

table of contents

- Part 1: Intro
- Part 3: Usage Fees (Possibility)
- Part 3: Application for use
- Part 4: Flow of Usage Fees
- Part 5: How to use representative services
 - Add an owner
 - Get started with OpenAI's services
 - Let's create a virtual machine
 - Use a notebook

Get started with OpenAI's services

This slide is created for the April 18, 2025 UTokyo Azure Usage Briefing.

Azure specifications change frequently, so if the following does not work, please refer to the Azure OpenAI Service documentation.

Purpose

- Two Options for Using Pretrained Models
 - **Azure OpenAI Service:** For immediate access to OpenAI models
 - **Azure AI Foundry:** For access to various models including OpenAI

• Roadmap for Azure OpenAI Service

1. Create a resource group
2. Select a model from the model catalog (e.g., choose “gpt-4o”)
3. Run chat sessions using the “Chat Playground”
4. Use the service via Web API from your code

Additional Topics

- 5.1 Pricing
- 5.2 Token limits per minute
- 5.3 Data, privacy, and security

• Using Other Pretrained Models (via Azure AI Foundry)

- Example: DeepSeek R1

1. Create a resource group

UTokyo Azure | utelecon

Home - Microsoft Azure

portal.azure.com/#home

本人確認を行ってください

Microsoft Azure

open AI

Copilot

(1) Open Azure Portal
<https://portal.azure.com/#home>

(2) Search with "OpenAI"

(3) Click "Azure OpenAI"

Create a resource

Azure OpenAI

SQL databases

More services

Resources

Recent

Favorite

Name

Services

Azure OpenAI

Factory Operations

Layer 2 Isolation Domains

Layer 3 Isolation Domains (Operator Nexus)

Marketplace

ChatBot with OPENAI-CHATGPT in Azure

Blip

Think Open AI Offer

Continue searching in Microsoft Entra ID

See more

Give feedback

App Services

Storage accounts

al nes

Last Viewed

Searching all subscriptions.

Azure AI services | Azure OpenAI

Azure AI services

Search Create Manage deleted resources Manage view Refresh Export to CSV Open query Assign tags Delete

Overview All Azure AI services Azure AI services Azure AI services Azure OpenAI AI Search Computer vision Face API Custom vision Speech service Language service Translator Document intelligence Bot services

Filter for any field... Subscription equals all Type equals all Resource group equals all Add filter More (1) No grouping List view

Showing 0 to 0 of 0 records.

Name ↑↓	Kind ↑↓	Location ↑↓	Custom Domain ... ↑↓	Pricing tier ↑↓	Status ↑↓
---------	---------	-------------	----------------------	-----------------	-----------

No Azure OpenAI to display

Perform a wide variety of natural language tasks.

Create Azure OpenAI Learn more ↗ Give feedback

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI ...

1 Basics 2 Network 3 Tags 4 Review + submit

Azure OpenAI Service provides access to OpenAI's powerful language models, including all the latest OpenAI models.

These models can be easily adapted to your specific tasks, including but not limited to:

summarization, image understanding, semantic search, and natural language translation.

Call Centers, Virtual Assistants, Accessibility, and more.

Assistants API, Fine Tuning capabilities and more.

The service can be scaled through Azure Functions and more.

Learn more

Project Details

Subscription * ⓘ

Resource group * ⓘ

A resource group is a container for all the resources required for an Azure solution.

