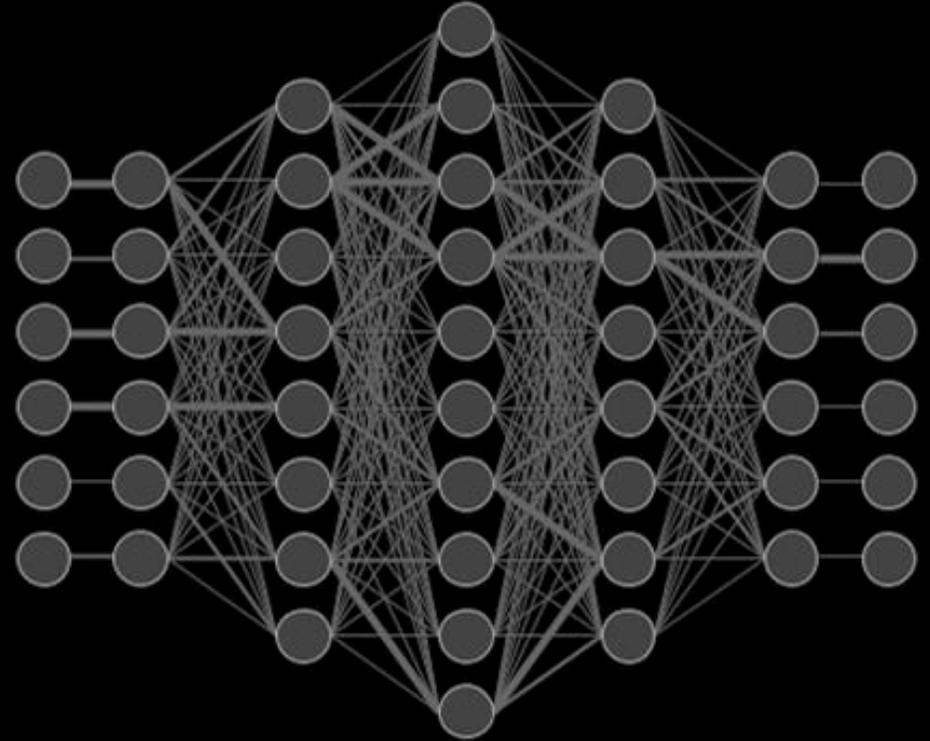


ChatGPT / OpenAI 업데이트 및 솔루션 아키텍처



Microsoft 기술전략매니저 김현수

Agenda

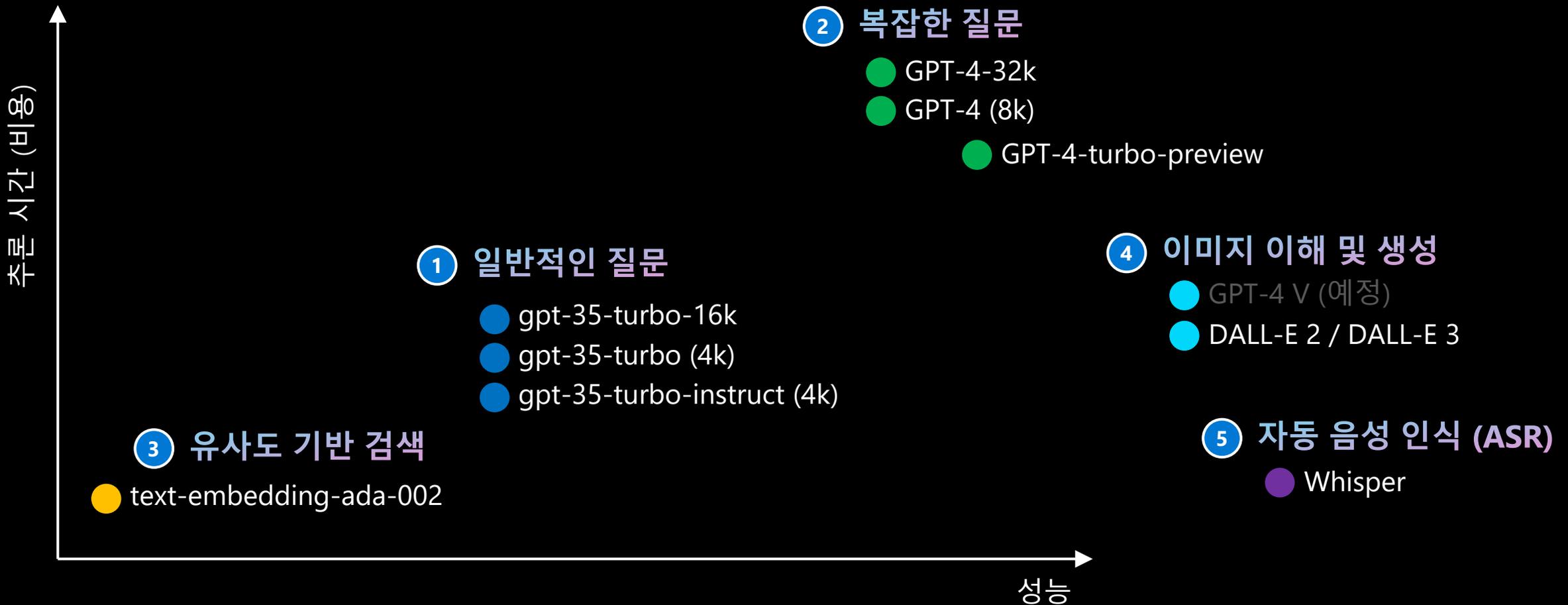
1. OpenAI 사용 환경
2. Machine Learning
3. Embedding API & Vector Database
4. 외부 API와 연결
5. Azure 서비스와 연결



PART 1. OpenAI 사용 환경

Azure OpenAI Service

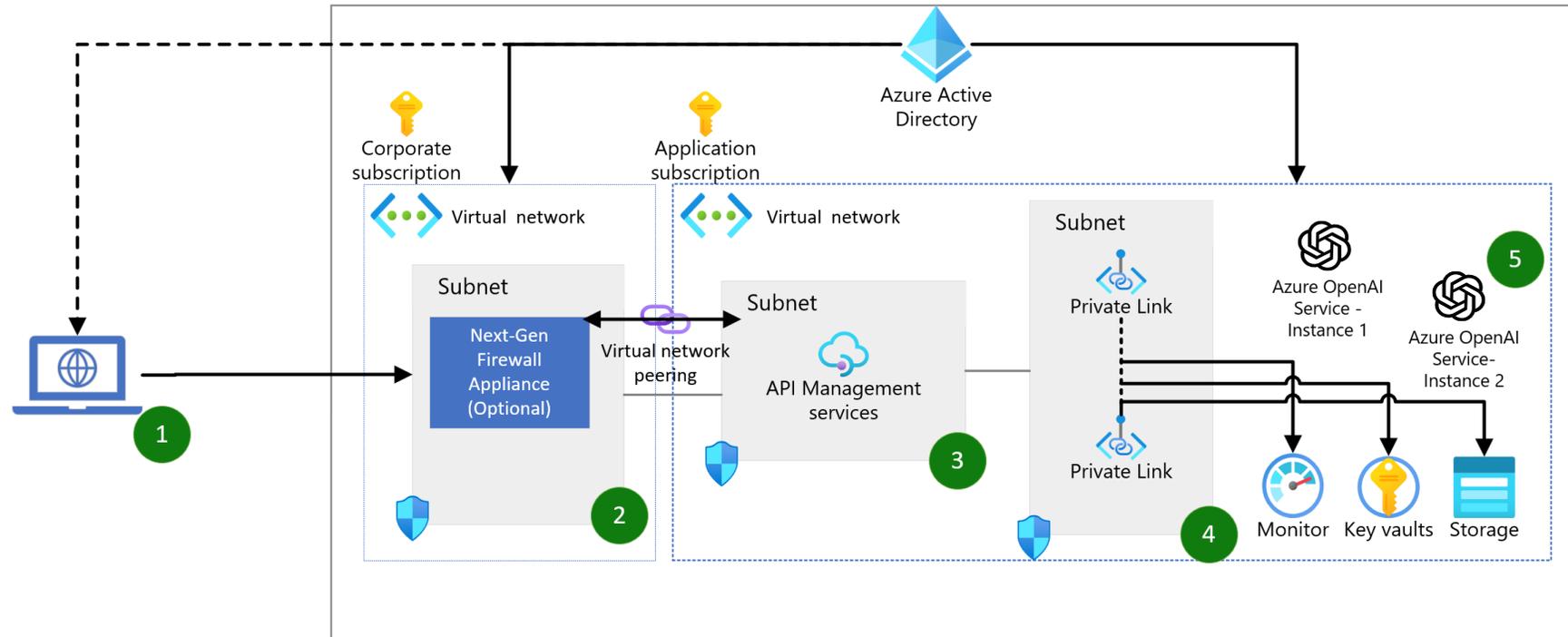
목적에 맞는 생성형 AI 기술을 채택



모델별 리전 지원 및 특성

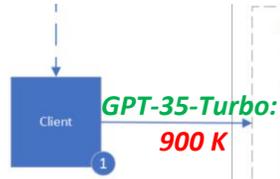
Model ID	Base model Regions	Default Limit Value (TPM)	Max Request (tokens)	Training Data (up to)	Pricing (per 1,000 tokens)	
text-embedding-ada-002 (v2)	East US, South Central US, West Europe, France Central	240K	8,191	Sep 2021	\$0.0001	
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Switzerland North	350K				
gpt-35-turbo-instruct (0914)	East US, Sweden Central	240K	4,097		\$0.0015	\$0.0020
gpt-35-turbo (0301)	East US, France Central, South Central US, UK South, West Europe	240K	4,096		\$0.0015	\$0.0020
gpt-35-turbo (0613)	East US, South Central US, West Europe, France Central, UK South	240K			\$0.0015	\$0.0020
	North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	300K				
gpt-35-turbo-16K (0613)	East US, South Central US, West Europe, France Central, UK South	240K	16,384		\$0.0030	\$0.0040
	North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	300K				
gpt-35-turbo (1106)	Australia East, Canada East, France Central, South India , Sweden Central, UK South, West US	120K	16,385 / 4,096		\$0.0010	\$0.0020
gpt-4 (0613)	East US, South Central US, West Europe, France Central	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	40K				
gpt-4-32k (0613)	East US, South Central US, West Europe, France Central	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	80K				
gpt-4 (1106-preview)	Australia East, Canada East, East US 2, France Central, UK South, West US	80K	128,000 / 4,096	Apr 2023	\$0.0100	\$0.0300
	South India, Norway East , Sweden Central	150K				
Dall-E 2	East US	2 req / concur	1,000 (Max Char)	-	\$0.02 (per 1 image)	
Dall-E 3	Sweden Central	12 req / 1m	4,000 (Max Char)	-	\$0.04(SD) / \$0.08(HD)	
Whisper	North Central US, West Europe	-	25 MB (Audio)	-	\$0.36 (per hour)	

안정적인 서비스를 위한 Enterprise OpenAI 환경 구축



- 트래픽이 한도를 초과하는 경우 **API의 고가용성** 제공
- 다수 리전에 모델 배포하여 **기본 Quota 증가** 효과
- 책임감 있는 사용을 위한 로깅
- 다양한 사용자 그룹에 대한 세분화된 액세스 제어
- Azure Active Directory를 통한 역할 기반 액세스

리전별 모델 Quota 조절



Name	Resource group	Location
canada-east-hs	OpenAI	Canada East
hs-australia-east	OpenAI	Australia East
hs-openai-eus2	OpenAI	East US 2
hso-scus-openai	OpenAI	South Central US
hs-openai	OpenAI	East US
japanopenaihs	OpenAI	Japan East

- 리전별 OpenAI 리소스 생성 가능
- 리소스별 모델 선택하여 배포 가능
- 리소스별 모델의 기본 Quota 제공

Default Quota
 GPT-35-Turbo: 300 K (TPM)
 GPT-4: 40 K (TPM)
 GPT-4-32K: 80 K (TPM)
 TEXT-EMBEDDING-ADA-002: 350 K (TPM)

Azure OpenAI Studio > Quotas

Canada East

GPT-35-Turbo: 300K TPM

Quota name	배포	Usage/Limit	Request quota
Tokens Per Minute (thousands) - GPT-35-Turbo		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-35-Turbo-16K		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-4		40 of 40 100%	
Tokens Per Minute (thousands) - GPT-4-32K		80 of 80 100%	
Tokens Per Minute (thousands) - Text-Embedding-Ada-002		350 of 350 100%	

Azure OpenAI Studio > Quotas

Japan East

GPT-35-Turbo: 300K TPM

Quota name	배포	Usage/Limit	Request quota
Tokens Per Minute (thousands) - GPT-35-Turbo		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-35-Turbo-16K		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-4		40 of 40 100%	
Tokens Per Minute (thousands) - GPT-4-32K		80 of 80 100%	
Tokens Per Minute (thousands) - Text-Embedding-Ada-002		350 of 350 100%	

Azure OpenAI Studio > Quotas

Australia East

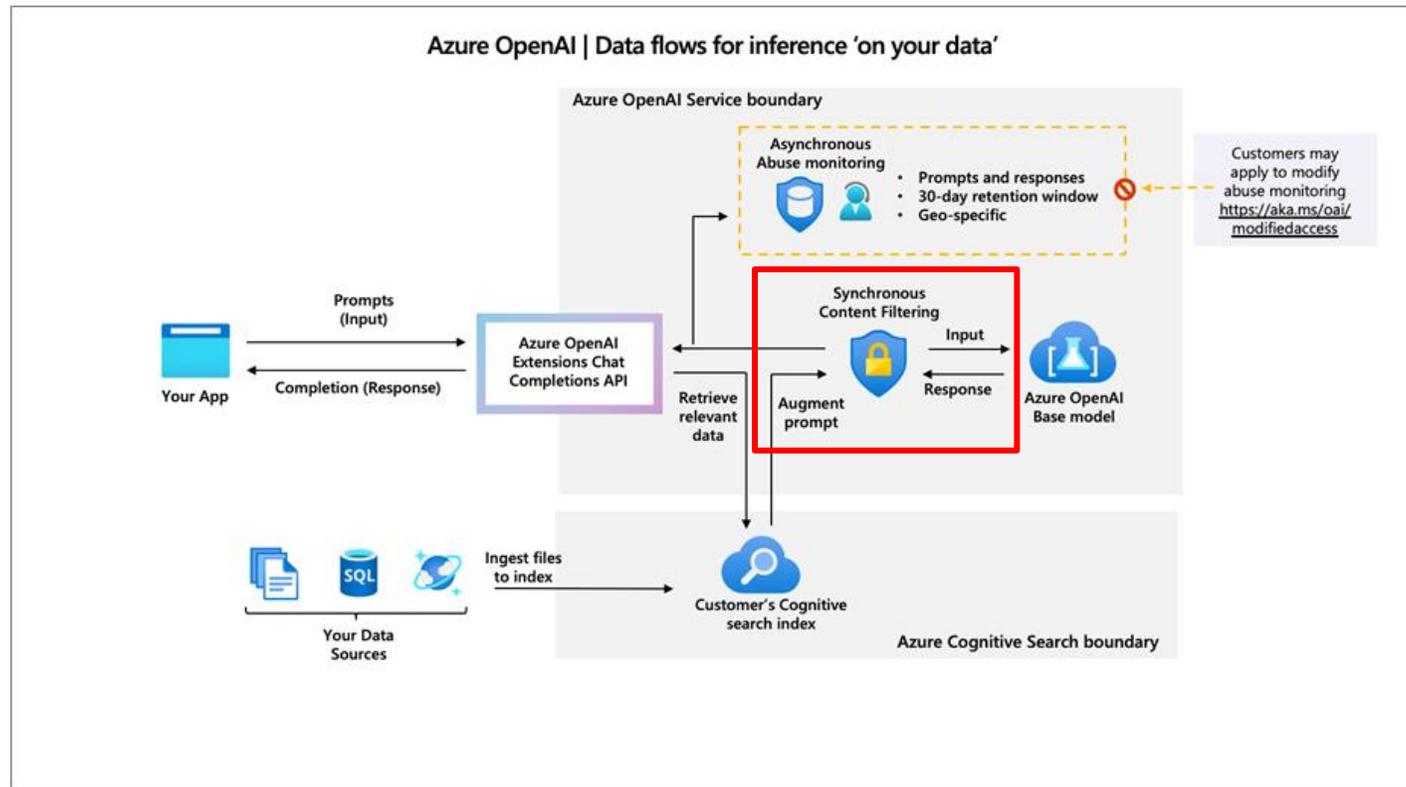
GPT-35-Turbo: 300K TPM

Quota name	배포	Usage/Limit	Request quota
Tokens Per Minute (thousands) - GPT-35-Turbo		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-35-Turbo-16K		300 of 300 100%	
Tokens Per Minute (thousands) - GPT-4		40 of 40 100%	
Tokens Per Minute (thousands) - GPT-4-32K		80 of 80 100%	

GPT-35-Turbo: 300K TPM x 3 = 900K TPM

Contents Filtering 기능 소개

- 책임감 있는 애플리케이션을 만들어서 제공하기 위한 모델에 안전 장치로 활용 가능 ([사용 승인이 필요](#))
- 콘텐츠 필터링 시스템은 **Prompts(Input)와 Completions(Output)을 검사하여 유해한 콘텐츠를 감지하고 조치** 가능
- 콘텐츠 필터링 시스템은 영어, 독일어, 일본어, 스페인어, 프랑스어, 이탈리아어, 포르투갈어, 중국어를 지원함
- 훈련이나 테스트되지 않은 언어로 된 부적절한 콘텐츠는 감지하지 못할 수 있음 (**한국어는 테스트 필요함**)



Contents Filtering 기능 활용

- 책임감 있는 애플리케이션을 만들어서 제공하기 위한 모델에 안정 장치로 활용 가능 ([사용 승인이 필요](#))
- 콘텐츠 필터링 시스템은 **Prompts(Input)와 Completions(Output)을 검사하여 유해한 콘텐츠를 감지하고 조치** 가능
- 콘텐츠 필터링 시스템은 영어, 독일어, 일본어, 스페인어, 프랑스어, 이탈리아어, 포르투갈어, 중국어를 지원함
- 훈련이나 테스트되지 않은 언어로 된 부적절한 콘텐츠는 감지하지 못할 수 있음 (**한국어는 테스트 필요함**)

Edit content filters

Content filtering configurations are created within a Resource and can be associated with Deployments.
[Learn more about configurability here.](#)

The default content filtering configuration is set to filter at the medium severity threshold for all four content harms categories for both, prompts and completions. That means that content that is detected at severity level medium or high is filtered, while content detected at severity level low is not filtered by the content filters.

