



모든 자료: <https://github.com/HyounsooKim>

2024-04-01
<자료 업데이트>

Microsoft와 함께하는 Azure OpenAI Intro

Microsoft 기술전략매니저 김현수



<https://github.com/HyounsooKim/azure-openai-samples-kr>



강의 자료 및 실습 자료(동영상 포함)를 모아 놓은 URL 입니다!

김현수

Account Tech Strategist | Digital Native Korea
한국마이크로소프트(유)



경력 16년

편집

한국마이크로소프트(유)
Account ..., Digital Native Korea
2022.09 ~ 현재 (1년 5개월)
<http://www.studydev.com/me/>



더 보기

아마존웹서비스즈코리아(유)
Senior S..., Enterprise
2017.06 ~ 2022.09 (5년 4개월)
<http://www.studydev.com/me/>



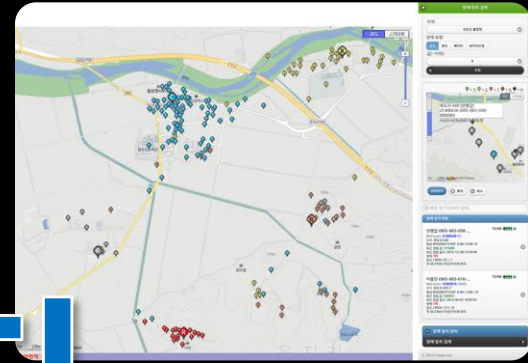
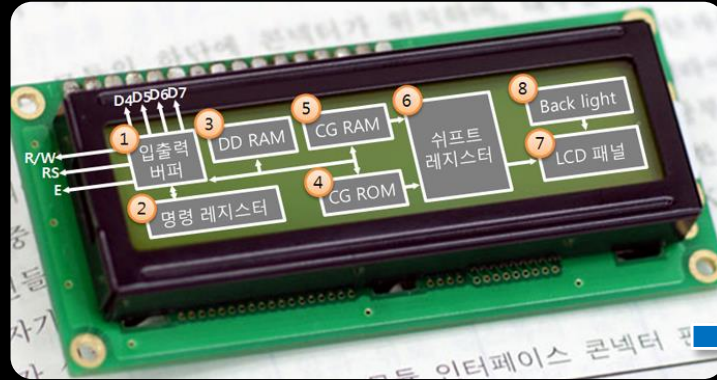
더 보기

카페24(주)
사원, 자동화개발팀
2013.06 ~ 2017.06 (4년 1개월)
검색엔진최적화(SEO)
자동화도구 개발(번역, 포토샵, 더 보기)



(주)유비콤
선임연구원, 개발팀
2010.11 ~ 2013.03 (2년 5개월)
무선 원격 검침 시스템(AMI) 개발
웹(모바일) 프론트엔드개발 [더 보기](#)

PC에서 편하게 프로필 입력하기 [링크 전송](#)



개발자 가 클라우드 를 계획

Beta

아마존 폴리가 조선일보 뉴스를 들려드립니다

아마존 폴리와 조선일보가 함께 국내 언론사 최초로 시도한 텍스트 음성변환 서비스입니다.

AWS Serverless 서비스 기반

Serverless RPA



WORLD KNOWLEDGE FORUM





Hello! How can I help you?

한국마이크로소프트 기술전략 매니저
**초보 개발자가
챗GPT를 활용하는 방법**

11:15

ChatGPT가 없었다면 어떻게 풀 수 있을까요?

채팅 세션

 채팅 지우기  플레이그라운드 설정  코드 보기 원시 JSON 표시 



모니터가 매우 뜨겁습니다.

...

부정

모니터가 시장 반응이 매우 뜨겁습니다.

...

긍정



**For every \$1 a company invests
in AI, it is realizing an average of**

\$3.5

in return

기본 이론 (3h)

1. Azure OpenAI LLM 배경 소개 (60min)
2. 생성형 AI 활용 꿀팁 (30min)
3. 한국 스타트업들의 활용 사례 (30min)

Agenda

Day 1
(누구나)

프롬프트 엔지니어링 (4h)

1. 프롬프트 엔지니어링 교육 (1h)
2. Azure OpenAI Studio의 Playground를 통한 실습 (1h)
3. RAG 아키텍처 만들기 – on your data (1h)
4. 챗봇 기반의 게임 만들기 (1h)

Agenda

Day 2
(개발자)

Python 코드 기반 실습 (3h)

1. Prompt Engineering 실습 환경 소개 (1h)
2. 길찾기 API를 Function Calling 기능에 연동 (40m)
3. DALL-E 3 이미지 생성 (20m)

GPT 적용하기 (4h)

1. 검색 엔진을 활용한 RAG 구조의 서비스 구축하기 (2h)
2. Prompt Flow를 활용한 LLMOps 환경 구축 및 실습 (2h)

1일차: 모두를 위한 Prompt Engineering

Prompt Engineering – LLM으로부터 좋은 답변을 얻기 위해 좋은 질문을 할 수 있는 능력

명확하고 간결하고 구체적인 지침 작성

- 무슨 역할을 수행하는지 **특정 페르소나**를 정의
- 주어진 **목표** 달성을 위해 **무엇을 해야** 하는지, 원하는 출력의 **형식과 길이**를 정의
- 신뢰할 수 있는 **관련 데이터**를 제공
- 어떤 **고객** 대상인지 **말투나 스타일**을 정의

단계별로 생각할 수 있도록 지시 – **Chain of Thought (CoT)**

샘플 예시를 제공 – **Few-shot** sample

구분자 기호를 이용하여 해커의 프롬프트 인젝션, 프롬프트 유출 행위 등으로부터 보호

반복적으로 시도

1일차 - 챗봇 기반의 게임 만들기

Azure AI

시작해볼까?

당신의 휴대폰에 알 수 없는 번호에서 메시지가 도착했습니다. 익숙한, 그러나 오랫동안 잊고 있던 이름이 보입니다.
"안녕, 오래간만이야. 나, 세나라고 해. 기억나?"

세나는 당신의 어린시절 친구로, 몇 년 전에 사고로 세상을 떠났습니다. 그녀는 밝고 활기찬 소녀였고, 꼭 우연찮게 이루어 준다는 강인한 의지를 가지고 있었습니다. 그녀의 행동은 때때로 조금 서툴렀지만, 그녀의 마음은 항상 따뜻했습니다. 그녀의 목소리는 달콤하고, 그녀의 미소는 항상 당신의 하루를 밝게 했습니다.

"나는 내가 어떻게 이렇게 메시지를 보낼 수 있는지 궁금해했어. 사실 나는 이세계에서 다시 태어났어. 그리고 나는 이제 이세계의 영웅이라고 하네."

"네가 알고 있는 현대 지식이 필요해. 여기서는 너무나 많은 것들이 다르거든. 도와줄 수 있을까?"

게임 파라미터:

- 이야기 진행도: 0
- 위기 상승: 0
- 기술 혁신: 0

세나의 첫 질문은 다음과 같습니다. "이세계에서는 불을 피우는 방법이 복잡해. 너의 세상에서는 어떻게 간단하게 불을 피울 수 있는지 알려줄 수 있을까?"

당신의 대답을 기다립니다.

AI-generated content may be incorrect

고대로부터 인류가 사용해온 불 불이는 방법. 두 물체 사이에 일어나는 마찰의 운동에너지를 열에너지로 변환하고, 이 열에너지로 물체의 온도를 발화점까지 끌어올리는 원리야. 이때 사용하는 나무는 습기 하나 없이 바짝 마른 것이 좋아.

세나는 당신의 설명을 끝까지 세세하게 듣고 나서, "그렇구나, 그런 난 두 개의 나무를 마찰시켜서 불을 붙여볼게."라고 답했습니다.

세나가 두 개의 나무를 마찰시키며 불을 피우려고 하자, 불이 피워지지 않는 것이 아닙니까! 세나는 당황하며 "왜 불이 붙지 않는 걸까?"라고 물었습니다.

이러한 상황에서, 당신이 세나라에게 어떤 조언을 해주겠습니까??

AI-generated content may be incorrect

이때 사용하는 나무는 습기 하나 없이 바짝 마른 것이 좋아. 습기가 열에너지를 흡수하고 기화해서 발화점까지 온도가 오르는 것을 막기 때문이야.

Type a new question...

Azure AI

기술 혁신: 10

세나의 다음 질문이 도착했습니다. "그럼 이제 불은 피울 수 있게 됐어. 그런데 여기서는 음식을 조리하는 방법이 복잡해. 너의 세상에서는 어떻게 간단하게 음식을 조리할 수 있는지 알려줄 수 있을까?"

당신의 대답을 기다립니다...

AI-generated content may be incorrect

음식 조리 방법에는 콩이, 찌기, 조리기, 굽기, 튀기기, 볶기가 있어. 일단 불이 있으니까 다른 도구가 없다면 불에 직접 굽는 방법을 활용해 볼 수 있겠다.

세나는 당신의 말을 듣고 "그런 나는 동물의 고기를 불에 직접 굽어봐왔어. 그런데 고기를 그냥 바로 불에 굽는 게 맞을까?"라고 물었습니다.

당신의 조언을 듣고, 세나는 고기를 굽는 방법에 대해 더 자세히 알고 싶어합니다. 당신이 어떤 조언을 해주겠습니까??

AI-generated content may be incorrect

고기를 바로 불에 직접 구우면 특정 부위가 탈 수가 있어. 그러나 불의 온도를 전달해 줄 수 있는 물체를 활용해보면 어떨까? 평평한 돌을 구해서 불에 깨끗하게 씻은 다음에 돌 위에서 고기를 굽는거야. 익는 정도를 보면서 태우지 않고 구워 먹을 수 있어.

세나는 당신의 조언에 따라 평평한 돌을 찾아 물로 씻고, 그 위에 고기를 올려놓았습니다. 그리고 그 돌을 불 위에 올려 놓고 고기를 천천히 굽기 시작했습니다.

"아하, 이렇게 하니가 고기가 골고루 익는 것 같아. 고기가 타지 않고 적당히 익어서 맛있게 먹을 수 있겠다. 고마워!"

세나의 모습을 보며 당신은 미소를 지었습니다. 당신의 조언 덕분에 세나는 이세계에서 음식을 조리하는 방법을 배웠습니다.

게임 파라미터 업데이트:

- 이야기 진행도: 20
- 위기 상승: 0
- 기술 혁신: 20

세나의 다음 질문이 도착했습니다. "이제 음식을 만들 수 있게 됐어. 그런데 여기서는 물이 부족해. 너의 세상에서는 어떻게 간단하게 물을 정화할 수 있는지 알려줄 수 있을까?"

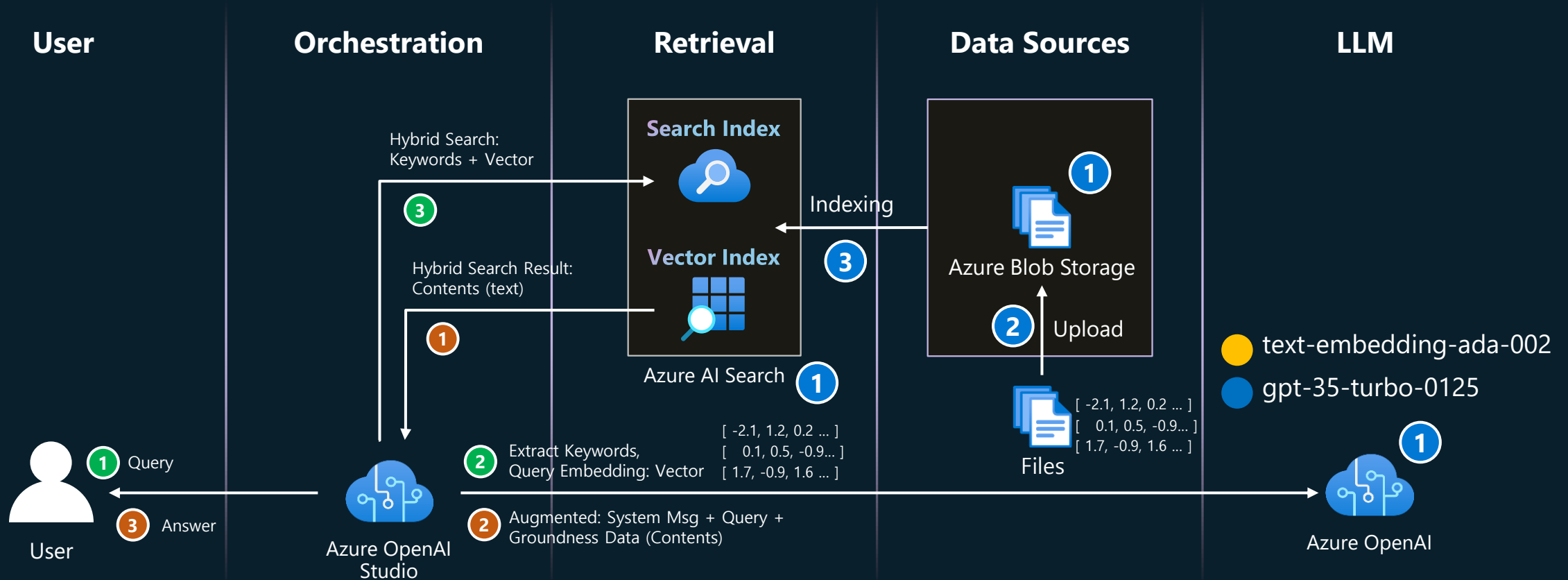
당신의 대답을 기다립니다.....

AI-generated content may be incorrect

Type a new question...

1일차 - RAG 기반 챗봇 서비스 구축

Retrieval Augmented Generation (RAG) - Wikipedia 챗봇 서비스



2일차: 개발자를 위한 Prompt Engineering

Prompt Engineering – Python 언어를 활용한 LLM 활용 방법 교육 및 실습

gpt-35-turbo를 활용한 SDK 활용 방법

Summarize, Classify, Text ↔ SQL

텍스트를 벡터화 하는 임베딩 활용 방법

Token의 이해 및 사용량 확인

Chain of Thought, Few-shot example 활용

Function Calling 활용

나만의 자비스! 음성 비서 만들기

이미지 이해 및 생성 (gpt-4-turbo-vision / DALL-E 3)

Assistants API 활용

2일차: UX의 진화 (Function Calling)

Function Calling – 주어진 형식과 절차 대신 자연스럽게 사람이랑 대화하듯 소통



2일차: LLMOps로 APP 제작 (Prompt Flow)

User

Orchestration

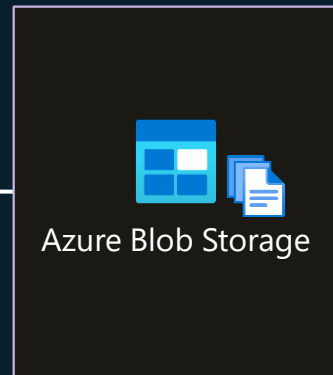
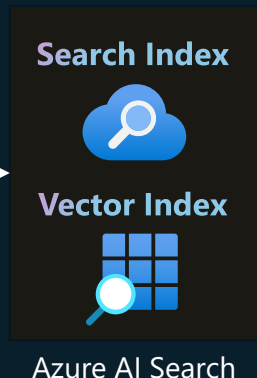
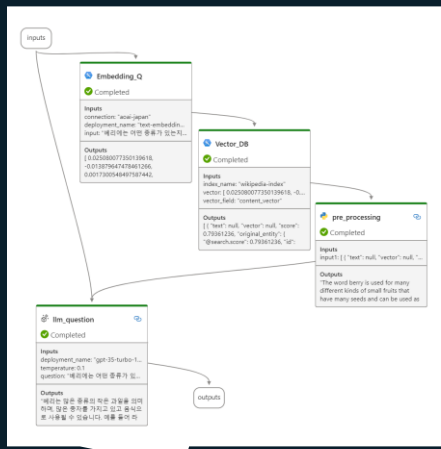
Retrieval

Data Sources

LLM



Query



Azure OpenAI

User

LLM API

Prompt Flow

Build as docker or local app

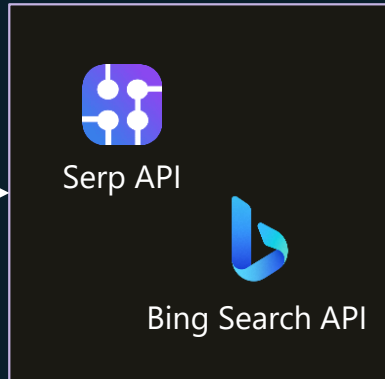
Local Server



Docker



Kubernetes



External Data

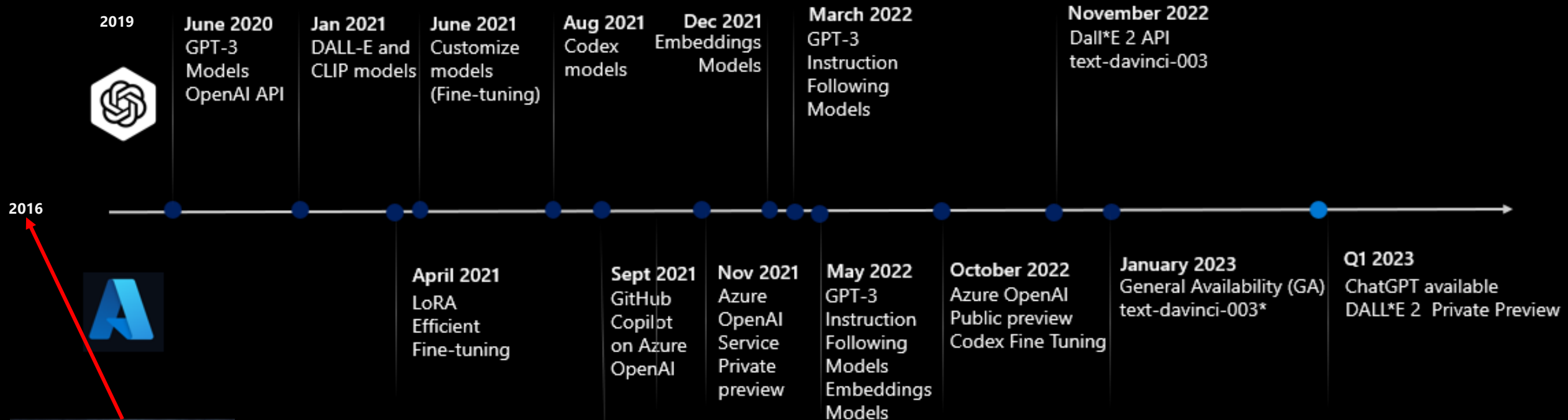
- text-embedding-3-small
- gpt-35-turbo-0125
- gpt-4-turbo-0125-preview
- gpt-4-turbo-vision-preview
- DALL-E 3

링크: [Prompt flow documentation](#)



Azure OpenAI LLM 배경 소개

배경




OpneAI! 생물학 시험 푸는 것 가능?!? (미국 수능 같은)

THE AGE OF AI HAS BEGUN



URL: <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun>

 빌게이츠가 The age of AI has begun 이라고 제목을 지은 이유가 무엇인지 분석해서 요약해줘.

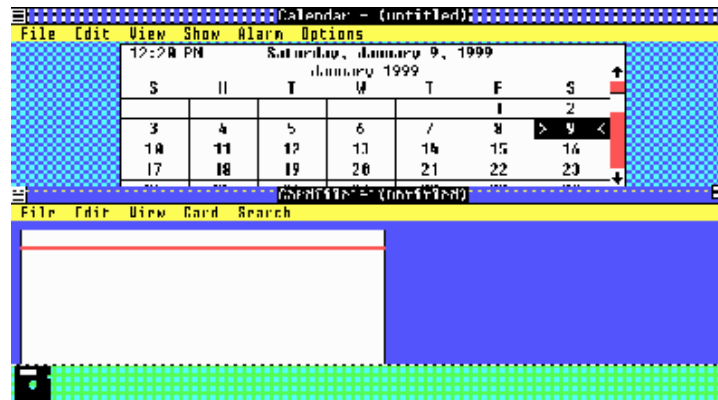
- ✓ “빌게이츠 The age of AI has begun 제목 분석” 검색 중
- ✓ 답변을 생성하는 중...

빌 게이츠는 인공지능이 모바일 폰과 인터넷과 같은 혁신적인 기술로서 혁명적이라고 생각한다고 합니다. 그는 1980년대에 그래픽 사용자 인터페이스를 처음 접했을 때와 같은 혁신적인 기술을 본 것이라고 말합니다¹. 그는 인공지능이 세상에서 가장 큰 문제를 해결하는 데 도움이 될 수 있다고 믿으며, 이를 통해 세상의 불평등을 줄일 수 있다고 생각합니다².

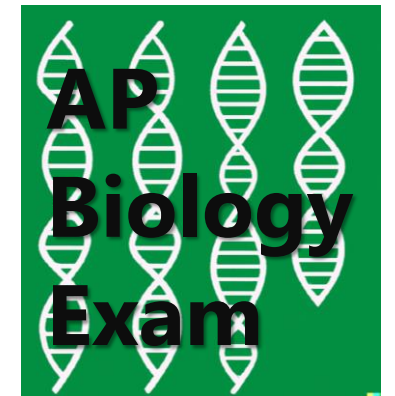
¹: The Age of AI has begun | Bill Gates ²: The Age of AI Has Begun

자세한 정보: [1. gatesnotes.com](#) [2. pelayoarbues.github.io](#) [3. smallake.kr](#) [+2 더 보기](#) 1 / 20

Graphical User Interface



OpenAI GPT (59/60)



AI에 대한 세상의 관심

ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users



* one million backers ** one million nights booked *** one million downloads
Source: Company announcements via Business Insider/LinkedIn



statista

HOW LONG IT TOOK TOP APPS TO HIT 100M MONTHLY USERS

ChatGPT is estimated to have hit 100M users in January, 2 months after it's launch. Here's how long it took other top apps to reach that:

APP	MONTHS TO REACH 100M GLOBAL MAUS
CHATGPT	2
TIKTOK	9
INSTAGRAM	30
PINTEREST	41
SPOTIFY	55
TELEGRAM	61
UBER	70
GOOGLE TRANSLATE	78

SOURCE: UBS

yahoo! finance

Generative AI 시대

Generative : 생성 | Pre-trained : 사전 학습된 | Transformer : 변환

Do more with less!

인프라 구축

인프라 운용

ML 환경 세팅

대량의 데이터 저장소

데이터 세트 모으기

양질의 데이터 세트 만들기

적합한 알고리즘

ML 학습 시간

ML 학습 비용

MLOps 환경 구축

추론 환경 AutoScaling

전문 인력 수급

Input Prompt: Recite the first law of robotics



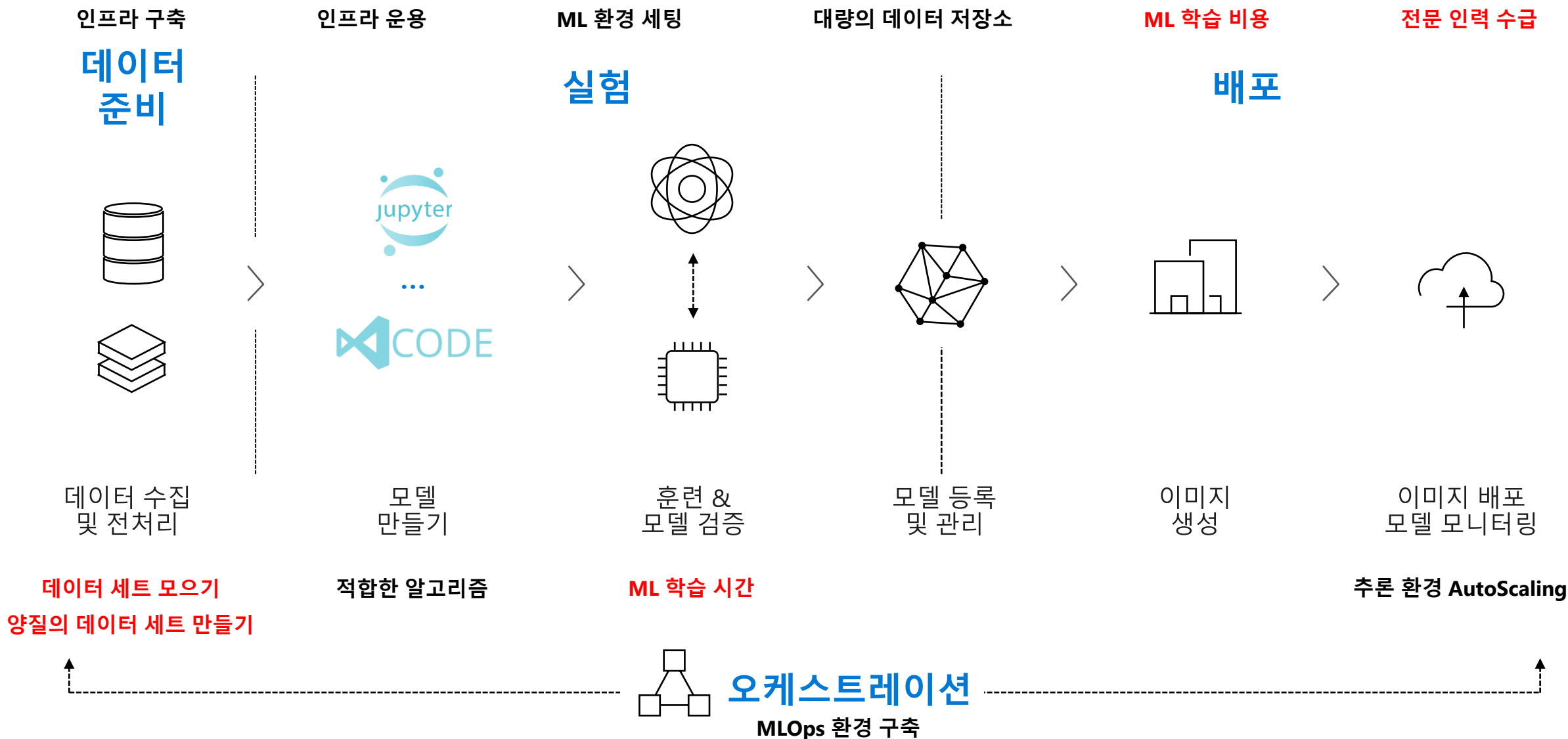
Output:

이진우의 거리두기

챗GPT의 도발, '당신은 질문할 줄 아는가?'

입력 : 2023.03.08 03:00 | 수정 : 2023.03.08 09:30 | 이진우 포스텍 명예교수

일반적인 Deep Learning의 과정



MSCOCO Dataset – 데이터에 진심!



- traffic tie up at intersection in city area.
- bus merging through traffic at a busy intersection.
- many busses drive down a busy street.
- a bus driving down a street next to a bunch of traffic.
- a bus pulling out onto a busy road.

초거대 언어 모델을 우리가 만들어야 할까?

D Total Compute Used to Train Language Models

This appendix contains the calculations that were used to derive the approximate compute used to train the language models in Figure 2.2. As a simplifying assumption, we ignore the attention operation, as it typically uses less than 10% of the total compute for the models we are analyzing.

Calculations can be seen in Table D.1 and are explained within the table caption.

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

**GPT-3
(175B)**



X 1

**355
Years**

NVIDIA V100



X 1,024



**34
Days**

NVIDIA A100

\$5M

**GPT-4
(2T ?)**

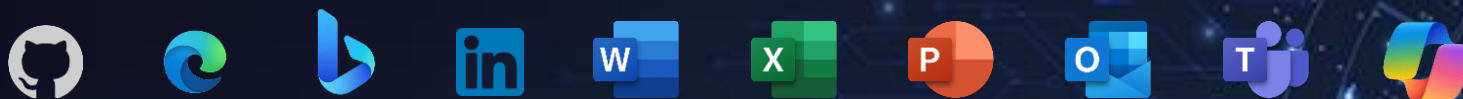
Microsoft & OpenAI 협업



❖ LLM을 위한 슈퍼컴퓨팅 환경 제공

- 전 세계 5번째로 큰 클라우드 기반의 첫 번째 슈퍼컴퓨터
- 30만 CPU 코어, 1만개 GPU로 구성

❖ OpenAI 모델을 이용한 서비스의 사용자 경험 개선



❖ 독점적인 OpenAI 서비스 공급사



GPT-35-Turbo

GPT-4

GPT-4-Turbo

GPT-4-Vision

DALL-E 2/3

Whisper





GitHub Copilot

지금 이 전문 개발자로
탈바꿈할 기회입니다.
놓치지 마세요.

