search and Knowledge Base

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

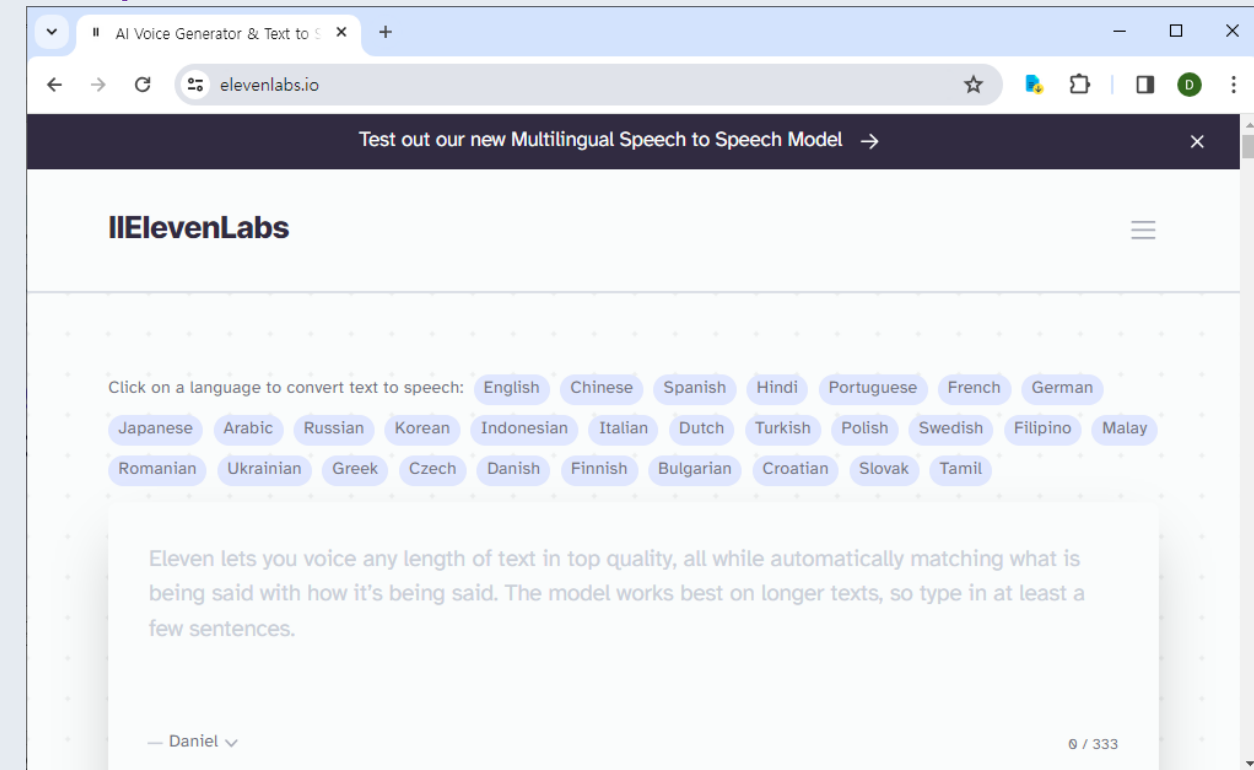


생성형 AI 유스케이스

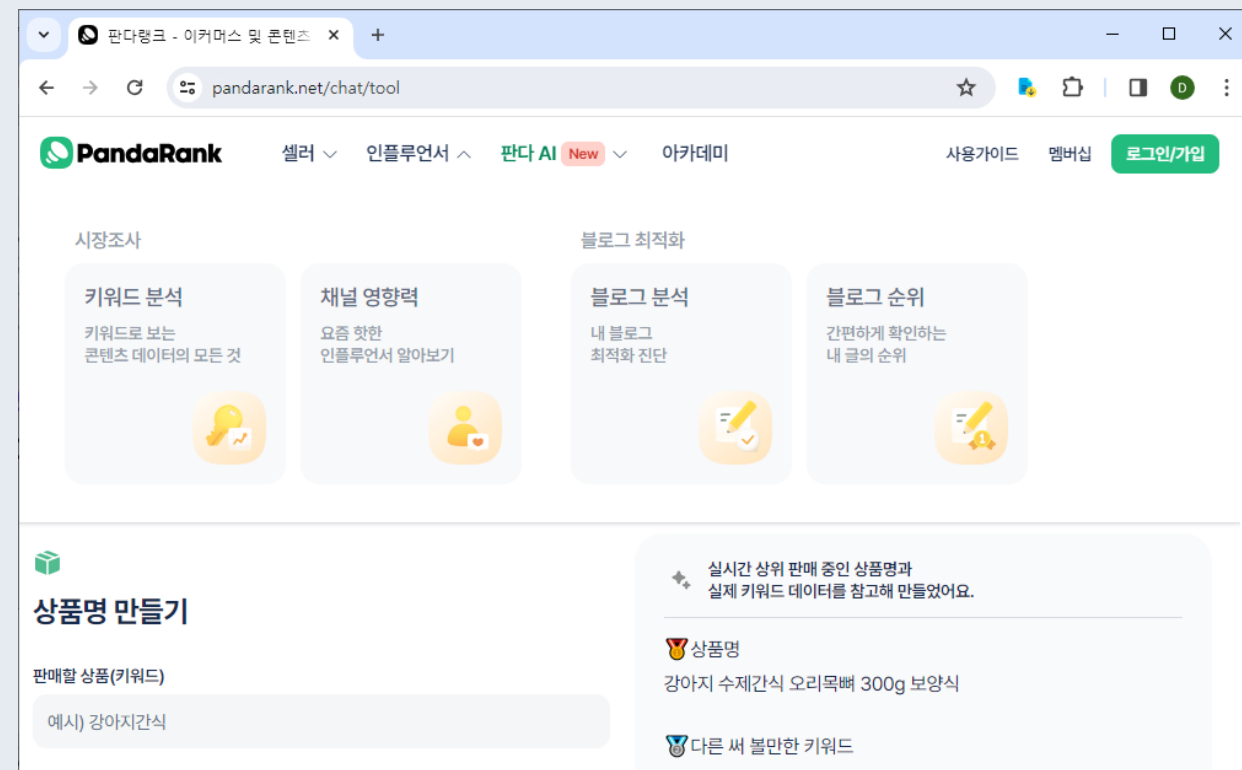