Name *

 ✓

OK Cancel

Create new

You can choose any name for the resource group.

i Azure AI services resource creation requires subscription registration, we detected that your selected subscription did not register Cognitive services resource type before, we will help you to register Cognitive services resource type when you select a subscription in subscription dropdown. Click to learn more how to check registration state for your selected

Previous

Next

Give feedback

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI

experiences. The service can be scaled through Standard (tokens) an

[Learn more](#)

Project Details

Subscription * ⓘ

1923306880-YamakataTraining

Resource group * ⓘ

(New) openai_test1

[Create new](#)

Instance Details

Region ⓘ

Japan East

Name * ⓘ

gprt-4o-test-1

Pricing tier * ⓘ

Standard S0

[View full pricing details](#)

Not only the compute resources for running the model, but also various other resources such as public IP addresses and disks are associated with the model. **If you want to completely stop billing, the easiest way is to delete the entire resource group.**

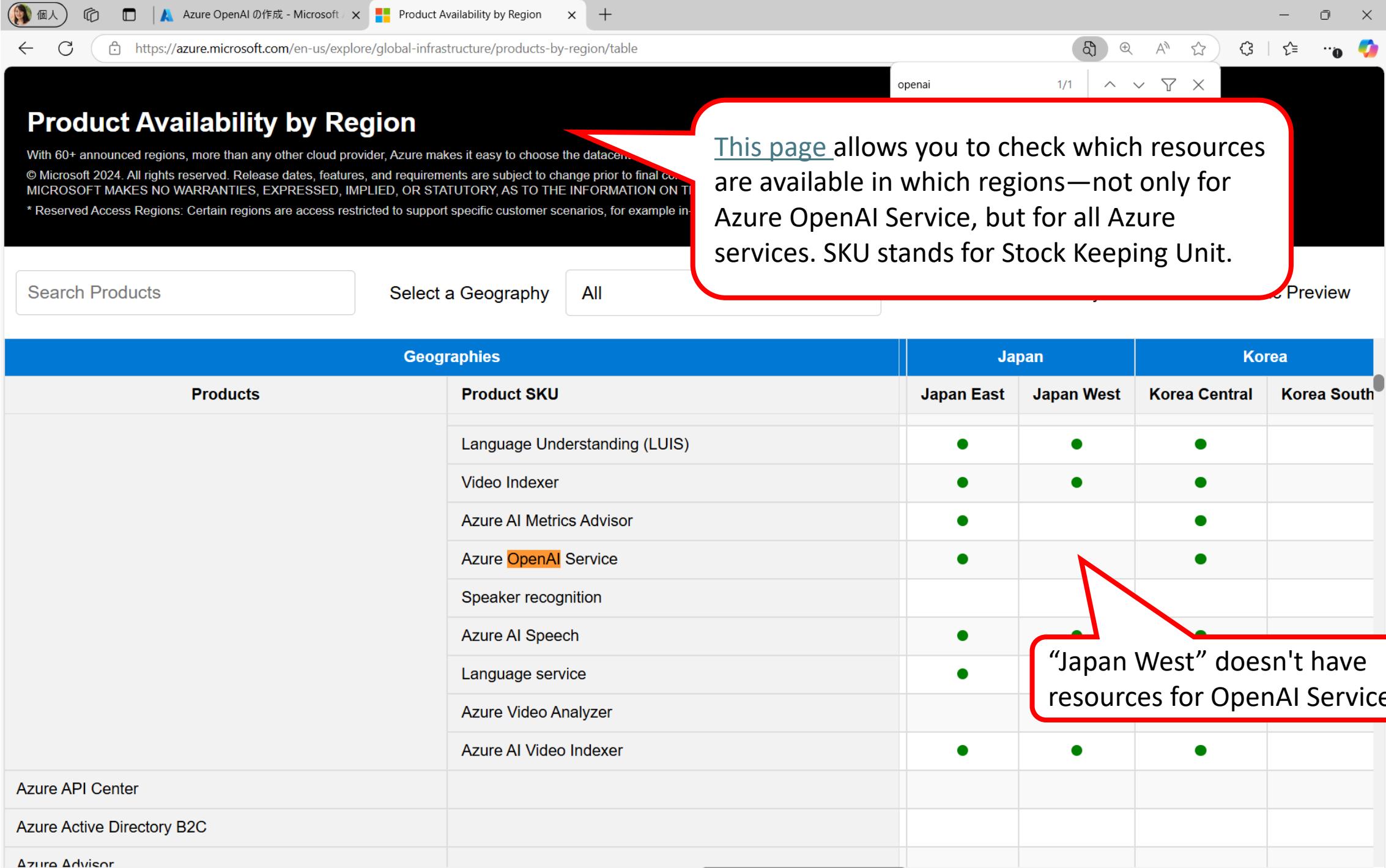
In other words, it's recommended to organize your resources into separate resource groups that can be safely deleted when they are no longer needed.