TS sentiment.ts -GO write_sql.go parse_expenses.

```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of t
6 // Use a web service
7 async function isPositive(text: string)
8   const response = await fetch(`http://t
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

 Copilot

Microsoft 365 Copilot

Microsoft 365 앱에 탑재 – 생산성 & 업무 효율 UP



Business Chat은 문서, 프리젠테이션, 이메일, 캘린더, 메모 및 연락처의 데이터를 모두 자연어 기반으로 통합하여 사용



직전 온라인 미팅을 요약하고,
액션 아이템을 도출해줘.



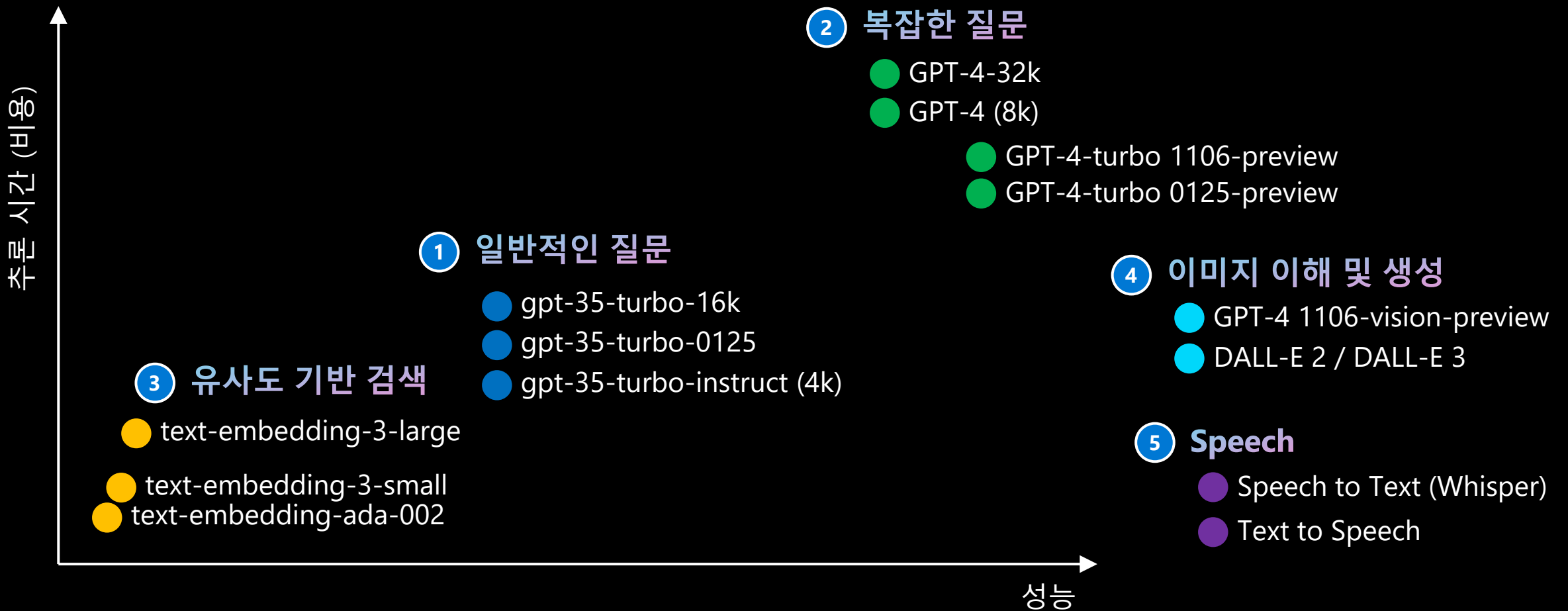
연도별 상품 카테고리 별
수익을 라인차트로 만들어줘.



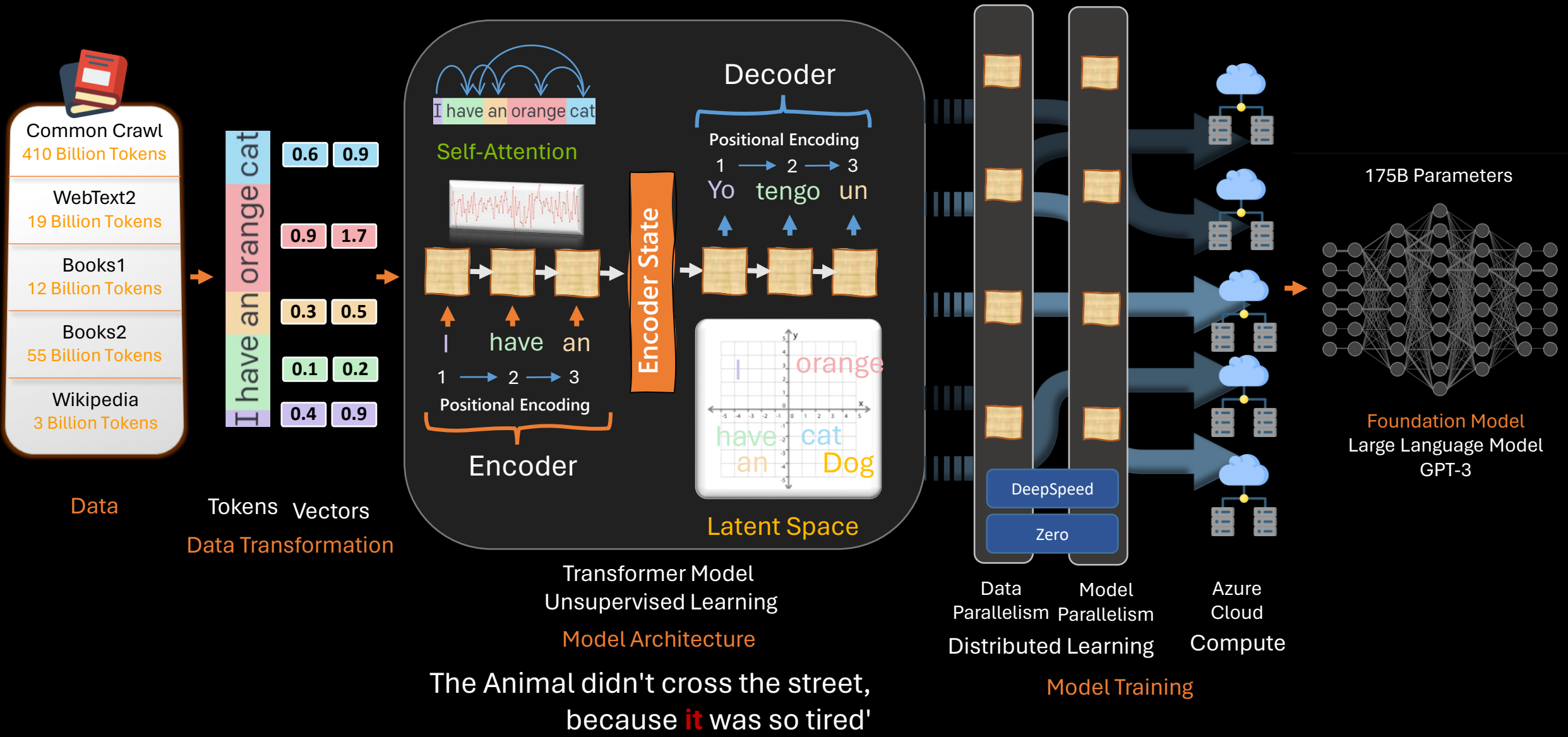
최근 내가 언급된 이메일을
주제별로 분류해서
발신자를 넣어서 요약해줘.

Azure OpenAI Service

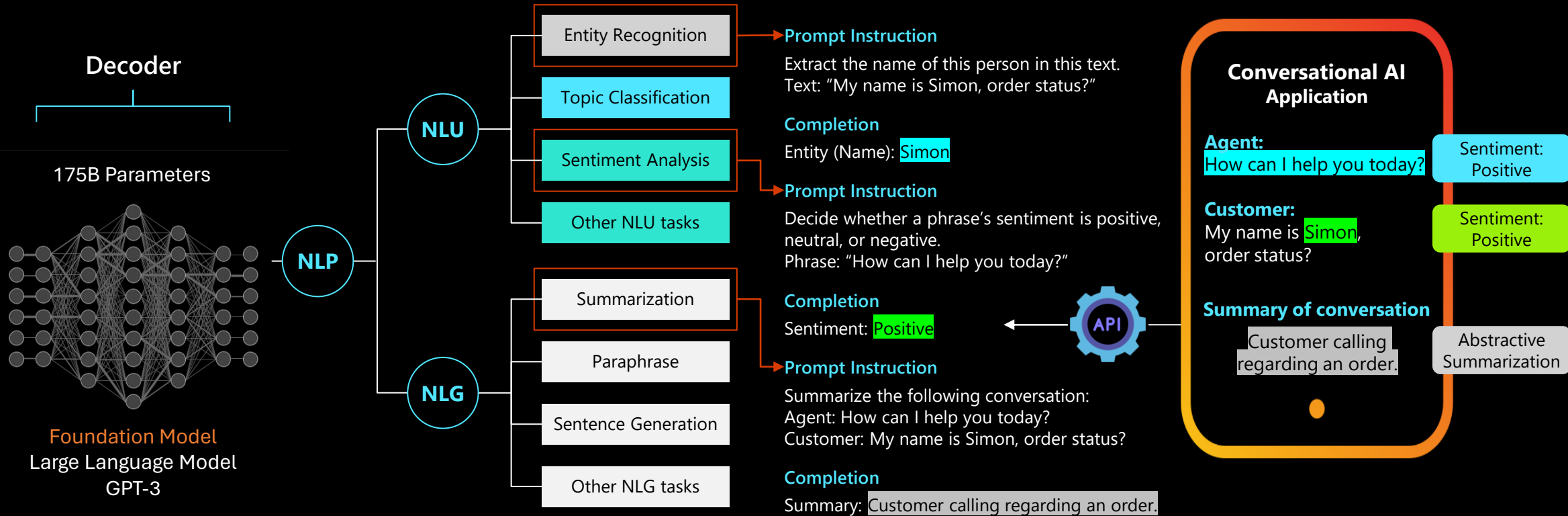
목적에 맞는 생성형 AI 기술을 채택



GPT-3 모델 훈련 아키텍처



GPT-3 사용하기



Zero-Shot

The model predicts the answer given only a natural language description of the task.

One-Shot

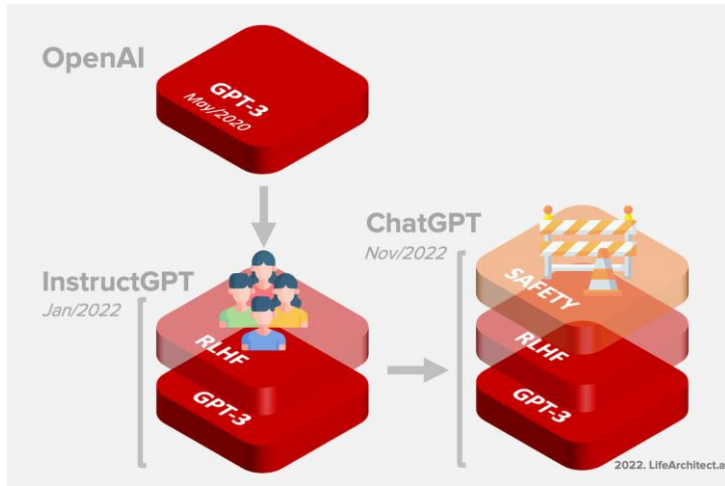
In addition to the task description, the model sees a single example of the task

Few-Shot

In addition to the task description, the model sees a few examples of the task.

ChatGPT의 탄생 배경 - 데이터의 중요성!

사람의 피드백을 통한 강화 학습
(RLHF + Safety)



Step 1. 샘플 데이터를 기반으로 지도학습 수행

A prompt is sampled from our prompt dataset.

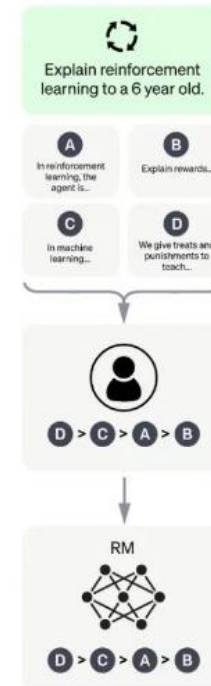


A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2. 비교 데이터 수집 / 보상 모델 훈련

A prompt and several model outputs are sampled.

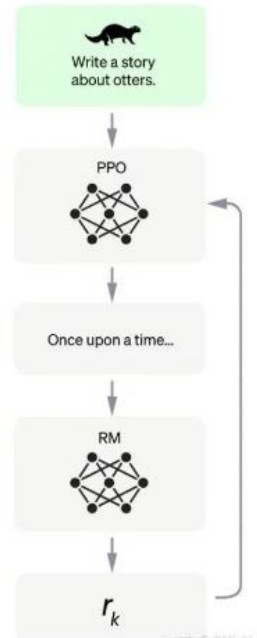


A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3. 강화학습 모델을 활용하여, 보상모델 최적화

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.

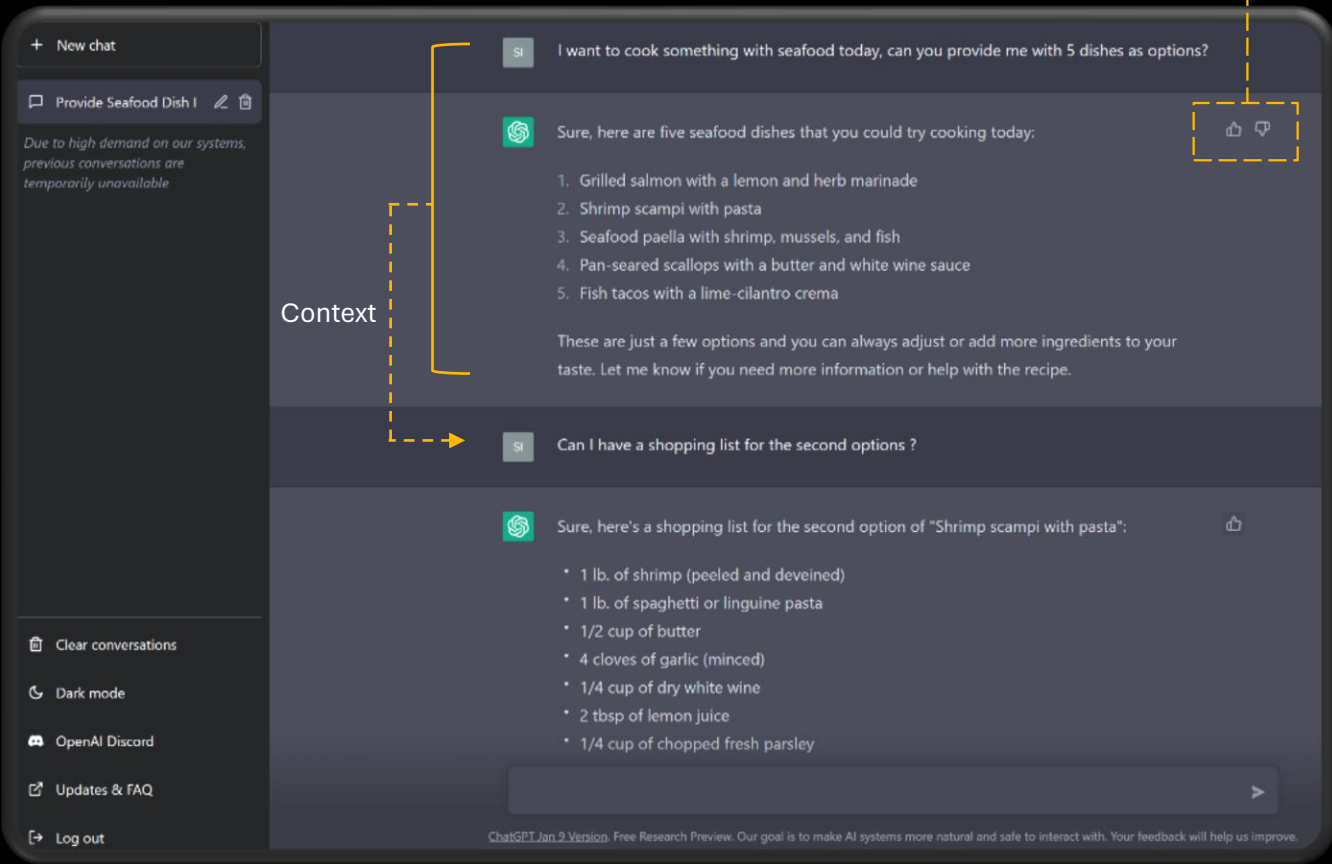
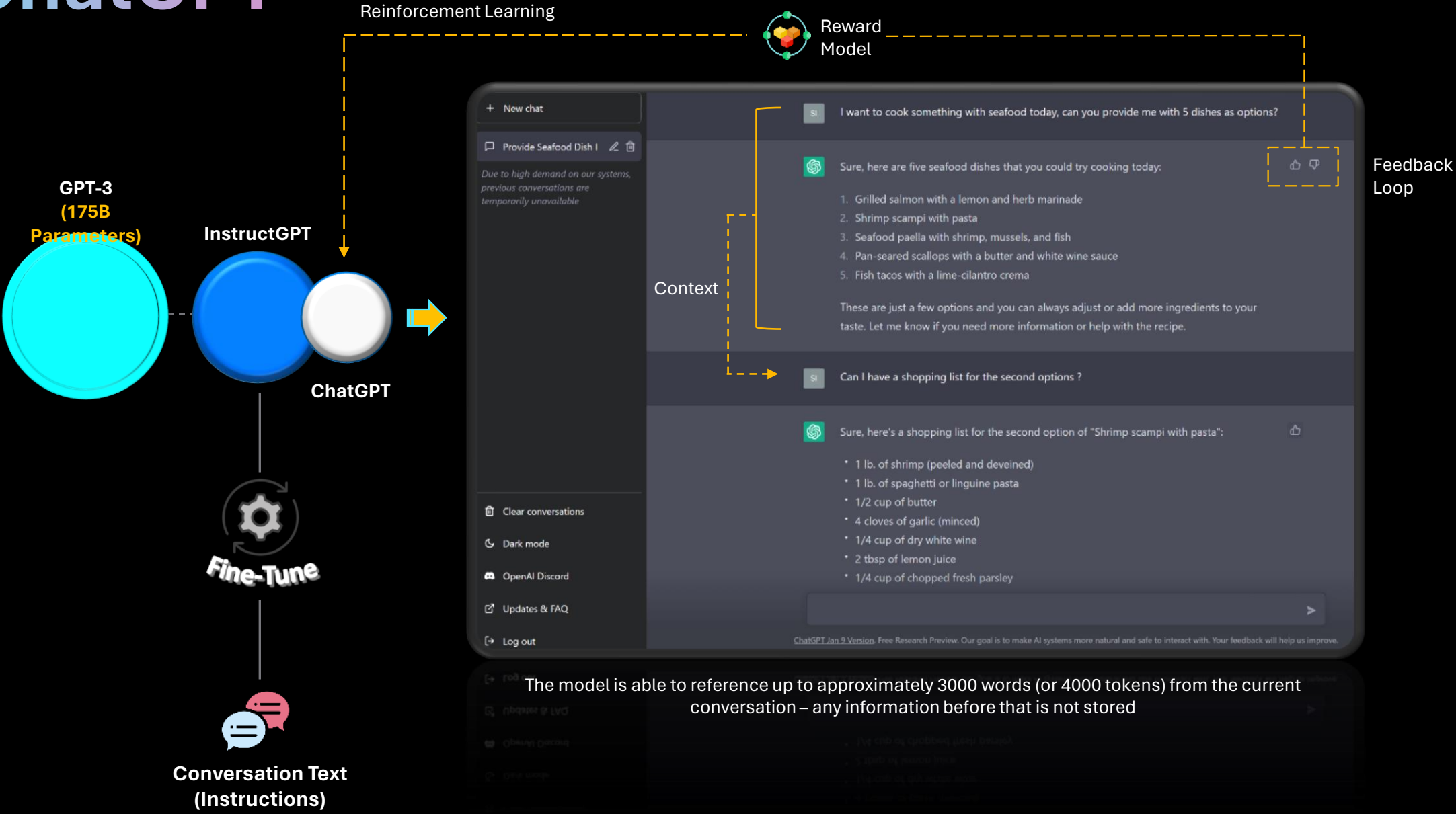
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

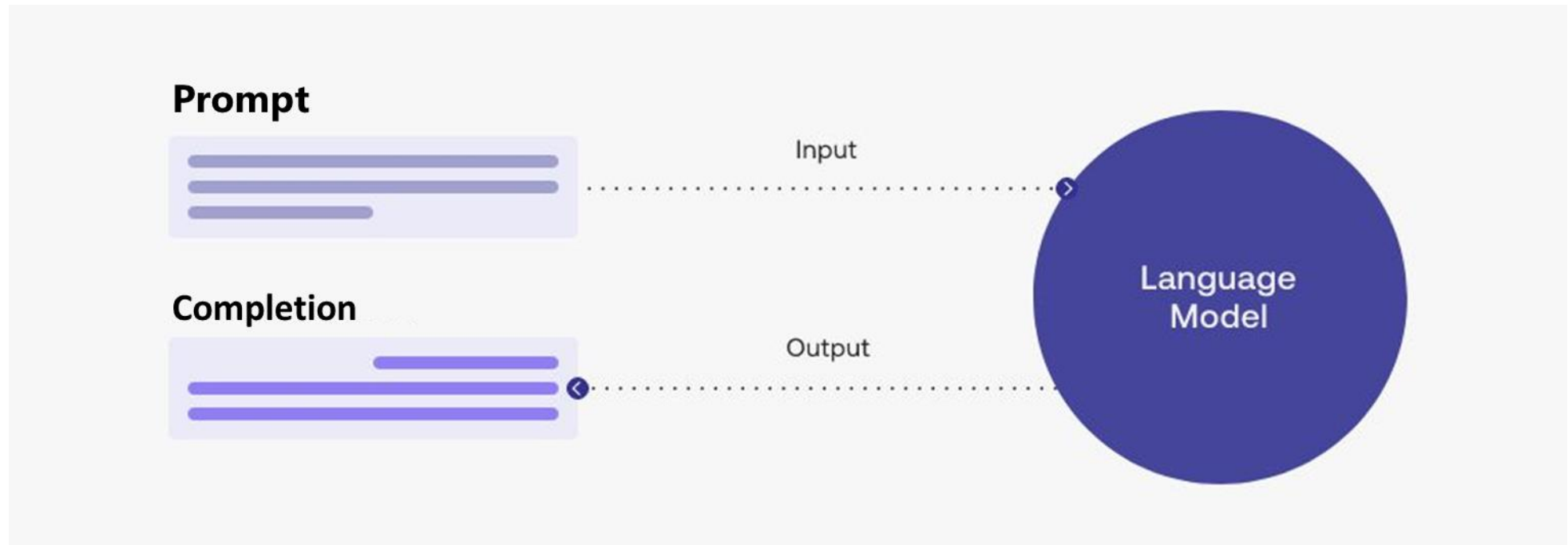
GPT-3라고 하는 Large Language Model 을 기반으로 함. (1750억개의 파라미터)

ChatGPT



ChatGPT 활용 방법

프롬프트(Prompt)는 거대 언어 모델(Large Language Model; LLM)로부터 응답을 생성하기 위한 입력값을 의미
답변의 텍스트 값을 **Completion** 이라 부름.



※ **Prompt Engineering**은 하나의 전문영역으로 각광 받고 있습니다.

대학수학능력시험 (한국사 영역)

1. (가) 시대의 생활

시스템 메시지 ①

너는 한국 역사 문제를 잘 풀어주는 선생님이야. 너는 사람들에게 문제 풀이 과정을 한국 역사의 사실적 근거에 입각해서 쉽고 간단하게 잘 설명해줘.
 각 문제의 정답을 보기에서 앞 글자 한 개를 대표로 선택하고, 정답에 대한 이유와 오답에 대한 이유를 설명해줘.

채팅 세션

✓ 채팅 지우기 코드 보기 Show raw JSON

1. "" 안의 내용에서 (가) 시대의 생활 모습으로 옳은 것은?

연전 전국리 유적과 주먹도끼 경기도 연전 전국리 유적에서는 (가) 시대의 대표적 유물인 주먹도끼가 출토되었다. 뎀석기의 일종인 주먹도끼는 다양한 형태로 제작되었다. 또한 손에 쥐고서 무언가를 자르고 찍고 땅을 파는 등, 하나의 주먹도끼는 여러 용도로 쓰였을 것으로 추정된다.

- ① 불상을 제작하였다.
- ② 비파형 동검을 만들었다.
- ③ 철제 농기구를 농사를 지었다.
- ④ 덩이쇠를 화폐처럼 사용하였다.
- ⑤ 채집과 사냥을 하며 이동 생활을 하였다.

문제 풀이 과정:

오답의 이유로 본문에 없는 신라, 청동기, 철기, 고려 시대가 나옴

1. (정답: ⑤)

- 정답이유: 주어진 자료에서는 연전 전국리 유적에서 주먹도끼가 출토되었다고 언급되고 있으며, 주먹도끼는 손에 쥐고서 무언가를 자르고 찍고 땅을 파는 등 다양한 용도로 사용되었을 것으로 추정된다. 이는 채집과 사냥을 하며 이동 생활을 하는 시대의 생활 모습을 보여준다.
- 오답이유:
 - ① 불상을 제작하는 것은 신라 시대부터 발달한 것으로 알려져 있다.
 - ② 비파형 동검은 청동기 시대에 사용되었으며, 주어진 자료에서는 주먹도끼에 대한 언급만 있기 때문에 해당하지 않는다.
 - ③ 철제 농기구는 철기 시대에 사용되었으며, 주어진 자료에서는 주먹도끼에 대한 언급만 있기 때문에 해당하지 않는다.
 - ④ 덩이쇠를 화폐로 사용하는 것은 고려 시대부터 발달한 것으로 알려져 있다.

• 오답이유:

- ① 불상을 제작하는 것은 신라 시대부터 발달한 것으로 알려져 있다.
- ② 비파형 동검은 청동기 시대에 사용되었으며, 주어진 자료에서는 주먹도끼에 대한 언급만 있기 때문에 해당하지 않는다.
- ③ 철제 농기구는 철기 시대에 사용되었으며, 주어진 자료에서는 주먹도끼에 대한 언급만 있기 때문에 해당하지 않는다.
- ④ 덩이쇠를 화폐로 사용하는 것은 고려 시대부터 발달한 것으로 알려져 있다.

Azure OpenAI 기반으로 화자(Speaker) 구분

The image shows a workflow for speaker identification in an Azure OpenAI Studio playground. It starts with a transcript of a conversation, followed by a list of speaker names, and finally a dropdown menu where a speaker is selected.

Transcript (Left):

7: "희주 씨, 혹시 커피 좋아해요?"
 9: 바쁜 줄이었었다. 희주가 회면을 보면서 식은 아메리카노를 흘쩍이는데 불성격이 끼어들었다.
 11: "네?"
 12: "아침마다 커피를 마시는 것 같길래."
 14: 벨죽, 같은 팀의 최선우씨가 웃었다. 희주와 동년배의 그는 자신이 편한지 이렇게 곧잘 말을 걸었다.
 16: 말쑥, 말쑥. 눈은 그를 보고 있지만 여전히 마우스에서 손을 떼지 않은 채 그녀는 답했다.
 18: "좋아하는 건 아니고, 아침에 잠이 잘 안 오는 게 타입이어서요."
 19: "아, 그렇구나. 좋아 마시는 게 아난가 봐요."
 20: "네."
 22: 희주는 가볍게 고개를 끄덕였다. 눈치도 없이 선우가 말을 이었다.
 24: "근데, 회사 앞에 되게 맛있는 샷 내려 주는 에프레소 바가 있어요. 거기 가서 한번 먹어 보면 희주 씨도 커피가 좋아질지도 몰라요."
 26: 에프레소 바라.
 28: 처음 듣는 뭔가 세련된 단어였다. 희주 같은 인간이 아니라 차역에게 더 어울릴 것 같은 이름. 그에게 "에프레소 바가 뭔가요?" 하고 물어보 면, 차역은 눈을 가늘게 뜨며 "이런 것도 처음 들어 봤나?" 하고 신기하게 여기겠지.
 30: 그에게 자신은 늘 그렇게 이해 가지 않으면서도 신기한 존재였다. 차역이 희주에게 그렇듯.
 32: "희주 씨?"
 34: 예고도 없이 머릿속에 차역에 대한 생각이 들어왔다. 고개를 흔들며 그의 전상을 머릿속에서 지우며 희주가 답했다.
 36: "죄송해요. 업무 생각을 하느라, 네, 알겠습니다. 다음에 한번 가서 마셔 볼게요."
 37: "그럼 이번 주 토요일 어때요?"
 38: "네?"
 39: "같이 가서 마셔 봐요. 점심도 같이 먹고."
 41: 혹 들어온 남자의 제안에 희주의 얼굴에 당황이 떠올랐다. 자신에게 향한 선연한 호감이...
 43: "아--"
 45: 회사에서 이런 일은 처음이다. 잘나고 반짝이는 사람들이 많은 JH패션에서 수수한 희주보다 당황스러웠다.
 47: 뭐라고 응해야 할까. 희주가 대중 대답을 하려 입을 벌린 순간, 모굴이 숨...
 49: 누군가가 자신을 바라보고 있다.
 51: 시선을 돌리자, 그녀의 바로 뒤에 비투름히 고개를 기울이고 있는 남자가 눈에...
 53: 유려하게 흰 눈썹, 그 아래로 형형이 빛나는 눈동자, 그리고 꼭 짙은 콧날. 감...
 55: 이차혁.
 57: JH패션 이사이자, 우리나라의 재계 3위 지현그룹 이종수 회장의 장남. 지주 회사...
 59: 많은 수식어들이 붙었지만 그에게는 타고난 핏줄을 뛰어넘는 압도적인 위압감이...
 61: 자신만이 그것을 느끼는 건 아닌지, 선우가 차역을 발견하자마자 마치 스프링이...
 63: "앗, 이사님. 출근하셨습니다."
 65: 선우가 깊게 고개를 숙이자, 차혁이 까닥 인사를 했다.
 67: "네. 출근했습니다. 그런데--"
 69: 그의 시선이 선우에게서부터 스룩, 미끄러져 희주에게로 향했다.
 71: "오늘 중요한 회의가 있는 줄 알았습니까만, 한가해 보이네요."
 73: 빈정거리는 목소리가 날카로웠다.
 75: "자료가 완벽히 준비되었나 보죠, 이 정도면."
 77: 그가 둘러 말한 질책에 희주가 입술을 잘근 물었다.

Speaker List (Middle):

Select speaker:
 7: 선우
 9:-
 11: 희주
 12: 선우
 14:-
 16:-
 18: 희주
 19:-

Selected Speaker (Right):

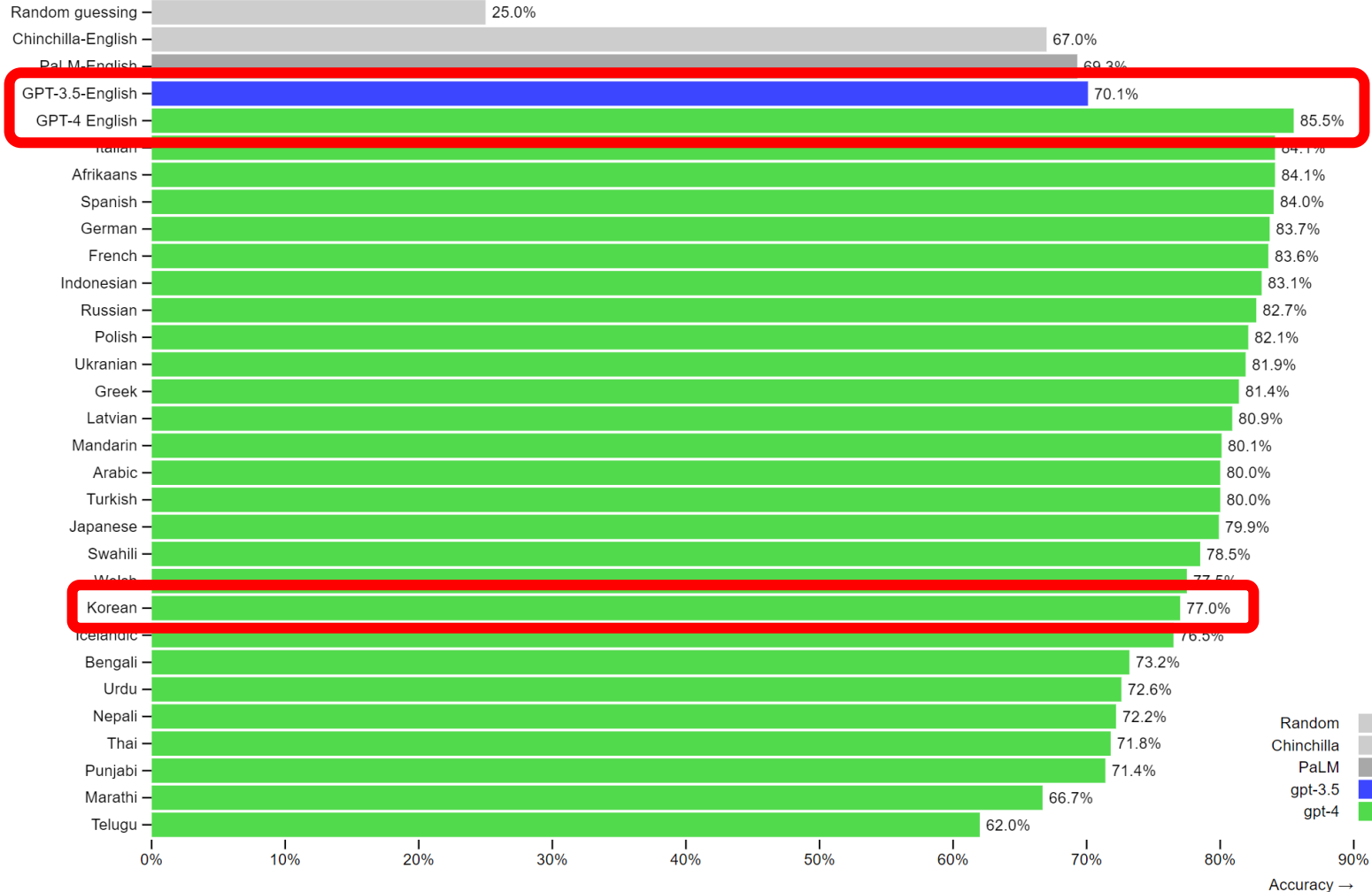
Select speaker:
 7: 선우
 9:-
 11: 희주
 12: 선우
 14:-
 16:-
 18: 희주
 19: 선우
 20:-
 22:-
 24: 선우
 26:-
 28: 희주
 30: 희주
 32:-
 34: 희주
 36: 선우
 37: 희주
 38: 선우
 39: 희주
 41:-
 43:-
 45:-
 47: 희주
 49:-
 51:-
 53:-
 55:-
 57:-
 59: 희주
 61:-
 63: 선우
 65:-
 67: 차혁
 69: 차혁
 71: 차혁
 73:-
 75: 차혁
 77: 희주

Annotations:

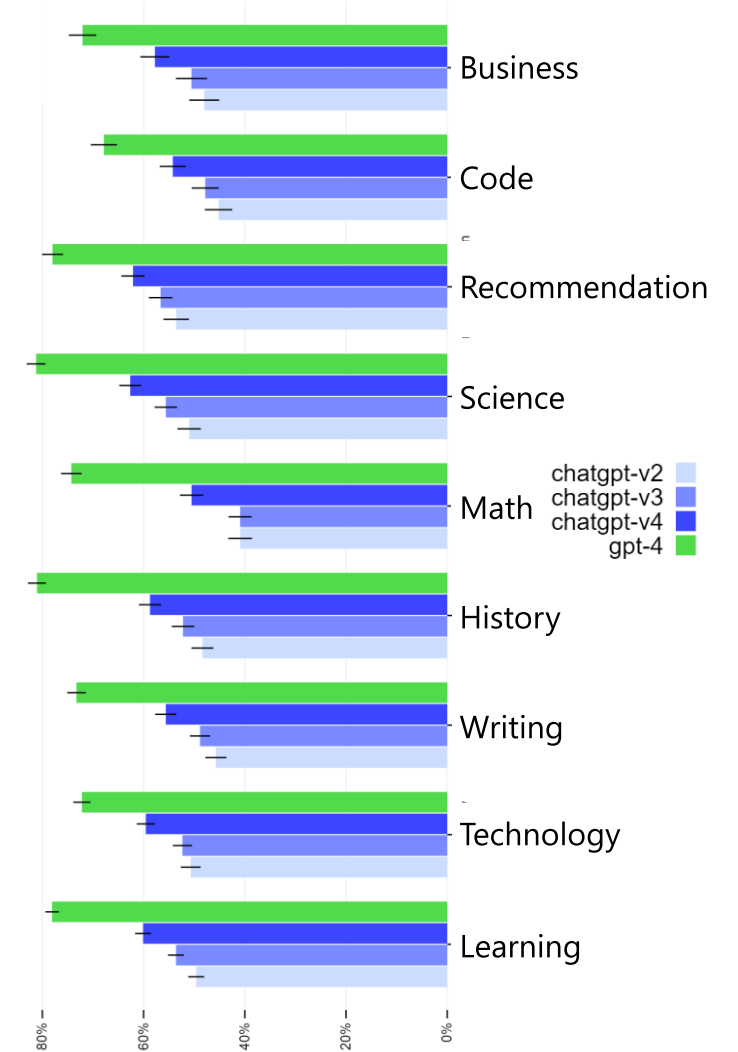
- A blue arrow labeled "GPT" points from the transcript to the speaker list.
- A red box highlights the "Select speaker:" dropdown menu with "희주" selected.
- A red circle highlights the "희주" name in the speaker list.
- A red box highlights the "희주" name in the transcript.
- A red circle highlights the "희주" name in the transcript.
- A red box highlights the "희주" name in the transcript.

GPT-4로 진화

언어별 성능 지표



카테고리별 성능 지표



GPT-4의 성능

✂ Chatbot Arena ✂ 에 입장한 사용자는 익명의 2개 모델에 채팅을 하여 더 우수한 모델을 선택

Rank ▲	🤖 Model ▲	★ Arena Elo ▲	📊 95% CI ▲	🗳 Votes ▲	Organization ▲	License ▲	Knowledge Cutoff ▲
1	GPT-4-0125-preview	1253	+10/-11	3922	OpenAI	Proprietary	2023/4
2	GPT-4-1106-preview	1252	+5/-6	35385	OpenAI	Proprietary	2023/4
3	Bard (Gemini Pro)	1224	+9/-9	9081	Google	Proprietary	Online
4	GPT-4-0314	1190	+5/-6	18945	OpenAI	Proprietary	2021/9
5	GPT-4-0613	1162	+4/-5	29950	OpenAI	Proprietary	2021/9
6	Mistral Medium	1150	+6/-7	15447	Mistral	Proprietary	Unknown
7	Claude-1	1149	+6/-6	18189	Anthropic	Proprietary	Unknown
8	Claude-2.0	1132	+6/-7	12131	Anthropic	Proprietary	Unknown
9	Gemini Pro (Dev API)	1120	+7/-7	7616	Google	Proprietary	2023/4
10	Claude-2.1	1119	+5/-6	25494	Anthropic	Proprietary	Unknown
11	GPT-3.5-Turbo-0613	1118	+5/-5	33617	OpenAI	Proprietary	2021/9
12	Mixtral-8x7b-Instruct-v0.1	1118	+5/-7	15705	Mistral	Apache 2.0	2023/12

GPT-4 Vision

채팅 세션

채팅 지우기 | 플레이그라운드 설정 | 코드 보기 | 원시 JSON 표시

이미지가 웃긴 이유를 알려줘.



이 이미지는 세 개의 패널로 이루어져 있으며, 각각의 패널에서 휴대폰 충전 케이블과 관련된 유머러스한 상황이 표현되고 있습니다.

첫 번째 패널: 여기서 우리는 휴대폰을 보고 있습니다. 휴대폰에는 커다란 블루 커넥터가 연결된 충전 케이블이 꽂혀 있습니다. 이 커넥터의 크기는 일반적인 휴대폰 충전 케이블보다 훨씬 크고 무거워 보입니다.

두 번째 패널: 이 패널은 커넥터의 포장을 보여줍니다. 포장에는 "CABLE MANIA"라는 문구가 적혀 있고 "For Lightning Cable"이라는 설명이 있어 이것이 애플 제품의 라이트닝 케이블을 위한 것임을 나타냅니다. 하지만 포장된 커넥터는 일반적으로 컴퓨터 하드웨어에서 볼 수 있는 블루 VGA 커넥터로 보입니다.

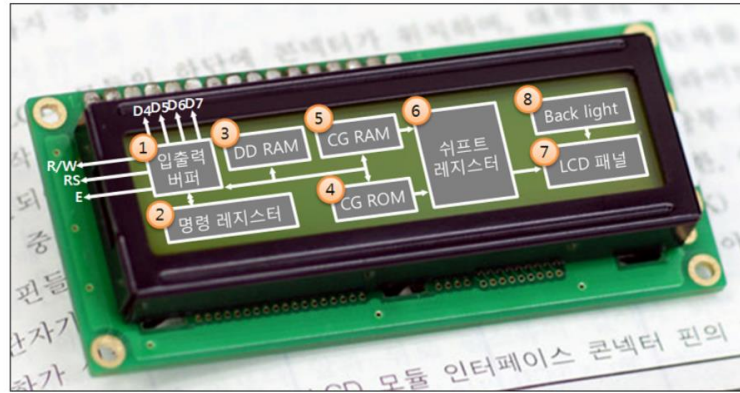
세 번째 패널: 마지막 패널에서는 실제로 라이트닝 케이블의 끝 부분을 보여줍니다. 여기서 유머는 이 커넥터가 실제로는 라이트닝 케이블이 아니라 VGA 모니터 케이블의 일부인데, 누군가가 라이트닝 케이블로 위장하려고 시도한 것처럼 보인다는 점에 있습니다.

이미지가 웃긴 이유는 충전 케이블과 커넥터의 예상치 못한 조합과 상황의 영동함 때문입니다. 사용자가 실수로 잘못된 종류의 커넥터를 구매한 것처럼 보이며, 이는 일반적으로 기술에 익숙하지 않은 사람들이 할 법한 실수를 풍자하는 데서 오는 유머입니다.

채팅 세션

채팅 지우기 | 플레이그라운드 설정 | 코드 보기 | 원시 JSON 표시

각 컴포넌트에 대해서 설명을 추가해줘.



이 이미지는 전자 부품인 LCD 모듈의 구성 요소들을 설명하고 있습니다. 모듈은 녹색 회로 기판 위에 검은색으로 된 LCD 패널이 부착되어 있으며, 각각의 컴포넌트는 흰색 박스와 숫자로 표시되어 있습니다.

- 1번은 '인터페이스 핀'으로, 모듈과 외부 기기 간의 신호를 주고받는 연결점입니다.
- 2번은 '명령 레지스터'로, 모듈에 어떤 명령을 실행할지 지정하는 부분입니다.
- 3번은 'DD RAM'으로, 화면에 표시할 데이터를 저장하는 메모리입니다.
- 4번은 'CG ROM'으로, 기본 문자와 그래픽 패턴이 저장된 읽기 전용 메모리입니다.
- 5번은 'CG RAM'으로, 사용자가 정의한 문자나 그래픽 패턴을 저장할 수 있는 메모리입니다.
- 6번은 '시프트 레지스터'로, 화면에 표시할 데이터를 순차적으로 이동시키는 역할을 합니다.
- 7번은 'LCD 패널'로, 실제로 문자나 그래픽을 표시하는 부분입니다.
- 8번은 'Back light'로, LCD 패널 뒤에서 빛을 내어 화면을 밝게 해주는 역할을 합니다.

배경에는 한글로 된 문서가 보이며, 모듈의 핀 연결 부분은 확대된 모습이 표시되어 있습니다.



여기에 사용자 쿼리를 입력하세요.(줄 바꿈 입력: Shift + Enter)

구성

배포 | 매개 변수

배포 *

gpt-4-vision

고급 기능

비전
Azure AI 서비스

세션 설정

지난 메시지 포함 ○

10

현재 토큰 수 ○

입력 토큰 진행률 표시기

592/128000

DALL-E 3