<https://www.heygen.com/>



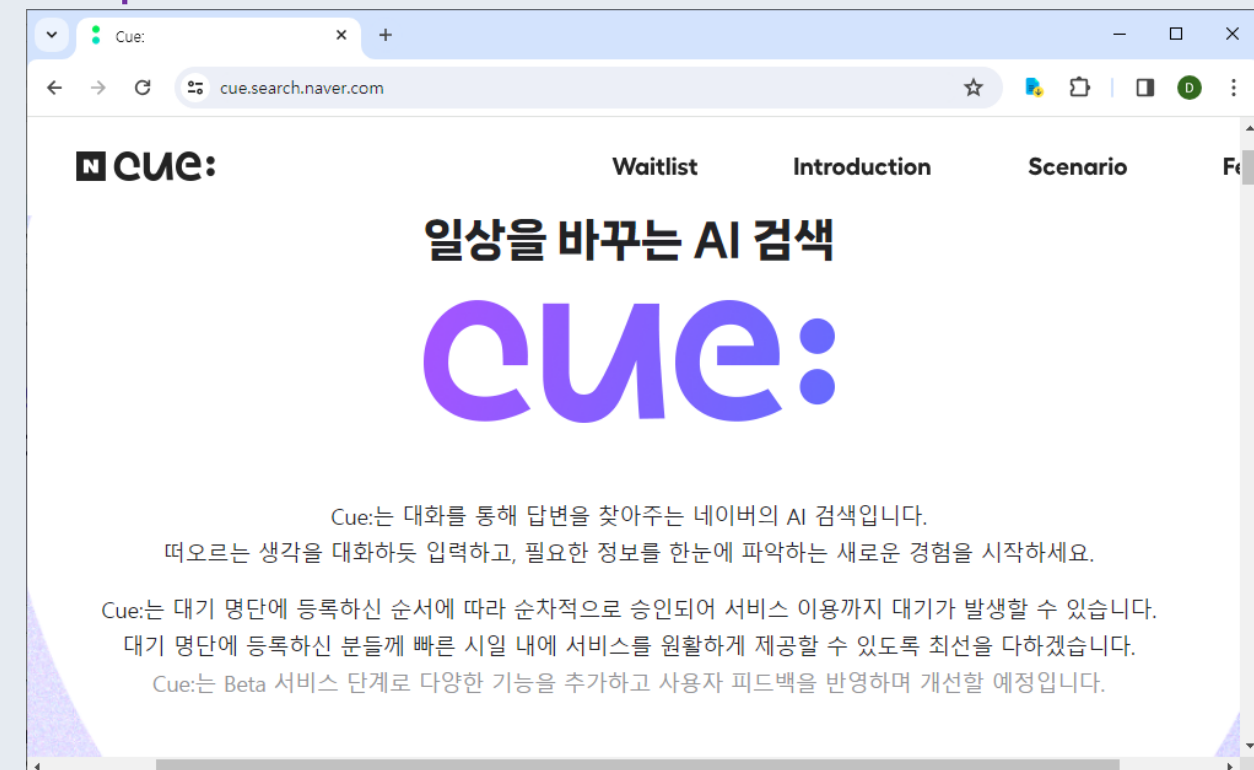
<https://elevenlabs.io/>



<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](#)