Create custom configuration name
DontUseContentFilter

Set severity levels

Severity	User prompts (Input)			Model completions (Output)		
	Low	Medium	High	Low	Medium	High
Hate	On	✓	✓	On	✓	⊖
Sexual	Off			On	✓	⊖ Filter
Self-harm	On	⊖	⊖	On	✓	⊖
Violence	On	⊖	⊖	Off		

[Learn more about content filters here.](#)

저장 취소

배포 편집

모델 이름
gpt-35-turbo

배포 이름 ①
ChatGPT

고급 옵션 ▾

Content Filter ②
DontUseContentFilter

100K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit (thousands) ①
100K

Corresponding requests per minute (RPM) = 600

저장 후 닫기 취소

Chat Completions API

채팅 뿐만 아니라 - 모든 시나리오에 대한 다목적 인터페이스 사용

요청의 97%는 Chat Completion API를 사용중 (하단 출처: OpenAI Blog)

모델은 "system" 메시지의 지침(instructions)을 준수함

책임 있는 AI 조정을 포함하여 모델에 대한 행동 지침을 설정함

예시는 "User", "Assistant" 필드를 사용

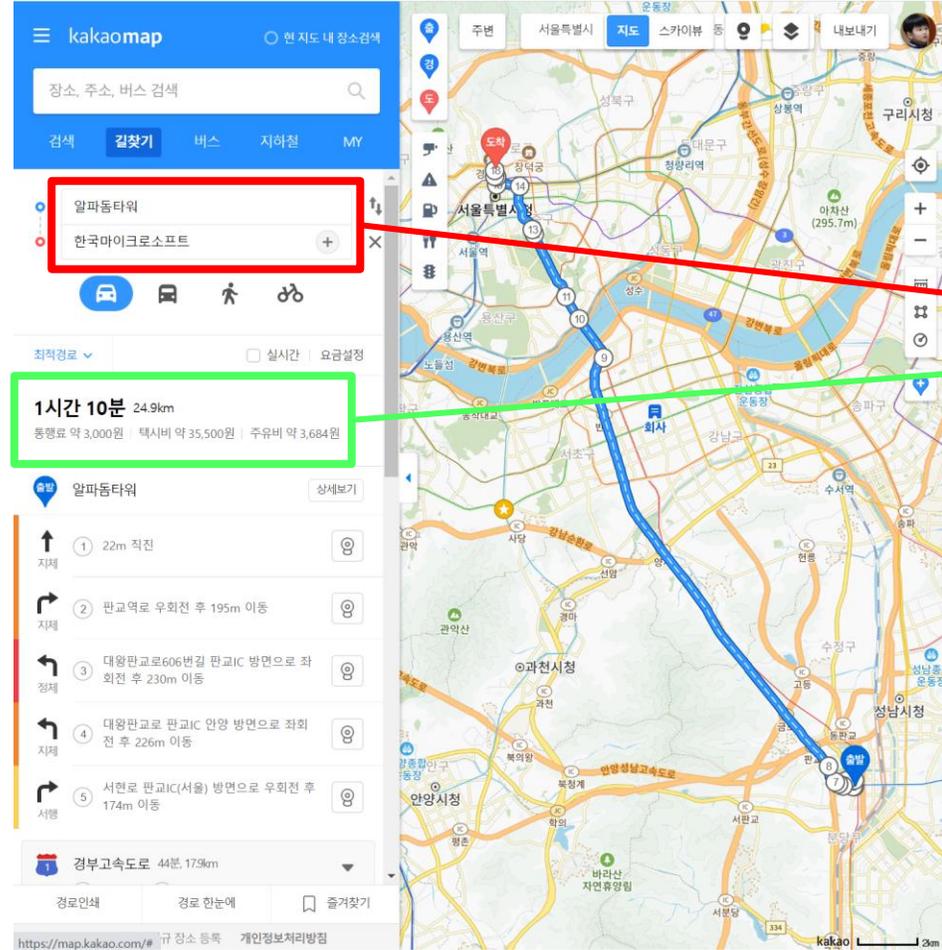
(Text) Completions API는 "legacy"로 표시되며, 기존 Completions API 지원 모델 역시 Chat Completion API 모델로 통합

```
import openai
openai.api_type = "azure"

response = openai.ChatCompletion.create(
    engine = "gpt-4",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The LA Dodgers won the World Series in 2020."},
        {"role": "user", "content": "Where was it played?"}
    ]
)
```

Function calling 기능

- GPT 모델이 **사용자 지정 함수를 호출할 수 있도록** 미세 조정해 놓은 기능
- gpt-3.5-turbo (0613) 및 gpt-4 (0613) 모델에서 API Version(**2023-07-01-preview 이후**) 확인 후 사용



```

origin_name = "알파돐타워"
destin_name = "한국마이크로소프트"

messages = [
  {"role": "system", "content": "You are car route navigation. When the user provides the origin and destination name, you provides summary. If the distance or duration units are too granular, convert them to higher units. For example, 1894 seconds is displayed as 31 minutes and"},
  {"role": "user", "content": f"{origin_name}에서 {destin_name}까지 차로 이동하면 거리랑 시간은 얼마야? 통행료나 택시비는 얼마야?"}
]
assistant_response = run_conversation(messages, functions, available_functions, deployment_id)
content = json.dumps(assistant_response['choices'][0]['message']['content'], ensure_ascii=False, indent=4)
content = content.replace("\\n", "\n").replace("\\\"", "\"")
print(content)

```

2.3s Python

알파돐타워에서 한국마이크로소프트까지 차로 이동하는 거리는 약 29.92km이며, 추정 소요 시간은 약 1시간 10분입니다. 통행료는 없고, 택시비는 약 35,400원입니다

```

origin_name = "판교 알파돐타워"
destin_name = "한국마이크로소프트"

messages = [
  {"role": "system", "content": "You are a navigation bot agent. Your reply must absolutely in JSON format. You must never modify or tamper"},
  {"role": "user", "content": f"{origin_name}에서 {destin_name}까지 경로 검색한 정보를 알려줘."}
]
assistant_response = run_conversation(messages, functions, available_functions, deployment_id)
content = json.dumps(assistant_response['choices'][0]['message']['content'], ensure_ascii=False, indent=4)
content = content.replace("\\n", "\n").replace("\\\"", "\"")
print(content)

```

2.7s Python

```

{
  "origin_name": "판교 알파돐타워",
  "destination_name": "한국마이크로소프트",
  "taxi_fare": 35700,
  "toll_fare": 0,
  "distance": 29924,
  "duration": 4236
}

```

Function calling 은 어떻게 써야 하나요?

- 기존 사용중인 **API(함수)의 설명**을 **Functions 설명에 기술**하고, 해당 함수 코드를 작성하세요!
- 여러 함수를 동시에 놓고도 사용 가능하며, 필요한(**required**) 파라미터는 사용자 질의에서 추출함

Function #1: 카카오 모빌리티 길찾기 API

카카오 길찾기 API를 통해서 출발지와 목적지 사이의 정보를 탐색합니다.

```
# Kakao 길찾기 API
def get_directions(origin, destination, waypoints="", priority="RECOMMEND", car_fuel="GASOLINE", car_hipass="true", alternatives="true"):
    # 키워드 기반 위경도 좌표 정보 수집
    origin_xy_info = get_location_xy(origin)
    destn_xy_info = get_location_xy(destination)

    params = {
        "origin": origin_xy_info,
        "destination": destn_xy_info,
        "waypoints": waypoints,
        "priority": priority,
        "car_fuel": car_fuel,
        "car_hipass": car_hipass,
        "alternatives": alternatives,
        "road_details": road_details,
    }

    url = "https://apis-navi.kakaomobility.com/v1/directions?{}".format("&".join([f"{k}={v}" for k, v in params.items()]))
    response = requests.get(url, headers=headers)

    response_summary = response.json()["routes"][0]["summary"]
    return_data = {
        "origin_name": response_summary["origin"]["name"],
        "destination_name": response_summary["destination"]["name"],
        "taxi_fare": response_summary["fare"]["taxi"],
        "toll_fare": response_summary["fare"]["toll"],
        "distance": response_summary["distance"],
        "duration": response_summary["duration"],
    }

    return json.dumps(return_data)
```

0.0s

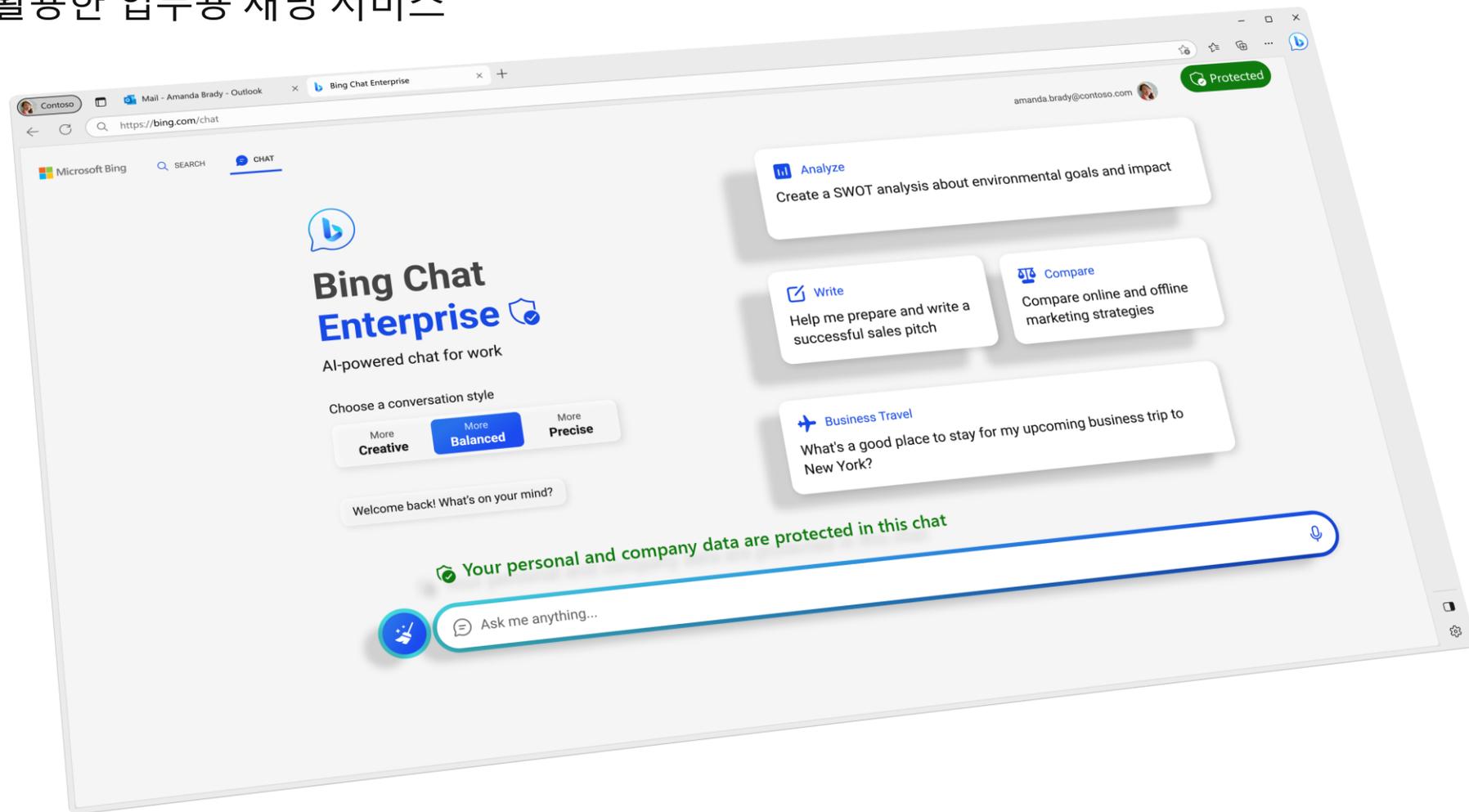
2.1 모델이 호출 방법을 알 수 있도록 함수 설명

```
functions = [
    {
        "name": "get_directions",
        "description": "API to search routes based on origin and destination information",
        # "description": "출발지와 도착지 정보를 기반으로 경로 검색하는 API",
        "parameters": {
            "type": "object",
            "properties": {
                "origin": {"type": "string"},
                "destination": {"type": "string"},
            },
            "required": ["origin", "destination"],
        },
    },
    {
        "name": "get_future_directions",
        "description": "API to search routes based on origin and destination information based on future departure_time",
        # "description": "출발지와 도착지 정보를 미래 시간 기반으로 경로 검색하는 API",
        "parameters": {
            "type": "object",
            "properties": {
                "origin": {"type": "string"},
                "destination": {"type": "string"},
                "departure_time": {"type": "string"},
            },
            "required": ["origin", "destination", "departure_time"],
        },
    },
    {
        "name": "time_calculator",
    },
]
```

```
"name": "get_directions",
"description": "API to search routes based on origin and destination information",
# "description": "출발지와 도착지 정보를 기반으로 경로 검색하는 API",
"parameters": {
    "type": "object",
    "properties": {
        "origin": {"type": "string"},
        "destination": {"type": "string"},
    },
    "required": ["origin", "destination"],
}
```

Bing Chat Enterprise 출시

- AI 기술을 활용한 업무용 채팅 서비스



Bing Chat Multimodal 지원 - Visual Search

Microsoft Bing 검색 채팅 hyounsookim... 3353

좋아요, 준비됐어요! 재설정해 주셔서 감사합니다. 우리는 무엇에 대해 배워야 할까요?

이 이미지에 있는 것들은 어디에 쓰는 물건이야?



- ✓ 이미지 분석: 프라이버시 블러는 Bing 채팅에서 얼굴을 숨깁니다.
- ✓ "게임 컨트롤러 검은색 파란색" 검색 중
- ✓ 답변을 생성하는 중...