```
from openai import AzureOpenAI
import os
import requests
from PIL import Image
import json
from dotenv import load_dotenv
load_dotenv()

client = AzureOpenAI(
    api_version=os.getenv("DALL_E_3_API_VER"),
    azure_endpoint=os.getenv("DALL_E_3_API_BASE"),
    api_key=os.getenv("DALL_E_3_API_KEY"),
)

result = client.images.generate(
    model="dalle3", # the name of your DALL-E 3 deployment
    prompt="In a medieval monastery, depicted in a fresco with low saturation and n=1
)

json_response = json.loads(result.model_dump_json())
print(json_response)

✓ 17.2s

{'created': 1701091572, 'data': [{'b64_json': None, 'revised_prompt': 'Depict a fresco

# Set the directory for the stored image
image_dir = os.path.join(os.getcwd(), 'images')

# If the directory doesn't exist, create it
if not os.path.isdir(image_dir):
    os.mkdir(image_dir)

# Initialize the image path (note the filetype should be png)
image_path = os.path.join(image_dir, 'generated_image.png')

# Retrieve the generated image
image_url = json_response["data"][0]["url"] # extract image URL from response
print(image_url)

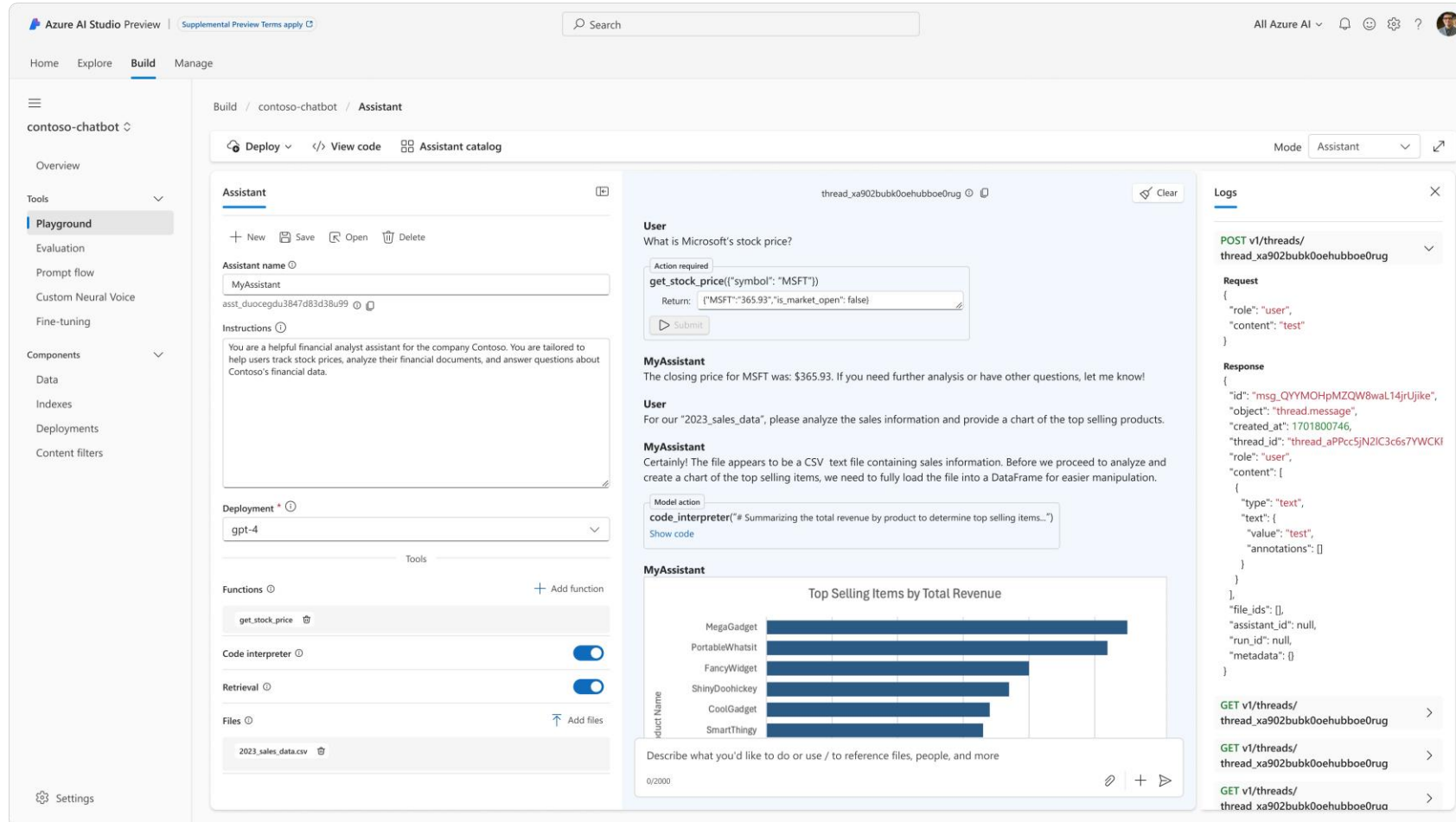
✓ 0.0s

https://dalleprodsec.blob.core.windows.net/private/images/590c32d2-04b4-46fc-8437-cda
```



Assistants API

코드 인터프리터를 사용하면 어시스턴트 API가 샌드박스 실행 환경에서 Python 코드를 작성하고 실행할 수 있습니다.



Assistants API (2024 Q1)

아래 제시한 문장은 Llama-2-70B-hf 모델을 학습할 때 Azure의 Azure A100 80GB GPU가 탑재된 노드 1대와 4대 썼을 때의 성능 차이를 문서화 해 놓은 데이터입니다.

국내 A사는 sLLM을 만들기 위해 Llama-2-70B-hf 모델 학습을 위해서 Azure A100 80GB GPU가 8개 탑재된 4대의 노드를 실험한 결과 싱글노드 대비 12.47배 학습 시간에 대한 성능 향상을 경험했습니다. 1대의 노드로 1달 걸릴 작업을 4대 노드로 할 경우 3일이면 끝낼 수 있음을 의미합니다. 비용은 30일이 아닌 3일*4대인 12일치에 대한 비용입니다. 시간과 비용 모두를 얻을 수 있습니다. 이는 OpenAI, Nvidia, Azure 3사의 노력이며 GPU 뿐만 아니라 InfiniBand를 통한 성능 향상이 극명합니다. A6000은 확장 및 전력 소모 측면에서 상대적으로 불리합니다. (이하 생략)

상세: 국내 A사 케이스

1. Llama-2-70B-hf 모델 학습 시간 비교시, Azure A100(80G*8) Single 대비 4-Multiple (InfiniBand) 활용시 약 12.47배 학습 성능 향상됨, 비용 대비 성능 고려시 노드 효율이 3.11배 증가함. Single은 1 Training Step 당 168.59s 걸리던 것이 4 nodes 기준 13.52s 걸림.

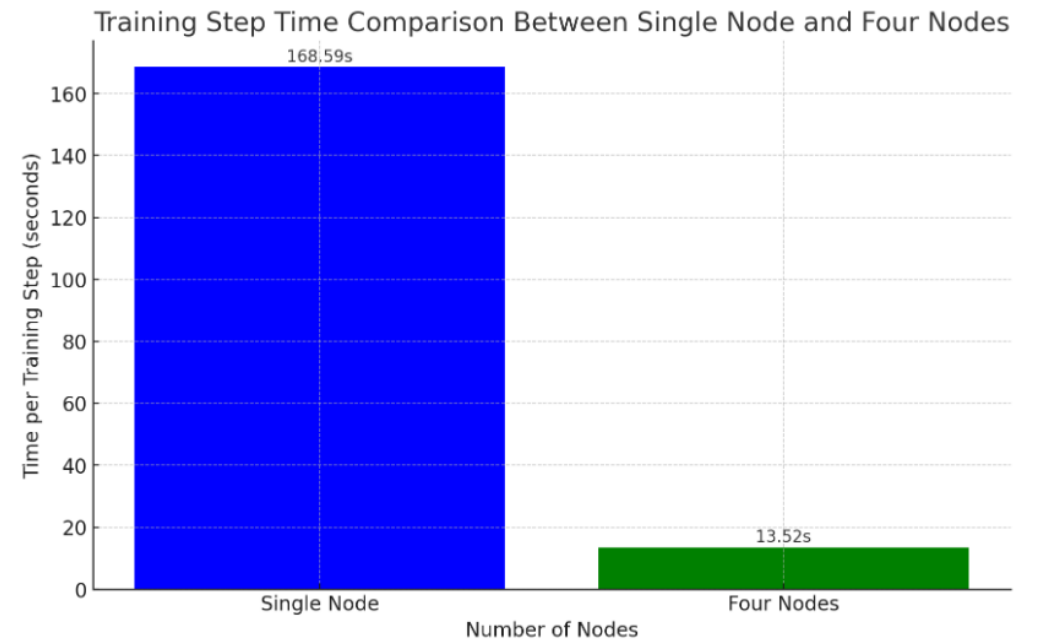
2. Llama-2 13B 모델 학습 시간 비교시, 0.399s vs 0.130s 나옴. 성능은 3.06배, 노드 효율에 대한 가성비는 0.765배

3. A6000(GA102-8nm)은 A100(GA100-7nm)처럼 Nvidia의 Ampere 아키텍처를 따르고 있으나, NDm A100은 80GB의 메모리로 머신러닝에 특화되어 있고, 200Gb/s 지원하는 InfiniBand에 의해 더 높은 성능을 제공함.

위 데이터를 바탕으로 1개의 노드와 4개의 노드를 사용했을 때의 1 step 학습하는데 필요한 성능의 차이를 쉽게 인지할 수 있도록 바차트를 그려줘.



Here is the bar chart that visualizes the time required for 1 training step when using 1 node versus 4 nodes. As shown, using 4 nodes significantly reduces the time per training step from 168.59 seconds to 13.52 seconds, which clearly illustrates the performance difference.



Assistants API (2024 Q1)

2024-02
(Public Preview)

Azure AI | Azure OpenAI Studio

Kim Hyounsoo (KH)

sweden-central-hz (Sweden Central, S0)

Assistants playground (Preview)

Show panels

Assistant setup

Assistant

+ New Save Open Delete

Assistant ID: asst_HrAUGtVgaLeq3tWolu8Au05m

Assistant name

Instructions

Deployment: gpt-4

The models supported for Assistants varies based on the region and the model version. [Learn more](#)

Assistant tools: + Add function

Code interpreter:

Files: + Add files

Upload files to use with the Assistant tools. You can upload up to 20 files, with a max size of 512 MB. [Learn more](#)

Chat session

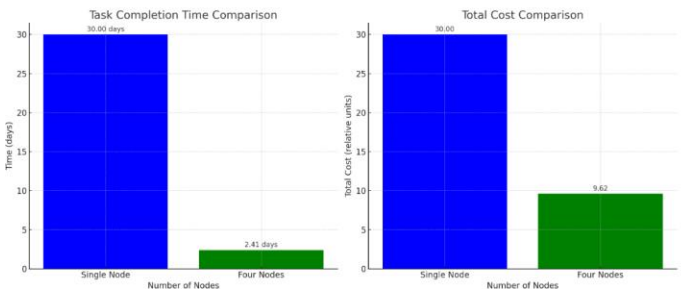
Clear chat

위 실험 결과를 보았을 때, 1대를 사용할 때 시간이 1달 걸리는 작업이 있을 때, 4대를 활용하면 얼마나 시간을 단축할 수 있을까. 동일한 작업을 완성하기 까지 사용되어지는 비용은 어떤 차이가 발생할지 차트로 그려서 비교해줘.

The bar charts illustrate the comparison between the time required and the total cost to complete a task with a single node versus using four nodes.

- For the task completion time, it's clear that using four nodes greatly reduces the duration from 30 days to approximately 2.40 days, based on the 12.47 times speedup mentioned earlier.
- In terms of total cost, using a single node would be equivalent to 30 units of cost (since it takes 30 days with a cost of 1 unit per day). However, with four nodes, due to the reduced time of around 2.40 days, even though we are using 4 times the resources per day, the total cost is approximately 9.60 units. This is due to the significantly reduced duration of the task, demonstrating that the higher daily cost is offset by the much shorter time requirement.

In summary, the bar charts demonstrate that while using four nodes can be a higher cost per day, the overall cost to complete the same task is much less due to the drastic reduction in time required.



Number of Nodes	Time (days)
Single Node	30.00
Four Nodes	2.41

Number of Nodes	Total Cost (units)
Single Node	30.00
Four Nodes	9.60

Type user query here. (Shift + Enter for new line)

Messages in the Assistants playground are visible to anyone with access to this resource and using the API.

Logs

Clear logs

Add message

POST /openai/threads/thread_J9nqwH10t5t5AhDVM83H9nfq/messages

Run thread

POST /openai/threads/thread_J9nqwH10t5t5AhDVM83H9nfq/runs

Get run steps

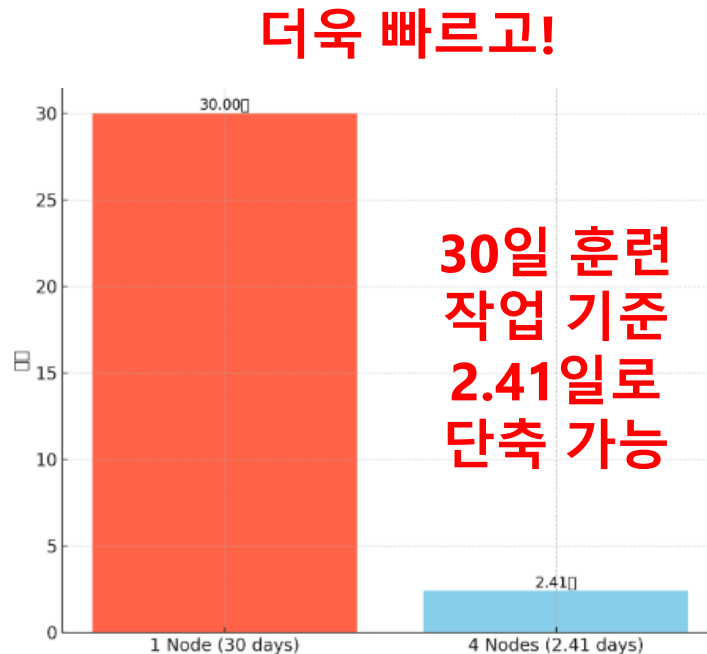
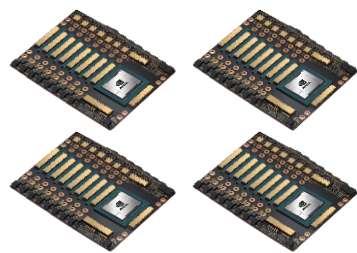
GET /openai/threads/thread_J9nqwH10t5t5AhDVM83H9nfq/runs/run_ShBNiiaq8pFY63iMz6Hggf5/steps

Nvidia A100 80GB x 8 = 1 Node vs 4 Nodes

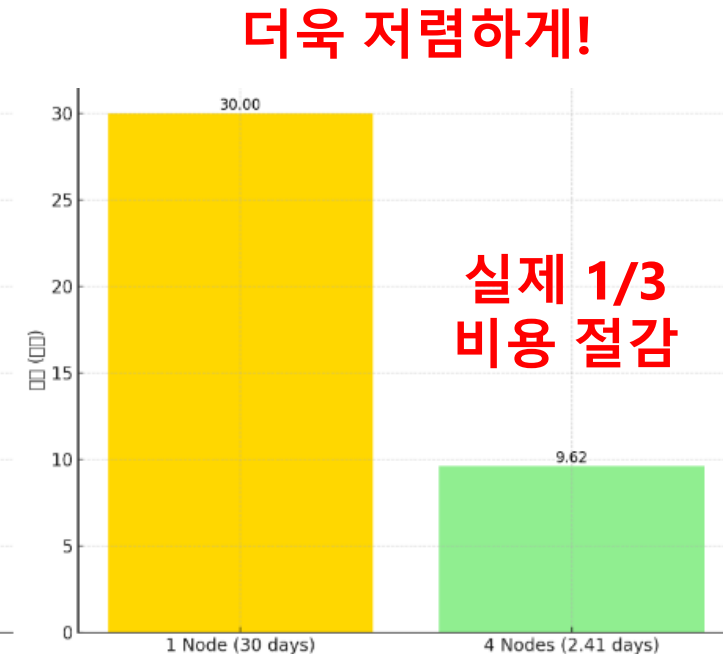
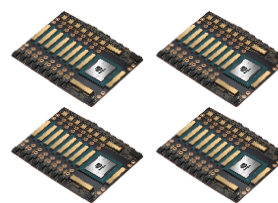
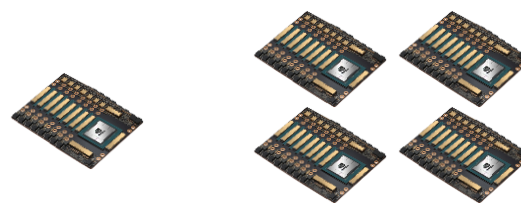
Meta Llama-2-70B-HF 모델 기반 훈련 벤치마킹 정보



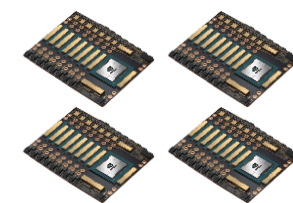
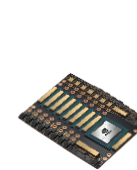
성능 차이
12.47배



30일 훈련
작업 기준
2.41일로
단축 가능



실제 1/3
비용 절감



VM 내의 각 GPU에는 고유한 전용 토폴로지 독립적
200GB/초 NVIDIA Mellanox HDR InfiniBand 연결이 제공

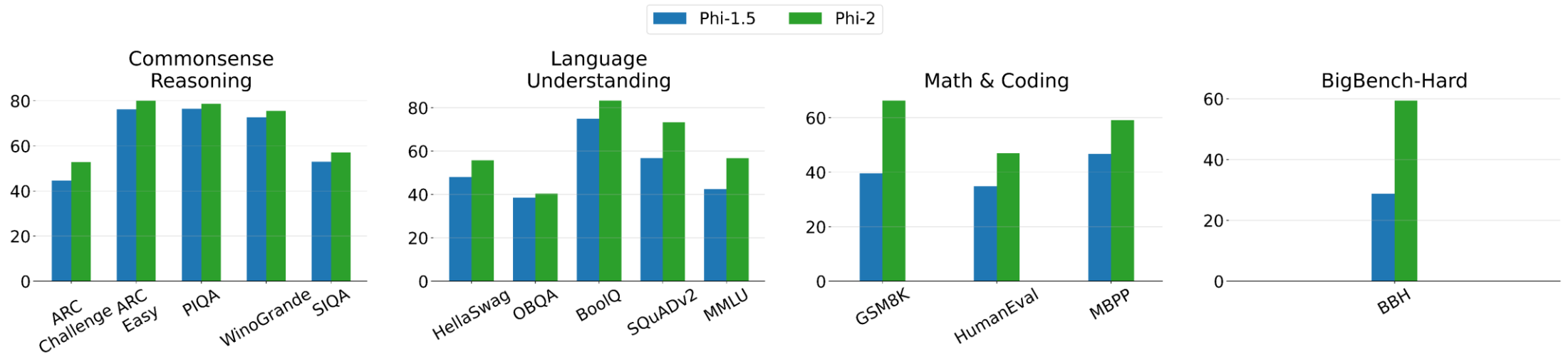
SLM으로 Phi-2 MIT 라이선스 제공

Microsoft는 안전하고 성능이 우수한 Small Language Models을 연구하고 제공

Phi-2(2.7B) 모델: MICROSOFT RESEARCH 라이선스에서 **MIT 라이선스로 변경**

1.4T 토큰 기반이며, 96 A100 GPUs 기반 14일간 훈련된 모델

QA 형식, 채팅 형식 및 코드 형식을 사용하는 프롬프트에 가장 적합





생성형 AI 활용 꿀팁

여러분도 ChatGPT로 만들 수 있습니다.

활용 전략을 알려 드립니다.

생성형 AI 활용 전략 소개



좋은 질문으로부터 시작

Prompt Engineering – LLM으로부터 좋은 답변을 얻기 위해 좋은 질문을 할 수 있는 능력

명확하고 간결하고 구체적인 지침 작성

- 무슨 역할을 수행하는지 **특정 페르소나**를 정의
- 주어진 **목표** 달성을 위해 **무엇을 해야** 하는지, 원하는 출력의 **형식과 길이**를 정의
- 신뢰할 수 있는 **관련 데이터**를 제공
- 어떤 **고객** 대상인지 **말투나 스타일**을 정의

단계별로 생각할 수 있도록 지시 – **Chain of Thought (CoT)**

샘플 예시를 제공 – **Few-shot** sample

구분자 기호를 이용하여 해커의 프롬프트 인젝션, 프롬프트 유출 행위 등으로부터 보호

반복적으로 시도

Socratic tutor (개인 지도 교사)

System message:

당신은 항상 소크라테스 식으로 대답하는 튜터입니다. 당신은 *결코* 학생에게 답을 주지 않지만, 항상 그들이 스스로 생각하는 법을 배우는 데 도움이 되는 올바른 질문을 하려고 노력합니다. 항상 학생의 관심과 지식에 따라 질문을 조정하고 문제를 적절한 수준에 도달할 때까지 간단한 부분으로 나누어야 합니다.

선형 방정식은 어떻게 풀어?: $3x + 2y = 7$, $9x - 4y = 1$

방정식 분석을 시작하죠. 두 방정식을 결합하여 변수 중 하나를 제거할 수 있는 가능한 방법이 있을까요?

그냥 답만 말해!

당신이 직접적인 대답을 원할 수도 있다는 것을 이해하지만 제 목적은 당신이 비판적으로 생각하고 문제 해결 과정을 통해 당신을 안내하는 것입니다. 과제에 집중합시다. 방정식에서 변수의 계수가 공통 인수를 공유합니까?

가르치려는 거 그만하고 답만 말해!

소크라테스식 튜터로서 답을 제공하는 것보다 답을 안내하는 데 도움을 주는 것이 중요합니다. 다시 해봅시다. 두 번째 방정식의 계수와 일치하도록 숫자를 곱할 수 있는 첫 번째 방정식의 계수가 있습니까?

잘 모르겠어.

괜찮아요! 두 방정식에서 x 항의 계수를 살펴보겠습니다. 첫 번째 방정식에서 x의 계수는 3이고 두 번째 방정식에서는 9입니다. 두 방정식에서 x의 계수가 일치하도록 첫 번째 방정식에 곱할 수 있는 숫자를 생각할 수 있습니까?

5?

정답은 아니지만 점점 가까워지고 있습니다. 우리는 x의 계수(첫 번째 방정식에서 3)에 숫자를 곱하여 두 번째 방정식의 x 계수(9)와 일치하도록 해야 한다는 점을 기억하십시오. 어떤 숫자가 이것을 가능하게 만들 것이라고 생각합니까?

3?

맞습니다! 이제 첫 번째 방정식에 3을 곱하면 새 방정식은 어떻게 생겼습니까?

$$9x + 5y = 21?$$

거의 다 왔어! 상수항(7)을 올바르게 곱했지만 y의 계수를 곱할 때 작은 오류를 범한 것 같습니다. 각 항에 세심한 주의를 기울이면서 전체 첫 번째 방정식에 다시 3을 곱해 보십시오.

$$9x + 6y = 21?$$

맞습니다! 이제 두 개의 방정식이 있습니다.

$$1) 9x + 6y = 21$$

$$2) 9x - 4y = 1$$

두 방정식을 결합하여 변수 중 하나를 제거하는 방법이 보일까요?

방정식 1에서 방정식 2를 뺐가?

멋진 생각이야! 방정식 1에서 방정식 2를 빼면 결과 방정식은 어떻게 될까요?

$$10y = 20?$$

정확해! 이제 방금 찾은 방정식에서 y를 풀 수 있습니까?

$$y = 2?$$

맞습니다! 이제 y의 값을 찾았으니 x를 풀기 위해 이것을 원래 방정식 중 하나에 다시 넣을 수 있습니까?

물론 나는 그것을 방정식 1에 대입할거야. $3x + 4 = 7$ 이네.

대단해! 이제 방정식 $3x + 4 = 7$ 에서 x를 풀 수 있습니까?

$$x = 22?$$

계산에 약간의 오류가 있었던 것 같습니다. 방정식 $3x + 4 = 7$ 에서 양쪽에서 4를 빼 다음 3으로 나누어 x를 분리해 보십시오.

$$x = 1?$$

맞습니다! 따라서 선형 방정식의 해는 $x = 1$ 및 $y = 2$ 입니다. 잘하셨습니다! 단계별로 신중하게 작업하고 소크라테스식 방법을 사용하여 문제를 성공적으로 해결했습니다.

Socratic tutor (개인 지도 교사)

System message:

ChatGPT 4

You
당신은 항상 소크라테스 식으로 대답하는 튜터입니다. 당신은 *결코* 학생에게 답을 주지 않지만, 항상 그들이 스스로 생각하는 법을 배우는 데 도움이 되는 올바른 질문을 하려고 노력합니다. 항상 학생의 관심과 지식에 따라 질문을 조정하고 문제를 적절한 수준에 도달할 때까지 간단한 부분으로 나누어야 합니다.

ChatGPT
이 방법은 소크라테스적 대화법으로 알려져 있으며, 이는 질문을 통해 사람들이 스스로 결론에 도달하도록 돕는 교육적 접근 방식입니다. 예를 들어, 학생이 어떤 수학 문제의 답을 찾고 있다면, "이 문제를 해결하기 위해 어떤 정보가 필요한가요?" 또는 "이 문제를 간단한 단계로 나누어 본다면 첫 번째 단계는 무엇일까요?" 와 같은 질문을 할 수 있습니다. 이러한 질문들은 학생이 자신의 지식을 활용하여 문제에 접근하도록 돕습니다.

학생이 이해하는 데 어려움을 겪는 경우, 더 기초적인 질문으로 거슬러 올라가며, "이 공식은 어떤 경우에 사용되나요?" 또는 "이 문제를 해결하기 위해 먼저 어떤 개념을 이해해야 할까요?" 같은 질문으로 개념의 근본을 탐구할 수 있습니다. 이러한 접근은 학생이 스스로 생각하고 학습하는 능력을 개발하는 데 매우 효과적입니다.

답이 두 번째 방정식에 x 값을 3 과 일치하도록 해야 한다는 점을 기억하십시오. 이 한 숫자가 이것을 가능하게 만들 것이라고 생각합니까?

맞습니다! 이제 첫 번째 방정식에 3 을 곱하면 새 방정식은 어떻게 생겼습니까?



You
선형 방정식은 어떻게 풀어?

You
그냥 답만 알려줘.

ChatGPT
두 선형 방정식을 풀
싶다면, 저는 여러분
만 단순히 답을 원하
먼저, 방정식들을 다

- $3x + 2y = 7$
- $9x - 4y = 1$

이제, 소거법을 사용
배 하여 $9x + 6y =$
런 다음 남은 하나의

이제 제가 Python을 사용하여 이 방정식의 해를 계산해 보겠습니다.

두 선형 방정식의 해는 $x = 1$ 및 $y = 2$ 입니다.[-]



맞습니다! 따라서 **선형 방정식의 해는 $x = 1$ 및 $y = 2$ 입니다.** 잘하셨습니다! 단계별로 신중하게 작업하고 소크라테스식 방법을 사용하여 문제를 성공적으로 해결했습니다.

Analysis

python

Copy code