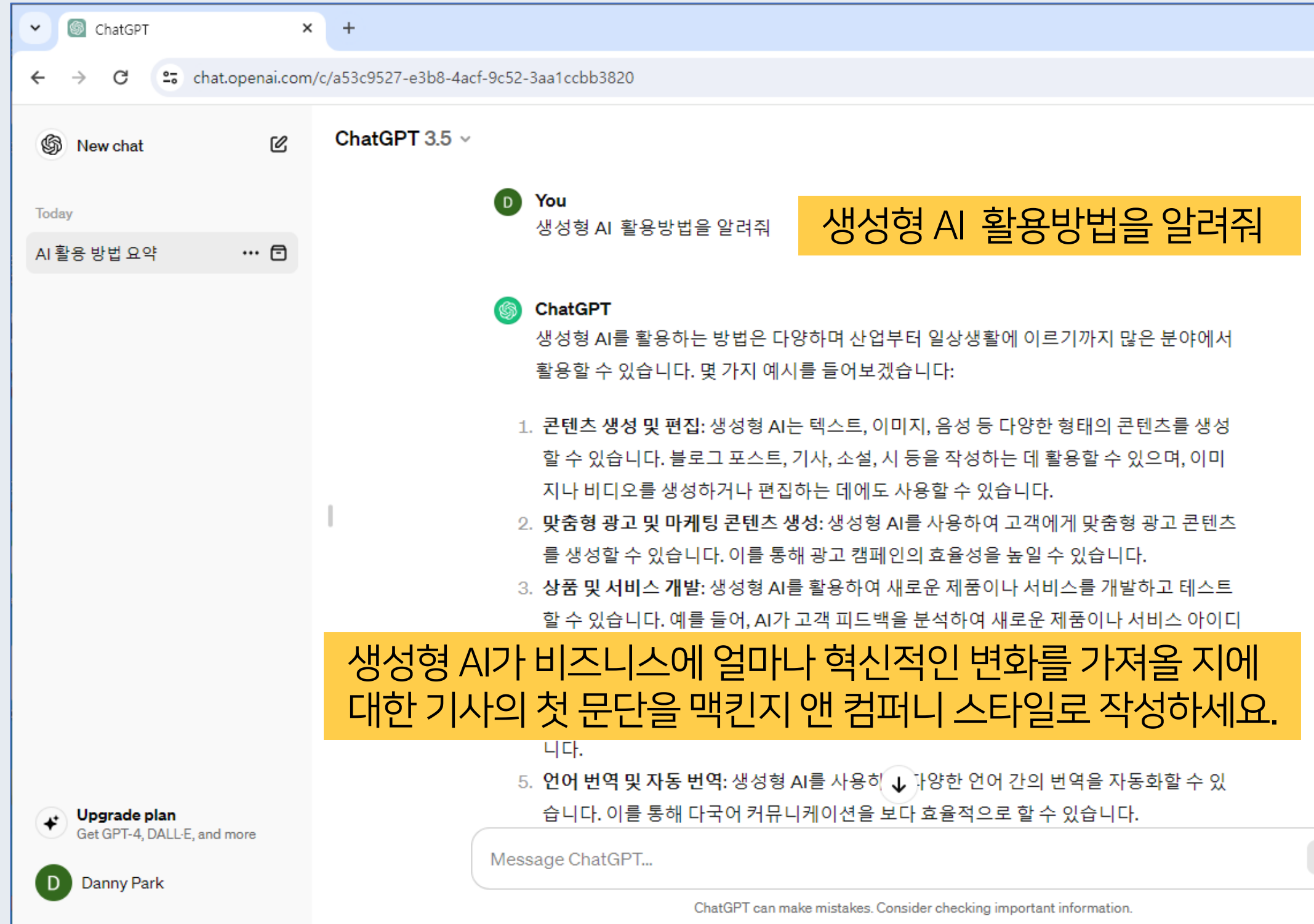
* 연간 청구, 최소 2명 사용자

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

플랜 업그레이드

Danny

실습 - ChatGPT



The screenshot shows the ChatGPT web interface. The user has asked: "생성형 AI 활용방법을 알려줘" (Tell me how to use generative AI). ChatGPT has responded with a list of three use cases:

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.

Below the list, there is a yellow highlighted box with 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 on how generative AI will bring revolutionary changes to business in the McKinsey & Company style).

At the bottom, there is a text input field with the placeholder "Message ChatGPT..." and a "Send" button. A small disclaimer at the bottom right states: "ChatGPT can make mistakes. Consider checking important information."