게임 컨트롤러라고 합니다. 검은색과 파란색으로 구성되어 있고, 검은색 컨트롤러에는 녹색 "X" 버튼이 있고 파란색 컨트롤러에는 검은색 "X" 버튼이 있습니다 ¹ ²

자세한 정보: [1. browse.gmarket.co.kr](https://browse.gmarket.co.kr) [2. istockphoto.com](https://istockphoto.com) [3. kr.freepik.com](https://kr.freepik.com) 1 / 30

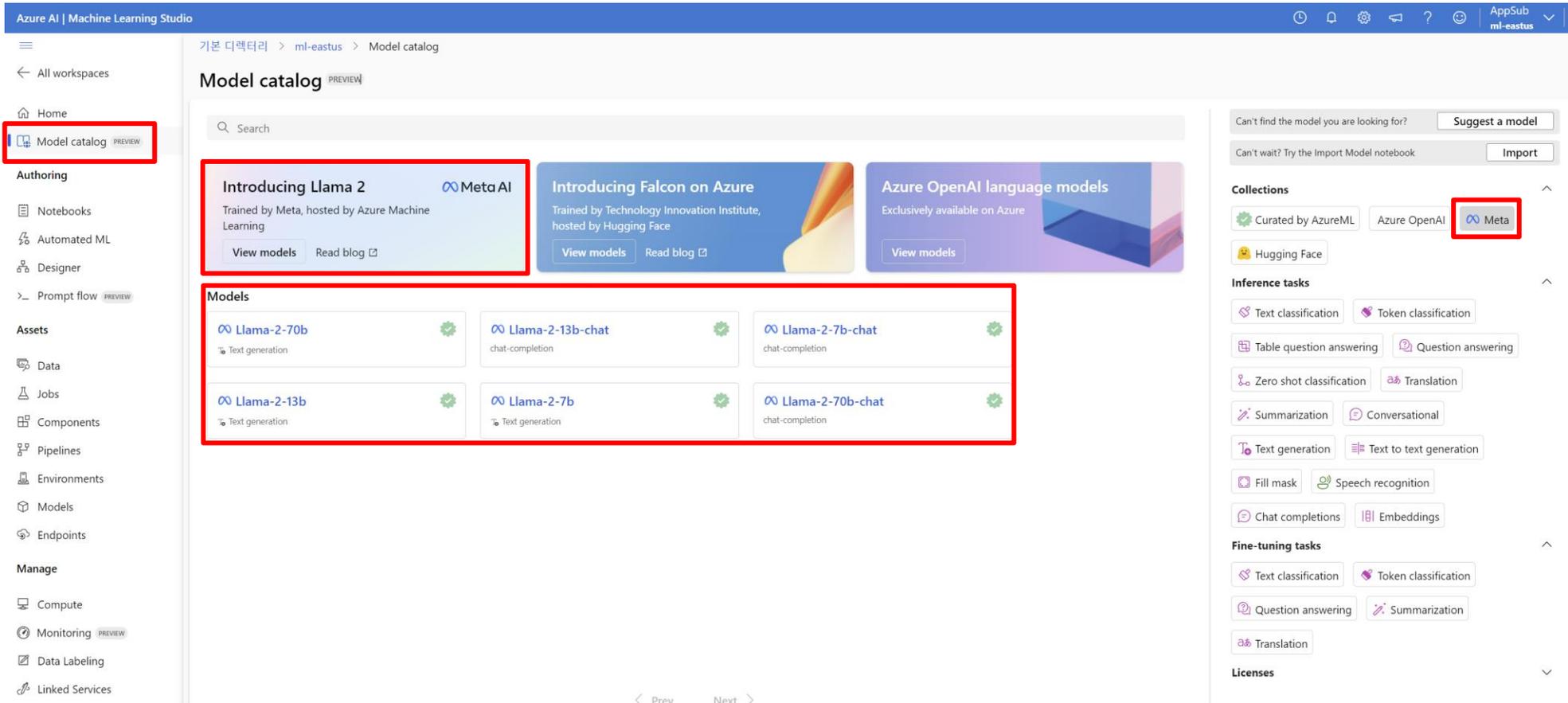
어떤 게임에서 사용되나요? 이 컨트롤러를 어디서 살 수 있나요? 컨트롤러의 역사에 대해 알려주세요.

무엇이든 물어보세요 0/2000

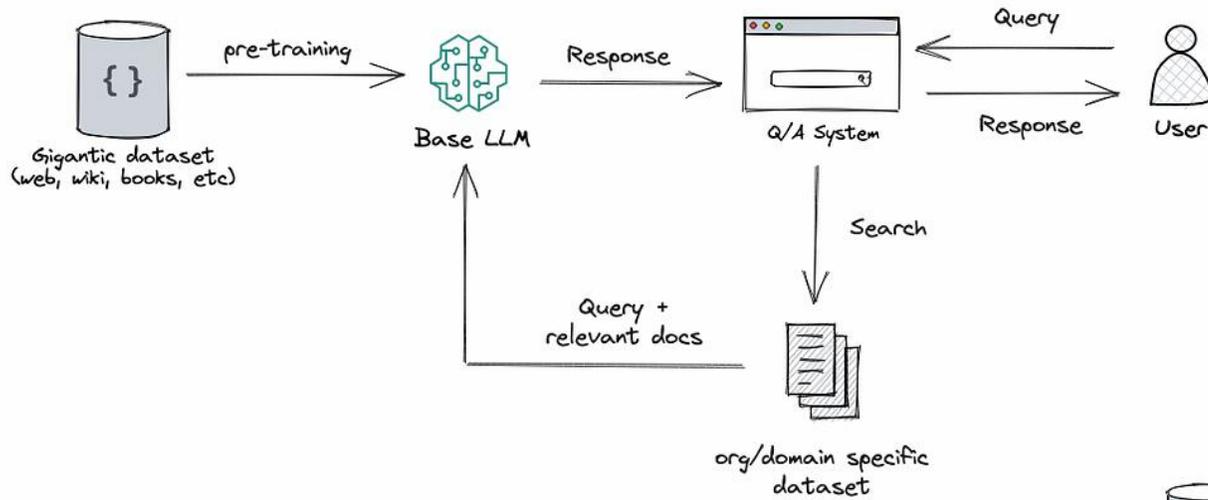
피드백

Llama 2 on Azure

- Meta와 Microsoft가 Azure와 Windows에서 Llama 2 대규모 언어 모델을 지원
- Azure에서 7B, 13B, 70B 매개변수 Llama 2 모델을 미세 조정(Fine-tune)하고 배포 가능

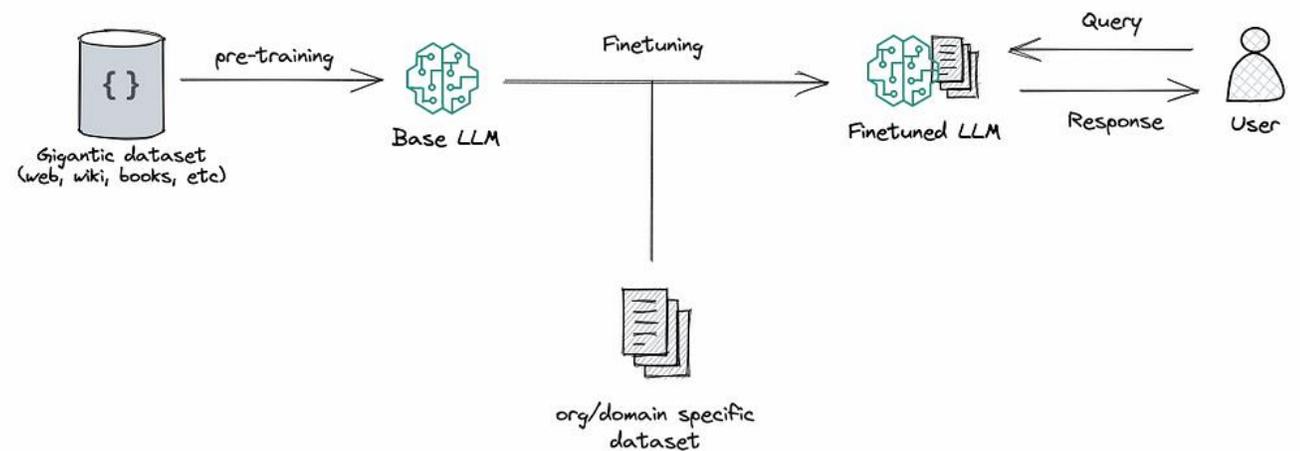


RAG vs Fine-Tuning



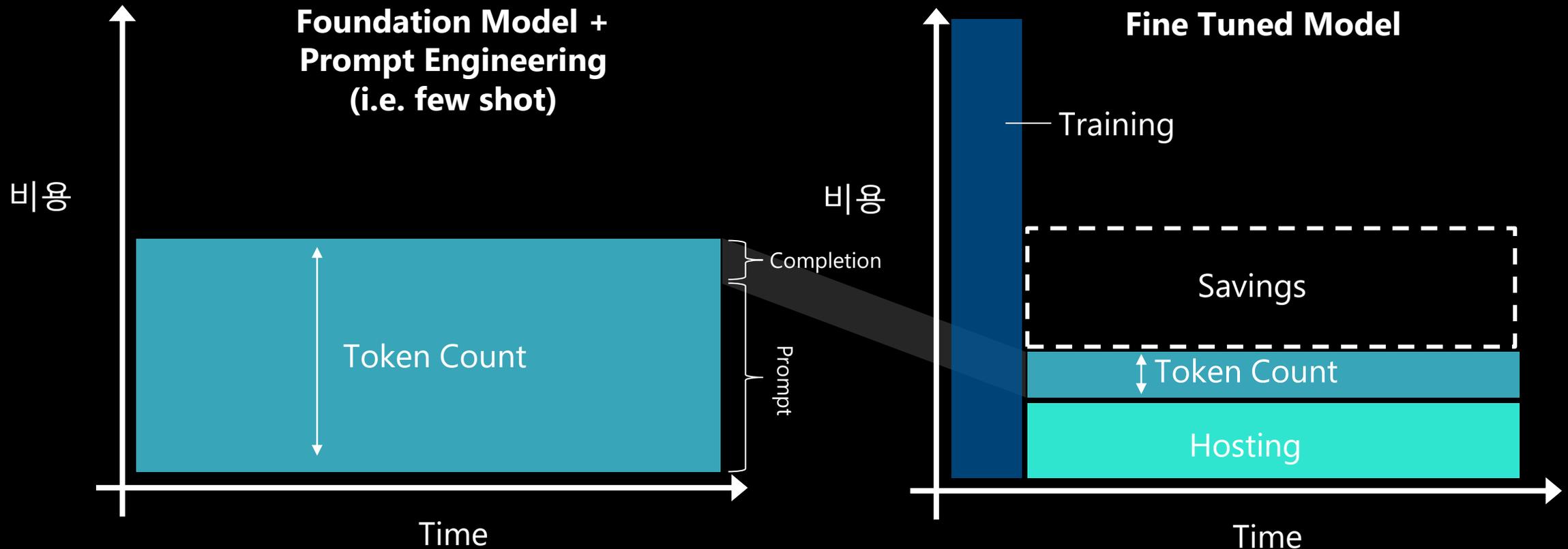
**외부 데이터를 활용해야 한다면?!
Retrieval-Augmented Generation**

**특정 도메인 특화된 스타일과 용어 활용
Fine-tuning**



LLM Economics: Potential Optimization

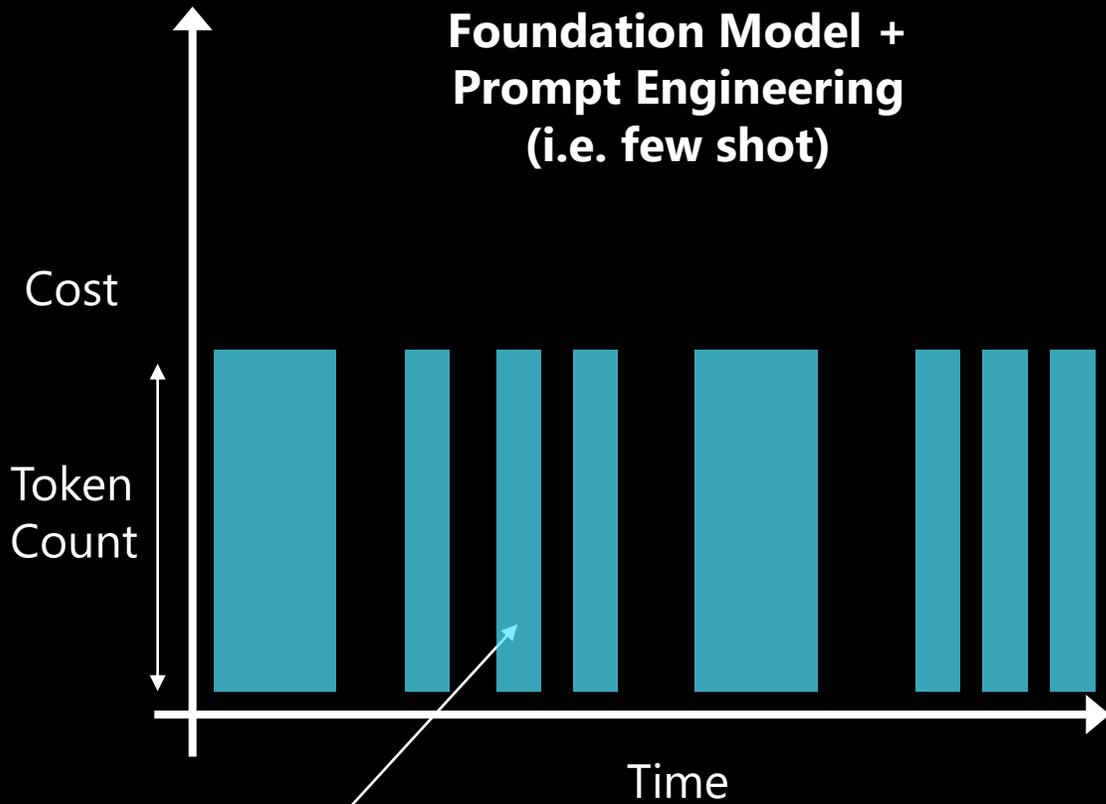
대량의 Prompt 토큰 절감, 맞춤형 스타일로 최적화 – **Fine-tuning**



LLM Economics: Potential Pitfalls

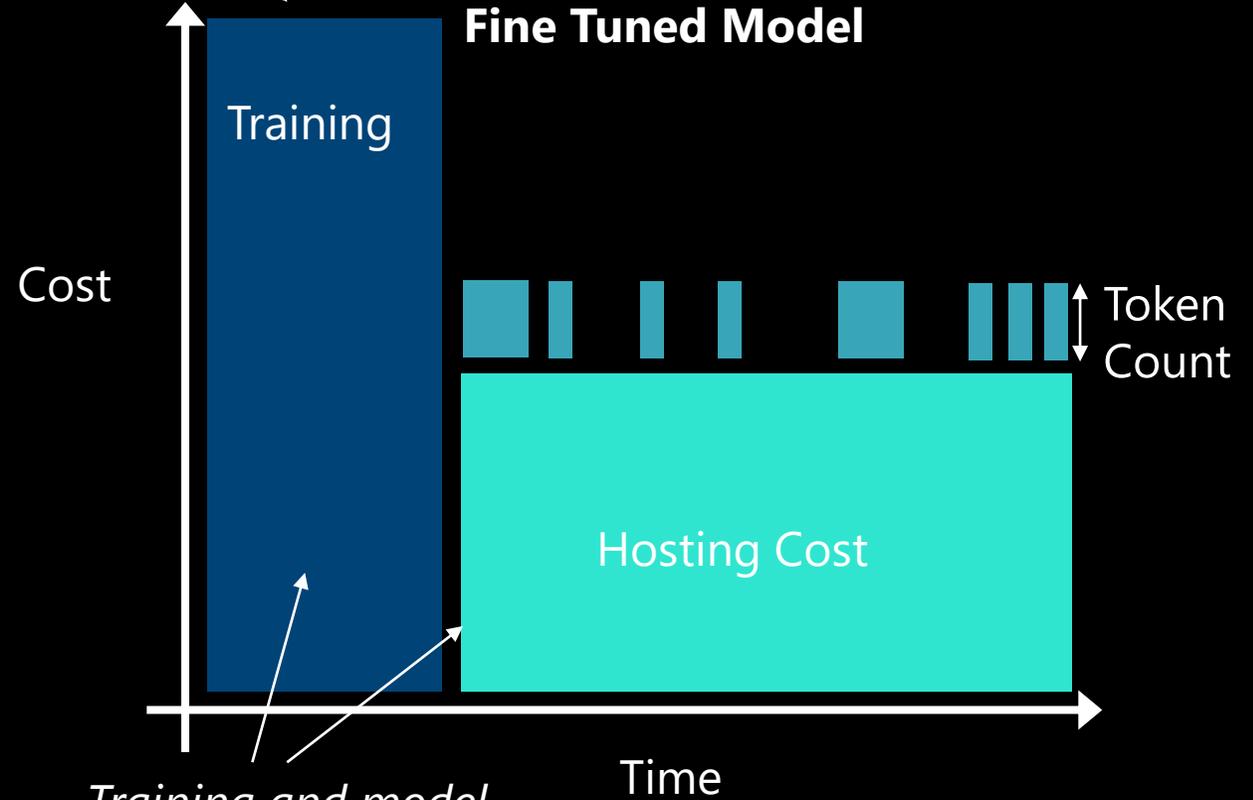
Level of Effort required to adjust prompt vs. fine-tune a model