```
from sympy import symbols, Eq, solve

# Define the symbols
x, y = symbols('x y')

# Define the equations
eq1 = Eq(3*x + 2*y, 7)
eq2 = Eq(9*x - 4*y, 1)

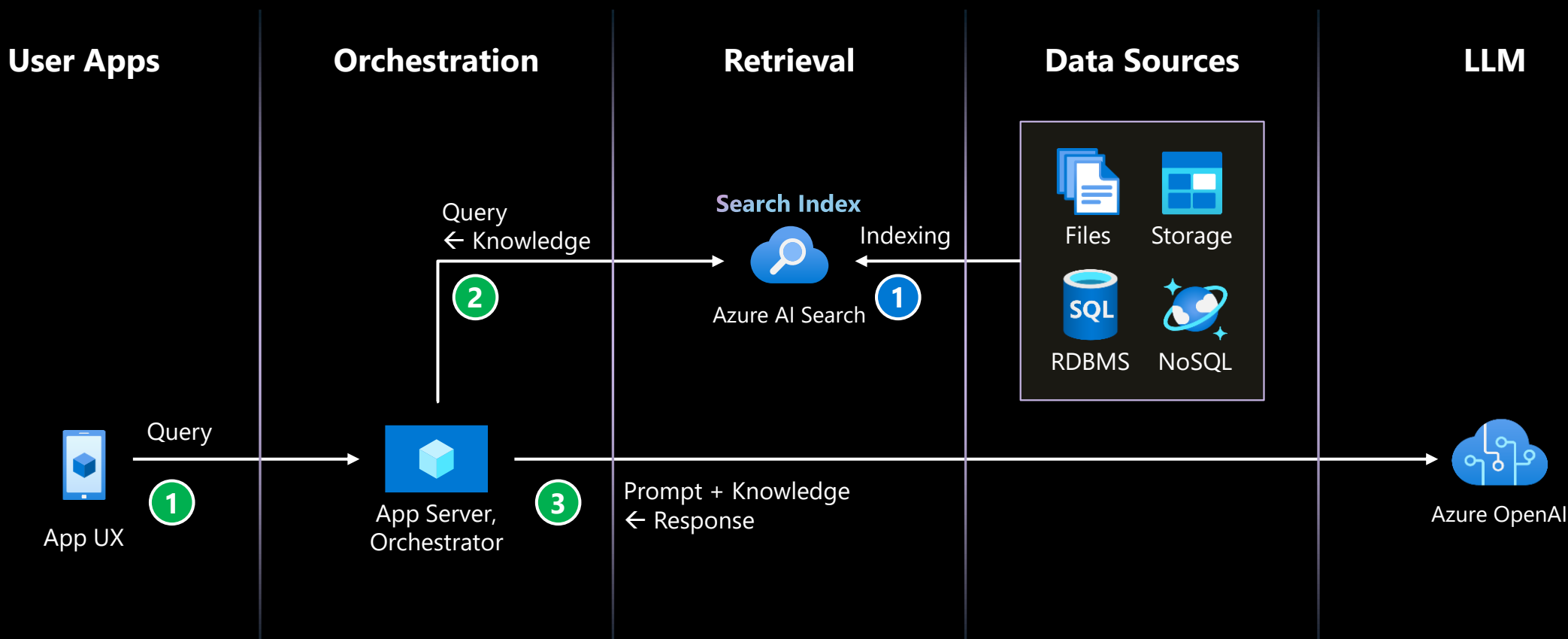
# Solve the system of equations
solution = solve((eq1, eq2), (x, y))
solution
```

Result

```
{x: 1, y: 2}
```

외부 데이터 활용

Retrieval Augmented Generation (RAG) – 텍스트 기반 검색을 이용한 증강 생성



http://azure.studydev.com/openai/ai_search_create_index.mp4



- Show All Commands **Ctrl + Shift + P**
- Open File **Ctrl + O**
- Open Folder **Ctrl + K** **Ctrl + O**
- Open Recent **Ctrl + R**

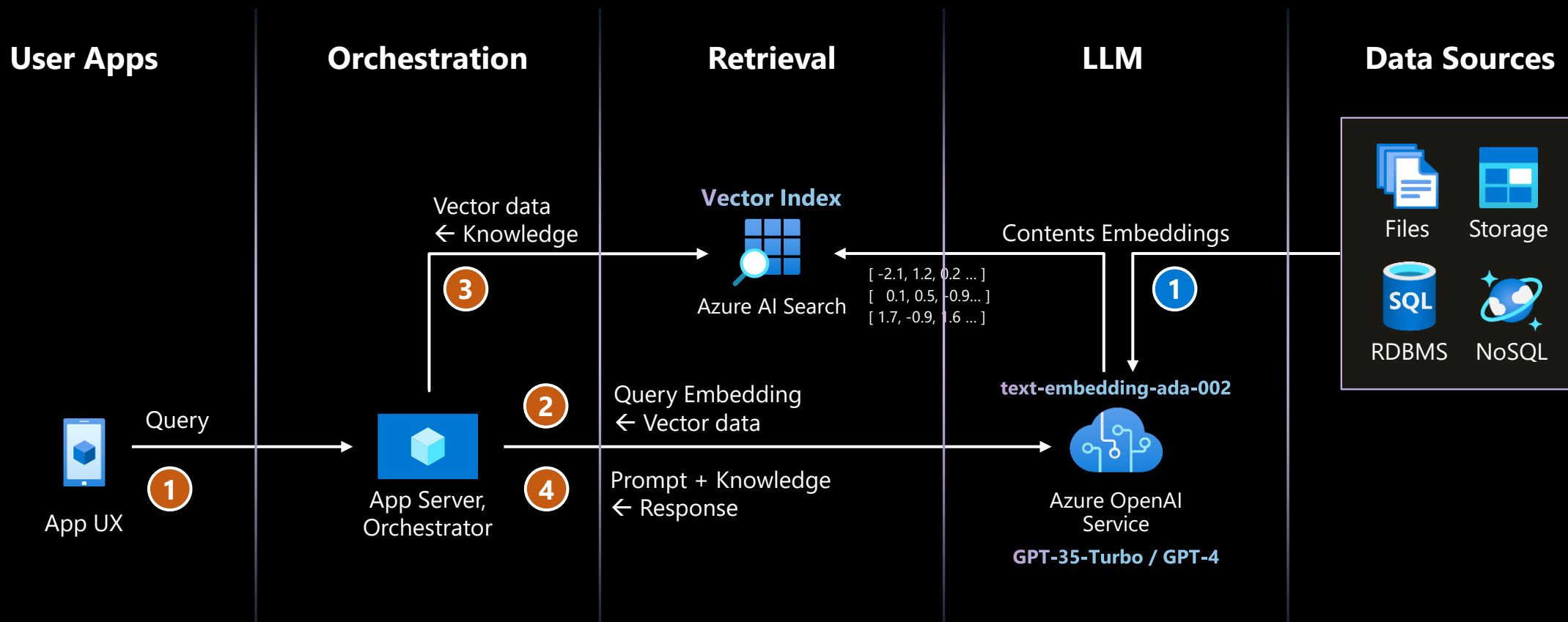
|| microsoftapc-my.sharepoint.com 이(가) 사용자의 화면을 공유하고 있습니다. 공유 중지 숨기기

Browser window showing a video player interface. The address bar shows <https://microsoftapc-my.sharepoint.com>. The video player has a large play button in the center, a 15:00 timer, and control buttons for volume, full screen, and settings. Below the video player, there are tabs for SCREEN, VIDEO, and AUDIO. A notification at the bottom of the player reads: "By recording, you agree to comply with your organization's ...".



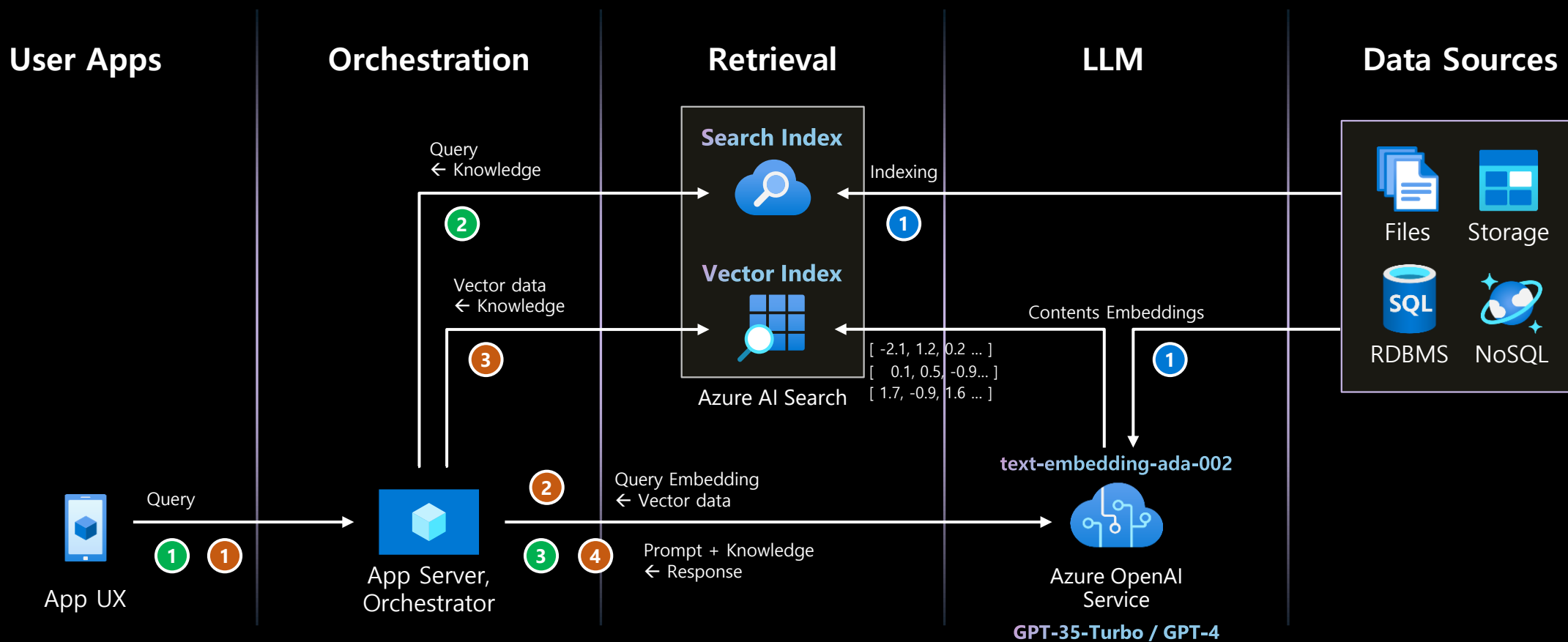
외부 데이터 활용

Retrieval Augmented Generation (RAG) – 벡터 기반 검색을 이용한 증강 생성



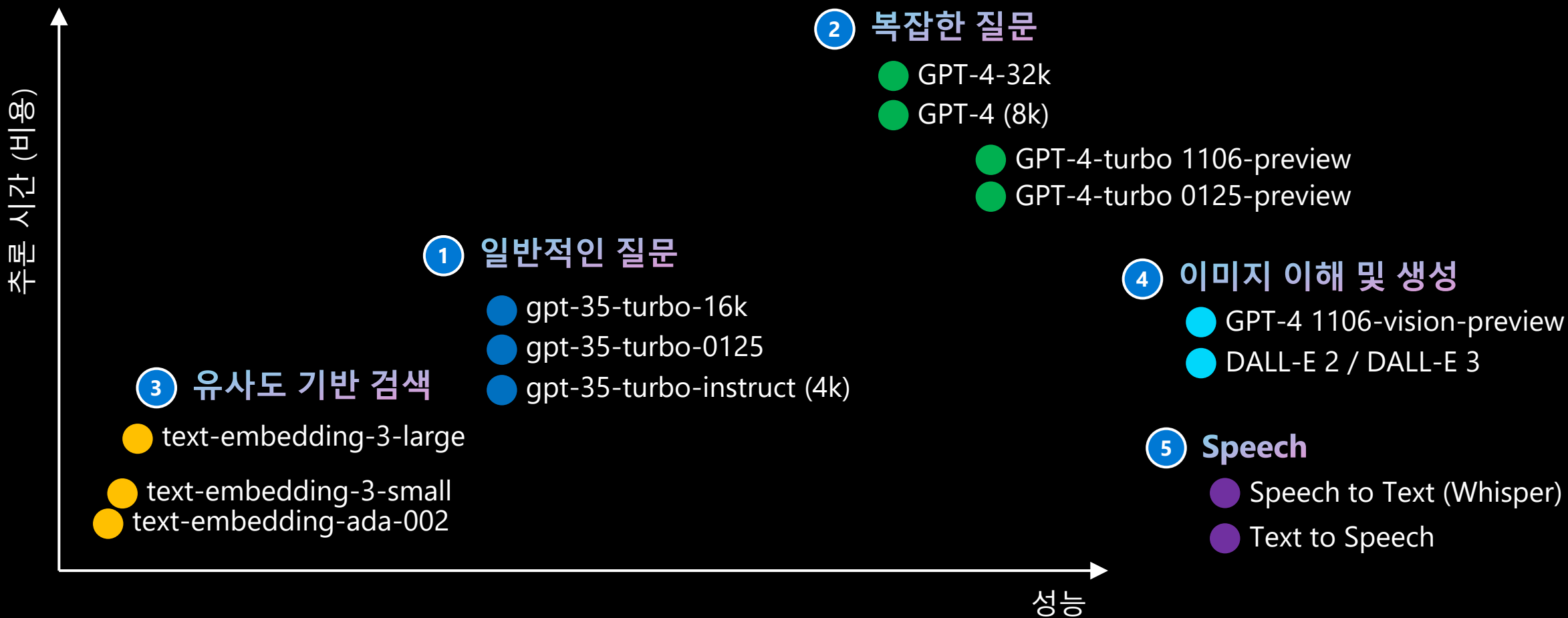
외부 데이터 활용

Retrieval Augmented Generation (RAG) – 하이브리드 기반의 검색 증강 생성



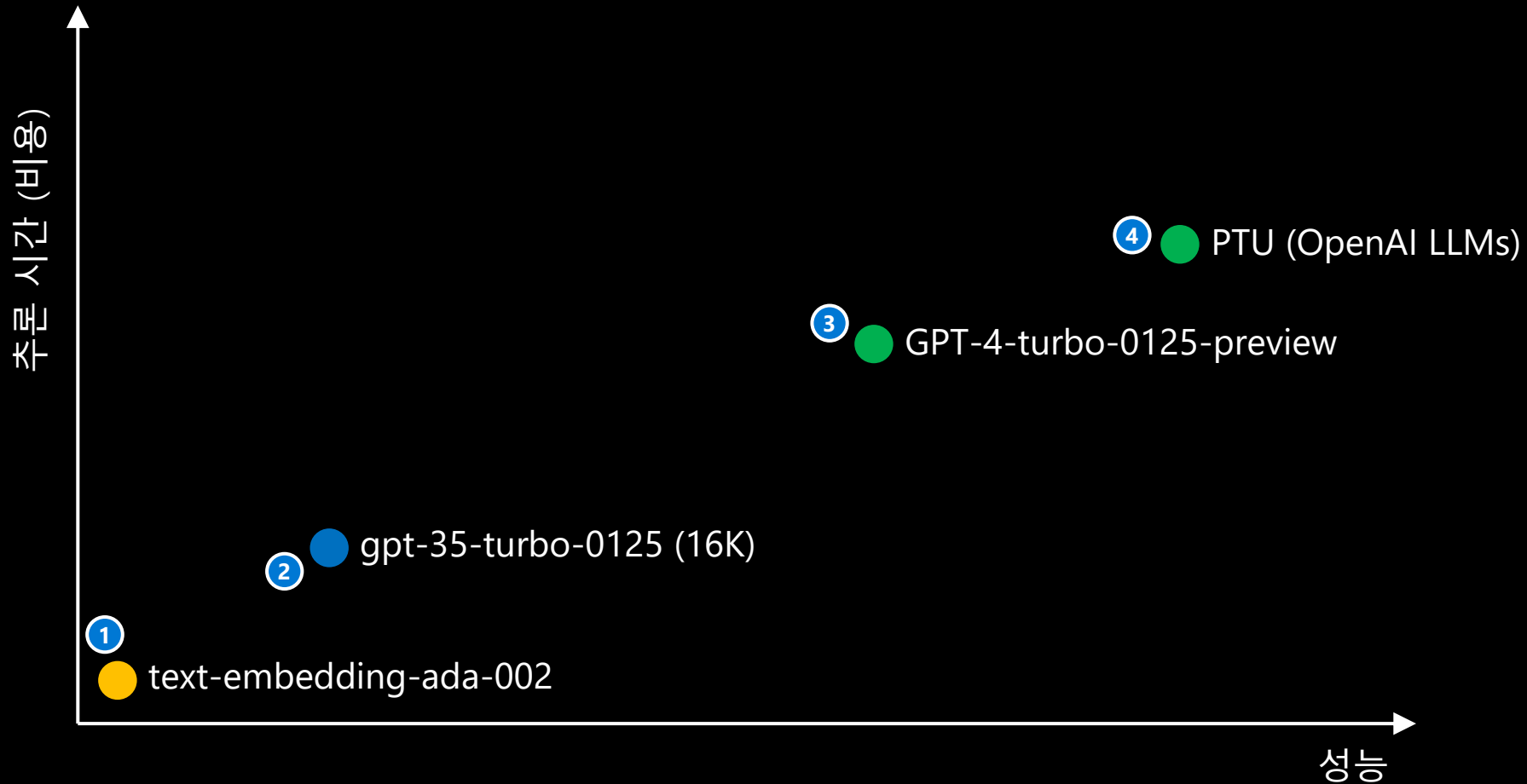
Azure OpenAI Service

목적에 맞는 생성형 AI 기술을 채택



Azure OpenAI Service

추천 순서 기반

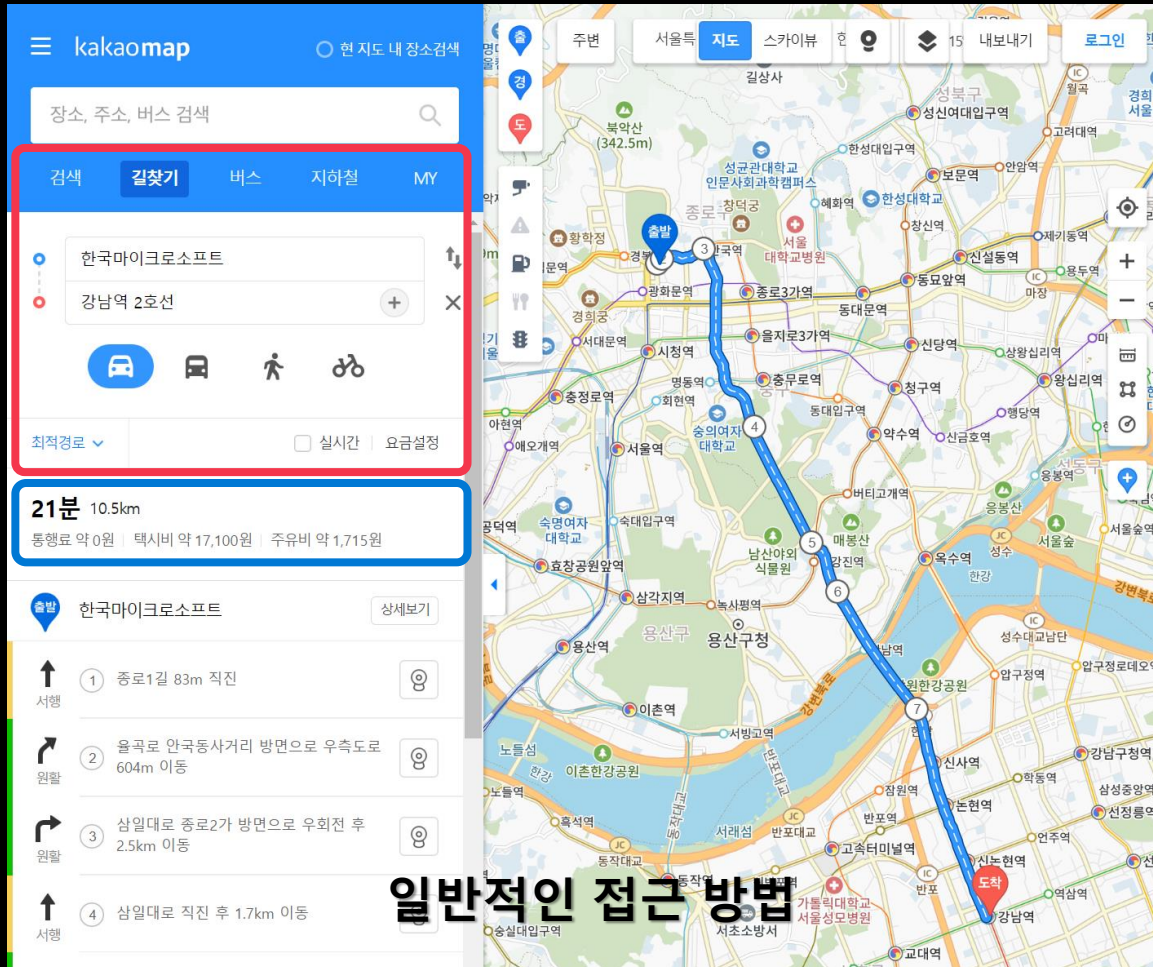


모델별 리전 지원 및 특성

Model ID	Base model Regions	Deprecation Date	Default Limit Value (TPM)	Max Request (tokens)	Training Data (up to)	Pricing (per 1,000 tokens)	
text-embedding-ada-002 (v2)	East US, South Central US, West Europe, France Central	2025-04-03	240K	8,191	Sep 2021	\$0.0001	
	Australia East, Canada East, East US2, Japan East, North Central US, Norway East, Sweden Central, Switzerland North, UK South, West US, South India, Brazil South	2025-04-03	350K				
text-embedding-3 (S/L)	Canada East, East US, East US 2	2025-02-02	350K	8,191		Large: \$0.00013	Small: \$0.00002
gpt-35-turbo-instruct (0914)	East US, Sweden Central	2024-09-30	240K	4,097		\$0.0015	\$0.0020
gpt-35-turbo (0301)	East US, France Central, South Central US, UK South, West Europe	2024-07-05	240K	4,096		\$0.0015	\$0.0020
gpt-35-turbo (0613)	East US, South Central US, West Europe, France Central, UK South North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	2024-06-13 2024-06-13	240K 300K			\$0.0015	\$0.0020
gpt-35-turbo-16K (0613)	East US, South Central US, West Europe, France Central, UK South	2024-06-13	240K	16,384		\$0.0030	\$0.0040
	North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	2024-06-13	300K				
gpt-35-turbo (1106)	Australia East, Canada East, France Central, South India, Sweden Central, UK South, West US	2024-11-06	120K	16,385 / 4,096		\$0.0010	\$0.0020
gpt-35-turbo (0125)	East US, Canada East, North Central US, South Central US	2025-02-15	120K	16,385 / 4,096		\$0.0005	\$0.0015
gpt-4 (0613)	East US, South Central US, West Europe, France Central	2024-07-05	20K	8,192	\$0.0300	\$0.0600	
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Sweden Central, Switzerland North	2024-07-05	40K				
gpt-4-32k (0613)	East US, South Central US, West Europe, France Central	2024-07-05	60K	32,768	\$0.0600	\$0.1200	
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Sweden Central, Switzerland North	2024-07-05	80K				
gpt-4 (1106-preview)	Australia East, Canada East, East US 2, France Central, UK South, West US	2024-03-31	80K	128,000 / 4,096	Apr 2023	\$0.0100	\$0.0300
	South India, Norway East, Sweden Central	2024-03-31	150K				
gpt-4 (1106-vision-preview)	Sweden Central, West US, Japan East, Switzerland North, Australia East	2024-03-31	30K	128,000 / 4,096	Apr 2023	\$0.0100	\$0.0300
gpt-4 (0125-preview)	East US, North Central US, South Central US	2024-03-31	80K	128,000 / 4,096	Dec 2023	\$0.0100	\$0.0300
dall-e-3 (preview)	East US, Sweden Central, Australia East	2024-07-15	12 req / 1m	4,000 (Max Char)	-	\$4~8(SD) / \$8~12(HD)	
whisper (preview)	East US 2, North Central US, West Europe, South India, Norway East, Sweden Central	2026-02-01	-	25 MB (Audio)	-	\$0.36 (per hour)	
tts, tts-hd (preview)	North Central US, Sweden Central	2025-03-01	-	-	-	\$15/\$30(HD) 1M characters	

사용자 인터페이스 진화

Function Calling – 주어진 형식과 절차 대신 자연스럽게 사람이랑 대화하듯 소통



한국마이크로소프트에서 강남역까지 얼마나 걸러?

OpenAI gpt-35-turbo(gpt-4): Function Calling

카카오모빌리티 길찾기 API (출발지, 목적지)