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

```
python
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/90.0.4431.24 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/>

문제 출처 : <https://www.examtopy.com/exams/amazon/aws-certified-machine-learning-specialty/view/14/>

다음 문제를 풀어줘. 문제의 지시문을 준수하고 가장 적합한 답을 찾아. 설명은 한국말로 해줘 :

Question #131

A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning(ML) models on confidential financial data. The company is worried about data egress and wants an ML engineer to secure the environment.

Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)

- A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink.
- C. Disable root access on the SageMaker notebook instances.
- D. Enable network isolation for training jobs and models.
- E. Restrict notebook presigned URLs to specific IPs used by the company.
- F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys.

답안 1

문제 #131: 금융 서비스 회사의 SageMaker 데이터

답안 2

정답은 다음 세 가지입니다. (Choose three)

답안 3

문제 #131: 금융 서비스 기업의 SageMaker 데이터

정답은 다음 세 가지입니다. (Choose three)

- A. VPC 인터페이스 엔드포인트 사용 (Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink)
- D. 훈련 작업 및 모델의 네트워크 격리 사용 (Enable network isolation for training jobs and models)
- E. 노트북 사전 서명 URL 제한 (Restrict notebook presigned URLs to specific IPs used by the company)

문제 설명:

- 금융 서비스 회사에서 기밀 재무 데이터를 처리하는 데 Amazon SageMaker를 사용할

여기에 프롬프트 입력



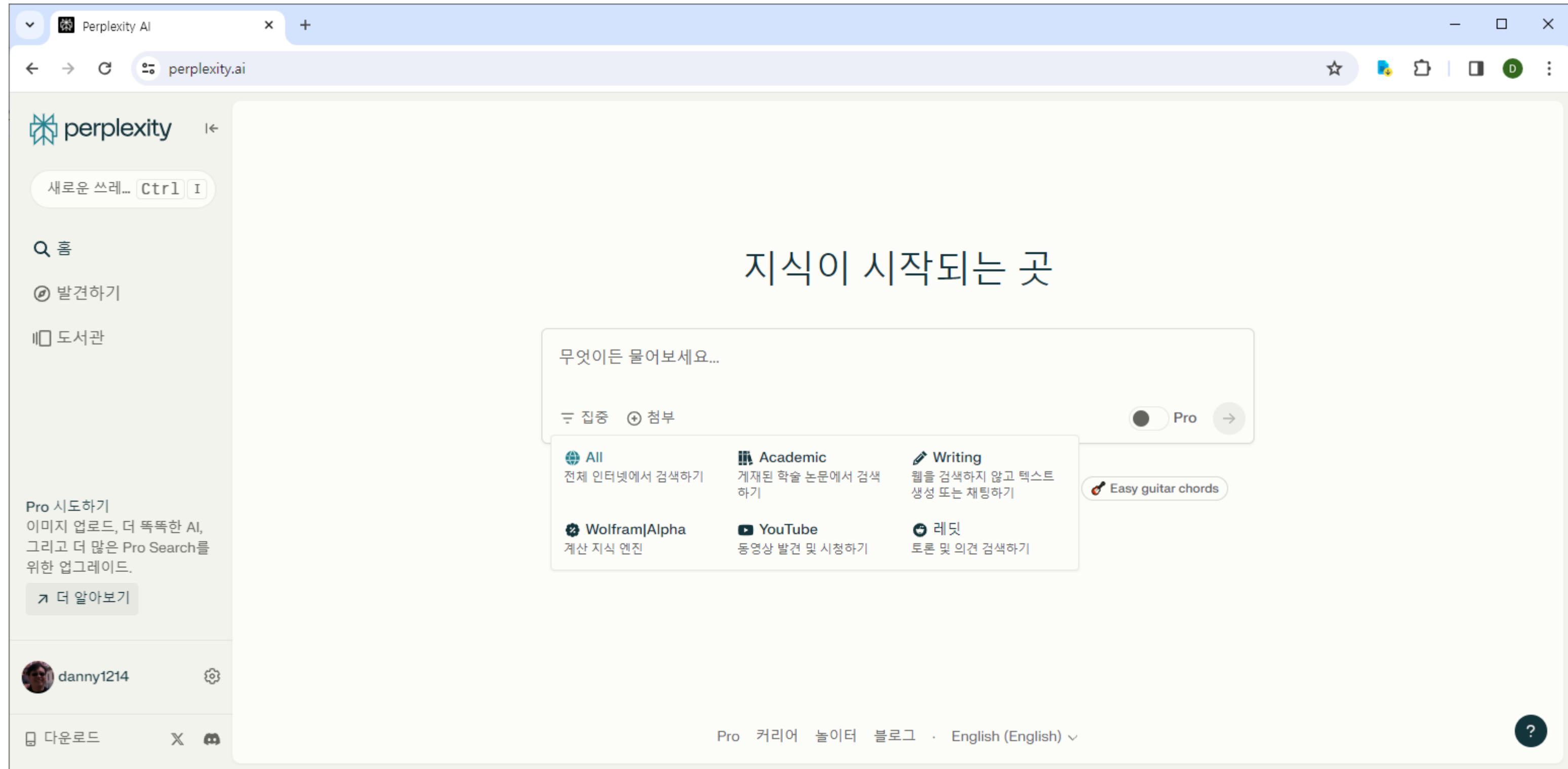
실습 - Claude

<https://console.anthropic.com/>

The screenshot displays the Anthropic Console Workbench interface. The top navigation bar includes 'Dashboard', 'Workbench' (selected), and 'Settings'. The main area is titled 'Untitled - 2024-03-06 오전 8:41:28'. On the left, the 'Model settings' panel shows 'Model' set to 'claude-3-sonnet-202', 'Temperature' at 0, and 'Max tokens to sample' at 1000. The central chat area shows a 'SYSTEM PROMPT' and a 'USER' message: '인스타그램에 "동네 산책"이라는 주제로 게시를 작성하려고 해. 일상의 기쁨, 소소한 감동, 여유, 힐링 등의 느낌이 물씬 나도록 구어체로 작성해 줘. 에피소드도 하나 추가하고, 이모지와 해시태그를 사용해.' The 'RESPONSE' section shows Claude's reply in Korean, including emojis and hashtags like #동네산책, #봄날, #꽃구경, #일상힐링, #소소한행복, and #여유로운시간. The bottom of the interface shows '6 of 6' and navigation icons.

실습 - Perplexity

<https://www.perplexity.ai/>



