

Azure OpenAI

인공지능 기술과 빅테크 전쟁



김영욱

Hello AI

Microsoft AI MVP

Microsoft Regional Director

youngwook@outlook.com



Hello AI

Azure OpenAI Service에 액세스

Azure OpenAI Service 리소스를 만들 때 구독 이름, 리소스 그룹 이름, 지역, 고유한 instance 이름을 제공하고 가격 책정 계층을 선택해야 합니다.

Home > Cognitive Services | Azure OpenAI >

Create Azure OpenAI

1 Basics 2 Tags 3 Review + submit

Enable new business solutions with OpenAI's language generation capabilities powered by GPT-3 models. These models have been pretrained with trillions of words and can easily adapt to your scenario with a few short examples provided at inference. Apply them to numerous scenarios, from summarization to content and code generation.

[Learn more](#)

Project Details

Subscription * ⓘ

Resource group * ⓘ [Create new](#)

Instance Details

Region ⓘ

Name * ⓘ ✓

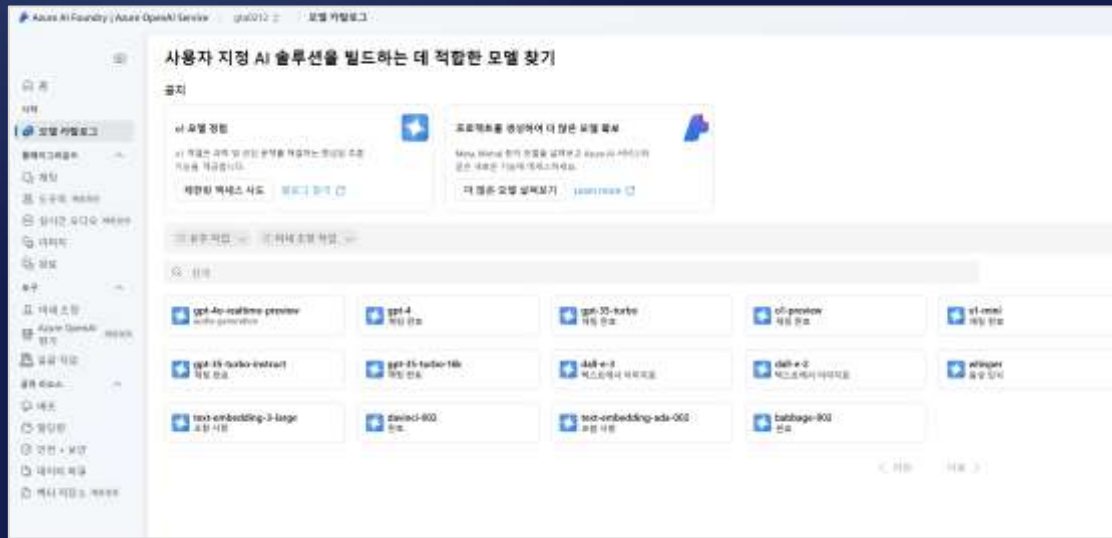
Pricing tier * ⓘ

지역별 가용성

Azure OpenAI Service는 다양한 유형의 모델에 대한 액세스를 제공합니다. 특정 모델은 일부 지역에서만 사용할 수 있습니다. 지역 가용성에 대한 Azure OpenAI 모델 가용성 가이드를 참조하세요. 지역당 두 개의 Azure OpenAI 리소스를 만들 수 있습니다.