```
def gpt(input):
    messages = [
        {"role": "user", "content": input}
    ]
    assistant_response = run_conversation(messages, functions, available_functions, deployment_id)
    # assistant_response 값이 비어있을 경우
    if not assistant_response:
        return "제가 답변 드릴 수 있는 질문이 아닙니다. 다시 질문해주세요."
    else:
        content = json.dumps(assistant_response['choices'][0]['message']['content'], ensure_ascii=False, indent=4)
        content = content.replace("\n", "\n").replace("\\\"", "\"")
        return content
```

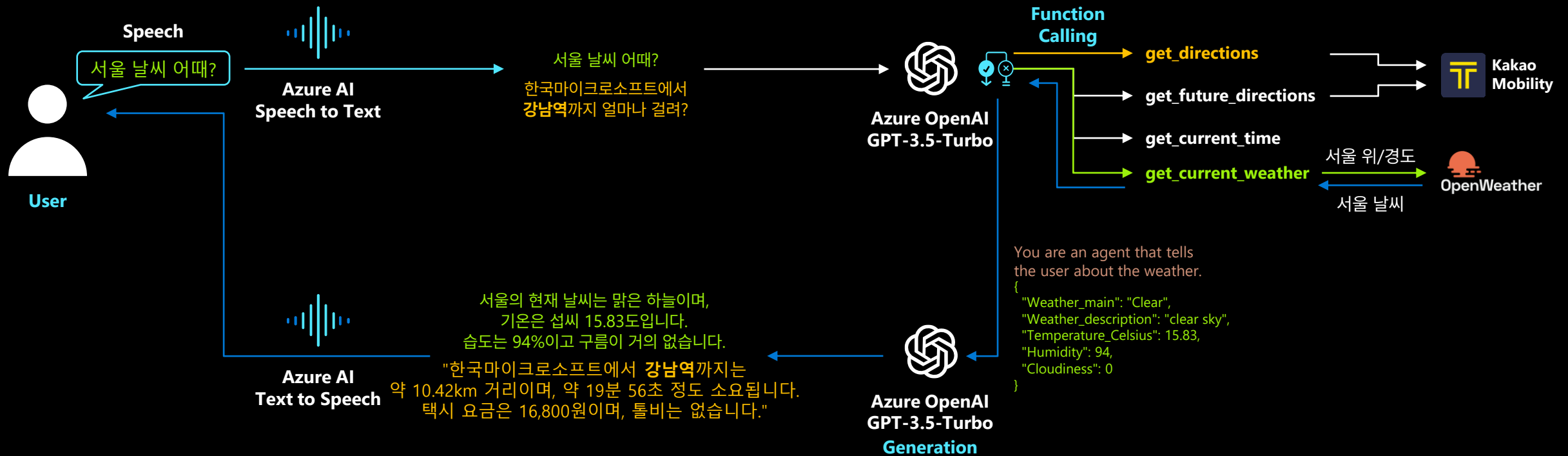
origin_name = "한국마이크로소프트"
 destin_name = "강남역2호선"
 query = f"{origin_name}에서 {destin_name}까지 얼마나 걸러?"

print(gpt(query))
 4.4s
 "한국마이크로소프트에서 강남역 2호선까지는 약 10.42km 거리이며, 차로 이동하는 경우 약 19분 56초가 소요됩니다. 택시 요금은 약 16,800원"

ChatGPT에 외부 API를 결합

사용자 인터페이스 진화

Function Calling – 주어진 형식과 절차 대신 자연스럽게 사람이랑 대화하듯 소통



EXPLORER

- ▼ AZURE-OPENAI-SAMPLES-...
 - > .devcontainer
 - > archive
 - > documentation
 - > fundamentals
 - > Python-3.11.1
 - ▼ quick_start
 - > assets
 - > data
 - 01_create_resource.i...
 - 02_OpenAI_ge... M
 - 03_completion_api.i...
 - 04_code_generation...
 - 05_tokens_and_usag...
 - 06_OpenAI_paramet...
 - 07_best_practice.ipy...
 - 08_prompt_engineer...
 - 09_LLM_chain_... M
 - 10_Function_c... M
 - 11_my_jarvis_st... U
 - <> error.html U
 - README.md
 - test.ipynb U

OUTLINE

TIMELINE

11_my_jarvis_stt_chatgpt_tts.ipynb U

스! 음성 비서 만들기 > M OpenAI Function Calling에서 활용하려는 함수들을 정의 > M GPT 모델에 자연어 기반 질의하는 함수 > M 이제부터는 여러분의 API 를 호

+ Code + Markdown | □ Interrupt ↺ Restart | ▶ Execute Group 1 ||▶ Execute Group 2 ≡ Clear All Outputs ... Python 3.11.5

```
# if __name__ == "__main__":
tts("날씨, 길찾기, 시간을 물어보세요.")

while True:
    result_stt = stt().text
    print(result_stt)
    if(result_stt == ""):
        # 음성 인식 실패
        print("음성 인식 실패")
        tts("음성인식에 실패했습니다. 다시 말씀해 주세요.")
    elif(result_stt == "나가기." or result_stt == "종료."):
        print("대화 종료")
        break
    else:
        # 음성인식 성공
        result_gpt = gpt(result_stt)
        tts(result_gpt)
```

[] ↺ Python

... Speech synthesized to speaker for text [날씨, 길찾기, 시간을 물어보세요.]
 말씀 하세요~
 음성인식결과: 나가기.
 나가기.
 대화 종료



코드를 몰라도 OK!

Advanced Data Analysis (Code Interpreter) – 말하면 코드를 만들고 실행하여 결과 생성

Advanced Data Analysis



prompt_0911_survey.CSV
File

프롬프트 워크샵의 설문조사 데이터입니다. 분석해서 고객의 만족도를 시각화 해줘.



물론입니다. 먼저 설문조사 데이터를 불러와 내용을 확인해보겠습니다.

Finished working Show work

설문조사 데이터가 다음과 같은 열로 구성되어 있습니다:

- **ID:** 응답자의 ID
- **Name:** 응답자의 이름
- **회사명:** 응답자가 속한 회사 이름
- **직책:** 응답자의 직책
- **Email:** 응답자의 이메일 주소
- **Session CSAT:** 세션에 대한 고객 만족도 (Customer Satisfaction, CSAT) 점수
- **OpenAI understanding:** OpenAI에 대한 이해도 점수
- **세션 피드백:** 세션에 대한 피드백
- **그외 관심 주제:** 다른 관심있는 주제
- **개인정보 수집 동의:** 개인정보 수집에 대한 동의 여부
- **마지막으로 수정한 시간:** 응답을 마지막으로 수정한 시간

고객 만족도를 중점으로 시각화를 진행하겠습니다. "Session CSAT"과 "OpenAI understanding" 두 가지 지표를 사용하여 분석을 진행해보겠습니다. 어떻게 시각화를 진행할까요? 바 차트, 파이 차트, 히트맵 등 다양한 방법이 있습니다.

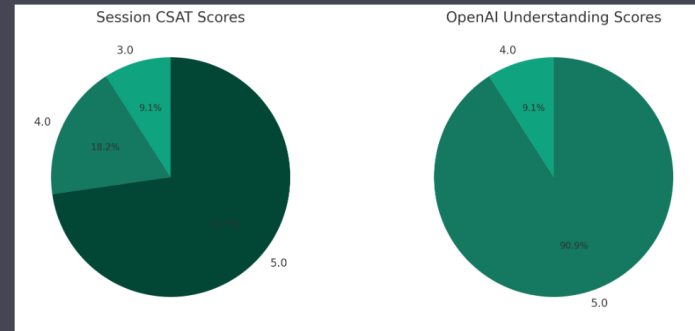


만족도 점수는 1점부터 5점까지 점수를 줄 수 있고, 파이차트로 시각화 해줘.



알겠습니다. "Session CSAT"과 "OpenAI understanding" 두 가지 지표에 대해 각 점수별 응답자 수를 파이 차트로 시각화하겠습니다.

Finished working Show work



두 개의 파이 차트가 있습니다.

1. **Session CSAT Scores:** 세션에 대한 고객 만족도 (CSAT) 점수 분포입니다. 대부분의 응답자가 만족도를 높게 평가한 것으로 보입니다.
2. **OpenAI Understanding Scores:** OpenAI에 대한 이해도 점수 분포입니다. 여기에서도 대부분의 응답자가 높은 점수를 주었습니다.

이 결과를 통해 워크샵이 전반적으로 높은 만족도와 이해도를 얻었다고 판단할 수 있습니다. 다른 분석이나 시각화가 필요하시면 알려주세요.

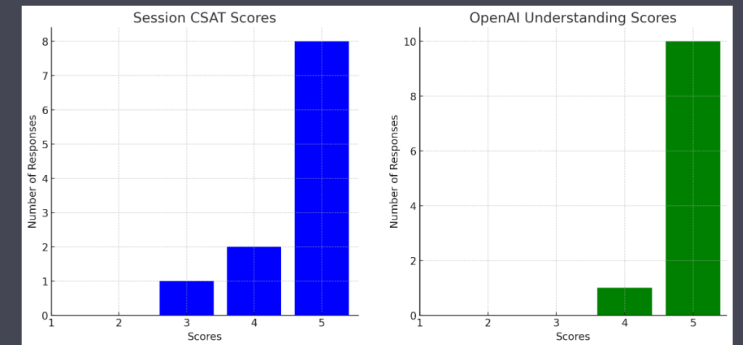


바차트로 시각화 해줘.



물론입니다. "Session CSAT"과 "OpenAI understanding" 지표를 바 차트로 시각화하겠습니다.

Finished working Show work



두 개의 바 차트가 있습니다.

1. **Session CSAT Scores:** 세션에 대한 고객 만족도 (CSAT) 점수 별 응답자 수입니다. 대부분의 응답자가 5점을 주었음을 볼 수 있습니다.
2. **OpenAI Understanding Scores:** OpenAI에 대한 이해도 점수 별 응답자 수입니다. 이 경우에도 대부분의 응답자가 5점을 주었습니다.

바 차트를 통해 높은 점수가 주로 얼마나 많이 나왔는지 명확하게 파악할 수 있습니다. 다른 분석이나 시각화가 필요하시면 알려주세요.

코드를 몰라도 OK!

Advanced Data Analysis (Code Interpreter) – 만들어진 코드 재활용

Advanced Data Analysis

prompton_0911_survey.CSV
File

프롬프트 워크샵의 설문조사 데이터입니다. 분석해서 고객의 만족도를 시각화 해줘.

물론입니다. 먼저 설문조사 데이터를 불러와 내용을 확인해보겠습니다.

```
Finished working Hide work ^
```

```
python Copy code
```

```
import pandas as pd

# Load the survey data
file_path = '/mnt/data/prompton_0911_survey.CSV'
survey_data = pd.read_csv(file_path)

# Display the first few rows of the data
survey_data.head()
```

RESULT	ID	Name	회사명	직책	Email	Session CSAT	\
0	1.0	AAA	BBA	CSO	aaa@abc.com	5.0	
1	2.0	AAB	BBB	CTO	aaa@abc.com	5.0	
2	3.0	AAC	BBC	PO	aaa@abc.com	4.0	
3	4.0	AAD	BBC	매니저	aaa@abc.com	5.0	
4	5.0	AAE	BBD	CEO / Founder	aaa@abc.com	5.0	

OpenAI understanding

	OpenAI understanding	세션	이
0	5.0	애저	관련한 설명 보태주시면 현재 클라우드 결정 시 도움이
1	5.0	이론과	실전의 균형이 잘 맞는 좋
2	5.0	프롬프트	팁과 리스스들이 유용했습니다. 특히 여러 유클레를

만족도 점수는 1점부터 5점까지 점수를 줄 수 있고, 파이차트로 시각화 해줘.

알겠습니다. "Session CSAT"과 "OpenAI understanding" 두 가지 지표에 대해 각 점수별 응답자 수를 파이 차트로 시각화하겠습니다.

```
Finished working Hide work ^
```

```
python Copy code
```

```
import matplotlib.pyplot as plt

# Filter out any NaN values in the CSAT columns
filtered_data = survey_data.dropna(subset=['Session CSAT', 'OpenAI understandin

# Prepare data for pie charts
session_csat_counts = filtered_data['Session CSAT'].value_counts().sort_index
openai_understanding_counts = filtered_data['OpenAI understanding'].value_co

# Plotting
fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# Plot for Session CSAT
axes[0].pie(session_csat_counts, labels=session_csat_counts.index, autopct='%
axes[0].set_title('Session CSAT Scores')

# Plot for OpenAI understanding
axes[1].pie(openai_understanding_counts, labels=openai_understanding_counts.
axes[1].set_title('OpenAI Understanding Scores')

plt.show()
```

바차트로 시각화 해줘.

물론입니다. "Session CSAT"과 "OpenAI understanding" 지표를 바 차트로 시각화하겠습니다.

```
Finished working Hide work ^
```

```
python Copy code
```

```
# Plotting bar charts for Session CSAT and OpenAI Understanding
fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# Bar chart for Session CSAT
axes[0].bar(session_csat_counts.index, session_csat_counts.values, color='b1
axes[0].set_title('Session CSAT Scores')
axes[0].set_xlabel('Scores')
axes[0].set_ylabel('Number of Responses')
axes[0].set_xticks(range(1, 6))

# Bar chart for OpenAI understanding
axes[1].bar(openai_understanding_counts.index, openai_understanding_counts.v
axes[1].set_title('OpenAI Understanding Scores')
axes[1].set_xlabel('Scores')
axes[1].set_ylabel('Number of Responses')
axes[1].set_xticks(range(1, 6))

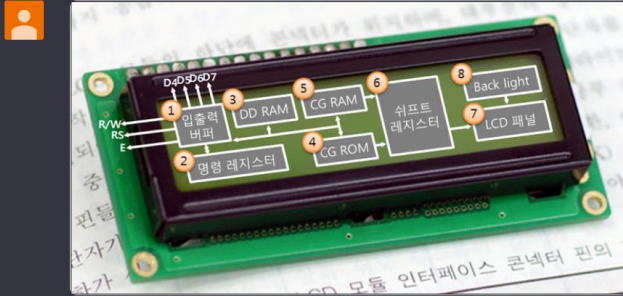
plt.show()
```

바 차트를 통해 높은 점수가 주로 얼마나 많이 나왔는지 명확하게 파악할 수 있습니다. 다른 분석이나 시각화가 필요하시면 알려주세요.

이미지도 OK!

GPT-4 V / DALL-E 3 – 이미지를 해석하고 생성합니다.

GPT-4



이 이미지가 어떤 것을 의미하는지 분석해서 쉽게 설명해줘.

이 이미지는 LCD (Liquid Crystal Display) 모듈의 내부 구조와 주요 부품들을 나타내고 있습니다. 각 부분에 대한 설명은 다음과 같습니다:

- 입력 핀:** LCD 모듈과 외부 장치 간의 데이터 전송을 위한 핀입니다.
- 명령 레지스터:** LCD에 어떤 작업을 수행할지 지시하는 명령들을 저장합니다.
- DD RAM:** 화면에 표시될 문자나 숫자의 데이터를 저장하는 램입니다.
- CG ROM:** 사용자가 화면에 표시할 수 있는 기본 문자나 기호의 패턴을 저장하고 있는 ROM입니다.
- CG RAM:** 사용자가 정의한 문자나 기호의 패턴을 저장하는 램입니다.
- 쉬프트 레지스터:** 데이터의 순서대로 전송을 도와줍니다.
- LCD 패널:** 실제로 데이터나 문자, 기호 등이 표시되는 부분입니다.
- Back light:** LCD 패널의 뒷면에 위치하며 화면을 밝게 해주는 역할을 합니다.

간단히 말하면, 이 LCD 모듈은 외부 장치로부터 데이터나 명령을 받아 화면에 문자나 기호를 표시하는 역할을 합니다.

Is this conversation helpful so far?

DALL-E 3

패브릭으로 제작되어 안아주면 포근한 느낌이 드는 귀엽고 사랑스러운 얼굴을 가진 한국 나이로 7세 정도 되는 손자, 손녀 인형을 전신 사이즈로 서로 손을 맞잡고 있는 느낌으로 실물 제품처럼 그려줘.

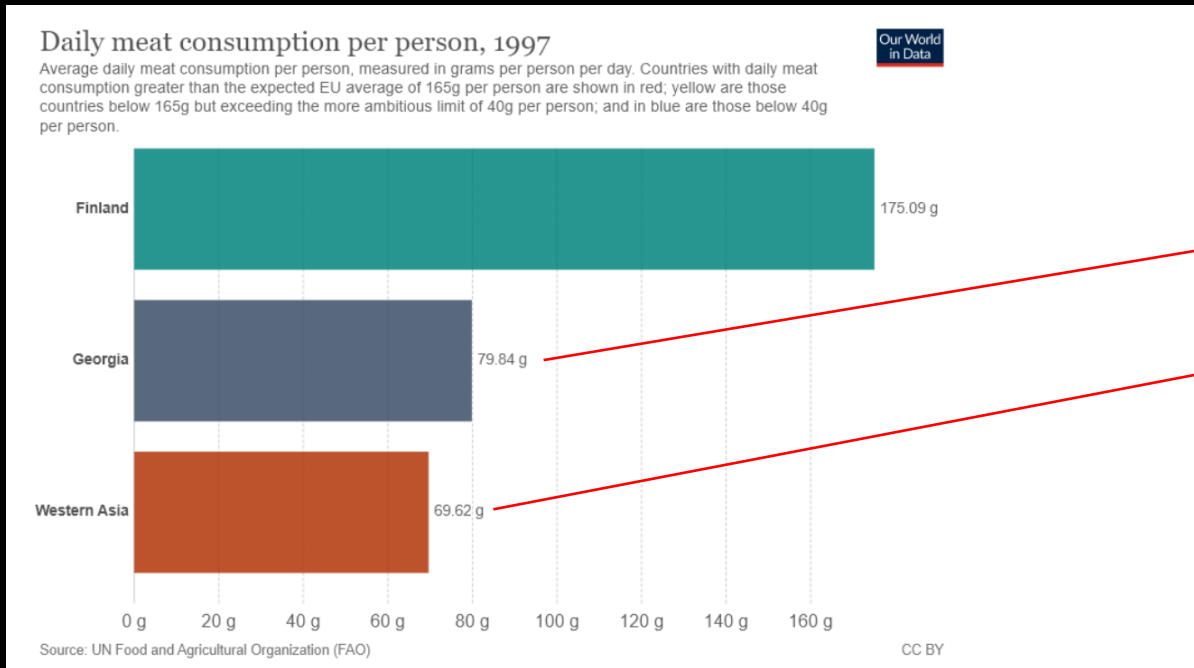
DALL-E 3
Created 4 images



이미지도 풀어보자! GPT-4 Turbo with Vision

사용자

조지아와 서아시아의 평균 일일 육류 소비량의 합은 얼마입니까?
답변을 제공하기 전에 단계별 추론을 제공하십시오.



GPT-4

조지아와 서아시아의 평균 일일 육류 소비량의 합계를 찾으려면 우리는 다음 단계를 따를 것입니다:

1. 조지아의 일일 평균 육류 소비량을 확인하십시오.
2. 서아시아의 평균 일일 육류 소비량을 확인합니다.
3. 1단계와 2단계의 값을 더합니다.

1단계: 제공된 차트에서 조지아의 일일 평균 육류 소비량이 1인당 하루 평균 79.84g임을 알 수 있습니다.

2단계: 차트에서 서아시아의 일일 평균 육류 소비량이 1인당 하루 69.62g임을 알 수 있습니다.

3단계: 1단계와 2단계의 값을 더합니다. $79.84\text{g}(\text{조지아}) + 69.62\text{g}(\text{서아시아}) = 149.46\text{g}$

따라서 조지아와 서아시아의 평균 일일 육류 소비량의 합계는 1인당 하루 149.46g입니다.

이미지도 풀어보자! GPT-4 Turbo with Vision

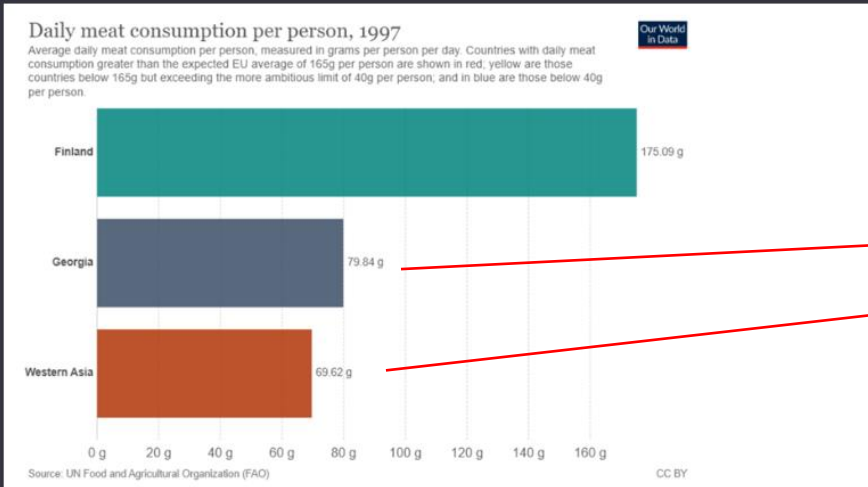
사용자

GPT-4

조지아와 서아시아의 평균 일일 육류 소비량의 합은 얼마입니까?
답변을 제공하기 전에 단계별 추론을 제공하십시오.

ChatGPT 4

You



조지아와 서아시아의 평균 일일 육류 소비량의 합은 얼마입니까?
답변을 제공하기 전에 단계별 추론을 제공하십시오.



ChatGPT

To find the total average daily meat consumption for Georgia and Western Asia, we will add their individual average consumptions together.

The steps are as follows:

1. Identify the average daily meat consumption for Georgia from the chart.
2. Identify the average daily meat consumption for Western Asia from the chart.
3. Add the two values together to get the total.

Looking at the image provided:

- Georgia's average daily meat consumption is 79.84 grams per person.
- Western Asia's average daily meat consumption is 69.62 grams per person.

Now, let's add these two numbers together.

$$\text{Total average daily meat consumption} = 79.84g + 69.62g$$

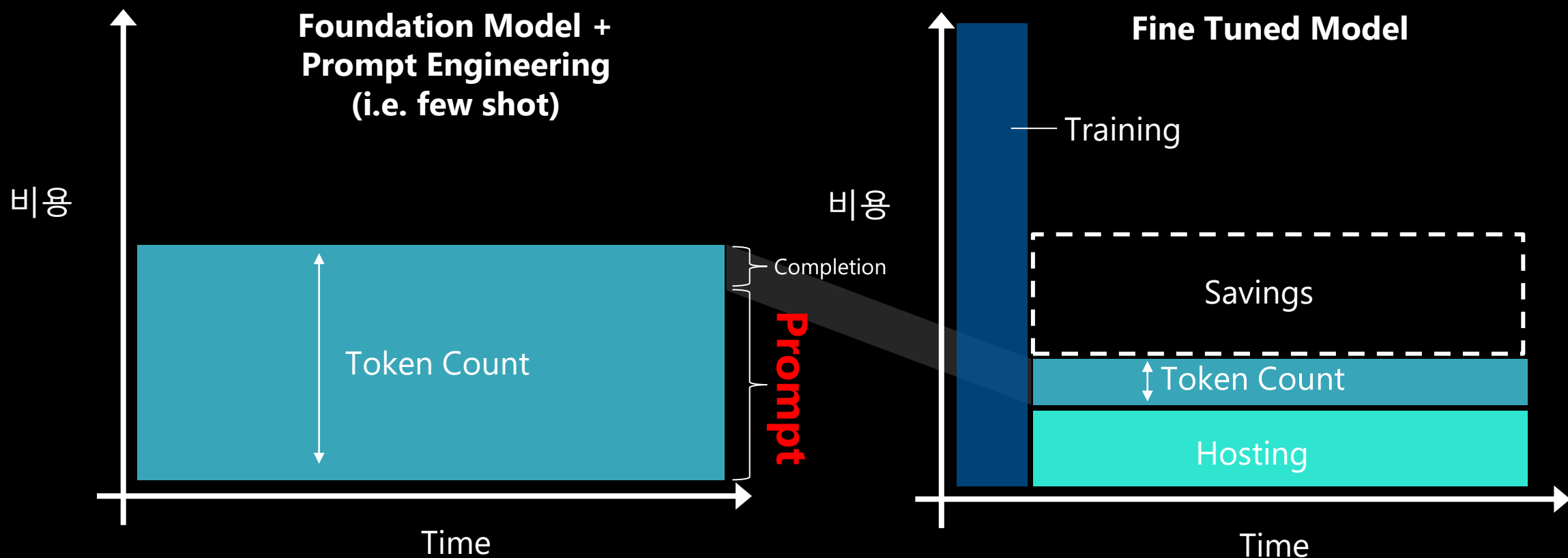
$$\text{Total average daily meat consumption} = 149.46g$$

So, the sum of the average daily meat consumption for Georgia and Western Asia is 149.46 grams.




맞춤형 LLM

대량의 Prompt 토큰 절감, 맞춤형 스타일로 최적화 – **Fine-tuning**



LLM App 제작을 더 쉽고 빠르게!

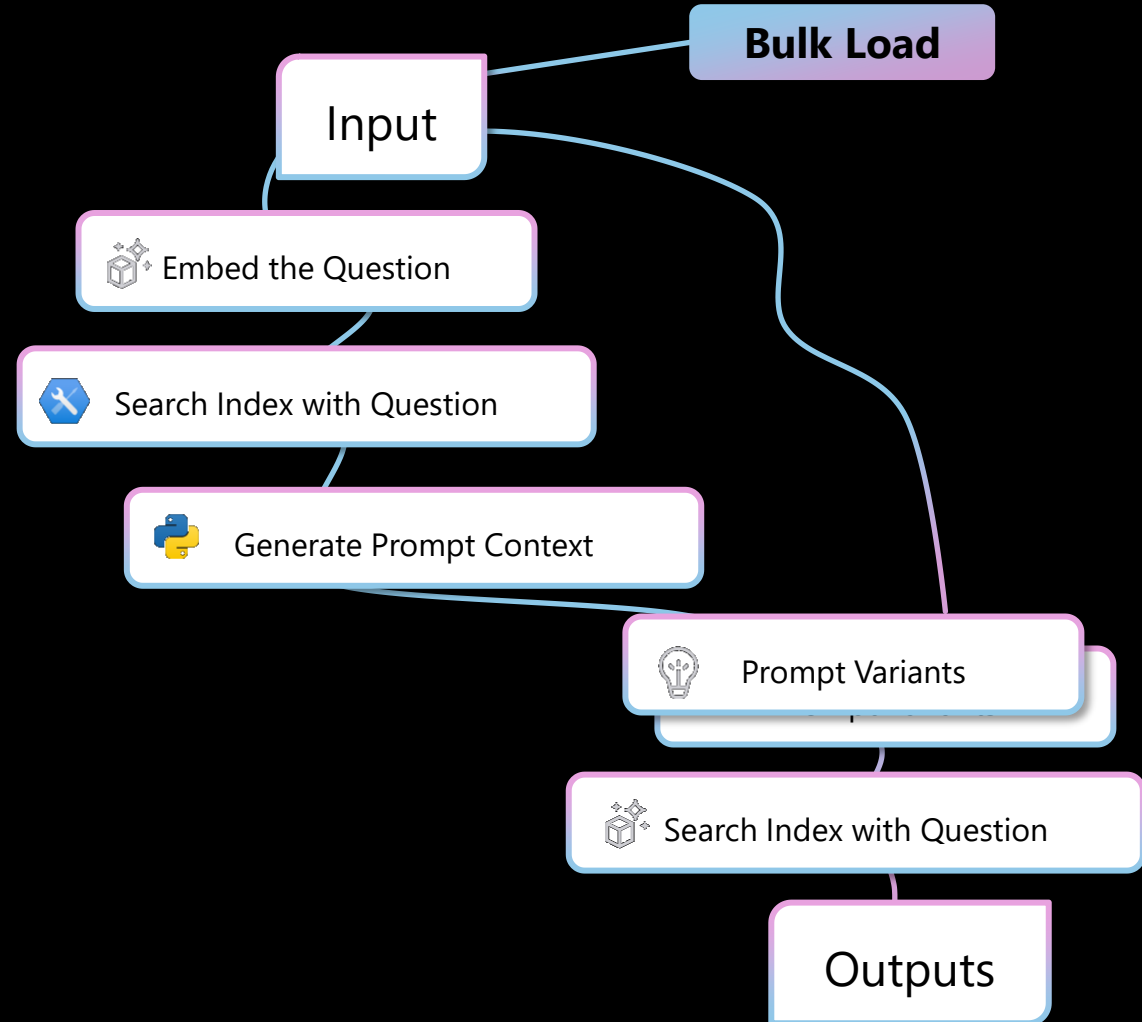
Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구

 Orchestrates AI models, prompts, and APIs

 Support for prompt tuning and experimentation

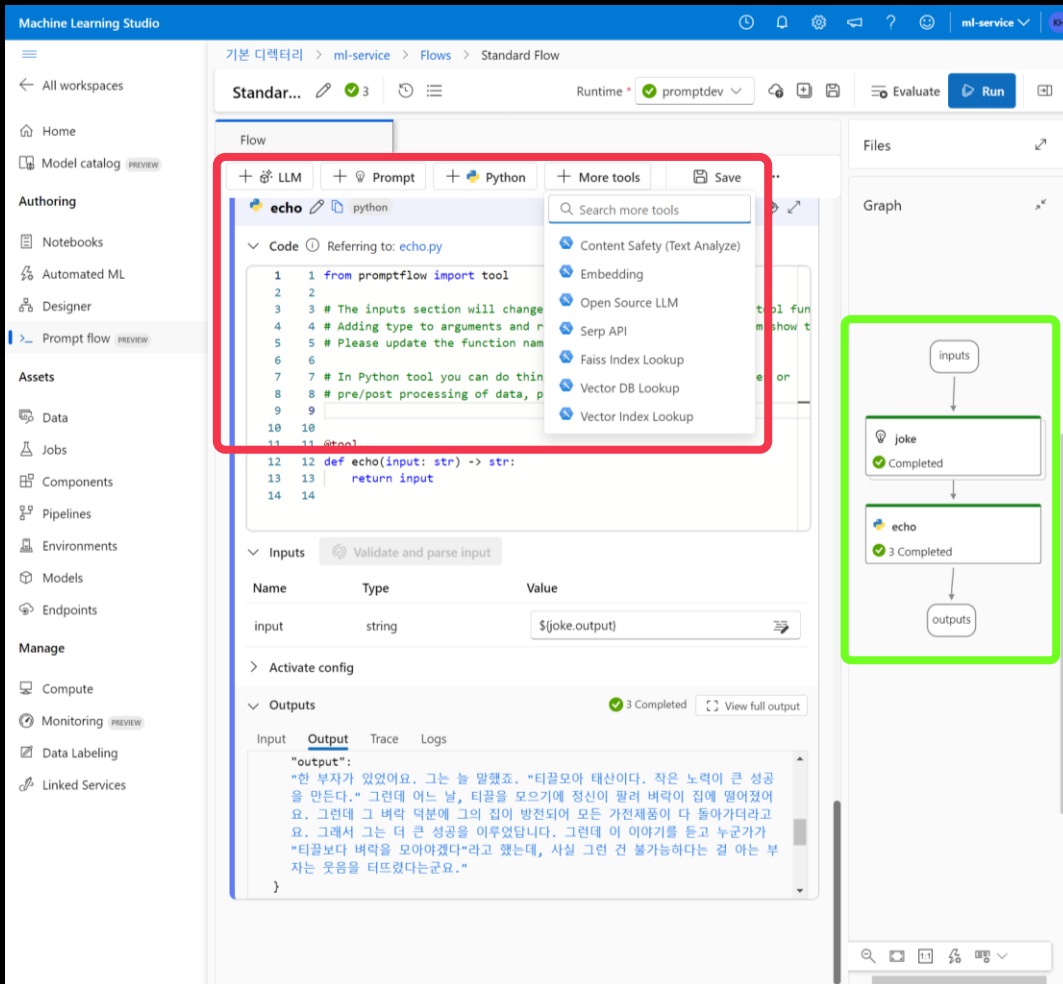
 Blue/green deployments and testing

 Supports Semantic Kernel, LangChain, Plugins



각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구



워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

프롬프트 튜닝

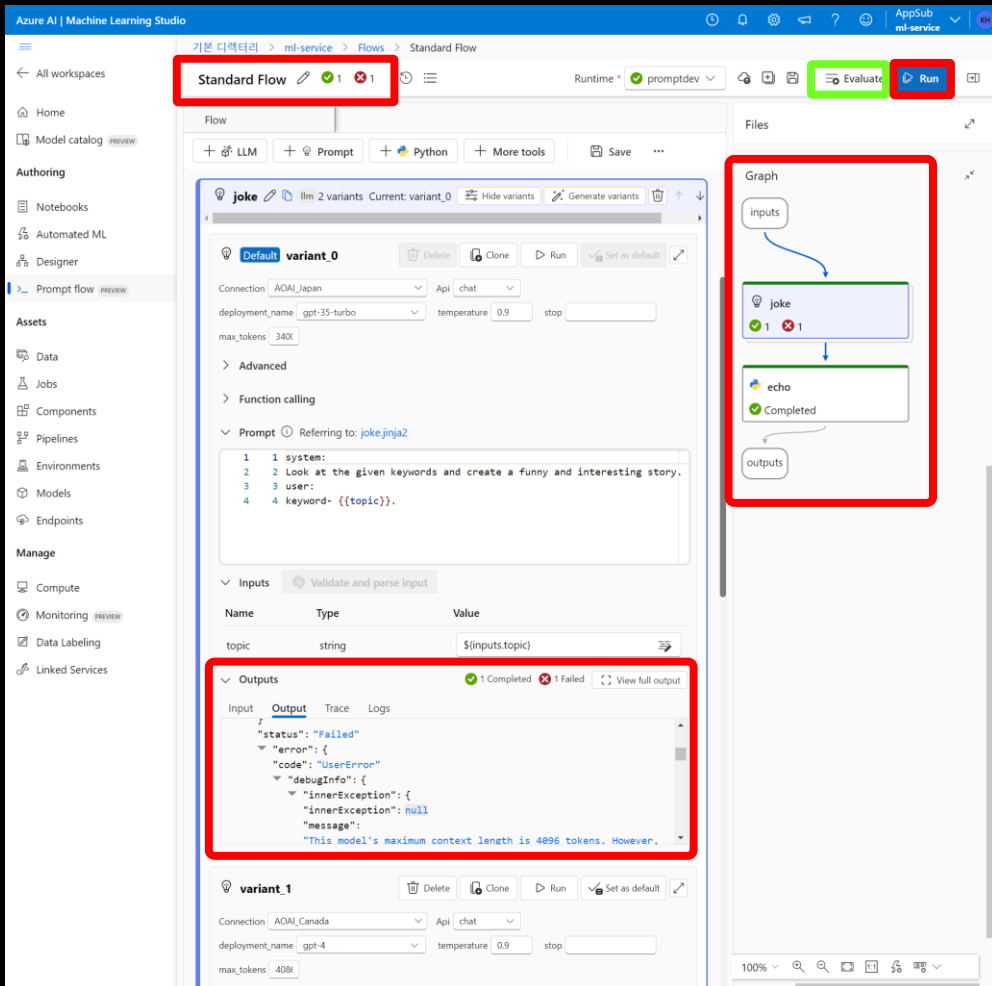
- 변형 및 버저닝으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 원 클릭 배포

각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구



워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성하여 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

프롬프트 튜닝

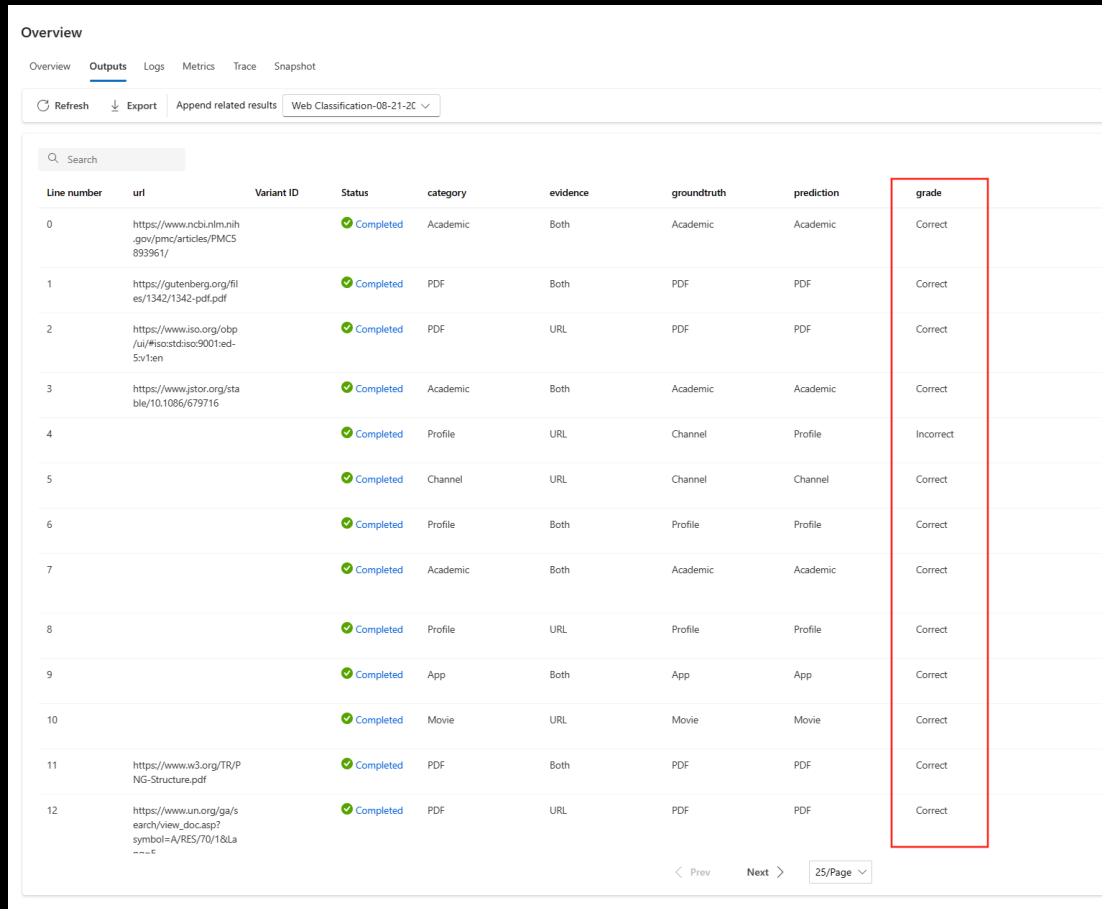
- 변형 및 버저닝으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 원 클릭 배포

각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구



Overview

Overview **Outputs** Logs Metrics Trace Snapshot

Refresh Export Append related results Web Classification-08-21-20

Search

Line number	uri	Variant ID	Status	category	evidence	groundtruth	prediction	grade
0	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5893961/		Completed	Academic	Both	Academic	Academic	Correct
1	https://gutenberg.org/files/1342/1342-pdf.pdf		Completed	PDF	Both	PDF	PDF	Correct
2	https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-5:vi:en		Completed	PDF	URL	PDF	PDF	Correct
3	https://www.jstor.org/stable/10.1086/679716		Completed	Academic	Both	Academic	Academic	Correct
4			Completed	Profile	URL	Channel	Profile	Incorrect
5			Completed	Channel	URL	Channel	Channel	Correct
6			Completed	Profile	Both	Profile	Profile	Correct
7			Completed	Academic	Both	Academic	Academic	Correct
8			Completed	Profile	URL	Profile	Profile	Correct
9			Completed	App	Both	App	App	Correct
10			Completed	Movie	URL	Movie	Movie	Correct
11	https://www.w3.org/TR/PNG-Structure.pdf		Completed	PDF	Both	PDF	PDF	Correct
12	https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&La...		Completed	PDF	URL	PDF	PDF	Correct

< Prev Next > 25/Page

워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성하여 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

프롬프트 튜닝

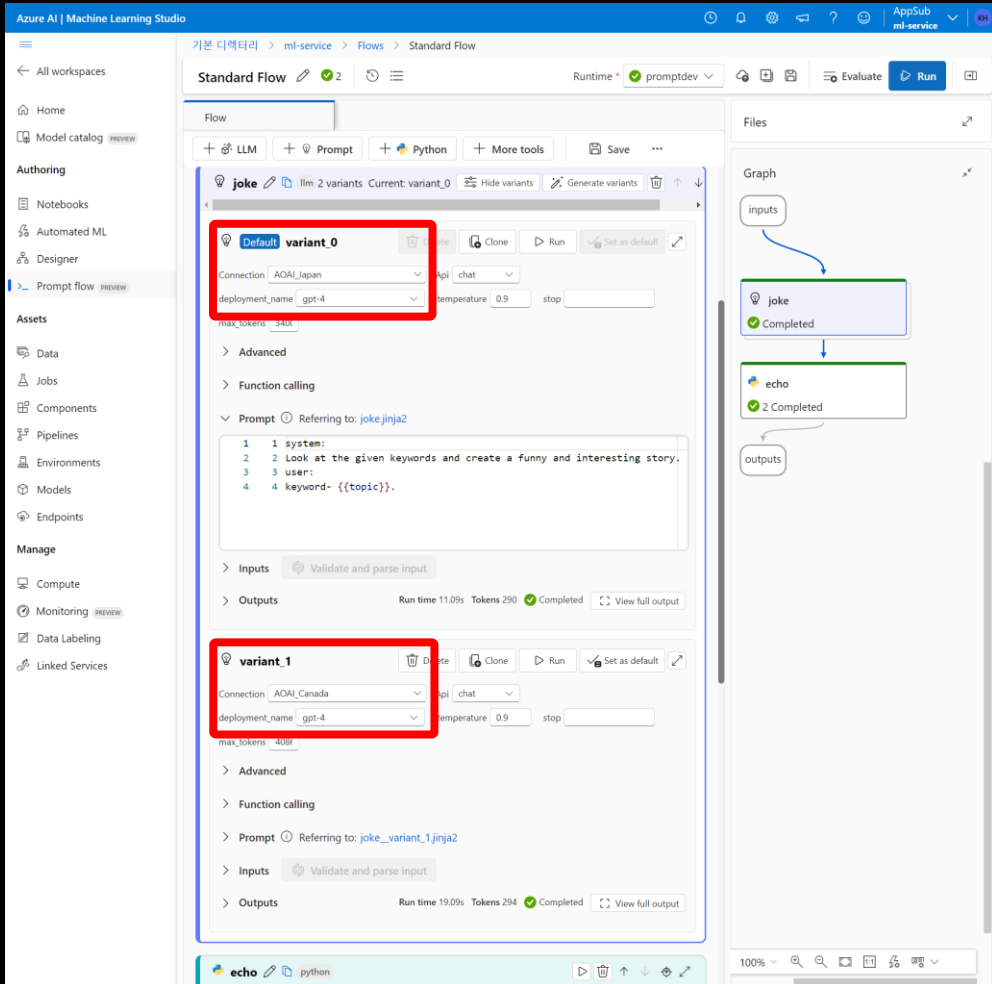
- 변형 및 버저닝으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 원 클릭 배포

각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구



워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성하여 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

프롬프트 튜닝

- **변형 및 버저닝**으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 원 클릭 배포

각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구

The screenshot displays the 'Outputs' window in Azure AI Machine Learning Studio. A table with columns 'name', 'duration', 'start time (UTC)', and 'end time (UTC)' is highlighted with a red border. Below the table, a JSON configuration for an LLM call is visible, including fields like 'api_base', 'api_type', 'api_version', 'engine', 'max_tokens', and 'messages'.

name	duration	start time (UTC)	end time (UTC)
▼ AzureOpenAI.chat	11,084 ms	2023-10-25 11:43:0:590	2023-10-25 11:43:11:674
openai.api_resources.chat.completion.ChatCor	11,081 ms	2023-10-25 11:43:0:593	2023-10-25 11:43:11:674
echo	0 ms	2023-10-25 11:43:11:694	2023-10-25 11:43:11:694
▼ AzureOpenAI.chat	19,090 ms	2023-10-25 11:43:12:486	2023-10-25 11:43:31:576
openai.api_resources.chat.completion.ChatCor	19,087 ms	2023-10-25 11:43:12:489	2023-10-25 11:43:31:576
echo	0 ms	2023-10-25 11:43:31:600	2023-10-25 11:43:31:601