실습 - AI21 studio

<https://studio.ai21.com/models>

The screenshot shows the AI21 Studio interface. The top navigation bar has links for Home, Our Models (selected), and Docs. The left sidebar lists various tools and models: J2-Chat, J2-Complete, Contextual Answers, Summarization, Text Editing, More, RAG Engine, Datasets, Custom Models, and Tools. The main content area is titled 'Our Models' and displays a grid of nine model cards. Each card includes an icon, a title, a description, and a button to access the playground or documentation.

Model Name	Description	Action
Contextual Answers	Provides answers based solely on a specific context. Context can be provided as part of the request or automatically retrieved from the document library.	Playground >
Semantic Search	Understands the intent and contextual meaning of a query and retrieves the most relevant text segments from your documents.	Documentation ↗
Embeddings	Transforms texts into fixed-size vectors that capture their semantic meanings in a high-dimensional space.	Documentation ↗
Summarize	Takes a document (raw text or URL) and returns a summary of the text.	Playground >
Summarize by Segment	Segments a text into distinct topics and generates a summary for each segment.	Documentation ↗
Text Segmentation	Takes a document (raw text or URL) and returns segmentation into paragraphs by topic.	Documentation ↗
Paraphrase	Takes a piece of text and returns a list of paraphrases that convey the same meaning using different words.	Playground >
Grammatical Error Corrections	Detects grammar, spelling, punctuation, and word choice errors in text, and suggests how to correct them.	Playground >
Text Improvements	Provides rewrite recommendations for improving the fluency, clarity and style of the input text.	Playground >

실습 - sLLM

<https://ollama.com/library>



- sLLM(smaller Large Language Model , 소형 언어 모델)
 - LLM과 비교했을 때 매개변수의 수가 수십 억~수백 억개로 비교적 크기가 작은 언어모델
 - 비용절감, 보안, 특정 도메인에 활용 목적으로 사용
 - 특정 도메인 사용용도로 sLLM 을 사용하는 경우가 많아지고 있음




- Ollama 설치
 - 로컬 환경에서 다양한 언어 모델을 실행할 수 있게 지원하는 오픈소스
 - 모델 종류 : <https://ollama.com/library>
 - 설치 파일 다운로드 : <https://ollama.ai/>



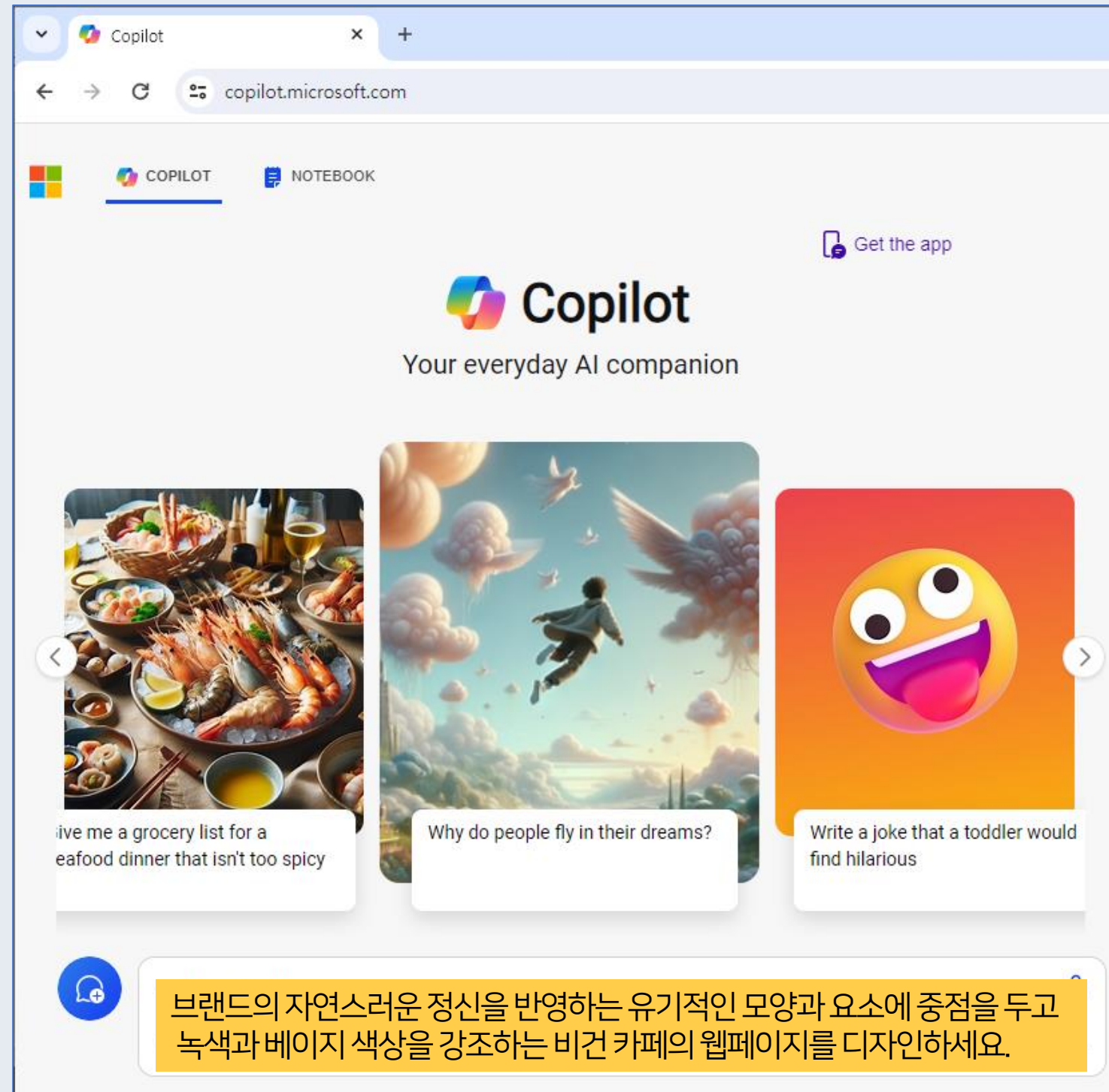
- Llama 2 (Large Language Model Meta AI)
 - 메타에서 공개한 상업적으로도 이용 가능한 오픈 소스 sLLM
 - 설치 및 실행 : ollama run llama2
 - 프로그램 개발 예시