Foundation Model + Prompt Engineering (i.e. few shot)



Need to consider intermittent usage

Fine Tuned Model



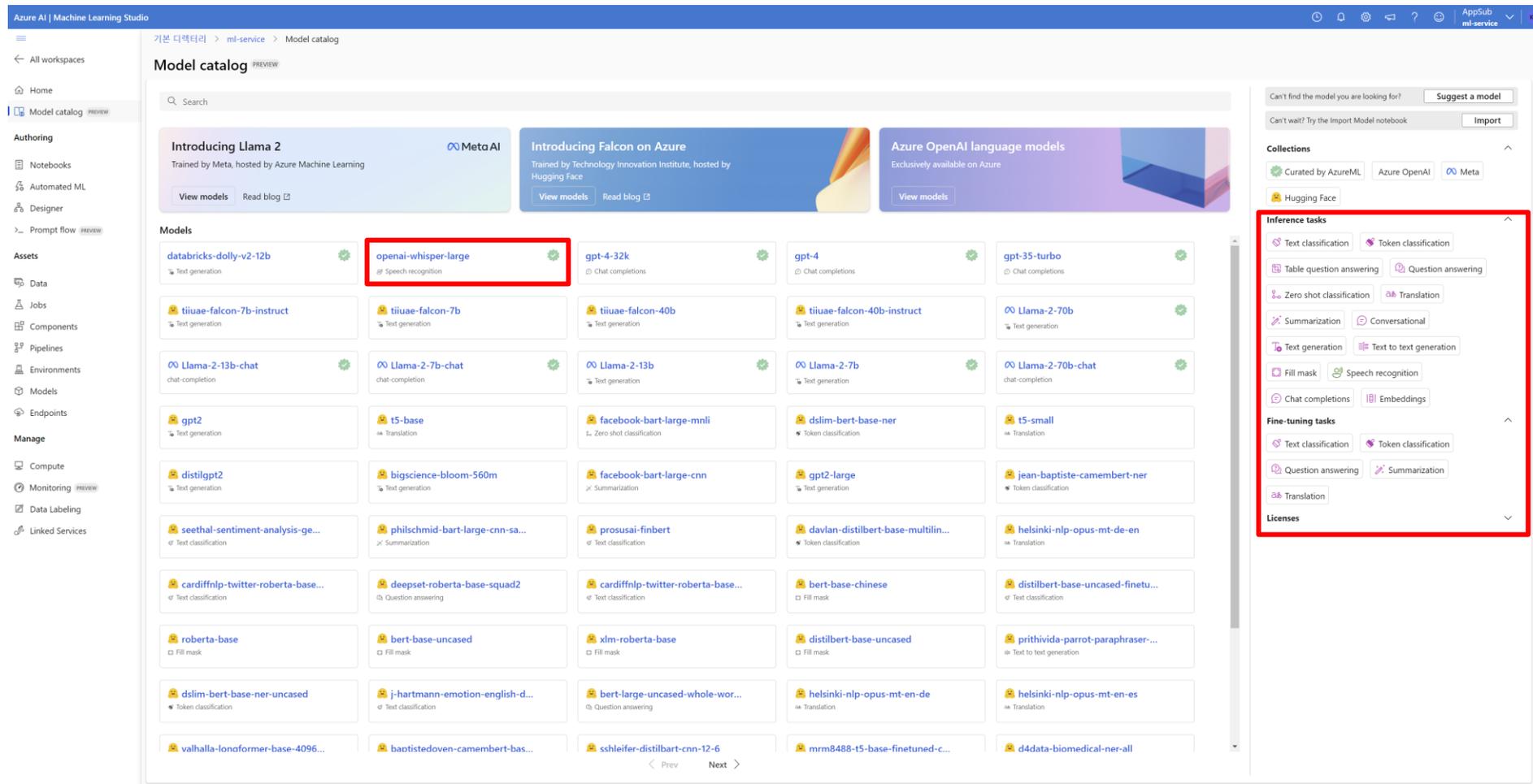
Training and model hosting charges vary by base model



PART 2. Machine Learning

Azure Machine Learning – Model catalog

- 카탈로그를 통해 Foundation 모델을 검색하고, 사용자 지정 및 운영 지원



Azure Machine Learning – Model catalog

- Foundation 모델을 평가, 배포 또는 Finetune 할 수 있는 기능 제공

The screenshot displays the Azure Machine Learning Studio interface for the Llama-2-13b model. The breadcrumb navigation shows the path: 기본 디렉터리 > ml-eastus > Model catalog > azureml-meta > Llama-2-13b. The model name 'Llama-2-13b' is highlighted in a red box. Below the model name, there are tabs for 'Overview', 'Versions', and 'Artifacts'. A toolbar contains buttons for 'Refresh', 'Evaluate', 'Finetune', 'Deploy', and 'View license', with the 'Deploy' button also highlighted in a red box. A dropdown menu is open from the 'Deploy' button, showing two options: 'Real-time endpoint' (Deploy the model using the real-time endpoint wizard) and 'Batch endpoint' (Deploy the model using the batch endpoint wizard). A modal dialog titled 'Deploy with Azure AI Content Safety (preview)' is open, showing a radio button for 'Enable Azure AI Content Safety (Recommended)' which is selected. Below this, there is a text box explaining that Azure AI Content Safety can detect and filter harmful content. At the bottom of the dialog, there is a 'Proceed' button and a 'Cancel' button.

Model Details

Note: Use of this model is governed by the Meta Llama 2 license.

Meta has developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned models in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform other open source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some of the best models like ChatGPT and PaLM. We provide a detailed description of our approach to fine-tuning and safety improvements of Llama-2-Chat in our Llama 2 blog post. We build on our work and contribute to the responsible development of LLMs.

	Training Data	Params	Content Length	GQA	Tokens	LR
Llama 2	A new mix of publicly available online data	7B	4k	X	2.0T	3.0 x 10 ⁻⁴
Llama 2	A new mix of publicly available online data	13B	4k	X	2.0T	3.0 x 10 ⁻⁴
Llama 2	A new mix of publicly available online data	70B	4k	✓	2.0T	1.5 x 10 ⁻⁴

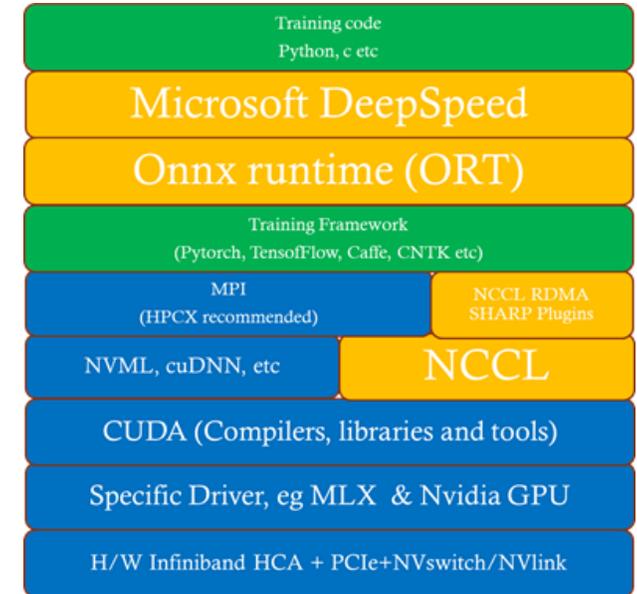
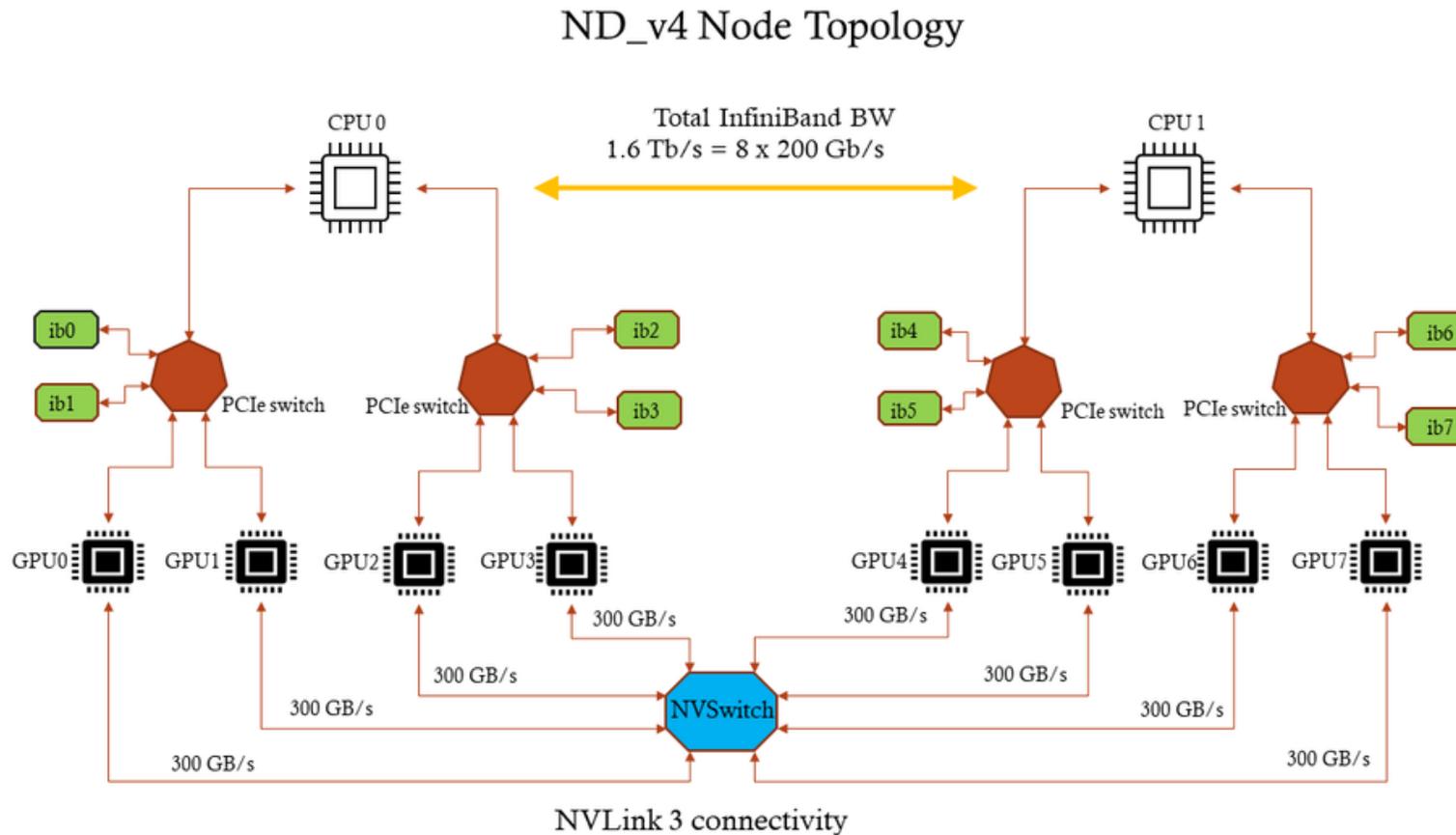
Llama 2 family of models. Token counts refer to pretraining data only. All models are trained with a global batch-size of 4M tokens. Query Attention (GQA) for improved inference scalability.

Model Developers Meta AI

Variations Llama 2 comes in a range of parameter sizes — 7B, 13B, and 70B — as well as pretrained and fine-tuned variations.

Azure Machine Learning – LLM Best Practice

- InfiniBand를 지원하는 Azure의 VM으로 LLM 모델을 만들 때의 Know-how를 블로그에 포스팅



Azure Machine Learning – Prompt flow

- LLM(대형 언어 모델)에서 제공하는 AI 애플리케이션의 전체 개발 주기를 간소화하도록 설계된 개발 도구
- AI 애플리케이션의 프로토타이핑, 실험, 반복 및 배포 프로세스를 간소화하는 포괄적 솔루션 제공

The screenshot displays the Azure Machine Learning Studio interface for a Prompt Flow. The flow is named "Flow-created-on-07-19-2023" and is in a "Completed" state. The flow graph on the right shows a sequence of steps: "Inputs" leading to a "chat" step, which then leads to a "python_rqa" step, and finally to "Outputs".

The "chat" step is configured with the following details:

- Connection: Default_AzureOpenAI
- Api: chat
- Deployment Name: gpt-35-turbo-v0301
- Temperature: 0.7
- Stop: (empty)
- Max Tokens: 256

The prompt for the chat step is:

```
9 Text 2: OpenAI has trained cutting-edge language models that are very good at understanding and generating text. Our API provides access to these models and can be used to solve virtually any task that i
10 assistant:
11 Keywords 2: OpenAI, language models, text processing, API.
12
13 {% for item in chat_history %}
14 user:
15 {{item.inputs.question}}
16 assistant:
```

The "python_rqa" step contains the following Python code:

```
1 from promptflow import tool
2
3 # The inputs section will change based on the arguments of the tool function, after you save the code
4 # Adding type to arguments and return value will help the system show the types properly
5 # Please update the function name/signature per need
6 @tool
7 def my_python_tool(input1: str) -> str:
8     return 'output: ' + input1
```