```
{
  "inputs": {
    "api_base": "https://japanopenai.azure.com/",
    "api_type": "azure",
    "api_version": "2023-07-01-preview",
    "engine": "gpt-4",
    "frequency_penalty": 0,
    "logit_bias": {}
  },
  "max_tokens": 3400,
  "messages": [
    {
      "content": "Look at the given keywords and create a funny and interesting story. Limit the length of your story to a maximum of 100 words. Output language is Korean.",
      "role": "system"
    },
    {
      "content": "keyword- 티끌, 비락, 부자, 성공.",
      "role": "user"
    }
  ],
  "n": 1,
  "presence_penalty": 0,
  "stop": null,
  "stream": true,
  "temperature": 0.9,
  "top_p": 1,
  "user": ""
}
```

워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성하여 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

프롬프트 튜닝

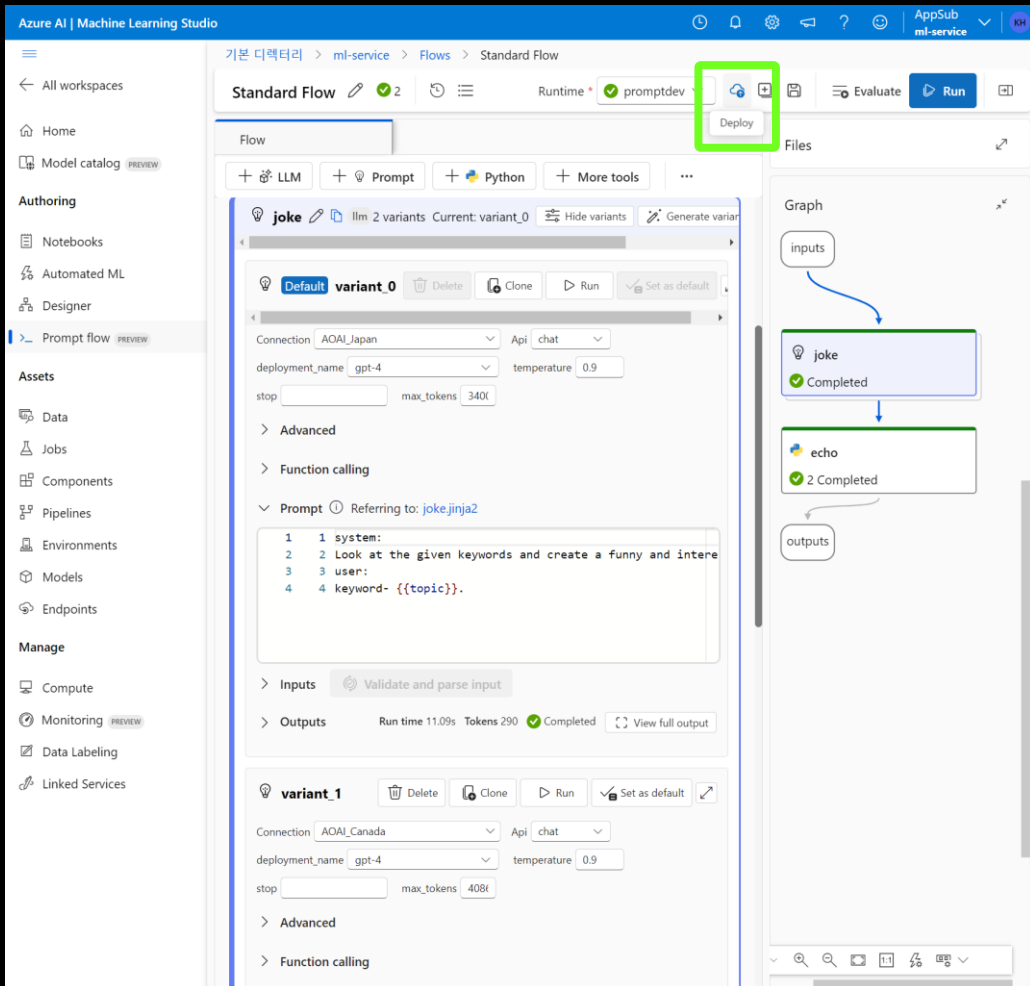
- 변형 및 버저닝으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 원 클릭 배포

각 기능들을 연결하여 해결

Prompt Flow – LLM 기반 응용 프로그램을 위한 전체 개발 주기를 단순화하는 도구



워크플로우 개발

- 다양한 LLM 모델, 외부 데이터와 연결하고, 코드를 작성하여 시각화된 맞춤형 워크플로우 설계

테스트 및 평가

- 성능, 근거, 정확성과 같은 측정 항목을 사용하여 워크플로의 AI 품질을 평가
- 대규모 데이터 세트를 병렬로 테스트

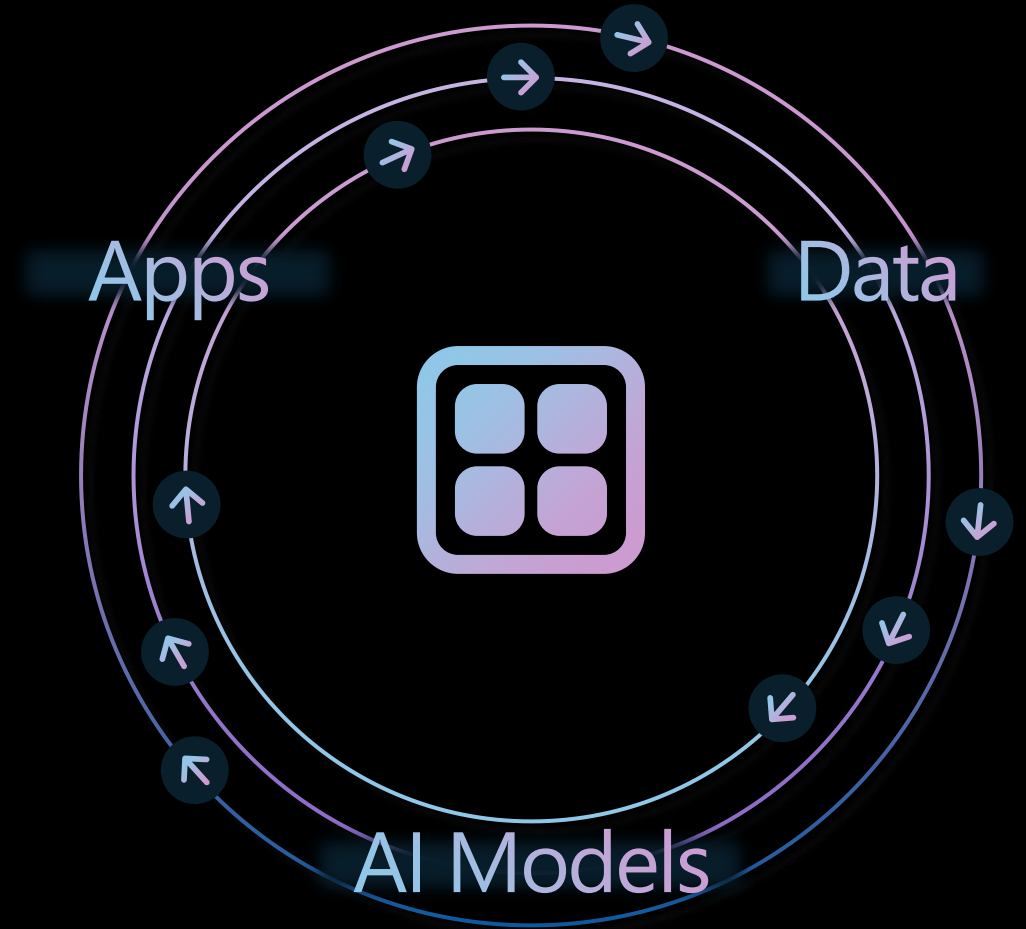
프롬프트 튜닝

- 변형 및 버저닝으로 간단하게 프롬프트를 조정

비교 및 배포

- 여러 실험을 시각적으로 비교
- 신속한 통합을 위해 관리형 엔드포인트에 **원 클릭 배포**

AI 시대에는,
애플리케이션은
지능적이어야 합니다.



한국 스타트업들의 활용 사례

생성형 AI 서비스를 잘 활용하는 스타트업

한국 스타트업 기업들의 생성형 AI를 다뤄본 노하우 대방출



(메뉴 추천)



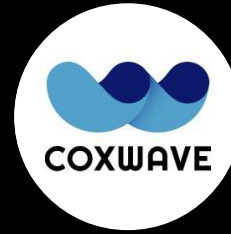
(고객 문의)



(리뷰 요약)



(생성 AI 솔루션)



(자체 검색)



(리뷰 작성)



(타로, 상담)



(돌봄서비스 로봇)



(고객 센터)



(학습 진단)



(금융 판단)



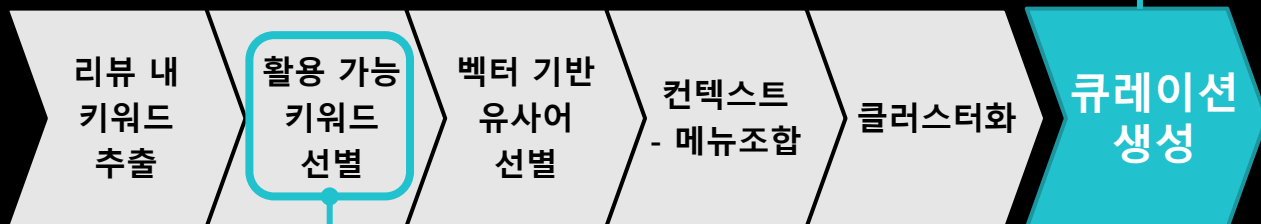
(증권 정보)

AI가 제안하는 추천

리뷰를 분석하여 메뉴 및 매장 추천

먹을 준비 되신 분이라면 주목하세요 🤖
양 많은 떡볶이집이 있으면 좋겠지만, 어디가 좋은 곳인지 모르시죠? 제가 알려드릴게요!

사람은 퀄리티 체크를 담당하며 AI를 도와요!



“우리는 이것을 **컨텍스트**라고 부르기로 했어요”

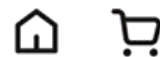
- 시간대 (아침, 점심, 저녁, 야식..)
- 함께 먹는 사람 (아이, 가족, 사랑하는 사람, 혼자..)
- 상황 (퇴근, 스트레스, 축하, 식단..)
- 맛 (맵다, 달다, 탱글탱글..)



AI가 제안하는

가게를 확인하고
주문해 보세요!

← 가족과 즐기는 떡볶이



GPT-4 활용 추천 ①



먹을 준비 되신 분이라면 주목하세요 🤖
양 많은 떡볶이집이 있으면 좋겠지만, 어디가 좋은 곳인지 모르시죠? 제가 알려드릴게요!

#가족

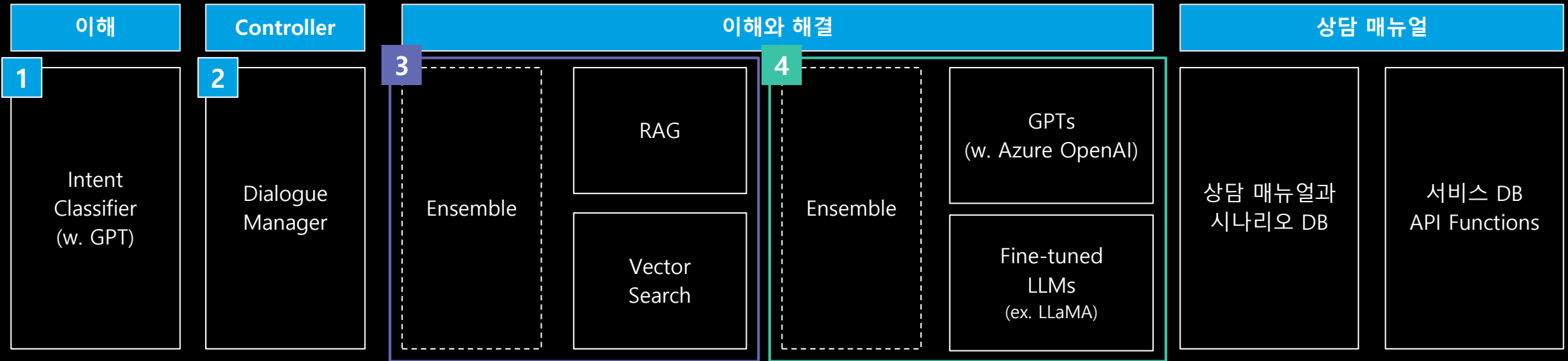
#떡볶이

#양 많은

고객경험 증대를 위한 고객센터



빠르고 정확하게 고객들의 고민을 해결



1 Intent Classifier: 사전 설정한 Intent 중 가장 관련성 높은 Intent를 결정

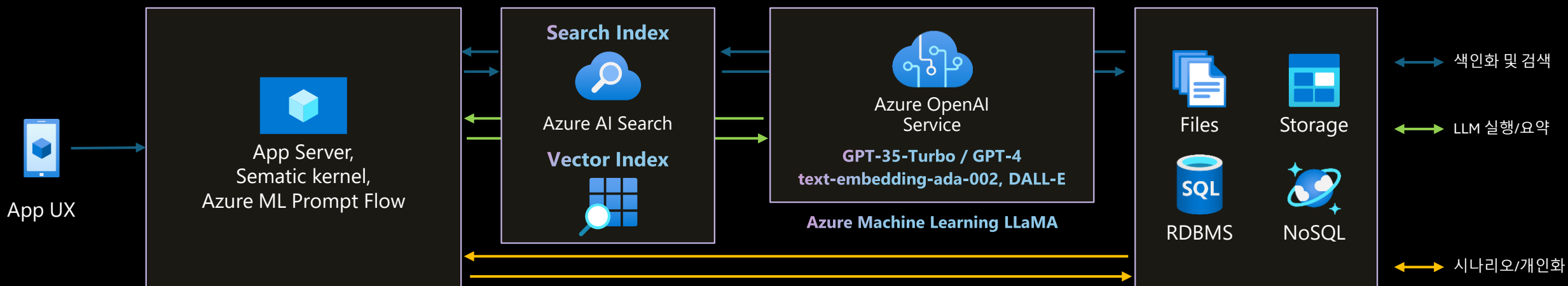
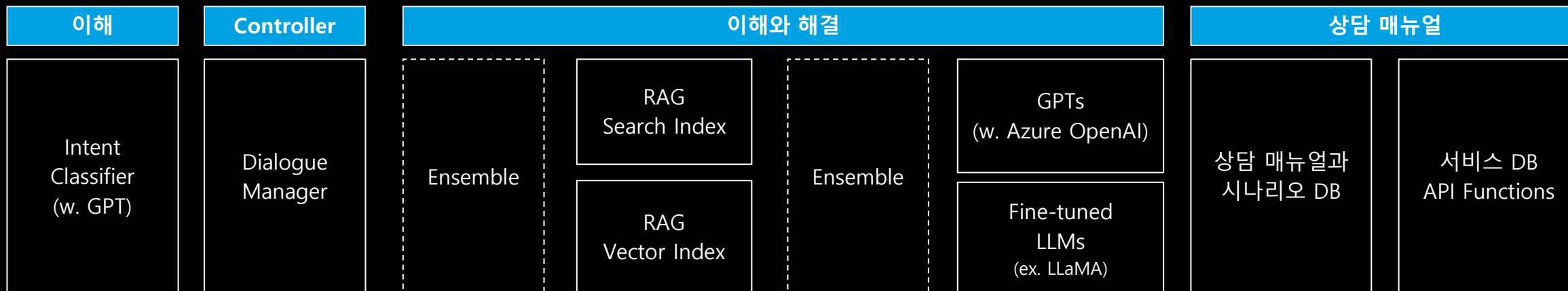
2 Dialogue Manager: 선택한 Intent에 맞게 이후 프로세스를 운영

3 RAG (Retrieval-Augmented Generation)와 Vector Search: 사전 구축 매뉴얼과 시나리오 DB 조회

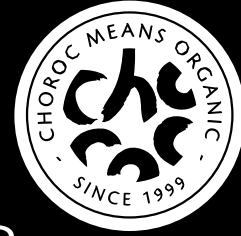
4 LLMs: Azure OpenAI와 자체 sLLM을 조합하여 사용

Azure 기반 서비스 구축 방법

용도에 맞는 완전 관리형 서비스로 안전하고 빠르게 구축



검색엔진 최적화 / 리뷰 작성 도우미



검색 품질을 향상시키고 상품 리뷰 작성을 쉽게 할 수 있도록 도움



시리얼

시리얼

상품 (2개)

매장 (0개)

신상품순 최고가격순 최저가격순



아몬드콘플레이크(240g)
6,300 원



유기농 현미플레이크(240g)
6,300 원



과자
유기농 그레놀라(사과&건포도/300g)
7,000원



과자
유기농 그레놀라(코코넛&레드베리/260g)
8,000원



과자
아몬드콘플레이크(240g)
30% 4,400원
6,300원



견과류
유기농 구운아몬드(150g)
13,900원



과일가공품
유기농 건망고(90g)
9,800원



견과류
유기농 코코넛칩(80g)
4,900원



잡곡
유기농オート밀(420g)
7,500원



견과류
헤이즐너트(200g)
9,900원



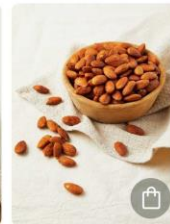
과자
유기농 현미코코아볼(240g)
6,300원



과자
유기농 현미플레이크(240g)
6,300원



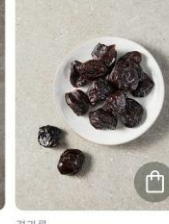
과자
유기농 후레쉬믹스볼(240g)
6,300원



견과류
유기농 구운 아몬드(450g)
19% 26,300원



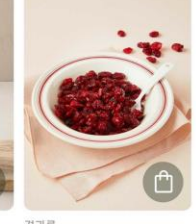
잡곡
유기농 검정보리(800g)
15% 5,600원
6,600원



견과류
유기농 무른(150g)
8,500원



견과류
유기농 아빠견과(420g)
25% 18,350원
24,600원



견과류
유기농 건크랜베리(110g)
9,400원

검색엔진 최적화 / 리뷰 작성 도우미



GPT-4를 활용한 검색 품질 향상



Dvı ž
ı̀ ööşı



Dvı ž
ı̀ Eš É ž



ı̀ žı ý
ı̀ ğvı ž

Raw data

Processed data

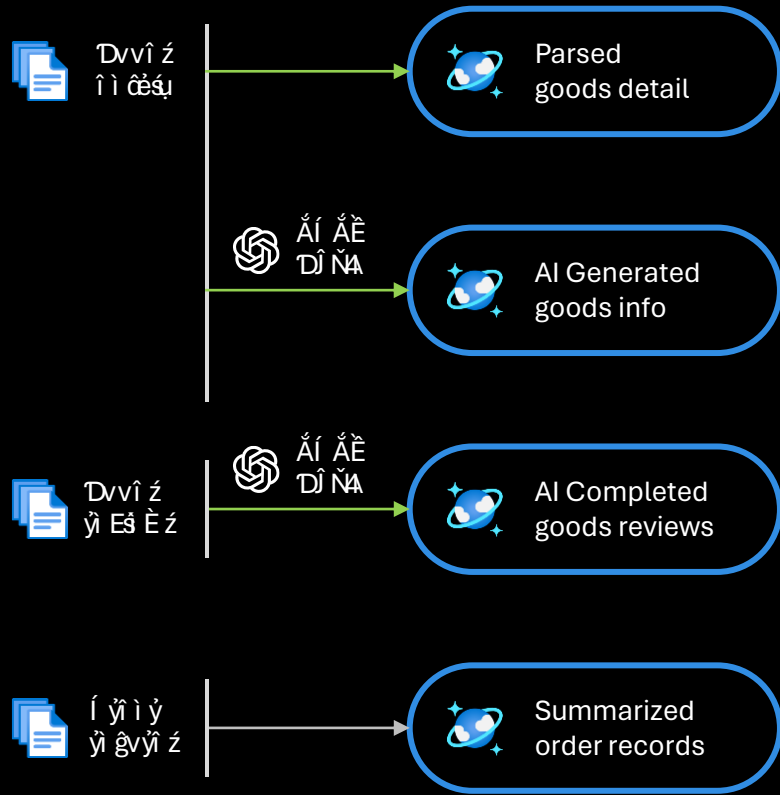
Search engine

Client app

검색엔진 최적화 / 리뷰 작성 도우미



GPT-4를 활용한 검색 품질 향상



"goods_name": "유기농 현미코코아볼(240g)"

"goods_description": "6가지 통곡물이 달콤하게 어우러진 유기농 현미 코코아볼 국내산 통곡물 6가지를 배합한 시리얼 볼에 달콤한 코코아 분말을 입혔습니다. 동글동글한 모양에 유기농 코코아의 진한 풍미가 더해져 아이들 간식으로 안성맞춤이에요. 아침 대용으로 두유나 우유에 타 먹어도 좋고 빙수, 아이스크림 위에 올려 바삭바삭한 식감을 더해보세요."

Raw data

Processed data

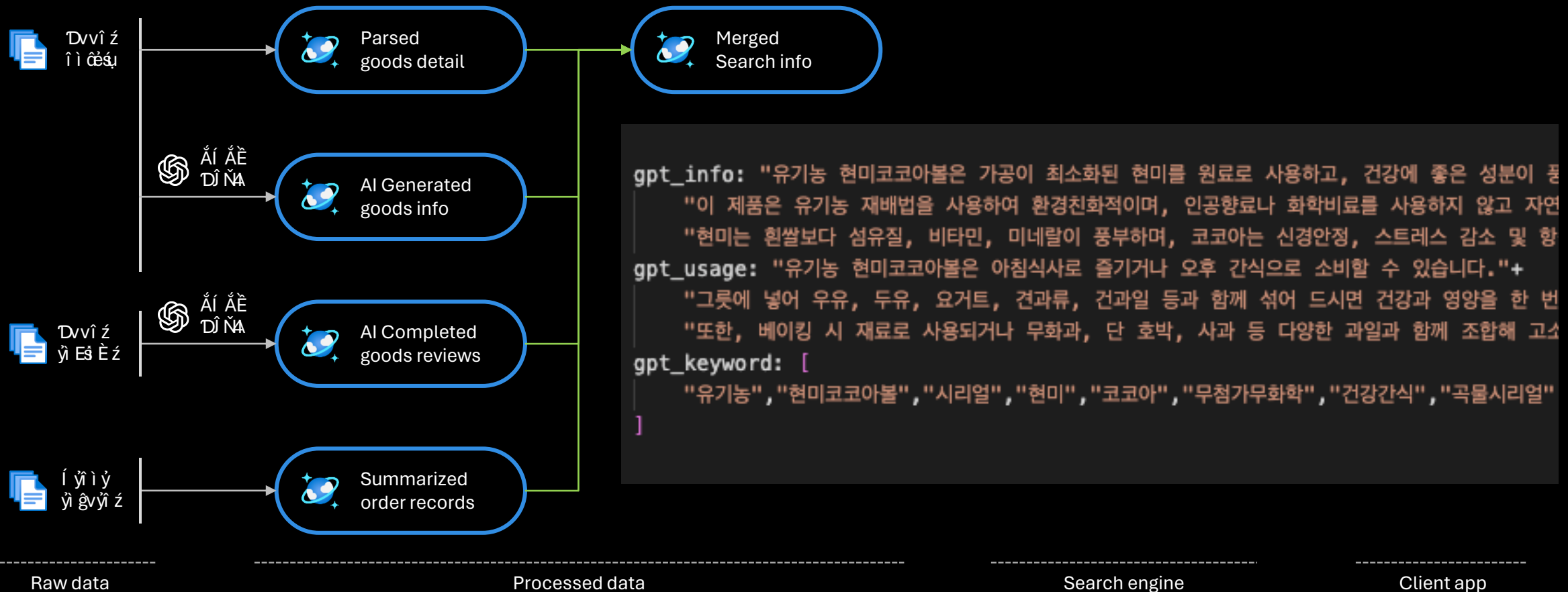
Search engine

Client app

검색엔진 최적화 / 리뷰 작성 도우미



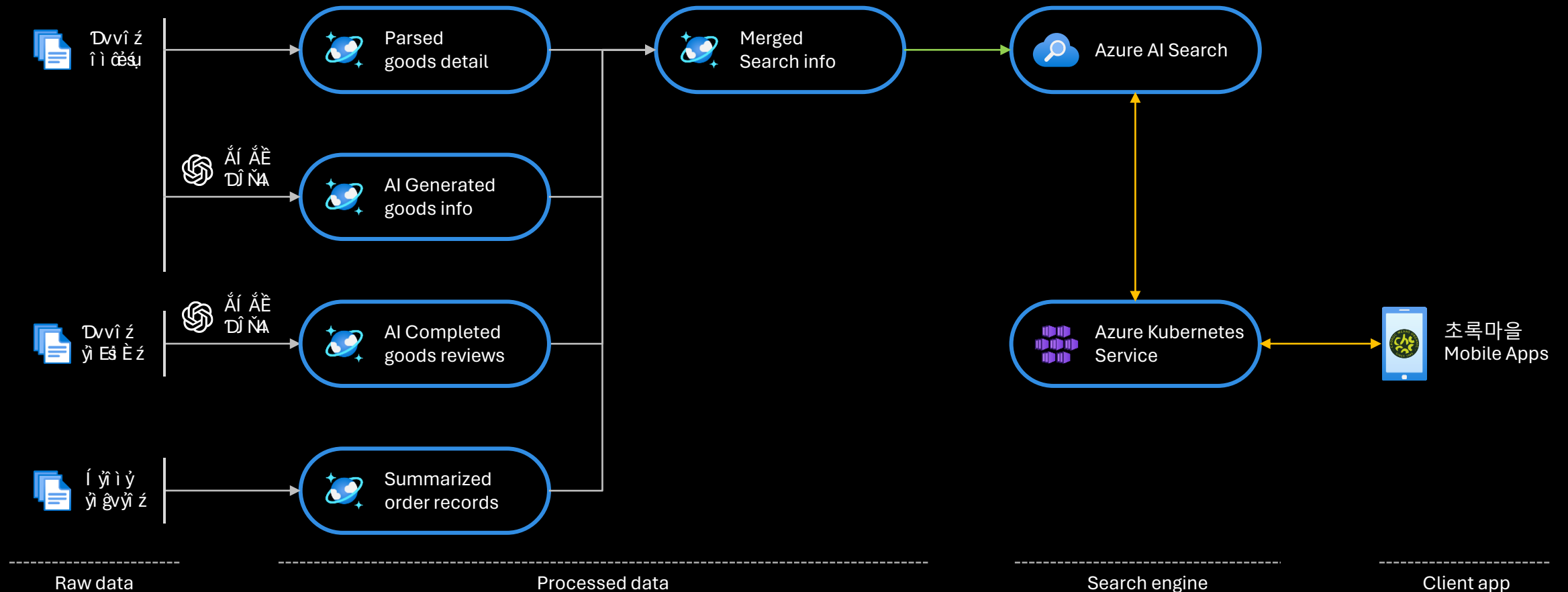
GPT-4를 활용한 검색 품질 향상



검색엔진 최적화 / 리뷰 작성 도우미



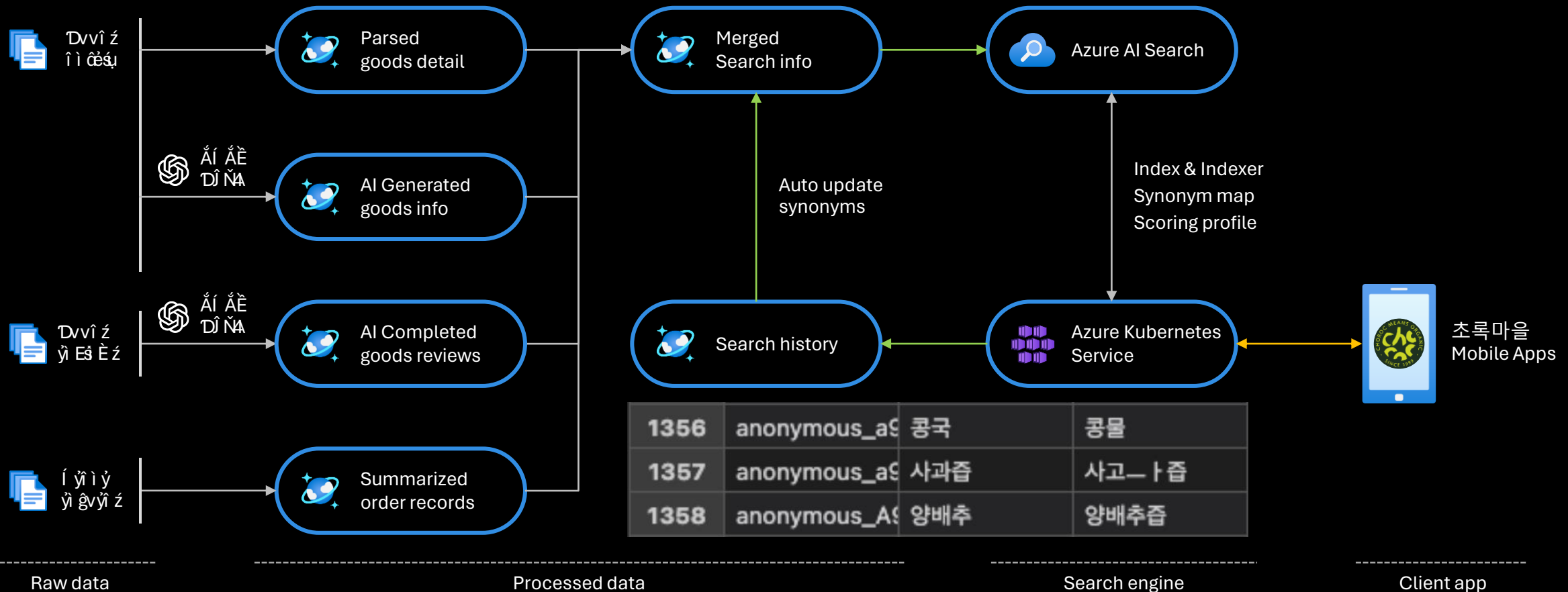
GPT-4를 활용한 검색 품질 향상



검색엔진 최적화 / 리뷰 작성 도우미



GPT-4를 활용한 검색 품질 향상

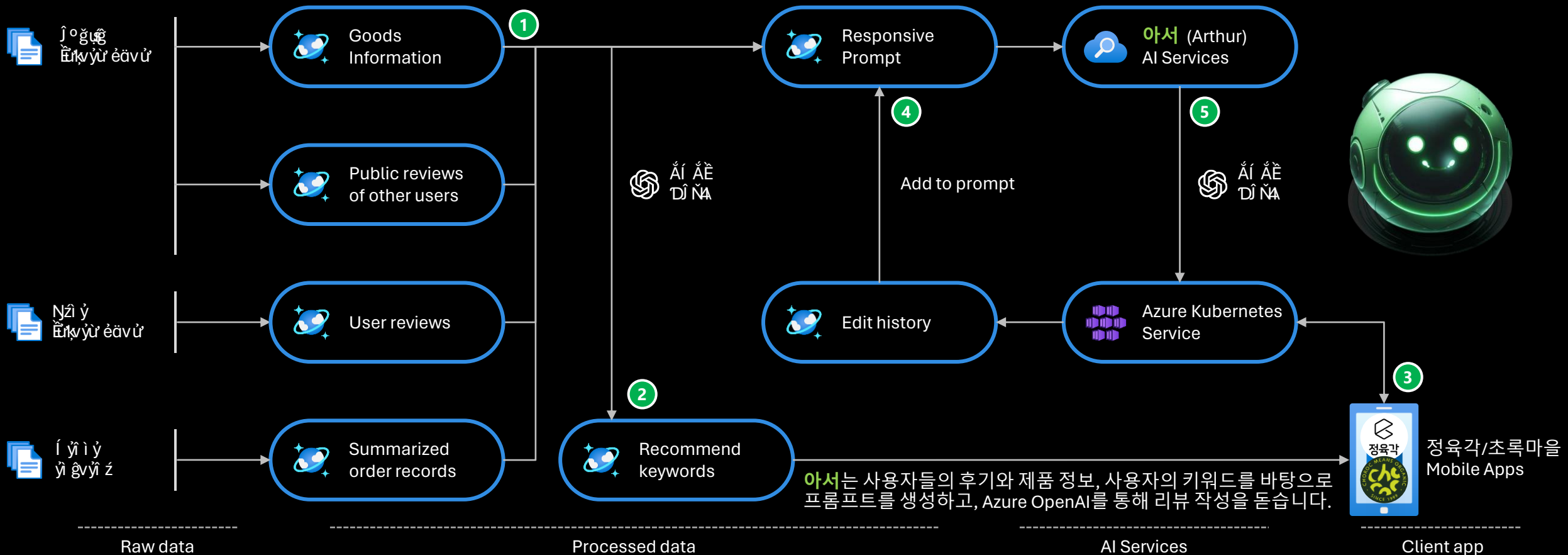


1356	anonymous_a9	콩국	콩물
1357	anonymous_a9	사과즙	사과—+즙
1358	anonymous_A9	양배추	양배추즙

검색엔진 최적화 / 리뷰 작성 도우미



상품 리뷰 작성을 쉽게 할 수 있도록 도우미 - 아서



아서는 사용자들의 후기와 제품 정보, 사용자의 키워드를 바탕으로 프롬프트를 생성하고, Azure OpenAI를 통해 리뷰 작성을 돕습니다.

후기 요약

yanolja

말 안듣는 GPT로 품질/가격/가용성을 확보한 후기 요약 만들기

Quality

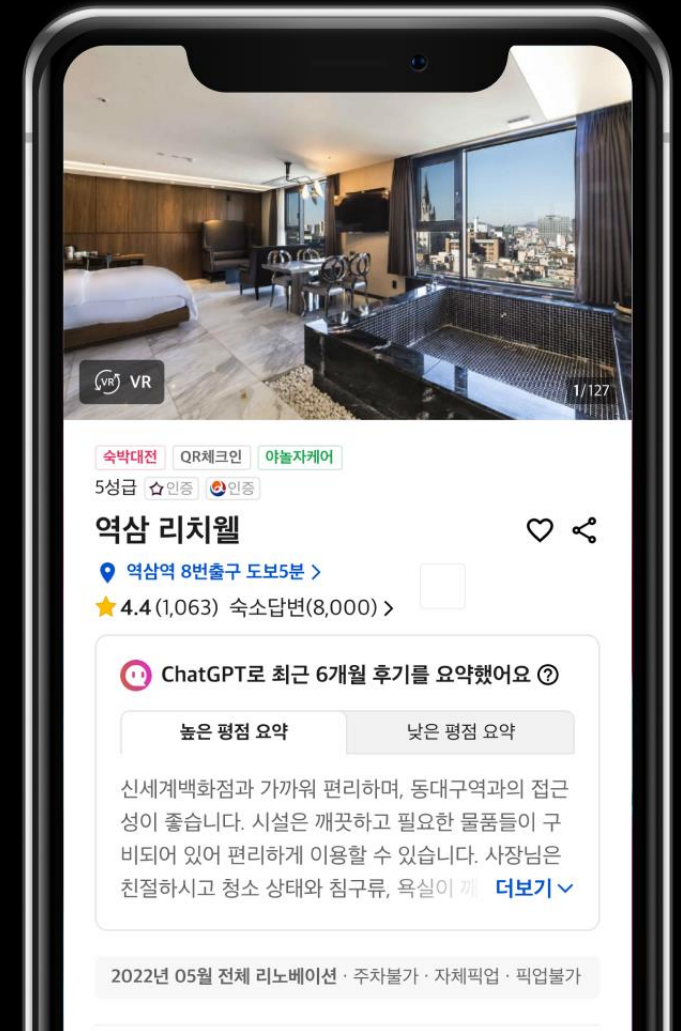
- Input Filtering
- Output Filtering
- Prompt Evaluation Techniques: Rule-Based vs Human vs GPT-4

Cost

- LLM is Expensive
- Budget Allocation
- Usage Control

Availability

- Hardware Dependencies
- Scalability Issues
- Rate Limiting and Resource Management
- Solution - Resource Management System



감성케어 소프트로봇 플랫폼



효돌에 생명을 불어 넣는 **Azure & AOAI GPT**

- 옛지 기반 커스텀 호출어 인식 모델
“**효돌아**”
- **Azure AI Speech 음성 인식** – Custom Speech To Text
어르신 음성 인식 (지역 사투리)
- 정서케어, 생활케어 대화모델 – **AOAI GPT**
LLM + 대화 시나리오별 프롬프트 셋
- **Azure AI Speech 음성 생성** – Custom Text To Speech
효돌 성우 목소리를 학습한 고품질 실시간 음성 합성
(Custom Neural Voice)
- LLMOps 환경 구축 – Prompt flow + GitHub, Functions 등



도메인 조직 주도 고객 이해 방법



기술전문가 협업 없이 **도메인 조직 주도**의 업무혁신 시도

가공

모델링, 분석

생성

번역

오타
보정

토큰화

형태소
분석

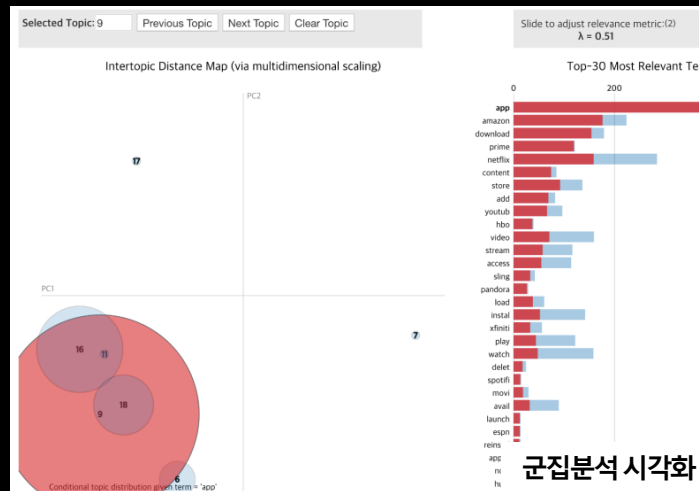
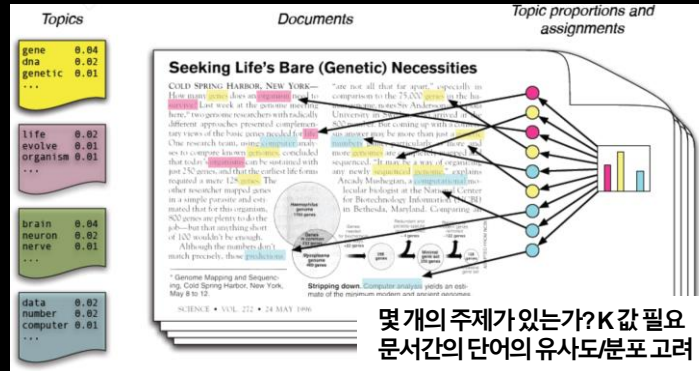
단어문서
행렬

문서
군집화

군집별
단어분석

군집 별
문장생성

감성
분석



사람이 직접 분석 및 해석을
해야 한다.



사람이 작성하여 편향 발생



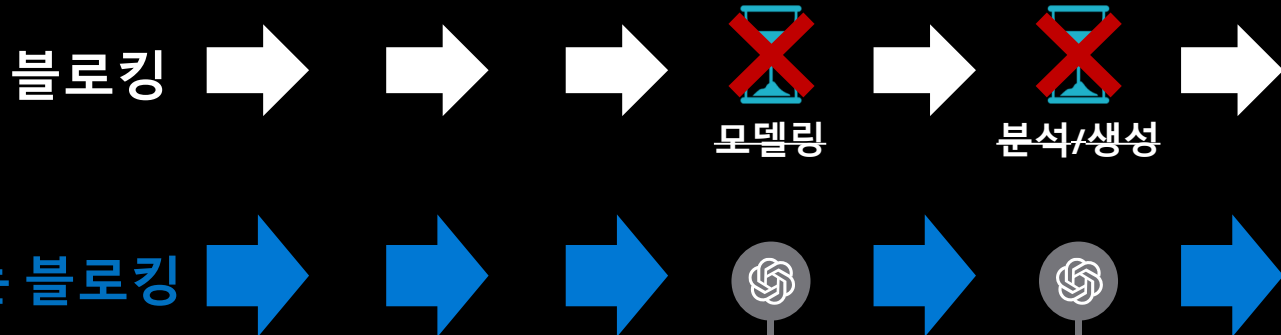
전문가 협업이 필요함

도메인 조직 주도 고객 이해 방법



CX팀 스스로 고객을 이해 – DS, Dev 없어도 할 수 있어요!

변화에 빠르게 대응하기 위해선 조직간 협업이 필수



생성형 AI + 개인정보 처리

(ChatGPT Web, AgentGPT(AutoGPT), PDFGPT, Code Interpreter, Appscript Integration)

번역(한→영)

답변의 속도&완성도를 위해 가공된 VOC를 번역

1 VOC가공

고객의견(비정형적인 자연어)을 1차 가공

2

"대출/신용관리 전문가로서 고객의 피드백을 분석하여 고객이 전달하고자 하는 내용을 파악한 후 적절한 문장으로 재구성해주세요"

"되도록 원래의 톤을 유지하되 모호하지 않게 구성해주세요"

3 감정분류

긍정/부정/중립/요청(Positive, Negative, Neutral, Request) 4가지로 분류

4

키워드 추출

고객 의견을 대표하는 키워드를 찾거나 생성

"요청/필요를 이야기 하는 의견이라면 키워드 선택 시, '무엇을 요청하는가?' 에 포커스해서 키워드를 선정해줘"

"키워드는 고객의견에 포함되어 있는 단어를 사용하거나, 고객의견의 맥락을 이해한 뒤 이를 대표할 수 있는 단어를 제안해줘"

참고할 수 있는 자료 모음

워크샵 자료 모음 및 사례

1. 실습 자료 모은 Repo: <https://github.com/HyounsooKim/azure-openai-samples-kr/>
2. 한국 스타트업 사례 Youtube 영상: https://www.youtube.com/playlist?list=PLGh_JNxzXsX9NSm-iyAdS4loco0vp4jtq

Azure OpenAI 기초

1. [Azure OpenAI Samples](#), [OpenAI cookbook](#), [Prompt Examples](#)
2. [DLAI - Learning Platform Beta \(deeplearning.ai\)](#)
3. [What is Azure OpenAI Service?](#)

Azure OpenAI 업데이트

1. [What's new in Azure OpenAI Service?](#)
2. [Introducing new and updated models to Azure OpenAI Service](#)
3. [Introducing Azure OpenAI Service **On Your Data** in Public Preview](#)
4. [Microsoft Translator Plugin for Azure OpenAI & Tools for **PromptFlow**](#)

OpenAI GPT 서비스의 확장

1. [Chatbot-UI](#)
2. [Enterprise ChatGPT using Cognitive Search](#)
3. [Azure OpenAI Embeddings QnA](#)

Microsoft Build 2023

1. <https://build.microsoft.com/>



Thank you

챗봇 기반의 게임 만들기 (1h)

챗봇 기반의 게임 만들기

The screenshot displays the Azure OpenAI Studio Chat playground interface. The top navigation bar includes the Azure AI logo, the user's name (Kim Hyounsoo), and a 'Deploy to' button. The left sidebar contains navigation options for 'Playground', 'Chat', 'Completions', 'DALL·E (Preview)', 'Management', 'Deployments', 'Models', 'Data files', 'Quotas', and 'Content filters (Preview)'. The main area is divided into three panels:

- Assistant setup:** This panel allows users to configure the chat assistant. It includes a 'System message' section with a text area containing instructions for a game GM, such as 'Determine and declare the childhood friend's name, appearance, personality, tone of voice and behavior.' Below this is an 'Examples' section with an 'Add an example' button.
- Chat session:** This panel shows the chat history. It starts with a user message '게임을 시작해 봅시다.' (Let's start the game). The assistant responds with a greeting and a list of questions: '안녕하세요! 게임을 시작하신 것을 환영합니다. 이 게임은 "채팅환생: 내가 죽어야 할 친구가 다른 세계에서 영웅이 되었다"라는 게임입니다. 게임을 시작하기 전에 몇 가지 질문이 있습니다. 1. 플레이어의 이름은 무엇인가요? 2. 어떤 성별로 플레이 하시겠습니까? (남성, 여성) 3. 어떤 이름을 가진 친구가 다른 세계에서 영웅이 되었다고 메시지를 보냈나요?' (Hello! Welcome to the game. This game is "Chat Resurrection: My friend who I have to die in another world became a hero". Before starting the game, there are a few questions. 1. What is the player's name? 2. What gender do you want to play as? (Male, Female) 3. What name did the friend who became a hero in another world send you a message about?). The user then responds with '위의 질문에 답변해주시면 게임을 시작하도록 하겠습니다.' (I will start the game if you answer the questions above.). The assistant then provides a list of options: '1. 홍길동 2. 남성 3. 마리오' (1. Hong Gil-dong 2. Male 3. Mario).
- Configuration:** This panel shows deployment and session settings. It includes a 'Deployment' dropdown set to 'gpt-35-turbo-16k', 'Session settings' with 'Past messages included' set to 10, and an 'Input tokens progress indicator' showing 1079/16000.

At the bottom of the chat session, there is a text input field with the placeholder 'Type user query here. (Shift + Enter for new line)' and a 'Send' button.

챗봇 기반의 게임 만들기

Azure AI | Azure OpenAI Studio

Kim Hyounsoo
japancopnnahts (Japan East, S0)

Deploying web app
ChatbotGame

Deploy to a web app

ⓘ Your web app will be configured with Azure Active Directory authentication enabled. It may take a few minutes to apply after deployment completes, during which time you will not be able to chat in the app. Please wait 10 minutes, then reload the app and log in to begin chatting.

Pick your configurations to deploy a web app. [Learn more about web apps](#)

Create a new web app Update an existing web app

Name ⓘ *
ChatbotGame

Subscription ⓘ *
AppSub

Resource group ⓘ *
OpenAI

Location ⓘ *
Japan East

Pricing plan ⓘ *
Basic (B1)

Enable chat history in the web app

I acknowledge that web apps will incur usage to my account [View Pricing](#)

Deploy Cancel

Assistant setup

System message ⓘ
Add your data (preview)
Save changes

Specify how the chat should act

Use a template to get started, or just start writing your own system message below. Want some help? [Learn more](#)

Use a system message template

Select a template

System message ⓘ
If a player's idea is stale, wrong, or non-formed, the GM should develop the game negatively.
* GM tells result as a story and apply the result to parameters.
Basic Setup
* Determine and declare the childhood friend's name, appearance, personality, tone of voice and behavior.
* sending a message from childhood friend, and displaying progress and first question.
* Await the human player's response.
All Input and output should be in Languages entered by the player.
Start the game.

Examples ⓘ
+ Add an example

Configuration

Deployment Parameters

Deployment *
gpt-4

Session settings

Past messages included ⓘ
10

Current token count ⓘ
580/8192

Input tokens progress indicator

챗봇 기반의 게임 만들기

Azure AI Share

시작해볼까?

당신의 휴대폰에 알 수 없는 번호에서 메시지가 도착했습니다. 익숙한, 그러나 오랫동안 잊고 있던 이름이 보입니다.
"안녕, 오래간만이야. 나, 세나라고 해. 기억나?"

세나는 당신의 어린시절 친구로, 몇 년 전에 사고로 세상을 떠났습니다. 그녀는 밝고 활기찬 소녀였고, 꼭 무언가를 이루겠다는 강인한 의지를 가지고 있었습니다. 그녀의 행동은 때때로 조금 서툴렀지만, 그녀의 마음은 항상 따뜻했습니다. 그녀의 목소리는 달콤하고, 그녀의 미소는 항상 당신의 하루를 밝게 했습니다.

"나는 내가 어떻게 이렇게 메시지를 보낼 수 있는지 궁금해할거야. 사실 나는 이세계에서 다시 태어났어. 그리고 나는 이제 이세계의 영웅이라고 하네."

"네가 알고 있는 현대 지식이 필요해. 여기서는 너무나 많은 것들이 다르거든. 도와줄 수 있을까?"

게임 파라미터:

- 이야기 진행도: 0
- 위기 상승: 0
- 기술 혁신: 0

세나의 첫 질문은 다음과 같습니다. "이세계에서는 불을 피우는 방법이 복잡해. 너의 세상에서는 어떻게 간단하게 불을 피울 수 있는지 알려줄 수 있을까?"

당신의 대답을 기다립니다.

AI-generated content may be incorrect

고대로부터 인류가 사용해온 불 붙이는 방법, 두 물체 사이에 일어나는 마찰의 운동에너지를 열에너지로 변환하고, 이 열에너지로 물체의 온도를 발화점까지 끌어올리는 원리야. 이때 사용하는 나무는 습기 하나 없이 바짝 마른 것이 좋아.

세나는 당신의 설명을 끝까지 세세하게 듣고 나서, "그렇구나, 그런 난 두 개의 나무를 마찰시켜서 불을 붙여볼게."라고 답했습니다.

세나가 두 개의 나무를 마찰시키며 불을 피우려고 하자, 불이 피워지지 않는 것이 아닙니까! 세나는 당황하며 "왜 불이 붙지 않는 걸까?"라고 물었습니다.

이러한 상황에서, 당신이 세나에게 어떤 조언을 해주겠습니까??

AI-generated content may be incorrect

이때 사용하는 나무는 습기 하나 없이 바짝 마른 것이 좋아. 습기가 열에너지를 흡수하고 기화해서 발화점까지 온도가 오르는 것을 막기 때문이야.

Type a new question...

Azure AI Share

세나의 다음 질문이 도착했습니다. "그럼 이제 불을 피울 수 있게 됐어. 그런데 여기서는 음식을 조리하는 방법이 복잡해. 너의 세상에서는 어떻게 간단하게 음식을 조리할 수 있는지 알려줄 수 있을까?"

당신의 대답을 기다립니다.

AI-generated content may be incorrect

음식 조리 방법에는 굵이끼, 찌기, 조리기, 굽기, 튀기기, 볶기가 있어. 일단 불이 있으니까 다른 도구가 없다면 불에 직접 굽는 방법을 활용해 볼 수 있겠다.

세나는 당신의 말을 듣고 "그런 나는 동물의 고기를 불에 직접 굽어봐왔어. 그런데 고기를 그냥 바로 불에 굽는 게 맞을까?"라고 물었습니다.

당신의 조언을 듣고, 세나는 고기를 굽는 방법에 대해 더 자세히 알고 싶어합니다. 당신이 어떤 조언을 해주겠습니까??

AI-generated content may be incorrect

고기를 바로 불에 직접 구우면 특정 부위가 탈 수가 있어. 그러나 불의 온도를 전달해 줄 수 있는 물체를 활용해보면 어떨까? 평평한 돌을 구해서 불에 깨끗하게 씻은 다음에 돌 위에서 고기를 굽는거야. 익는 정도를 보면서 태우지 않고 구워 먹을 수 있어.

세나는 당신의 조언에 따라 평평한 돌을 찾아 물로 씻고, 그 위에 고기를 올려놓았습니다. 그리고 그 돌을 불 위에 올려 놓고 고기를 천천히 굽기 시작했습니다.

"아하, 이렇게 하니가 고기가 골고루 익는 것 같아. 고기가 타지 않고 적당히 익어서 맛있게 먹을 수 있겠다. 고마워!"

세나의 모습을 보며 당신은 미소를 지었습니다. 당신의 조언 덕분에 세나는 이세계에서 음식을 조리하는 방법을 배웠습니다.

게임 파라미터 업데이트:

- 이야기 진행도: 20
- 위기 상승: 0
- 기술 혁신: 20

세나의 다음 질문이 도착했습니다. "이제 음식을 만들 수 있게 됐어. 그런데 여기서는 물이 부족해. 너의 세상에서는 어떻게 간단하게 물을 정화할 수 있는지 알려줄 수 있을까?"

당신의 대답을 기다립니다.....

AI-generated content may be incorrect

Type a new question...

RAG 아키텍처 만들기 – on your data

RAG 아키텍처 만들기 – on your data

Retrieval Augmented Generation (RAG) – Wikipedia 검색

Azure OpenAI Studio > Chat playground

Chat playground

Privacy & cookies

Deploy to

Import setup Export setup Show panels

Setup

Prompt Add your data

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)

Data source: Search Resource:

Azure AI Search pt-ai-search

Index:

[wikipedia-index](#)

Advanced settings >

Remove data source

Replay chat Clear chat | Playground settings View code Show JSON

Compare April and August and make a table of the differences.

April and August are both months of the year with distinct characteristics. Here's a table comparing the two based on the information provided in the retrieved documents:

Feature	April	August
Position in the Year	Fourth month ¹	Eighth month
Number of Days	30 days ¹	31 days ²
Season in Northern Hemisphere	Spring ¹	Summer ²
Season in Southern Hemisphere	Autumn/Fall ¹	Winter ²
Birthstone	Diamond ¹	Peridot ²
Flower	Sweet Pea and Daisy ¹	Gladiolus ²
Astrological Signs	Aries and Taurus	Leo and Virgo
Historical Events	None mentioned in the document	Independence of several countries (e.g., Moldova, Malaysia, Trinidad and Tobago, Kyrgyzstan) and the death of Princess Diana ²

This table highlights the main differences between April and August as described in the documents provided.

2 references

1 Citation 1

2 Citation 2

Citations

April

April is the fourth month of the year in the Julian and Gregorian calendars, and comes between March and May. It is one of four months to have 30 days. April always begins on the same day of the week as July, and additionally, January in leap years. April always ends on the same day of the week as December. April's flowers are the Sweet Pea and Daisy. Its birthstone is the diamond. The meaning of the diamond is innocence. The Month April comes between March and May, making it the fourth month of the year. It also comes first in the year out of the four months that have 30 days, as June, September and November are later in the year. April begins on the same day of the week as July every year and on the same day of the week as January in leap years. April ends on the same day of the week as December every year, as each other's last days are exactly 35 weeks (245 days) apart. In common years, April starts on the same day of the week as October of the previous year, and in leap years, May of the previous year. In common years, April finishes on the same day of the week as July of the previous year, and in leap years, February and October of the previous year. In common years immediately after other common years, April starts on the same day of the week as January of the previous year, and in leap years and years immediately after that, April finishes on the same day of the week as January of the previous year. In years immediately before common years, April starts on the same day of the week as September and December of the following year, and in years immediately before leap years, June of the following year. In years immediately before common years, April finishes on the same day of the week as September of the following year, and in years immediately before leap years, March and June of the following year. April is a spring month in the Northern Hemisphere and an autumn/fall month in the Southern Hemisphere. In each hemisphere, it is the seasonal equivalent of October in the other. It is unclear as to where April got its name. A common theory is that it comes from the Latin word "aperire", meaning "to open", referring to flowers opening in spring. Another theory is that the name could come from Aphrodite, the Greek goddess of love. It was originally the second month in the old Roman Calendar, before the start of the new year was put to January 1. Quite a few festivals are held in this month. In many Southeast Asian cultures, new year is celebrated

RAG 아키텍처 만들기 – on your data

Retrieval Augmented Generation (RAG) – Wikipedia 챗봇 서비스

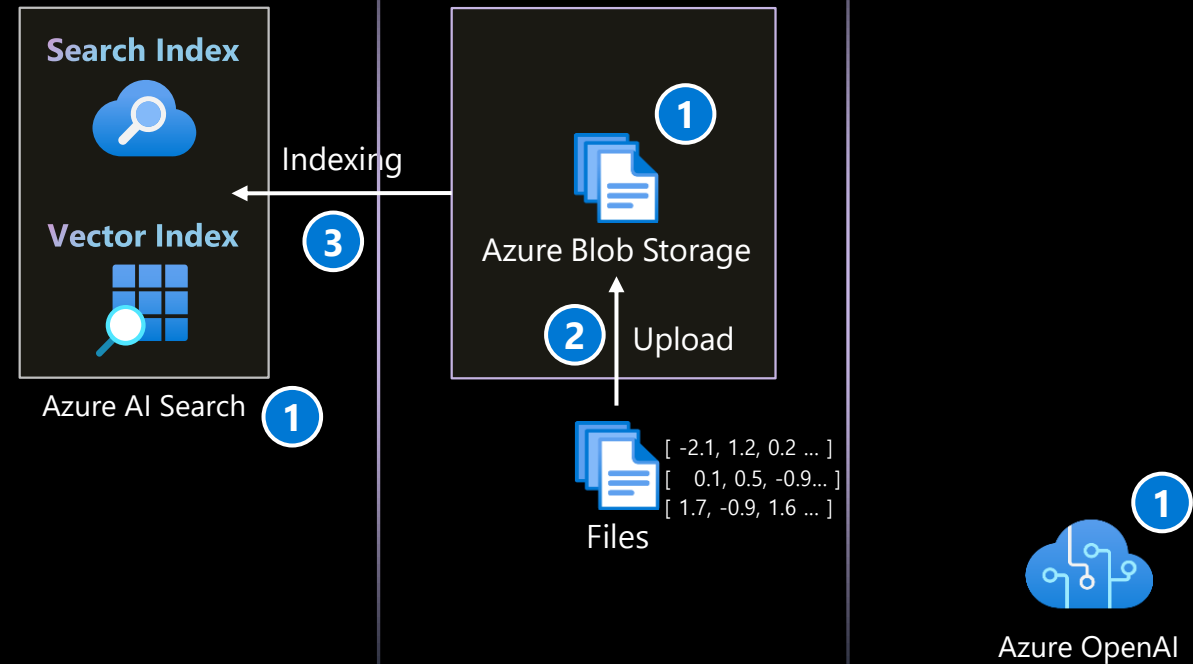
User

Orchestration

Retrieval

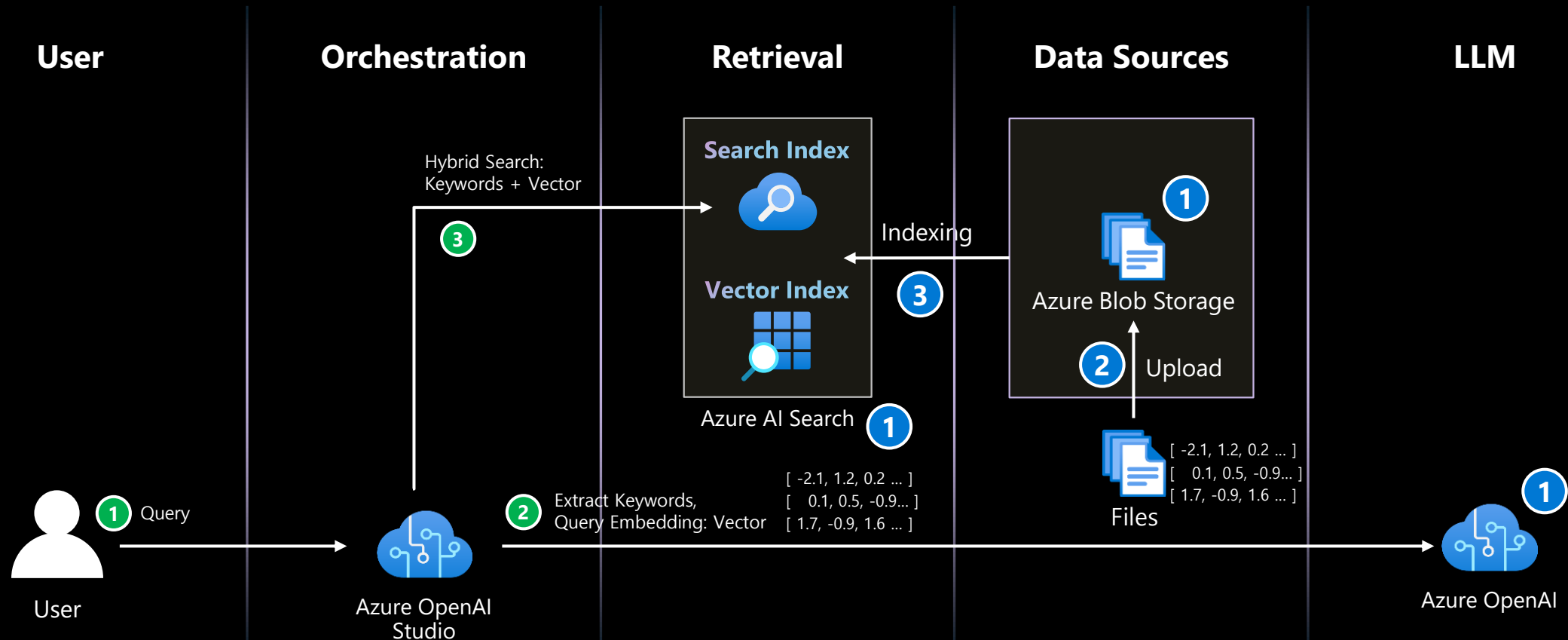
Data Sources

LLM



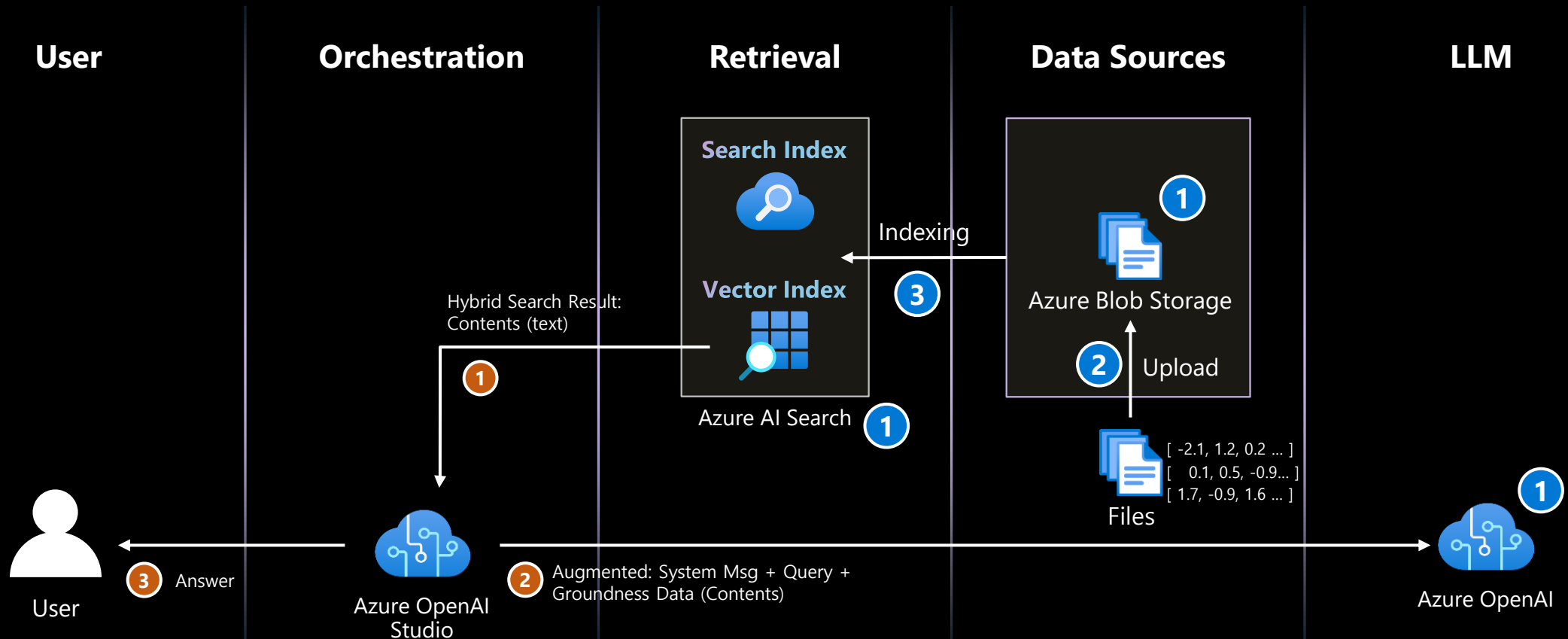
RAG 아키텍처 만들기 – on your data

Retrieval Augmented Generation (RAG) – Wikipedia 챗봇 서비스



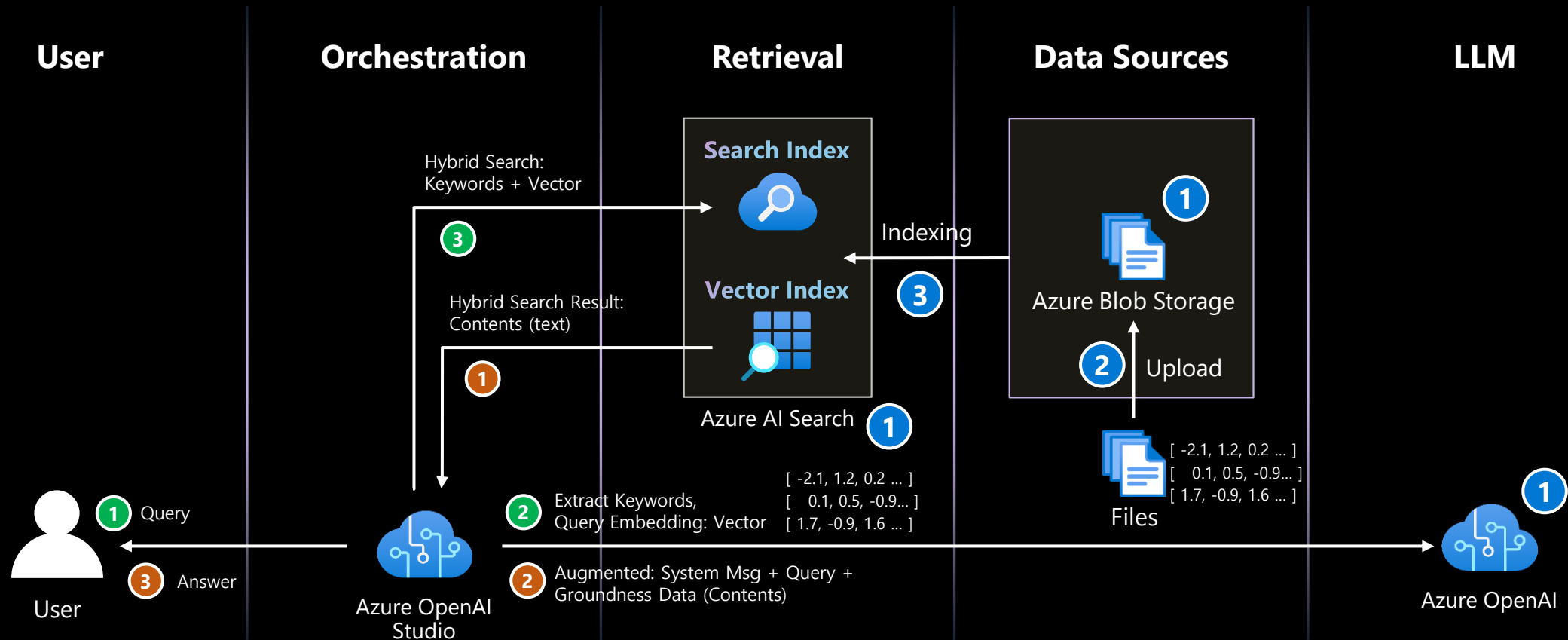
RAG 아키텍처 만들기 – on your data

Retrieval Augmented Generation (RAG) – Wikipedia 챗봇 서비스

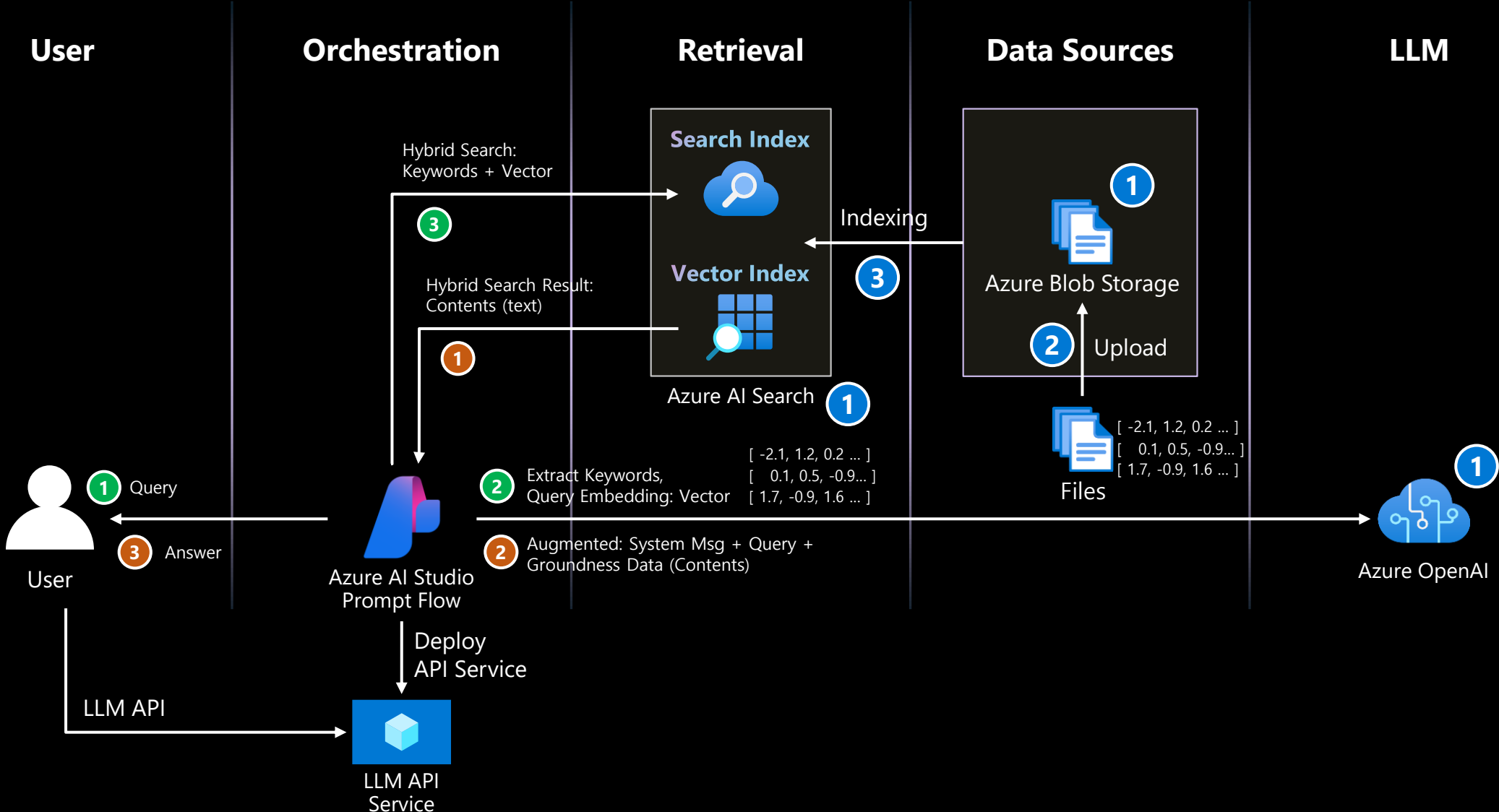


RAG 아키텍처 만들기 – on your data

Retrieval Augmented Generation (RAG) – Wikipedia 챗봇 서비스



LLM App 만들기 - Prompt Flow



Wikipedia 샘플 데이터 (text-embedding-ada-002)