Azure AI Foundry

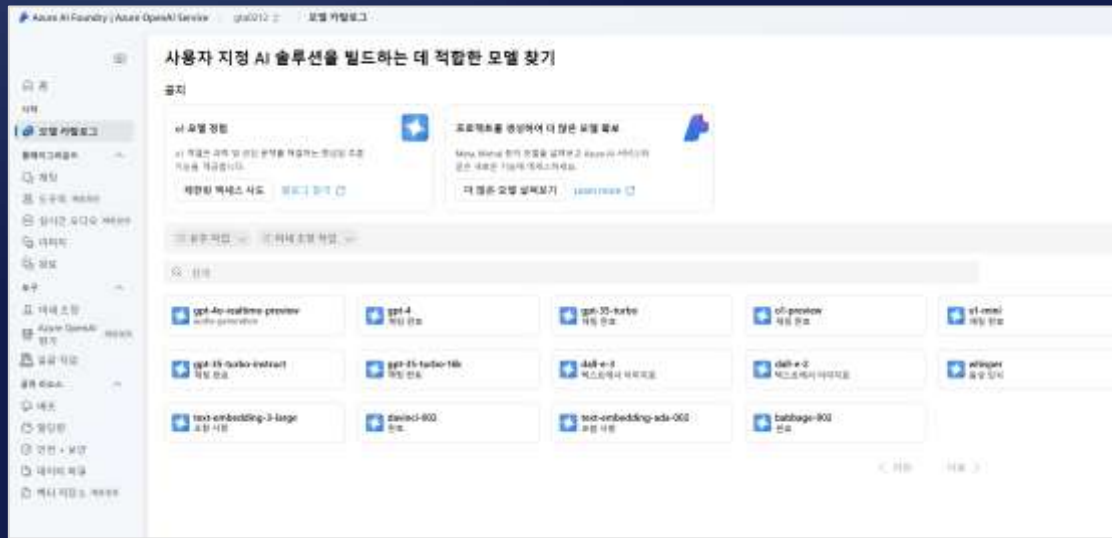
Azure OpenAI Service 리소스를 만들 때 구독 이름, 리소스 그룹 이름, 지역, 고유한 instance 이름을 제공하고 가격 책정 계층을 선택해야 합니다.



- 모델 및 미세 조정 기능의 전체 카탈로그를 사용한 프로 코드 개발
- 클라우드 인프라를 완전히 제어하는 PaaS 서비스
- 프롬프트 및 모델 오케스트레이션
- 성능, 안정성, 확장성, 책임 있는 AI 안전을 테스트하는 평가 엔진
- 사용자 지정 앱 및 서비스에서 사용하기 위해 Azure에 엔드포인트로 배포

Azure AI Foundry

Azure OpenAI Service 리소스를 만들 때 구독 이름, 리소스 그룹 이름, 지역, 고유한 instance 이름을 제공하고 가격 책정 계층을 선택해야 합니다.



- 모델 및 미세 조정 기능의 전체 카탈로그를 사용한 프로 코드 개발
- 클라우드 인프라를 완전히 제어하는 PaaS 서비스
- 프롬프트 및 모델 오케스트레이션
- 성능, 안정성, 확장성, 책임 있는 AI 안전을 테스트하는 평가 엔진
- 사용자 지정 앱 및 서비스에서 사용하기 위해 Azure에 엔드포인트로 배포

모델 할당량(Quotas) 확인

Quotas

View your quota by subscription and region and track usage across your deployments. Quota is required to create deployments and allows you to flexibly size them according to your traffic needs.

[Learn more](#)

Subscription *

Visual Studio Enterprise... ▾

Region *

SOUTHCENTRALUS ▾

PTU Available for reservation: 300

Standard

Provisioned

Global-Standard

Other

Provisioned throughput has been improved. [Learn about the new features](#)

[Discount hourly/reservation deployments](#) [Manage Commitments](#) [Capacity calculator](#) [Filter by resource](#) [Refresh](#)