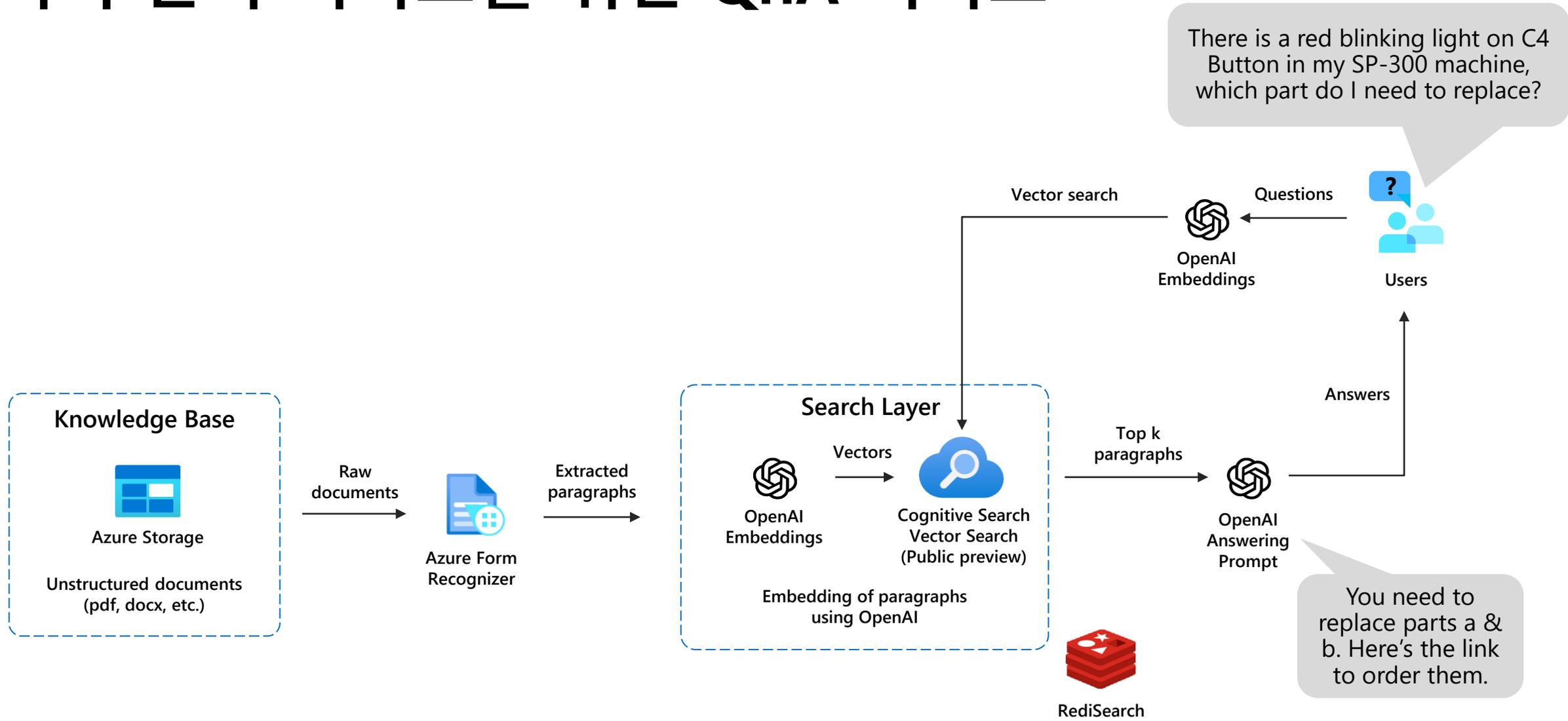
The execution results for the "python_rqa" step show the following output:

```
{
  "system_metrics": {
    "duration": 0.000534
  },
  "output": "output: Keywords 3: OpenAI, Microsoft, partnership, Investaent, independent research, AI, safe, useful, powerful."
}
```

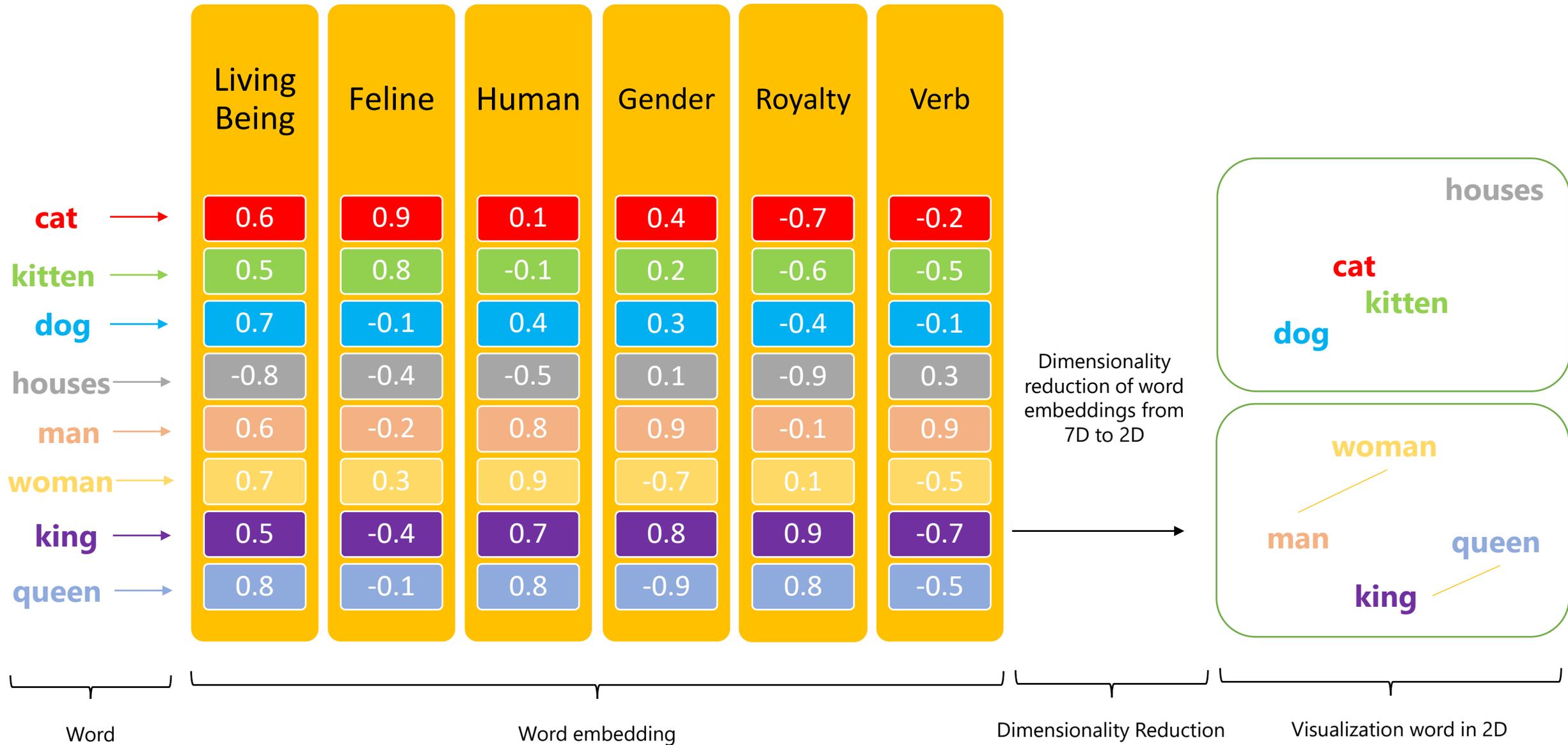


PART 3. Embedding API & Vector Database

지식 관리 서비스를 위한 QnA 서비스

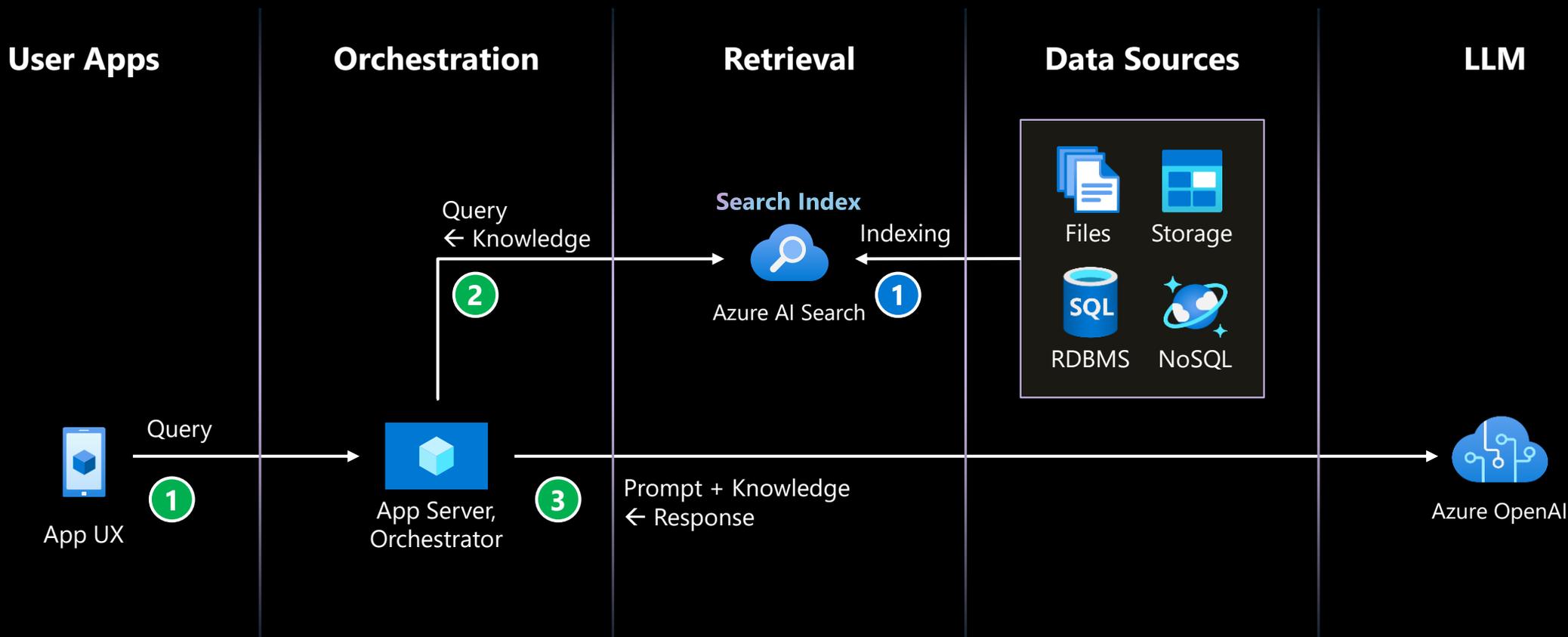


임베딩 (검색, 분류, 비교)



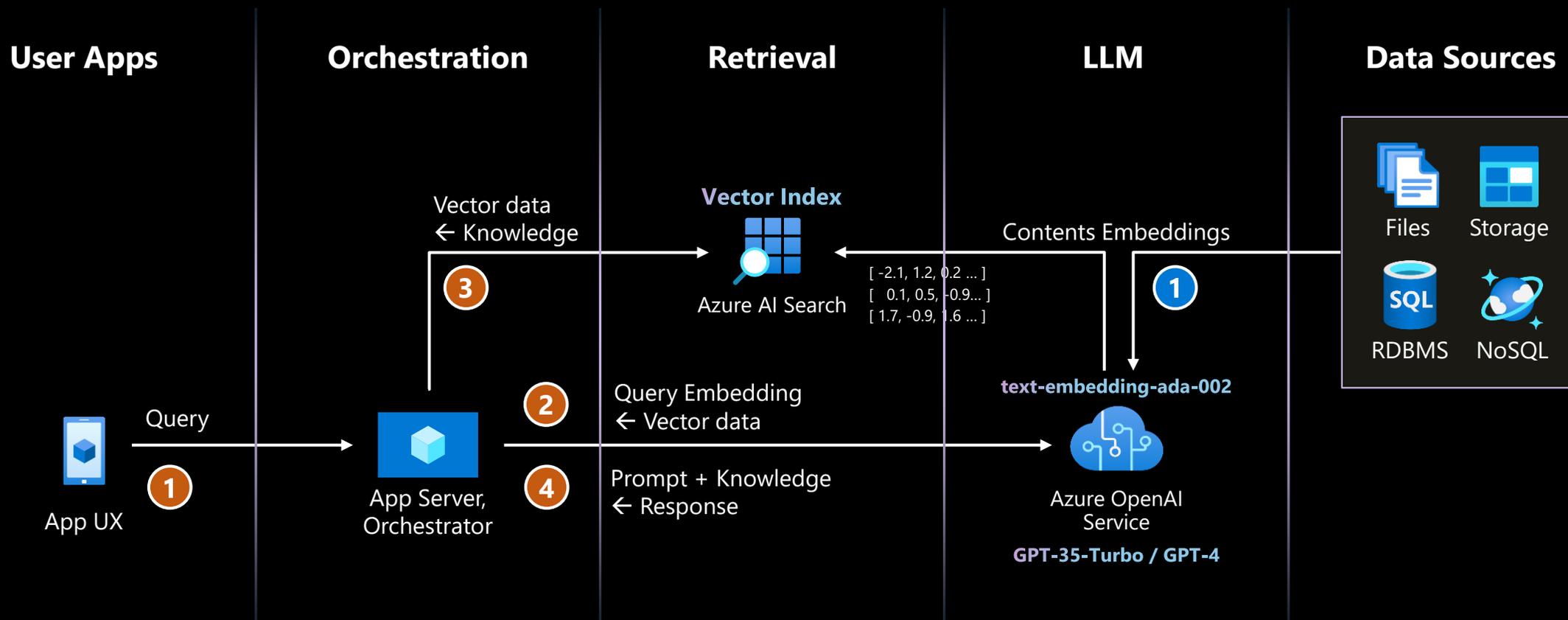
외부 데이터 활용

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



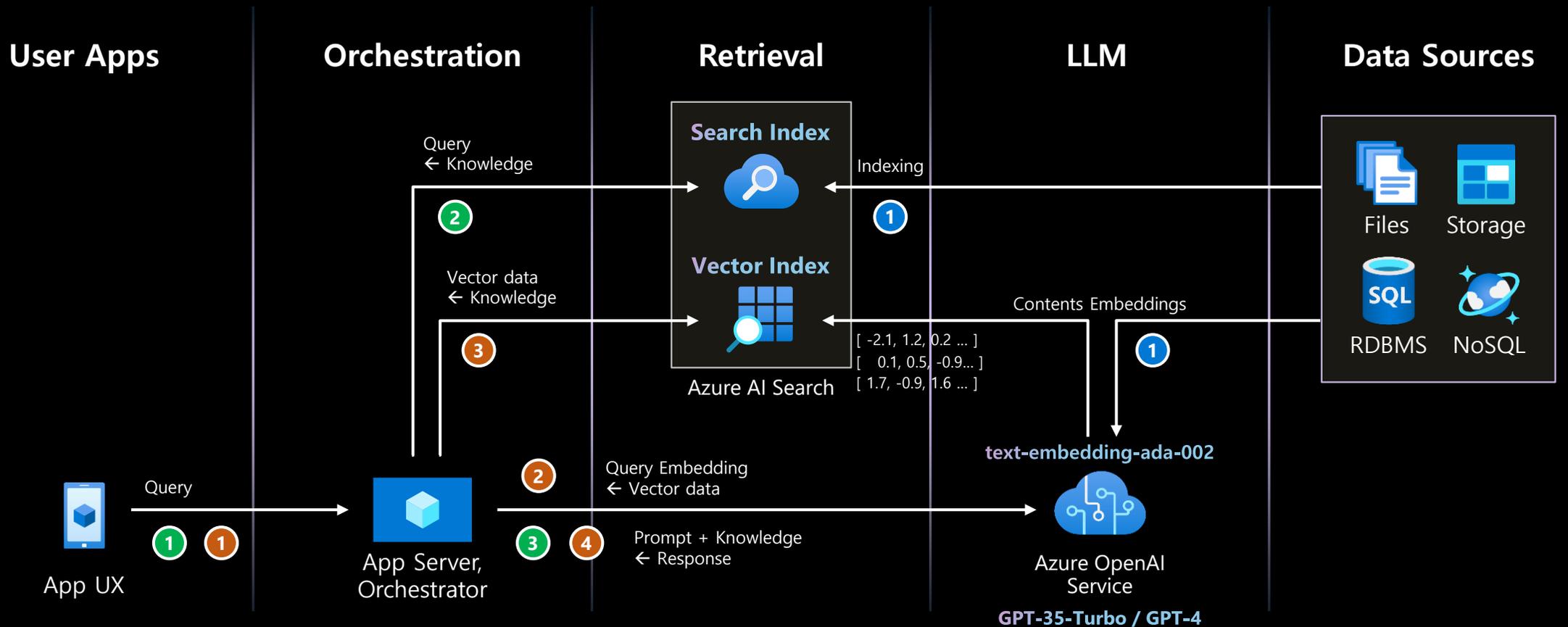
외부 데이터 활용

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



외부 데이터 활용

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



나의 데이터를 쉽게 통합하여 조회하고 요약 예시

도우미 설정

시스템 메시지 **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.](#)

Limit responses to your data content

Data Source: Azure Cognitive Search
Search Resource: search-for-rag
Index: openai-prompt

Remove data source

채팅 세션

채팅 지우기 코드 보기 Show raw JSON

한글로 출력
Prompting 잘 하는 방법을 한글로 번역해서 알려줘.

Prompting은 사용자가 원하는 정보를 추출하기 위해 질문을 만드는 것입니다. 이를 위해서는 사용자가 원하는 정보와 관련된 키워드를 파악하고, 그것을 기반으로 질문을 만들어야 합니다. 또한, 질문은 명확하고 구체적이어야 하며, 사용자가 이해하기 쉬운 언어로 작성되어야 합니다. [doc1]

Text Classification 하는 방법과 예시를 알려줘.