You can only select the region where the resources for that service are located. (For example, you can't choose Japan West.)

[Previous](#)

[Next](#)

[Give feedback](#)



Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI

A green circular icon containing a white checkmark symbol, indicating a completed or correct section.

Basics

2 Network

3 Tag

④ Review + submit

 Configure network security for your Azure AI services resource

Type *

- All networks, including the internet, can access this resource.
 - Selected networks, configure network security for your Azure AI services resource.
 - Disabled, no networks can access this resource. You could configure private endpoint connections that will be the exclusive way to access this resource.

[Previous](#)

Next

 Give feedback

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI

...

 Basics  Network  Tags  Review + submit

Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. [Learn more about tags](#)

Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.

Name 	Value 	Resource
<input type="text"/>	:	<input type="text"/> Azure AI services

Previous

Next

 Give feedback

UTokyo Azure | utelecon x Create Azure OpenAI - Microso x +

portal.azure.com/#create/Microsoft.CognitiveServicesOpenAI

Microsoft Azure Search resources, services, and docs (G+) Copilot 1 🔍 ⚙️ 🌐 🌐 2573969490@utac.u-to... THE UNIVERSITY OF TOKYO (UNI... 🌐

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI

...

View automation template

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the [Azure Marketplace Terms](#) for additional details.

Basics

Subscription	1923306880-YamakataTraining
Resource group	openai_test1
Region	Japan East
Name	gppt-4o-test-1
Pricing tier	Standard S0

Previous Next Create Give feedback



Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview

Deployment



Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name : Microsoft.CognitiveService... Start time : [REDACTED]
Subscription : 1923306880-YamakataTrai... Correlation ID : 02dab288-1f64-4c47-9722...
Resource group : openai_test1

> Deployment details

▽ Next steps

Go to resource

Give feedback

Tell us about your experience with deployment



Cost management

Get notified to stay within your budget and prevent unexpected charges on your bill.

[Set up cost alerts >](#)



Microsoft Defender for Cloud

Secure your apps and infrastructure

[Go to Microsoft Defender for Cloud >](#)

Free Microsoft tutorials

[Start learning today >](#)

Work with an expert

Azure experts are service provider partners who can help manage your assets on Azure

Home > Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview >

openai_test1

Resource group



Create Manage view



Delete resource group



Refresh



Export to CSV



Open query



Assign tags



Essentials

JSON View

Resources

Recommendations

Filter for any field...

Type equals all

Location equals all

Add filter

Showing 1 to 1 of 1 records.

Show hidden types

No grouping

Name ↑↓

Type ↑↓

gppt-4o-test-1

Azure OpenAI

Japan East

...

Switch between a list view of your resources and a summary chart view of resource counts.

CLOSE

Overview

Activity log

Access control (IAM)

Tags

Resource visualizer

Events

Settings

Cost Management

Monitoring

Automation

Help

 gprt-4o-test-1   ...