Quota name ▾

Deployment

Model name

Model version

Payment model

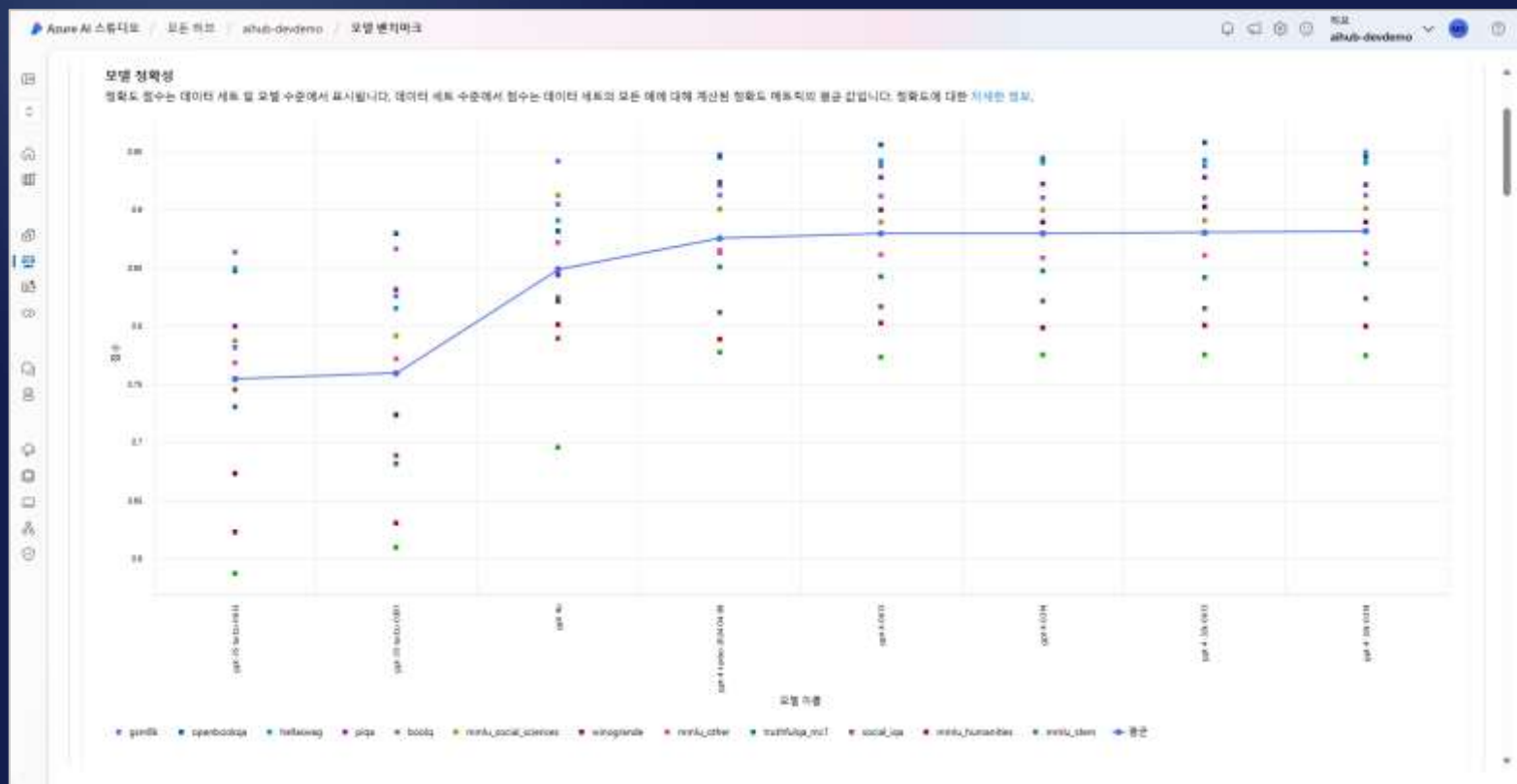
Resource type

Usage/Limit

> Provisioned Managed Throughput Unit

300 of 4900 6%

간 모델 별로 품질을 평가할 수 있는 지표들을 제공한다.



모델 배포

1. Azure AI Foundry 포털에 로그인합니다.
2. 프로비전된 배포에 사용하도록 설정된 구독을 선택하고 할당량이 있는 지역에서 원하는 리소스를 선택합니다.
3. 왼쪽 탐색 메뉴의 관리에서 배포를 선택합니다.
4. 새 배포 만들기를 선택하고 다음 필드를 구성합니다. 고급 옵션 드롭다운 메뉴를 확장합니다.
5. 각 필드에 값을 작성합니다. 예를 들면 다음과 같습니다.

Deploy model

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

ⓘ Selected model version does not have a standard deployment type.

Select a model ⓘ
gpt-4

Model version ⓘ
0613 (Default)

Deployment name ⓘ
gpt-4

ⓘ Advanced options ▾

Content Filter ⓘ
Default

Deployment type ⓘ
Provisioned-Managed

ⓘ 4300 Provisioned Throughput Units available for deployment.

Provisioned throughput units (PTU) ⓘ
100

Create Cancel

모델 배포

항목	설명	예시
모델 선택	배포하려는 특정 모델을 선택합니다.	GPT-4
모델 버전	배포할 모델 버전을 선택합니다.	0613
배포 이름	배포 이름은 코드에서 클라이언트 라이브러리 및 REST API를 사용하여 모델을 호출하는 데 사용됩니다.	gpt-4
콘텐츠 필터	배포에 적용할 필터링 정책을 지정합니다. <u>콘텐츠 필터링</u> 방법에 대해 자세히 알아봅니다.	기본값
배포 유형	이는 처리량과 성능에 영향을 미칩니다. 배포에 대한 배포 대화 상자 드롭다운에서 전역 프로비전된 관리형, DataZone 프로비전 관리 또는 프로비저닝된 관리형 선택	프로비전-관리
프로비전된 처리량 단위	배포에 포함할 처리량을 선택합니다.	100