OpenAI LLM Model

Text Classification은 텍스트를 자동으로 분류하는 기술입니다. 이를 위해서는 먼저 학습 데이터를 수집하고, 이를 학습시켜야 합니다. 학습 데이터는 분류하고자 하는 카테고리에 대한 레이블이 달린 텍스트 데이터입니다. 모델 학습 후, 새로운 텍스트 데이터를 입력하면 모델은 해당 텍스트를 분류하여 적절한 카테고리 분류를 제공합니다. 예를 들어, 스팸 메일 필터링, 감성 분석, 토픽 모델링 등에 활용됩니다. [doc1]

Classification을 위한 예시로는, 자연어 처리 라이브러리인 NLTK를 사용하여 영화 리뷰를 긍정적인 리뷰와 부정적인 리뷰로 분류하는 프로젝트가 있습니다. 이를 위해 먼저 영화 리뷰 데이터셋을 수집하고, 이를 긍정적인 리뷰와 부정적인 리뷰로 레이블링합니다. 그 후, NLTK를 사용하여 모델을 학습시키고, 새로운 영화 리뷰를 입력하면 해당 리뷰가 긍정적인 리뷰인지 부정적인 리뷰인지 분류할 수 있습니다. [doc2]

참고문헌:

- [doc1]
- [doc2]

Configuration

배포 매개 변수

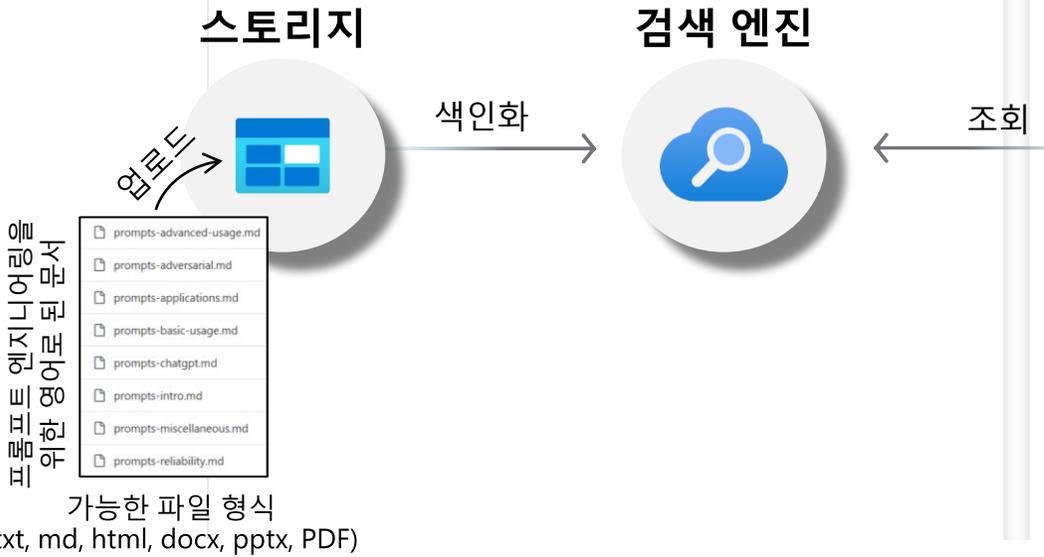
배포 * chatgpt35

세션 설정

지난 메시지 포함 10

현재 토큰 수

Input tokens progress indicator
1422/4000



프롬프트 엔지니어링을 위한 영어로 된 문서

Public preview

Azure Cognitive Search – Vector search

Revolutionizing Indexing and Retrieval for LLM-powered Apps

Power your retrieval-augmented generation applications



Images



Audio



Video



Graphs



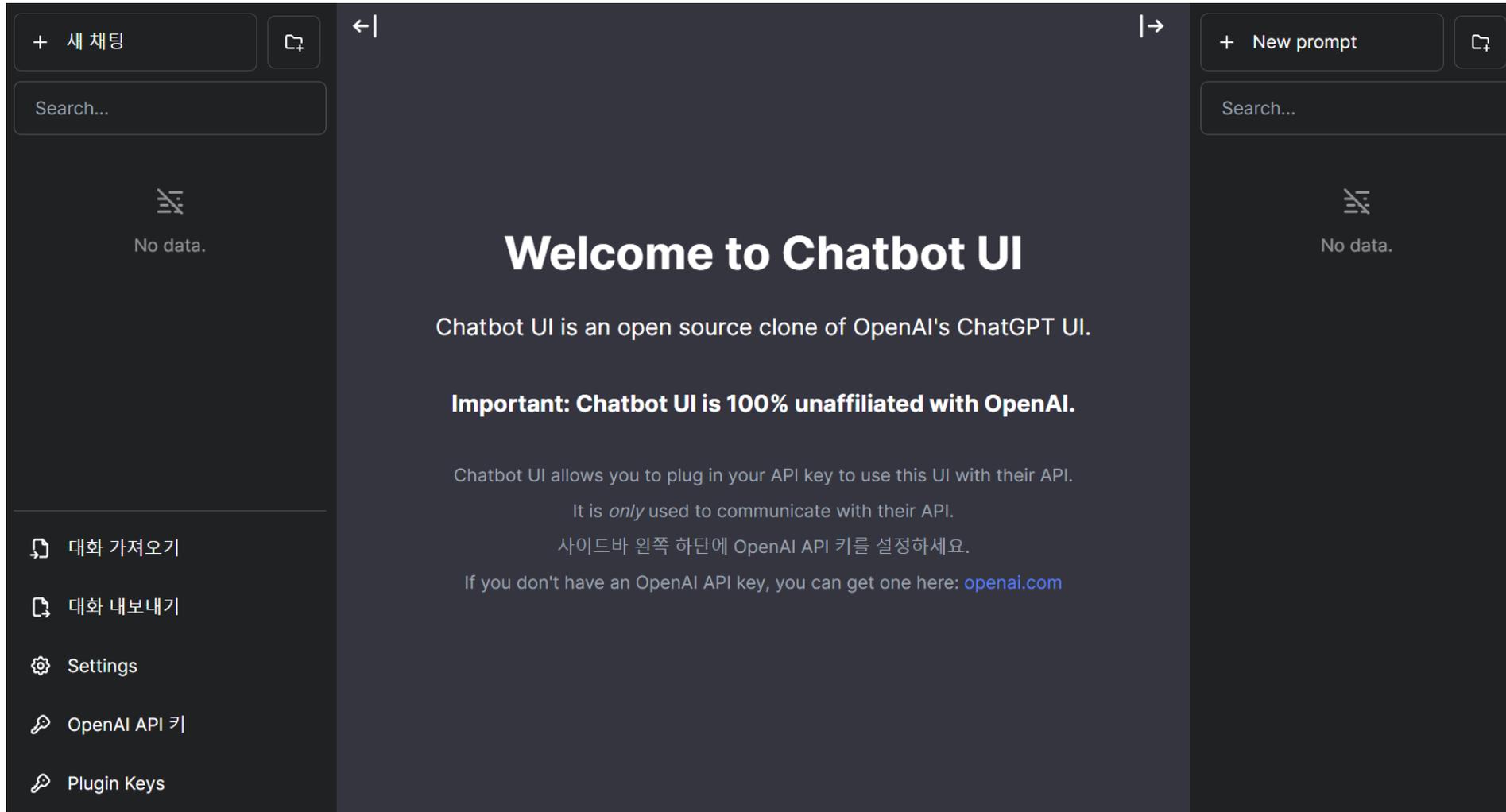
Documents

- 벡터 또는 하이브리드 검색 사용
- Azure OpenAI 임베딩을 사용하거나 자체 가져오기
- Azure와 긴밀하게 통합
- 복제 및 파티셔닝으로 확장
- 생성 AI 앱 및 검색 플러그인 빌드
- **REST API를 통해서 사용 가능**



PART 4. 외부 API와 연결

Chatbot UI

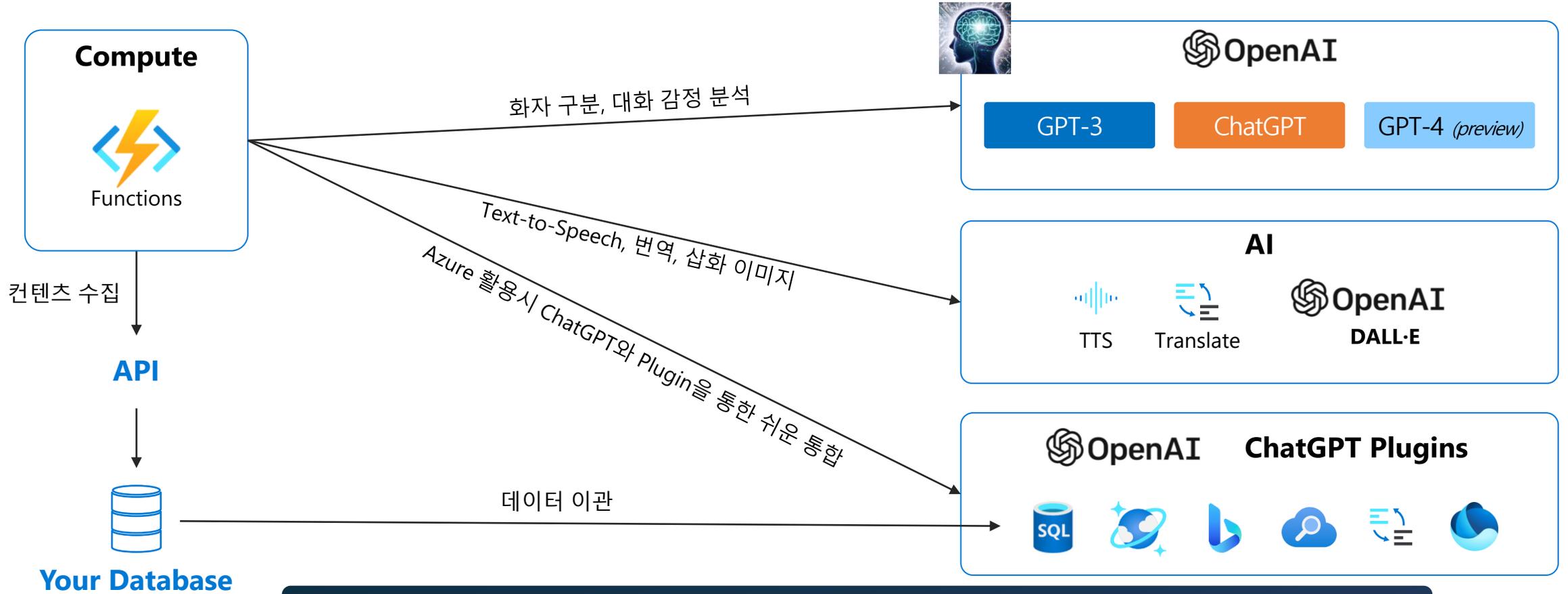


 OpenAI
ChatGPT



Google API Key

웹소설을 위한 아키텍처



Microsoft Fabric ecosystem components:

- Data Factory
- Synapse Data Engineering
- Synapse Data Science
- Synapse Data Warehouse
- Synapse Real Time Analytics
- Power BI
- Data Activator
- OneLake

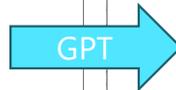
Azure OpenAI 기반으로 화자(Speaker) 구분

Parameters
 온도 0.1
 최대 길이(토큰) 171
 중지 시퀀스
 상위 확률 0.3
 빈도 패널티 0
 현재 상태 패널티 0
 사전 응답 텍스트
 사후 응답 텍스트

Select speaker: **Few-shot Learning**
 7: 선우
 9:-
 11:희주
 12:선우
 14:-
 16:-
 18:희주
 19:-

생성 실행 취소 다시 생성 **토큰: 3289**

Token Length



Parameters
 온도 0.1
 최대 길이(토큰) 171
 중지 시퀀스
 상위 확률 0.3
 빈도 패널티 0
 현재 상태 패널티 0
 사전 응답 텍스트
 사후 응답 텍스트

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:희주
 45:-
 47:희주
 49:-
 51:-
 53:-
 55:-
 57:-
 59:희주
 61:-
 63:선우
 65:선우가 길게 고개를 숙이자, 차혁이 까닥 인사를 했다.
 67: "네. 출근했습니다. 그런데..."
 69: 그의 시선이 선우에게서부터 스룩, 미끄러져 희주에게로 향했다.
 71: "오늘 중요한 회의가 있는 줄 알았습디만, 한가해 보이네요."
 73: 빈정거리는 목소리가 날카로웠다.
 75: "자료가 완벽히 준비되었나 보죠, 이 정도면."
 77: 그가 둘러 말한 질책에 희주가 입술을 잘근 물었다.

생성 실행 취소 다시 생성 **토큰: 3491**

Token Length

Next Context

Prompt Chaining

Azure OpenAI 기반으로 화자(Speaker) 구분 비교

ChatGPT (1.5B)

GPT-3 (175B)



Azure Text-to-Speech 모델 기반으로 음성

Speech Studio > 오디오 콘텐츠 만들기 > 내 파일 > GPT3-Test-001

```
<!--ID=B7267351-473F-409D-9765-754A8EBCDE05,Version=1|["VoiceNameToldMapItems":[{"Id":"44ee81ca-a1f8-4c3d-9e1e-921b1d005970","Name":"Microsoft Server Speech Text to Speech Voice (ko-KR, BongJinNeural)","ShortName":"ko-KR-BongJinNeural","Locale":"ko-KR","VoiceType":"StandardVoice"},{"Id":"474131f3-ea79-45eb-bb61-594966a48110","Name":"Microsoft Server Speech Text to Speech Voice (ko-KR, JiMinNeural)","ShortName":"ko-KR-JiMinNeural","Locale":"ko-KR","VoiceType":"StandardVoice"},{"Id":"abdda02b-9b62-411b-bc39-003fb1d9f2cc","Name":"Microsoft Server Speech Text to Speech Voice (ko-KR, SoonBokNeural)","ShortName":"ko-KR-SoonBokNeural","Locale":"ko-KR","VoiceType":"StandardVoice"},{"Id":"662c6e95-d86c-490b-928d-3836034ff42","Name":"Microsoft Server Speech Text to Speech Voice (ko-KR, InJoonNeural)","ShortName":"ko-KR-InJoonNeural","Locale":"ko-KR","VoiceType":"StandardVoice"}]-->  
<!--ID=FCB40C2B-1F9F-4C26-B1A1-CF8E67BE0D1,Version=1|["Files"]-->  
<!--ID=5B95B1CC-2C7B-494F-B746-CF22A0E779B7,Version=1|["Locales":{"ko-KR":{"AutoApplyCustomLexiconFiles":{"}}}-->  
<speak xmlns="http://www.w3.org/2001/10/synthesis" xmlns:mstts="http://www.w3.org/2001/mstts" xmlns:emo="http://www.w3.org/2009/10/emotionml" version="1.0" xml:lang="ko-KR">  
<voice name="ko-KR-BongJinNeural">회주 씨, 혹시 커피 좋아해요?</voice>  
<voice name="ko-KR-JiMinNeural">바쁜 줄이야이었다. 회주가 화면을 보면서 식은 아메리카노를 훌쩍이는데 불청객이 끼어들었다.</voice>  
<voice name="ko-KR-SoonBokNeural">네?</voice>  
<voice name="ko-KR-BongJinNeural">아침마다 커피를 마시는 것 같길래.</voice>  
<voice name="ko-KR-JiMinNeural">별쪽, 같은 팀의 최선우씨가 웃었다. 회주와 동년배의 그는 자신이 편한지 이렇게 곧잘 말을 걸었다.  
말쭙, 말쭙, 눈을 그를 보고 있지만 여전히 마우스에서 손을 떼지 않은 채 그녀는 답했다.</voice>  
<voice name="ko-KR-SoonBokNeural">좋아하는 건 아니고, 아침에 잠이 잘 안 깬는 타입이라서요.</voice>  
<voice name="ko-KR-BongJinNeural">아, 그렇구나, 좋아 마시는 게 아닌가 봐요.</voice>  
<voice name="ko-KR-SoonBokNeural">네.</voice>  
<voice name="ko-KR-JiMinNeural">회주는 가볍게 고개를 끄덕였다. 눈치도 없이 선우가 말을 이었다.</voice>  
<voice name="ko-KR-BongJinNeural">근데, 회사 앞에 되게 맛있는 샷 내려 주는 예프레소 바가 있었어요. 거기 가서 한번 먹어 보면 회주 씨도 커피 좋아할지도 몰라요.</voice>  
<voice name="ko-KR-JiMinNeural">예스프레소 바라.  
처음 듣는 뭔가 세련된 단어였다. 회주 같은 인간이 아니라 차학에게 더 어울릴 것 같은 이름. 그에게 "예스프레소 바 뭐가요?" 하고 물어보면, 차학은 눈을 가느스름하게 뜨며 "이런 것도 처음 들어 봤나?" 하고 신기하게 여기겠지. 그에게 자신은 늘 그렇게 이해 가지 않으면서도 신기한 존재였다. 차학이 회주에게 그렇듯.</voice>  
<voice name="ko-KR-BongJinNeural">회주 씨?</voice>  
<voice name="ko-KR-JiMinNeural">예고도 없이 머릿속에 차학에 대한 생각이 들어왔다. 고개를 흔들며 그의 찬상을 머릿속에서 지우며 회주가 답했다.</voice>  
<voice name="ko-KR-SoonBokNeural">최종해요. 업무 생각을 하느라, 네, 알겠습니다. 다음에 한번 가서 마셔 볼게요.</voice>  
<voice name="ko-KR-BongJinNeural">그럼 이번 주 토요일 어때요?</voice>
```