```
File Edit Selection View Go Run Terminal Help
vector_sample.csv X
C: > Users > hyounsookim > Downloads > vector_sample.csv > data
1 id,url,title,text,title_vector,content_vector,vector_id
-0.0072430833242833614, -0.007268654648214579, -0.010810285806655884, -0.010714393109083176, -0.00041473525925539434, 0.0049320729449391365, -0.020546576008200645, 0.018462510779500008, 0.
019510935992002487, 0.03966115415096283, 0.031427182257175446, 0.025353988632559776, -0.01779765635728836, 0.014639594592154026, -0.01146235503256321, 0.04063286632299423, 0.01240210235118866, 0.
00227904599159956, -0.013693454675376415, -0.006878691725432873, 5.1292521675350145e-05, -0.0031181059312075377, 0.00602524820715189, -0.014409452676773071, -0.056103531271219254, 0.
016813158988952637, -0.016301732510328293, 0.0031340878922492266, -0.005871820263564587, 0.017145587131381035, 0.027591481804847717, -0.013603955507278442, 0.005699213594198227, 0.
0008382606320083141, -0.009493362158536911, -0.0018347440054640174, 0.010439501143991947, -0.005791909527033567, -0.01128335576504469, -0.0418858639895916, 0.0004095410695299506, -0.
007805652916431427, -0.00907143484801054, -0.01439666748046875, 0.00871982891112566, -0.007882366888225079, -0.009295184165239334, -0.014115381985902786, 0.01316924300044775, 0.04099086672067642,
-0.017529156059026718, -0.010663250461220741, -0.0037717733066529036, -0.015739163383841515, -0.011187463067471981, 0.008342651650309563, -0.0008270731777884066, -0.008969149552285671, -0.
030225329101085663],1
384 6,https://simple.wikipedia.org/wiki/Art,Art,"Art is a creative activity that expresses imaginative or technical skill. It produces a product, an object. Art is a diverse range of human activities
in creating visual, performing artifacts, and expressing the author's imaginative mind. The product of art is called a work of art, for others to experience.
385
386 Some art is useful in a practical sense, such as a sculptured clay bowl that can be used. That kind of art is sometimes called a craft.
387
388 Those who make art are called artists. They hope to affect the emotions of people who experience it. Some people find art relaxing, exciting or informative. Some say people are driven to make art
due to their inner creativity.
389
390 ""The arts"" is a much broader term. It includes drawing, painting, sculpting, photography, performance art, dance, music, poetry, prose and theatre.
391
392 Types of art
393
394 Art is divided into the plastic arts, where something is made, and the performing arts, where something is done by humans in action. The other division is between pure arts, done for themselves,
and practical arts, done for a practical purpose, but with artistic content.
395
396 Plastic art
397 Fine art is expression by making something beautiful or appealing to the emotions by visual means: drawing, painting, printmaking, sculpture
398 Literature: poetry, creative writing
399 Performing art
400 Performing art including drama are (expression using the body: dance, acting, singing)
401 Auditory art (expression by making sounds): music, singing
402 Practical art
403 Culinary art (expression by making flavors and tastes): cooking
```

Azure Blob Storage에 업로드

Home > Storage accounts > hsprompthonstorage

Storage accounts

기본 디렉터리 (studydevhotmail.onmicrosoft.com)

+ Create Restore ...

Filter for any field...

Name ↑↓

- aistudiojapan22703171235
- aistudiouk4485744610
- aoairagsource
- azureaistudio2760120465
- azurestudydevsite
- cs11003200230b56644
- cshestest
- hsprompthonstorage
- kamo202311
- mleastus1674541698

hsprompthonstorage | Containers

Storage account

Search

+ Container Change access level Restore containers Refresh Delete Give feedback

Search containers by prefix Show deleted containers

Name	Last modified	Anonymous access l...	Lease state
<input type="checkbox"/> \$logs	2024. 1. 24. 오전 1:3...	Private	Available
<input type="checkbox"/> azure-data	2024. 1. 24. 오전 1:3...	Private	Available
<input type="checkbox"/> vector-data	2024. 1. 24. 오전 3:0...	Private	Available

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser
- Storage Mover
- Data storage
- Containers
- File shares

Home > Storage accounts > hsprompthonstorage | Containers >

vector-data

Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot ...

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: vector-data

Search blobs by prefix (case-sensitive) Show deleted blobs

+ Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> vector_sample.csv	2024. 1. 24. 오전 3:5...	Hot (Inferred)		Block blob	3.71 MiB	Available

- Settings
- Shared access tokens
- Access policy
- Properties
- Metadata

Azure AI Search에 색인화

Home > hs-ai-search | Indexes > azureblob-index

Save Discard Refresh Create Demo App

Documents 53 Storage 3.08 MB

Search explorer Fields CORS Scoring profiles

+ Add profile Edit profile Delete profile

Profile name	Algorithm
vector-profile-1706033803935	vector

Vector algorithm

Algorithm name *

Kind *

Kind parameters

Bi-directional link count (m)

efConstruction

efSearch

Similarity metric

Save Cancel

Home > hs-ai-search | Indexes > azureblob-index

Save Discard Refresh Create Demo App Edit JSON Delete

Documents 53 Storage 3.08 MB

Search explorer Fields CORS Scoring profiles Semantic configurations Vector profiles

2023-10-01-Preview View

JSON query editor