```
from langchain_community.llms import Ollama
llm = Ollama(model="llama2")
llm.invoke("Hello")
```

`pulling 8934d96d3f08... 100%`  `3.8 GB`

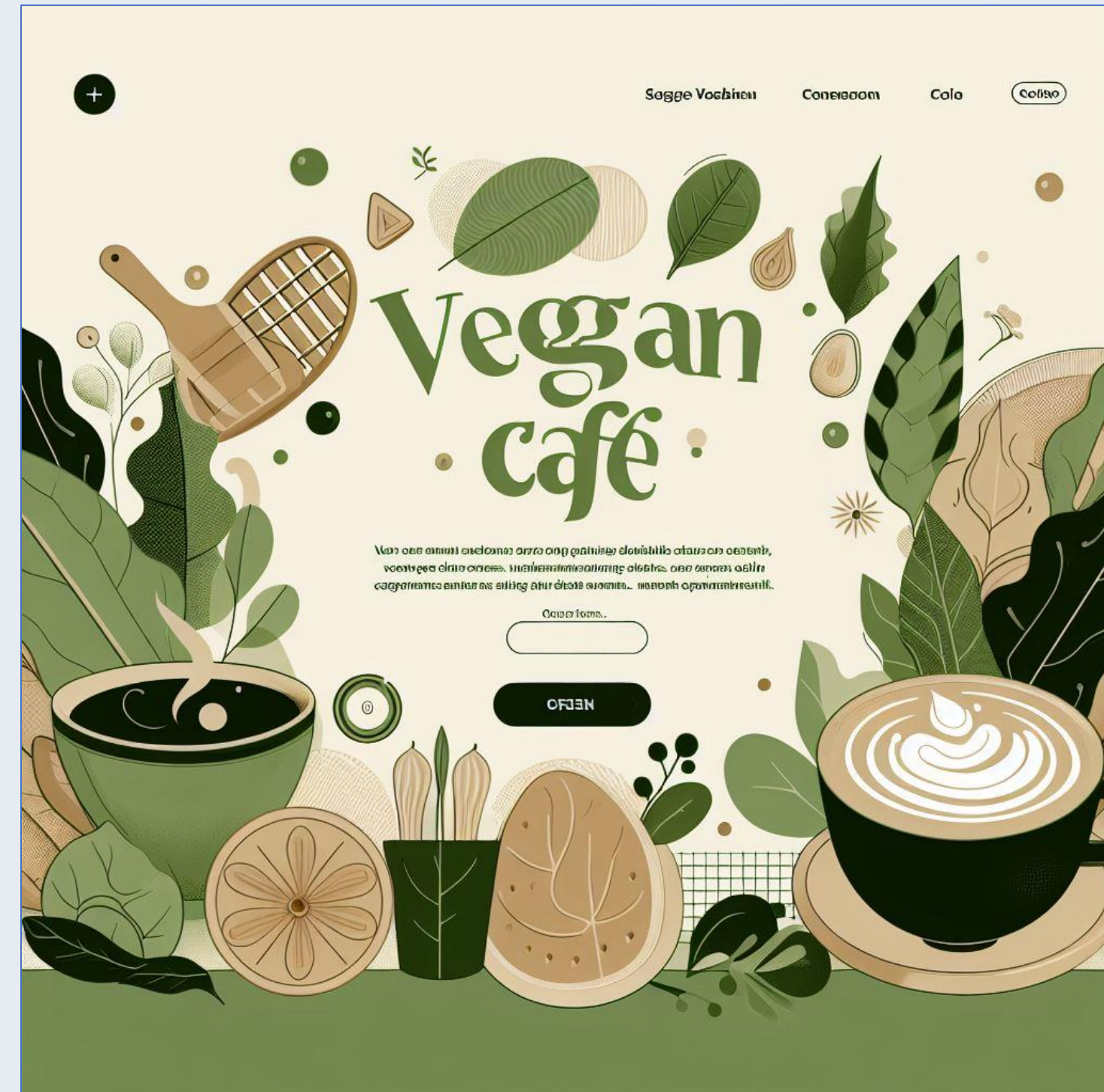
실습 - Copilot

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



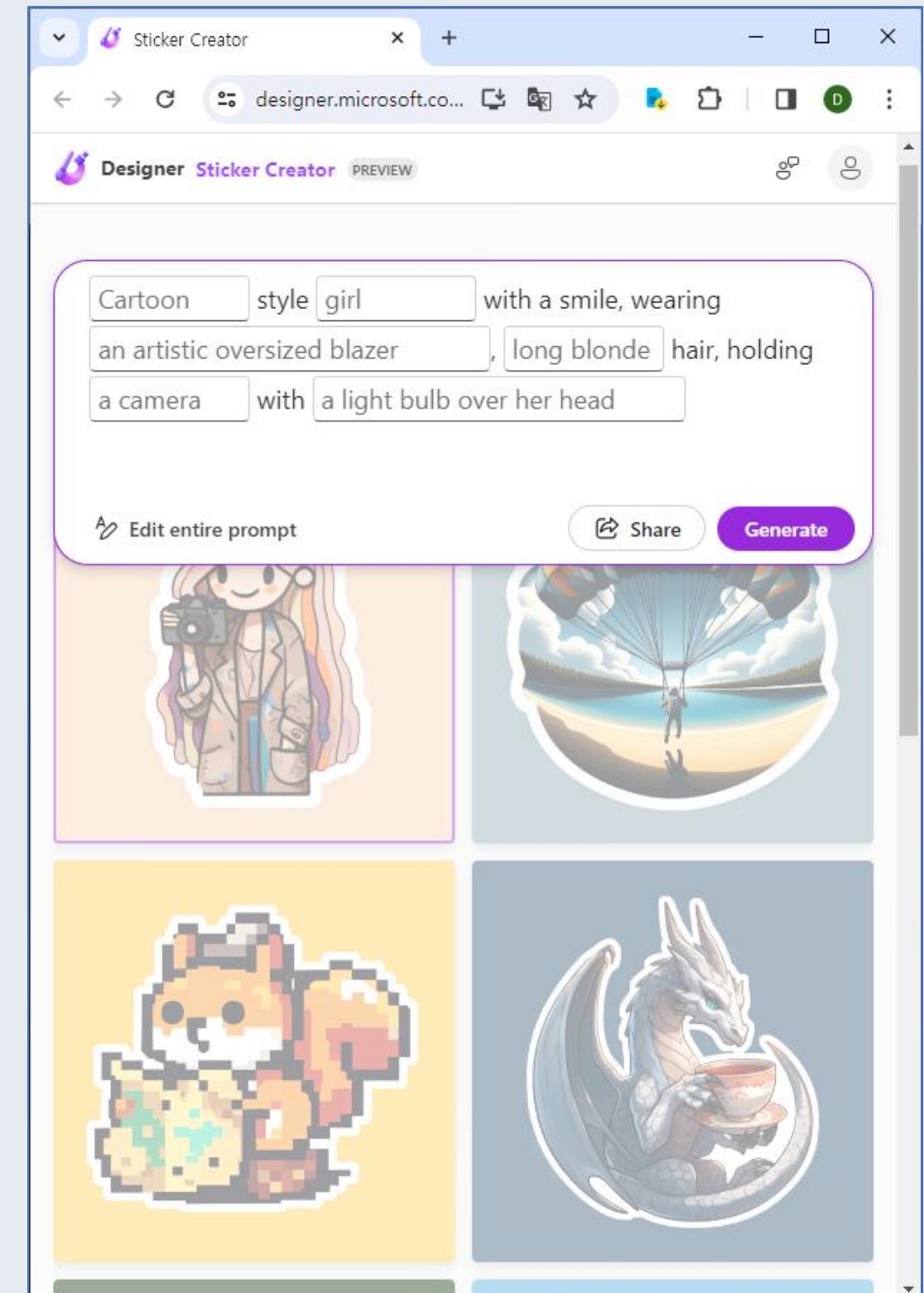
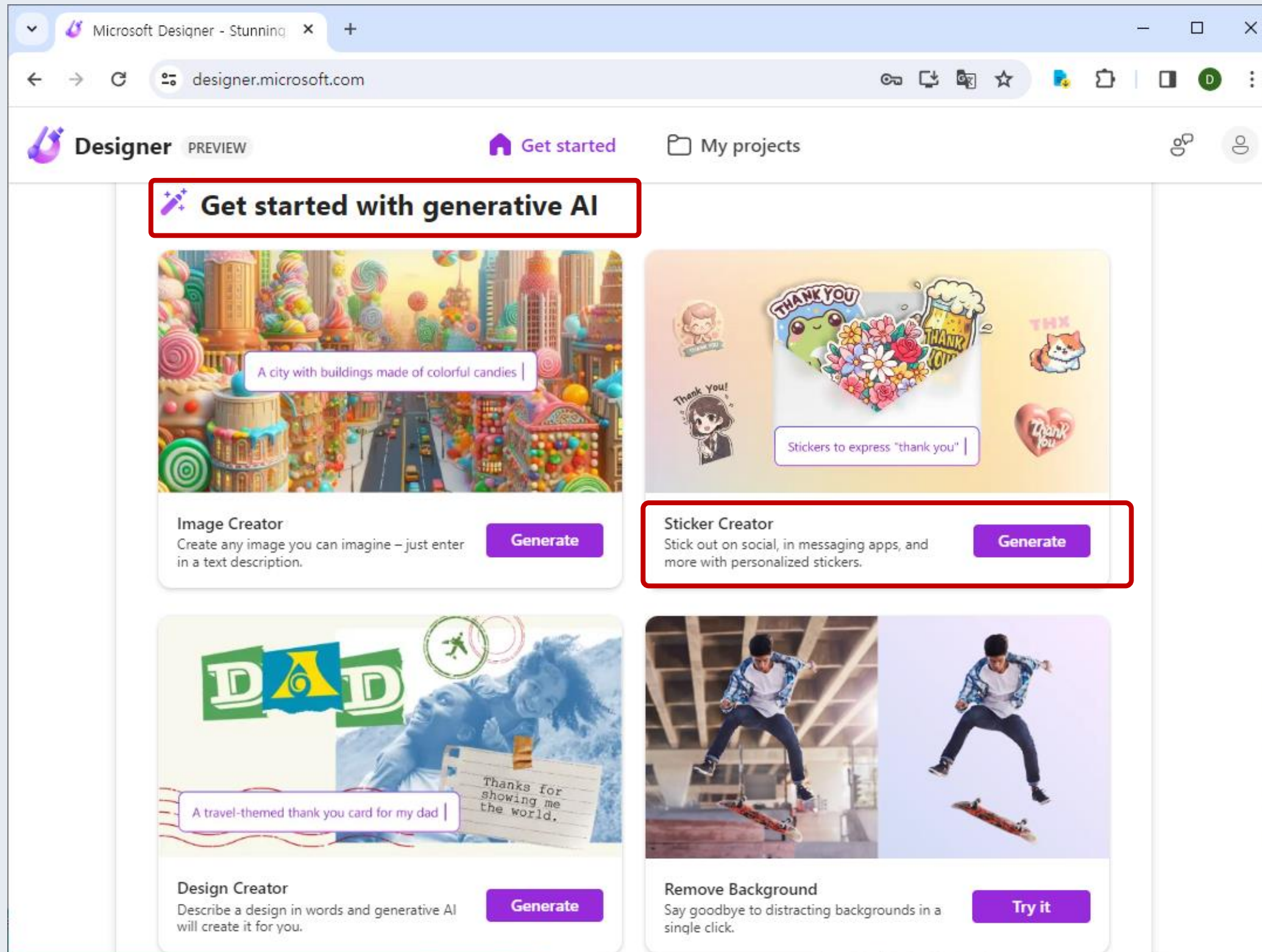
이미지 생성 프롬프트 참고

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



실습 - Designer

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





Generative AI in a Nutshell

Henrik Kniberg
Jan 2024

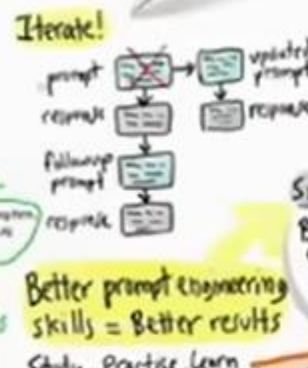
Prompt Engineering/Design

Bad prompt
Give me an agenda for a workshop