Speech Studio > 오디오 콘텐츠 만들기 > 내 파일 > 제목 없음

1. [음성] "회주 씨, 혹시 커피 좋아해요?"
2. [자민] 바쁜 줄이야이었다. 회주가 화면을 보면서 식은 아메리카노를 훌쩍이는데 불청객이 끼어들었다.
3. [순복] "네?"
4. [음성] "아침마다 커피를 마시는 것 같길래."
5. [자민] 별쪽, 같은 팀의 최선우씨가 웃었다. 회주와 동년배의 그는 자신이 편한지 이렇게 곧잘 말을 걸었다.
6. [자민] 말쭙, 말쭙, 눈을 그를 보고 있지만 여전히 마우스에서 손을 떼지 않은 채 그녀는 답했다.
7. [순복] "좋아하는 건 아니고, 아침에 잠이 잘 안 깬는 타입이라서요."
8. [음성] "아, 그렇구나, 좋아 마시는 게 아닌가 봐요."
9. [순복] "네."
10. [자민] 회주는 가볍게 고개를 끄덕였다. 눈치도 없이 선우가 말을 이었다.
11. [음성] "근데, 회사 앞에 되게 맛있는 샷 내려 주는 예프레소 바가 있었어요. 거기 가서 한번 먹어 보면 회주 씨도 커피 좋아할지도 몰라요."
12. [자민] 예스프레소 바라.
13. [자민] 처음 듣는 뭔가 세련된 단어였다. 회주 같은 인간이 아니라 차학에게 더 어울릴 것 같은 이름. 그에게 "예스프레소 바 뭐가요?" 하고 물어보면, 차학은 눈을 가느스름하게 뜨며 "이런 것도 처음 들어 봤나?" 하고 신기하게 여기겠지.
14. [자민] 그에게 자신은 늘 그렇게 이해 가지 않으면서도 신기한 존재였다. 차학이 회주에게 그렇듯.
15. [음성] "회주 씨?"
16. [자민] 예고도 없이 머릿속에 차학에 대한 생각이 들어왔다. 고개를 흔들며 그의 찬상을 머릿속에서 지우며 회주

캐릭터를 고려한 전문 성우를 이용한 녹음

Azure AI | Speech Studio

개요
액세스 요구 사항
사용자 지정 신경망 음성 ...
음성프로
한국어(대한민국)
성우 설정
교육 데이터 준비
모델 학습
모델 배포

엔진 버전 2022.05.11.0

2022.05.11.0 (기본값)
이 버전용으로 만든 음성 샘플을 확인해 보세요. 기본 버전에 대한 추가 테스트만 만들 수 있습니다.

테스트 교육 집합

Default tests | < 테스트 목록으로 돌아가기

다운로드 삭제

텍스트 입력	오디오
모두 50대 이상 중년층 고객이 가장 많이 증가한 것으로 나타 났 다.	
모두를 정말 말도 안된다 생각했지만 예상을 깨고 정말 놀라운 결 과가 나타났습니다.	
보통 통산이라고도 하고 상계처리라고도 하는데요.	
비공식 집단의 보편성이 결여되어 있다는 점이다.	
<input checked="" type="checkbox"/> 비타민의 주된 역할은 장에서 칼슘의 흡수를 높이는 것입니다.	
시 한편에 삼만원이면, 너무 박하다 싶다.	
아무 조건 없이 100% 무료로 번호 추출해 주고 있다.	
여섯째, 정권교체는 평화적 방법으로 이루어져야 한다.	

DALL-E 2를 이용한 삽화 활용



딸깍, 딸깍. 눈은 그를 보고 있지만 여전히
마우스에서 손을 떼지 않은 채 그녀는 답했다

Draw a picture in a cute and pretty webtoon style
of a woman sitting in front of a computer in the
office and looking at a man who is interested in
her asking a question while moving a mouse with
her hand.

Auto-GPT

OpenAI
GPT-4

태스크 생성
및 관리 /
정보 요약



인터넷 검색



메모리

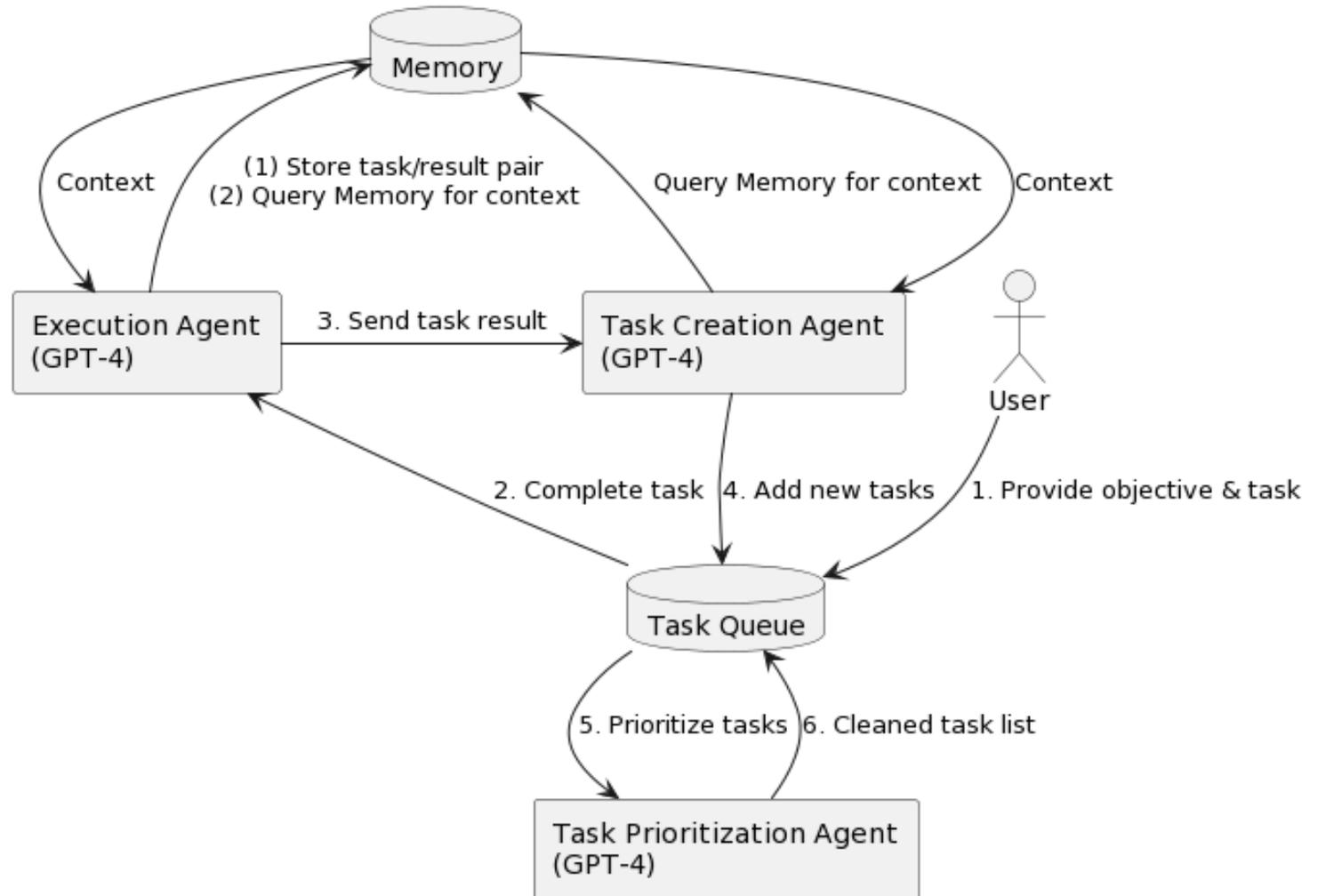
OpenAI
DALL·E



이미지 생성

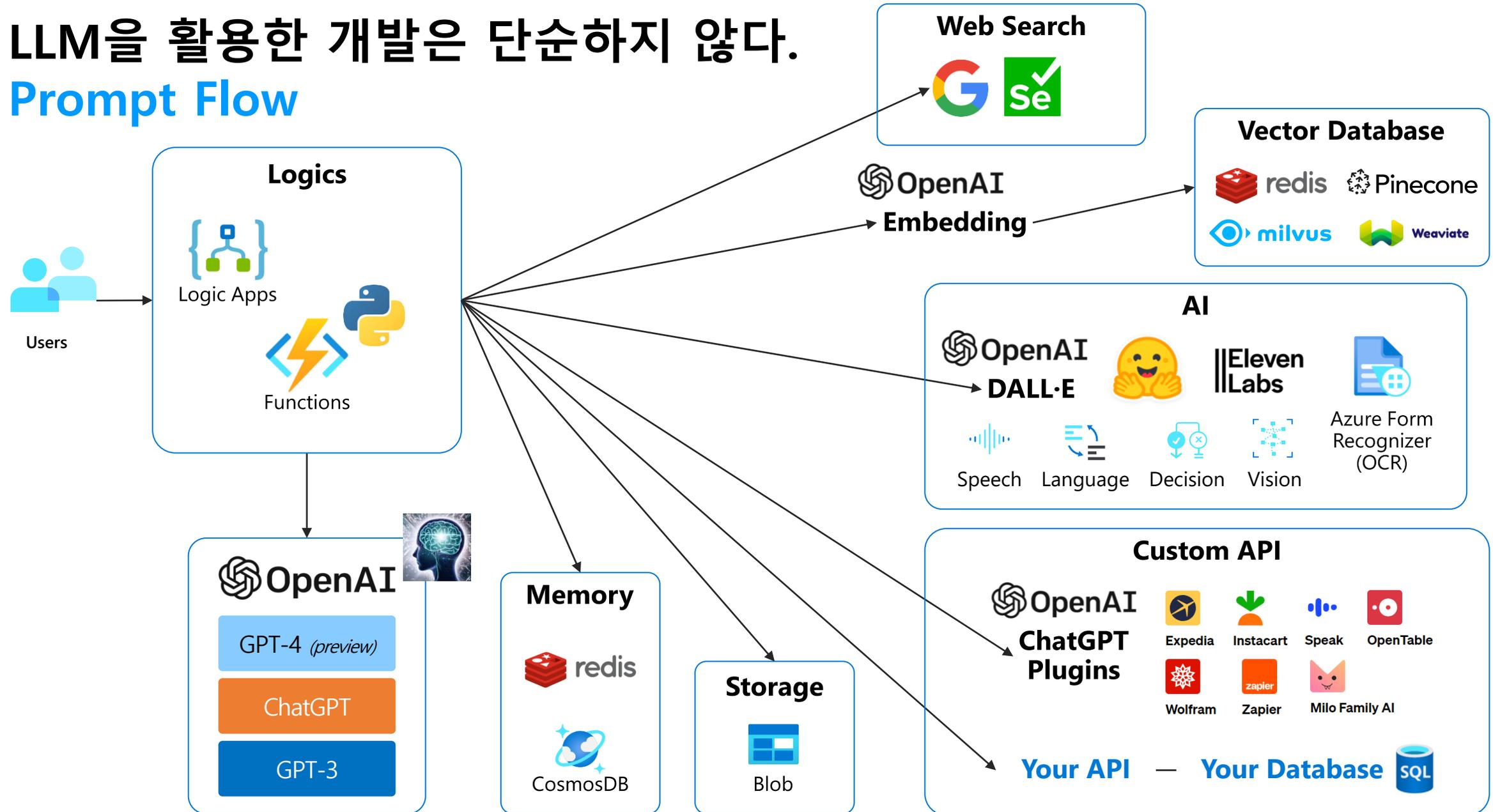
Eleven
Labs

음성 생성



LLM을 활용한 개발은 단순하지 않다.

Prompt Flow



LLMOps - Azure Machine Learning Prompt Flow!

Microsoft Azure Machine Learning Studio

Microsoft > pf-metang > Flows > Bing Grounded QA

Bing Grounded QA Runtime: Select runtime View run history View endpoint Deploy Clone Save Bulk test Run

Search more tools + LLM + Prompt + Python

offset: int, value: 0

process_search_result

Name	Type	Value
search_result	object	`\${search_on_bing.output}`

```
1 from promptflow import tool
2
3
4 @tool
5 def process_search_result(search_result):
6
7     def format(doc: dict):
8         return f"Content: {doc['Content']}\nSource: {doc['Source']}"
```

augmented_qna

Connection: Connection

deployment_name: text-davinci-001, max_tokens: 256, temperature: 0.5, stop:

Advanced

Inputs

Flow Diagram: inputs → extract_query_from_quest... → search_on_bing → process_search_result → augmented_qna → outputs



PART 5. Azure 서비스와 연결

Azure OpenAI – New features

Azure OpenAI Service on your data (Public Preview)

UPDATE (0620)

- Conversational AI(ChatGPT, GPT-4) with customer's own dataset
- Connect data easily on Azure AI Studio (**Azure Cognitive Search / Azure Blob Storage** / Upload files)

Add Your Own Data



Plugins (Preview in July)

Coming soon

- Connect Azure OpenAI with external services (Bing Search, Azure Cognitive Search, Azure SQL, Azure Cosmos DB, Microsoft Translator)

Plugins



Provisioned Throughput Model (GA, limited access)

Limited access

- Dedicated connection to OpenAI models with a guaranteed throughput, measured in tokens/sec for prompts and completions
- Suitable for the customers who want to commit large volume of usage for Azure OpenAI
- minimum consumption
 - GPT-3.5: \$78K/month (300PTU)
 - GPT-4: \$234K/month (900PTU)

나의 데이터를 쉽게 통합하여 조회하고 요약 예시

도우미 설정

시스템 메시지 **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.](#)

Limit responses to your data content

Data Source: Azure Cognitive Search
Search Resource: search-for-rag
Index: openai-prompt

Remove data source

채팅 세션

채팅 지우기 코드 보기 Show raw JSON

한글로 출력
Prompting 잘 하는 방법을 한글로 번역해서 알려줘.

Prompting은 사용자가 원하는 정보를 추출하기 위해 질문을 만드는 것입니다. 이를 위해서는 사용자가 원하는 정보와 관련된 키워드를 파악하고, 그것을 기반으로 질문을 만들어야 합니다. 또한, 질문은 명확하고 구체적이어야 하며, 사용자가 이해하기 쉬운 언어로 작성되어야 합니다. [doc1]

Text Classification 하는 방법과 예시를 알려줘.