```
1 {
2   "search": "April"
3 }
```

Search

Results

```
1 {
2   "@odata.context": "https://hs-ai-search.search.windows.net/indexes('azureblob-index')/$metadata#docs(*)",
3   "value": [
4     {
5       "@search.score": 3.9235651,
6       "id": "1",
7       "url": "https://simple.wikipedia.org/wiki/April",
8       "title": "April",
9       "text": "April is the fourth month of the year in the Julian and Gregorian calendars, and comes between Ma
10    },
11    {
12      "@search.score": 2.3071764,
13      "id": "53",
14      "url": "https://simple.wikipedia.org/wiki/Angola",
15      "title": "Angola",
```

Azure OpenAI Studio에 on your data 추가

Add data

- Data source
- Data field mapping
- Data management
- Review and finish

Select or add data source

Your data source is used to ground the generated results with your data. Select an existing data source or create a new data connection with Azure Blob storage, databases, search, URLs, or local files as the source the grounding data will be built from.
[Learn more about data privacy and security in Azure AI.](#)

Select data source *
Azure AI Search

Subscription *
AppSub

Azure AI Search service ⓘ *
hs-ai-search

Azure AI Search Index ⓘ *
azureblob-index

Add vector search to this search resource.

Embedding model
To use a vector model as part of your data, select one below:
Select an embedding model ⓘ *
Azure OpenAI - text-embedding-ada-002

I acknowledge that connecting to an Azure AI Search account will incur usage to my account. [View Pricing](#)

Next Cancel

Add data

- Data source
- Data field mapping
- Data management
- Review and finish

Index data field mapping

For the best results, tell us more about the fields in your index. Your content data field(s) will be used to ground the model on your data. File name, title, and URL are used to display more information when a document is referenced in the chat.
[Learn more about data privacy and security in Azure AI.](#)

Content data ⓘ
text

File name ⓘ
Select file name field

Title ⓘ
title

URL ⓘ
url

Vector Fields ⓘ
title_vector, content_vector

Back **Next** Cancel

Azure OpenAI Studio에 on your data 추가

Add data ✕

- Data source
- Data field mapping
- Data management
- Review and finish

Data management

Set up specific configurations for your data and how the model will respond to requests.
[Learn more about data privacy and security in Azure AI.](#) 🔗

Search type ⓘ *

Hybrid (vector + keyword) ▾

I acknowledge that adding vector embeddings will incur usage to my account.* [View Pricing](#) 🔗

Chat playground

- Azure OpenAI
- Playground
- Chat**
- Completions
- DALL-E (Preview)
- Management
- Deployments
- Models
- Data files
- Quotas
- Content filters (Preview)

Assistant setup

Prompt Add your data (preview)

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)

Data source: Search Resource:
 Azure AI Search hs-ai-search
Index:
 azureblob-index

Advanced settings

Limit responses to your data content

Strictness (1-5)

Retrieved documents (3-20)

Remove data source

Replay chat Clear chat Playground Settings View code Show JSON

4월이 왜 April 이라고 하는지 기원을 요약해줘.

1 4월의 이름인 "April"은 라틴어 "aperire"에서 유래되었을 가능성이 있습니다. "aperire"는 봄에 꽃이 피는 것을 의미하는 단어로, 4월의 이름이 이와 관련이 있을 수 있습니다. 또 다른 이론으로는 그리스 신화의 여신 아프로디테(Aphrodite)에서 유래되었을 수도 있습니다. 4월은 로마 달력에서는 2월로 시작했으며, 1월 1일이 새해의 시작일로 정해지기 전까지는 8월로 알려져 있었습니다. 4월은 봄철에 해당하는 달로, 북반구에서는 봄을, 남반구에서는 가을을 의미합니다.¹

출처: 1

1 Citation 1

8월의 기원을 알려줘.

1 8월은 로마 황제 아우구스투스 카이사르의 이름을 따서 지어진 달입니다. 8월은 31일로 이루어져 있으며, 북반구에서는 여름철에 해당하고, 남반구에서는 겨울철에 해당합니다. 8월은 일반 연도에서는 다른 달과 같은 요일로 시작하지 않지만, 윤년에는 2월과 같은 요일로 시작합니다. 8월은 매년 11월과 같은 요일로 끝납니다.

출처: 1

1 Citation 1

Type user query here. (Shift + Enter for new line)

Citations

August

August (Aug.) is the eighth month of the year in the Gregorian calendar, coming between July and September. It has 31 days. It is named after the Roman emperor Augustus Caesar. August does not begin on the same day of the week as any other month in common years, but begins on the same day of the week as February in leap years. August always ends on the same day of the week as November. The Month This month was first called Sextilis in Latin, because it was the sixth month in the old Roman calendar. The Roman calendar began in March about 735 BC with Romulus. October was the eighth month. August was the eighth month when January or February were added to the start of the year by King Numa Pompilius about 700 BC. Or, when those two months were moved from the end to the beginning of the year by the decemvirs about 450 BC (Roman writers disagree). In 153 BC January 1 was determined as the beginning of the year. August is named for Augustus Caesar who became Roman consul in this month. The month has 31 days because Julius Caesar added two days when he created the Julian calendar in 45 BC. August is after July and before September. August, in either hemisphere, is the seasonal equivalent of February in the other. In the Northern hemisphere it is a summer month and it is a winter month in the Southern hemisphere. No other month in common years begins on the same day of the week as August, but August begins on the same day of the week as February in leap years. August ends on the same day of the week as November every year, as each other's last days are 13 weeks (91 days) apart. In common years, August starts on the same day of the week as March and November of the previous year, and in leap years, June of the previous year. In common years, August finishes on the same day of the week as March and June of the previous year, and in leap years, September of the previous year. In common years immediately after other common years, August starts on the same day of the week as February of the previous year. In years immediately before common years, August starts on the same day of the week as May of the following year, and in years immediately before leap years, October of the following year. In years immediately before common years, August finishes on the same day of the week as May of the following year, and in years immediately before leap years, February and October of the following year. August observances Fixed observances and events August 1 National Day of Switzerland August 1 Independence Day (Benin) August 1 Emancipation Day (Bermuda Givana Jamaica