실습: Azure OpenAI 호출



```
import os from openai import AzureOpenAI client =  
AzureOpenAI( azure_endpoint =  
os.getenv("AZURE_OPENAI_ENDPOINT"),  
api_key=os.getenv("AZURE_OPENAI_API_KEY"), api_version="2024-10-  
21" ) response = client.chat.completions.create( model="gpt-4", #  
model = "deployment_name". messages=[ {"role": "system", "content":  
"You are a helpful assistant."}, {"role": "user", "content": "Does Azure  
OpenAI support customer managed keys?"}, {"role": "assistant",  
"content": "Yes, customer managed keys are supported by Azure  
OpenAI."}, {"role": "user", "content": "Do other Azure services support  
this too?"}] ) print(response.choices[0].message.content)
```

Azure AI Foundry



Copilot Studio



Visual Studio



GitHub



Azure AI
Foundry SDK



Model Catalog

Foundational models

Open-source models

Task models

Industry models



Azure
OpenAI Service



Azure
AI Search



Azure AI
Agent Service



Azure AI
Content Safety



Azure Machine
Learning

Evaluations

Customization

Governance

Monitoring

Observability

Now 1800 + Frontier, task, and Open Models



OpenAI
Model Family
(available day 1)



Phi SLM
Model Family



Mistral AI
Model Family



Meta Llama 2
Model Family



Jais G42
Model Family



Cohere
Model Family



Databricks
Model Family

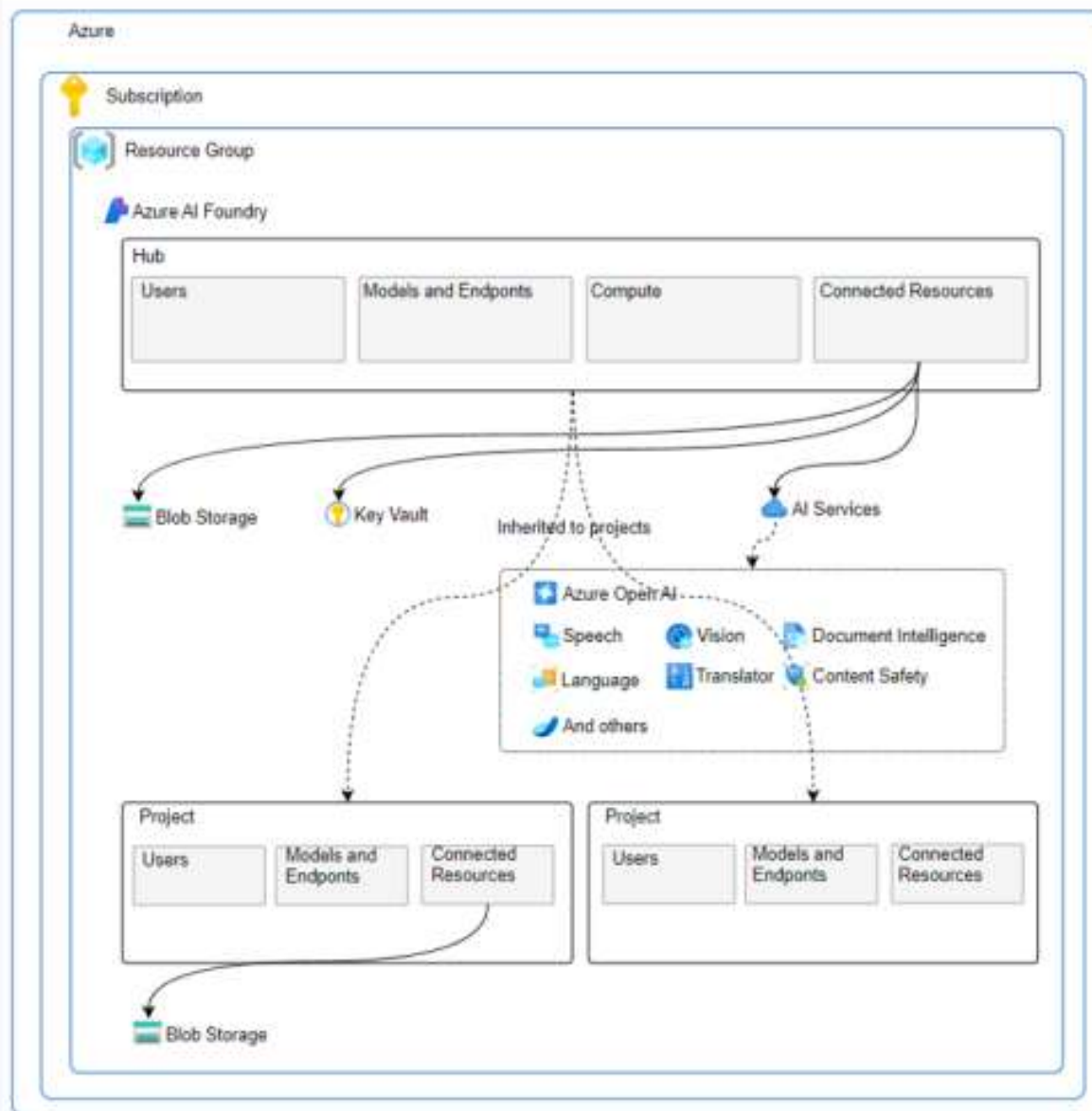


Hugging Face
Collection

- 통합 개발 환경: Azure AI Foundry는 웹 포털, SDK, CLI를 통해 AI 모델을 빌드, 평가, 배포할 수 있는 통합 환경을 제공합니다.
- 다양한 AI 모델 지원: Microsoft, OpenAI, Hugging Face, Meta, Mistral 등 다양한 공급업체의 프런티어 및 오픈 소스 모델을 포함하는 광범위한 모델 카탈로그를 제공합니다.
- 에이전트 관리 및 오케스트레이션: 다양한 AI 에이전트를 통합 관리할 수 있는 도구를 제공하여 반복적인 작업을 자동화하고 사용자 생산성을 향상시킵니다.