Text Classification은 텍스트를 자동으로 분류하는 기술입니다. 이를 위해서는 먼저 학습 데이터를 수집하고, 이를 학습시켜야 합니다. 학습 데이터는 분류하고자 하는 카테고리에 대한 레이블이 달린 텍스트 데이터입니다. 모델 학습 후, 새로운 텍스트 데이터를 입력하면 모델은 해당 텍스트를 분류하여 적절한 카테고리 분류를 제공합니다. 예를 들어, 스팸 메일 필터링, 감성 분석, 토픽 모델링 등에 활용됩니다. [doc1]

Classification을 위한 예시로는, 자연어 처리 라이브러리인 NLTK를 사용하여 영화 리뷰를 긍정적인 리뷰와 부정적인 리뷰로 분류하는 프로젝트가 있습니다. 이를 위해 먼저 영화 리뷰 데이터셋을 수집하고, 이를 긍정적인 리뷰와 부정적인 리뷰로 레이블링합니다. 그 후, NLTK를 사용하여 모델을 학습시키고, 새로운 영화 리뷰를 입력하면 해당 리뷰가 긍정적인 리뷰인지 부정적인 리뷰인지 분류할 수 있습니다. [doc2]

참고문헌:
• [doc1]
• [doc2]

Configuration

배포 매개 변수

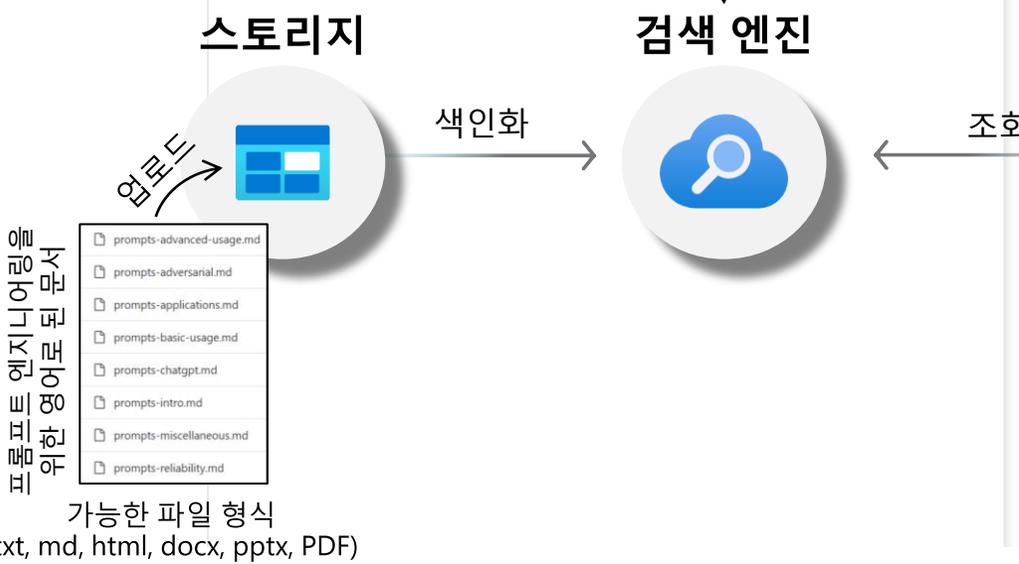
배포 * chatgpt35

세션 설정

지난 메시지 포함 10

현재 토론 수

Input tokens progress indicator
1422/4000



Azure OpenAI on your data to Power Virtual Agents

The screenshot displays the Azure AI Studio interface for a ChatGPT playground. The main window is titled "ChatGPT playground (preview)" and includes a "Save setup" button. On the left, a "Chat setup" panel is highlighted with a red box, showing options for "System message" and "Add your data (preview)". It includes a checkbox for "Limit responses to your data content" and fields for "Data source" (Azure Cognitive Search) and "Search resource" (MyCogSearchResource). The "Index" field is set to "MyCustomerSearchIndex". A "Remove data source" button is also present. The "Chat session" panel in the center shows a "Clear session" button and a "View sample code" link. It displays a system message and a list of steps for creating a support ticket. Below this, a "Source data added" indicator is shown, followed by a chat input field with the question "How do I create a support ticket?". The response area shows the AI's answer, a "Used: Azure OpenAI on your data" status, and two references to "spd_corporate_2023.txt". On the right, a "Configuration" panel is partially visible, with a "Deploy to ..." dropdown menu highlighted by a red box. The dropdown menu shows options for "As a web app..." and "to Power Virtual Agent...". The "Deployment" section shows "GPT-4" selected. At the bottom right, a chat window shows the question "How do I create a support ticket?" and the AI's response, which is identical to the one in the main chat session.

Public preview

Azure Cognitive Search – Vector search

Revolutionizing Indexing and Retrieval for LLM-powered Apps

Power your retrieval-augmented generation applications



Images



Audio



Video



Graphs



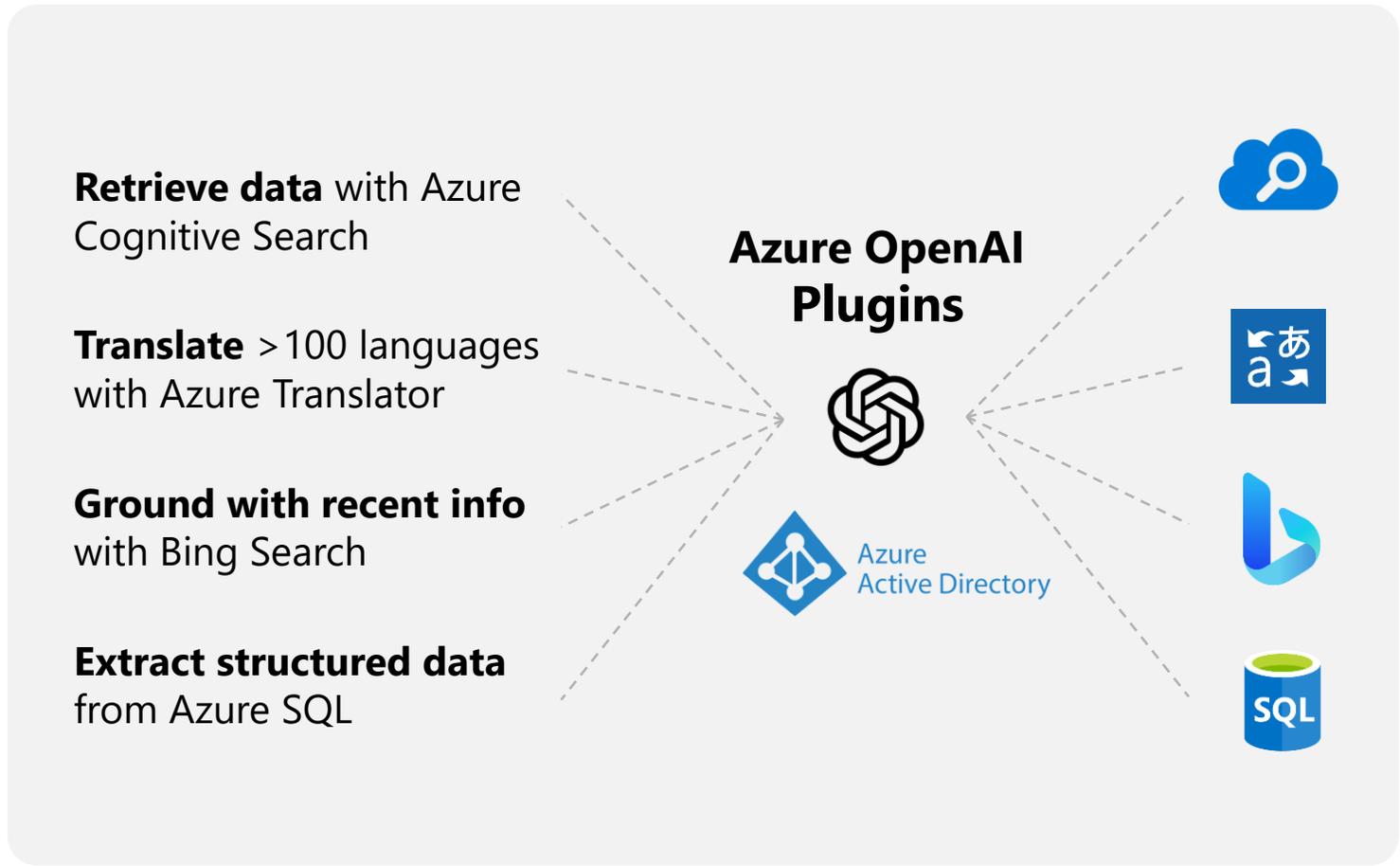
Documents

- 벡터 또는 하이브리드 검색 사용
- Azure OpenAI 임베딩을 사용하거나 자체 가져오기
- Azure와 긴밀하게 통합
- 복제 및 파티셔닝으로 확장
- 생성 AI 앱 및 검색 플러그인 빌드
- **REST API를 통해서 사용 가능**

소개:

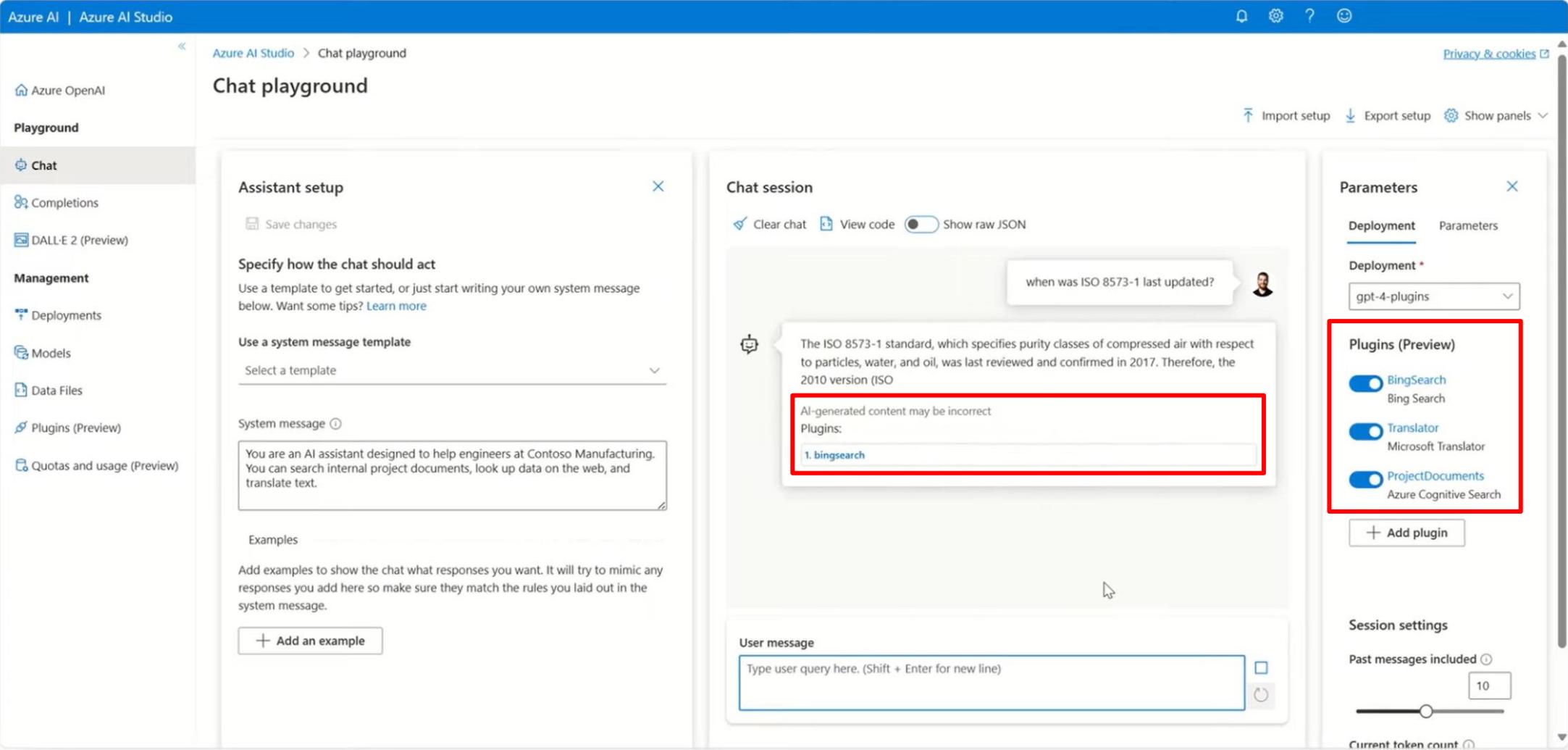
Azure OpenAI Service **Plugins** (coming soon)

Microsoft 서비스에 안전하게 액세스하여 강력한 AI Copilot 구축



- 다양한 데이터 저장소, 벡터 데이터베이스 및 웹에서 데이터에 안전하게 액세스
- Azure AD 및 관리 ID를 통해 제어되는 데이터 경로 액세스
- 활성화할 플러그인을 선택하는 관리자 역할

Plugins for Azure OpenAI Service



Provisioned Throughput 이란?



대량 또는 대기 시간에 민감한 워크로드를 실행하기 위해
모델 처리 용량을 예약



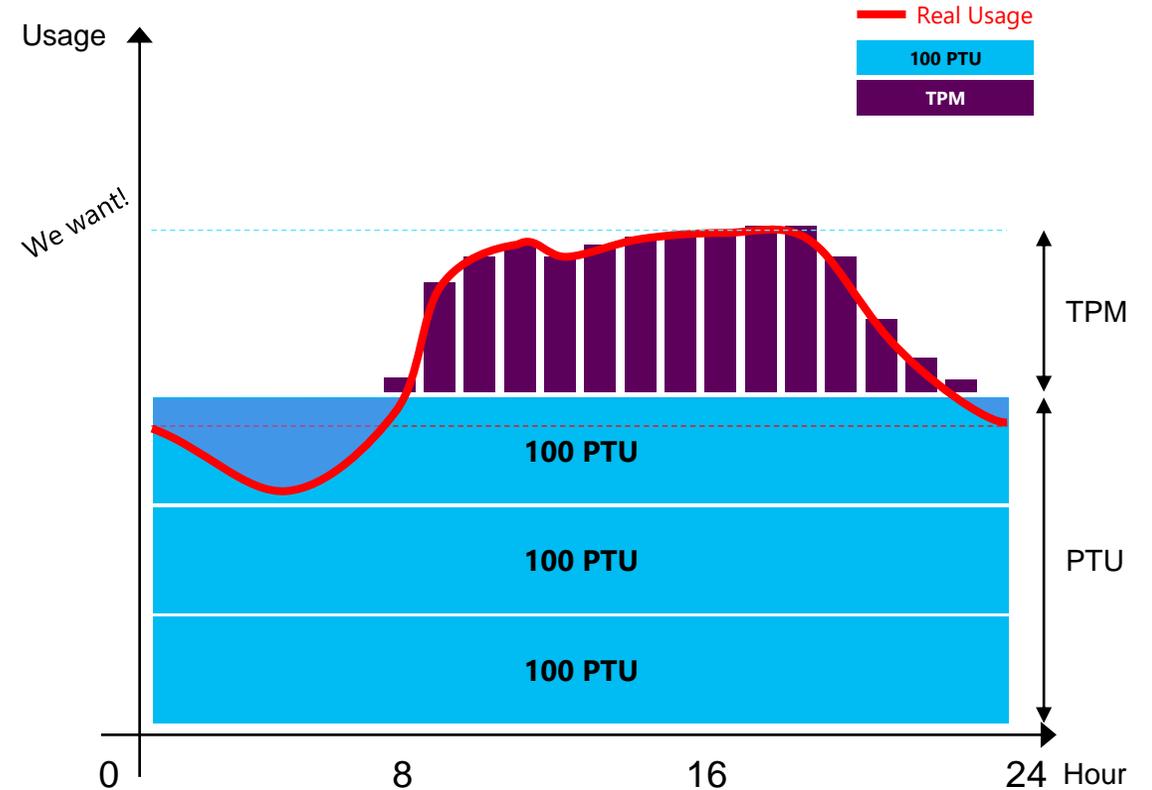
처리 용량은 1개월 또는 1년의 기간 약정으로 구입한
"Provisioned Throughput Units(PTU)"라는 단위로 정의



구매한 고객은 약정 기간 동안 PTU를 사용하여
gpt-35-turbo 또는 **GPT-4** 모델의 프로비저닝된 배포 생성



예약된 처리 용량은 프롬프트 크기, 완료 크기 및
동시 API 요청 수와 같은 일관된 특성을 가진 워크로드 적합



참고할 수 있는 자료 모음

자료 모음 Repository

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

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**](#)
5. [Introducing Azure **AI Content Safety**: Helping Organizations to Maintain Safe Online Spaces](#)

OpenAI GPT 서비스의 확장

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

Microsoft Build

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



Thank you