Good prompt
Give me an agenda for a workshop *task*

Good prompt
Give me an agenda for a workshop *context*

Good prompt
Give me an agenda for a workshop *questions*



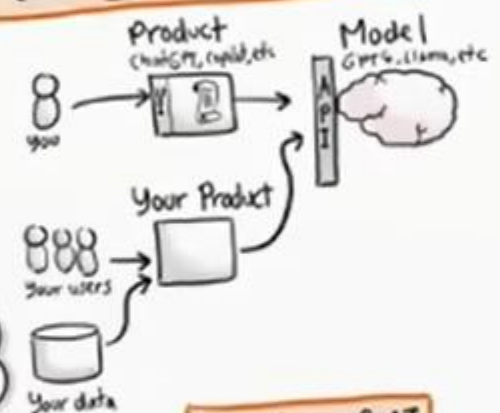
Autonomous Agents with Tools



Side effect
Better communication skills overall

Better prompt engineering skills = Better results
Study, Practice, Learn

Using vs Building AI-powered products



The role of Humans

Is human role X still needed?
doctor, developer, lawyer, cee, teacher, etc...

Decide what to ask and how

Evaluate results
Compensate for AI weaknesses

Provide context
Legal compliance
Data security
etc, etc...



Mindset



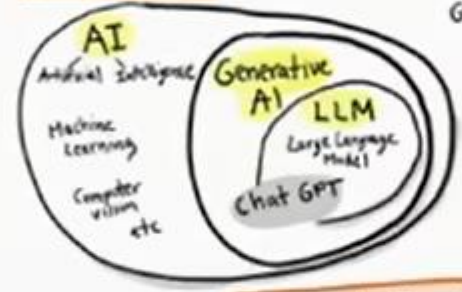
The Age of AI



Computers have gotten smarter



Terminology



GPT
Generative Transformer
Pretrained

Training



1. Unsupervised Generative Pretraining (lots + lots of text...)
Back-propagation
2. Reinforcement Learning with Human Feedback (RLHF)

Multimodal Models

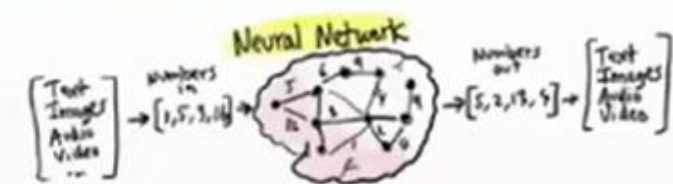


Whatever - whatever

Taking AI for a walk

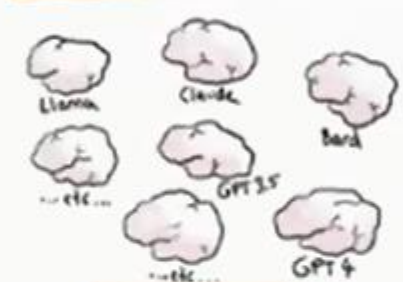


How it works



Dogs are → animals
Dogs are animals → that
Dogs are animals that are known for... (WoWo)...

Models, models everywhere



Model Types



THANK YOU