- 개발자 친화적인 SDK: AI 애플리케이션과 에이전트를 맞춤화, 테스트, 배포할 수 있는 강력한 기능을 제공하는 SDK를 제공합니다.

Azure AI Foundry Hubs and Projects

Walkthrough of my implementation of a hub and projects



실습: Azure Foundry 실습



```
import os from openai import AzureOpenAI client =  
AzureOpenAI( azure_endpoint =  
os.getenv("AZURE_OPENAI_ENDPOINT"),  
api_key=os.getenv("AZURE_OPENAI_API_KEY"), api_version="2024-10-  
21" ) response = client.chat.completions.create( model="gpt-4", #  
model = "deployment_name". messages=[ {"role": "system", "content":  
"You are a helpful assistant."}, {"role": "user", "content": "Does Azure  
OpenAI support customer managed keys?"}, {"role": "assistant",  
"content": "Yes, customer managed keys are supported by Azure  
OpenAI."}, {"role": "user", "content": "Do other Azure services support  
this too?"}] ) print(response.choices[0].message.content)
```




IT만담러가 푸는 이야기 영욱스튜디오



영욱 스튜디오 (YOUNGWOOK Studio)

@youngwook 구독자 9,32만명 동영상 350개

IT와 관련된 이야기를 쉽게 풀어드리는 IT만담러의 방송입니다.

채널 맞춤설정

동영상 관리

홈

동영상

라이브

채널특목

커뮤니티

채널

검색

🔍



환산: 올해 출원 분기전 보고하세요~

조회수 2,387회 · 5개월 전

IT만담러의 IT 이야기 ▶ 모두 재생

IT만담러와 함께하는 인기되고 재밌는 IT 이야기



Microsoft 365 Copilot, 내
동료가 되라! (MS 365 Copilot...)

영욱 스튜디오 (YOUNGWOOK Stud...
조회수 8,821회 · 2주 전



이제 공공한 것 치환에 물어와!
ChatGPT에 대해 알아보자

영욱 스튜디오 (YOUNGWOOK Stud...
조회수 8,971회 · 2개월 전



IT Trend 2023 | 이 기술, 이 사람
들을 주목하자!

영욱 스튜디오 (YOUNGWOOK Stud...
조회수 1,831회 · 2개월 전



메타버스 어디로 가는가?
영욱 스튜디오 (YOUNGWOOK Stud...

조회수 2,637회 · 3개월 전



IT-12세대가 이 가계의 찾아? 고성
의 열악시작2 의상

영욱 스튜디오 (YOUNGWOOK Stud...
조회수 6,821회 · 5개월 전



영욱닷컴 리브(리브) | 책을 이쁘게/
대술리/무명무

영욱 스튜디오 (YOUNGWOOK Stud...
조회수 1,131회 · 6개월 전