Search Go to Azure AI Foundry portal Delete

Overview

JSON View

-  Activity log
-  Access control (IAM)
-  Tags
-  Diagnose and solve problems
-  Resource visualizer
- > Resource Management
- > Security
- > Monitoring
- > Automation
- > Help

Resource group ([move](#))
[openai_test1](#)

API Kind
OpenAI

Status
Active

Pricing tier
Standard

Location
Japan East

Endpoints
[Click here to view endpoints](#)

Subscription ([move](#))
[1923306880-YamakataTraining](#)

Manage keys
[Click here to manage keys](#)

Subscription ID
a600356d-63a3-483e-b4bb-59629c01a212

Tags ([edit](#))
[Add tags](#)

Get Started Develop Monitor

Build your own secure copilot and generative AI applications with Azure OpenAI Service

Deploy an Azure OpenAI model and start making API calls. Connect your own data, call functions, and improve workflow with Azure OpenAI language, image and speech models. You can access the service through

UTokyo Azure | utelecon x | A gppt-4o-test-1 - Microsoft Azur x Chat playground - Azure Open/ +

ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/... ☆ 本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service / gppt-4o-test-1 / Chat playground All resources 5 🔍 ⚙️ 😊 gppt-4o-test-1 (japaneast, S0) 8

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment

No deployment exists

Parameters

Chat history

Start with a sample prompt

Marketing slogan
Create a catchy marketing slogan for a new eco-friendly product.

Character development
Describe a day in the life of a superhero who has lost their powers.

Poetry generation
Compose a poem about the beauty of nature in autumn.

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

Chat playground help

Relevant resources Chat Quickstart

2. Select a model
from the model catalog

UTokyo Azure | utelecon | gppt-4o-test-1 - Microsoft Azur | Chat playground - Azure Open/ +

ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/... ☆ 本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service / gppt-4o-test-1 / Chat playground All resources 5 🔍 ⚙️ 😊 gppt-4o-test-1 (japaneast, S0) 8

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment No deployment exists Hide

Parameters

Chat history

Start with a sample prompt

Marketing slogan Create a catchy marketing slogan for a new eco-friendly product.

Character development Describe a day in the life of a superhero who has lost their powers.

Poetry generation Compose a poem about the beauty of nature in autumn.

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

Microphone icon Share icon

Chat playground help

Relevant resources Chat Quickstart

Model catalog

Home Get started Playgrounds Chat Assistants PREVIEW Language PREVIEW Real-time audio PREVIEW Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions PREVIEW Batch jobs Metrics Shared resources Deployments

UTokyo Azure | utelecon | Feedback - Microsoft Azure | Model catalog - Azure OpenAI | +

ai.azure.com/resource/models?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/gpr... 本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service / gpert-4o-test-1 / Model catalog All resources 5 gear smiley gpert-4o-test-1 (japaneast, S0) 8

Find the right model to build your custom AI solution

Announcements

Responses API with CUA Create your agent with computer-use-preview  

Introducing GPT-4.5 Preview The latest GPT model that excels at diverse text and image tasks  

4o Mini Audio is here! Advanced audio at a fraction of the cost  

< ● ● ● >

Inference tasks ▾ Fine-tuning tasks ▾

Search Models 27

 gpt-4.5-preview Chat completion	 o3-mini Chat completion	 computer-use-preview Responses
 gpert-4o-mini-audio-preview Audio generation	 gpert-4o-mini-realtime-preview Audio generation	 o1 Chat completion
 o1-mini Chat completion	 gpert-4o Chat completion	 gpert-4o-mini Chat completion
 gpert-4o-audio-preview	 gpert-4o-realtime-preview	 o1-preview

Models help

How to use the model catalog

Search by name, filter, or browse to find the right model for your use case. Click to see model details and specifications from the publisher and to deploy a model.

Relevant resources

Available models as of February 2025 [Azure documentation]

Models	Description
computer-use-preview	An experimental model trained for use with the Responses API computer use tool.
GPT-4.5 Preview	The latest GPT model that excels at diverse text and image tasks.
o-series models	Reasoning models with advanced problem-solving and increased focus and capability.
GPT-4o & GPT-4o mini	The latest capable Azure OpenAI models with multimodal features which can accept both text and images as input.
GPT-4 Turbo	
GPT-4 audio	GPT-4o audio "speech in generation"
GPT-4	A set of models that improve on GPT-3.5 and can understand and generate natural language and code.
GPT-3.5	A set of models that improve on GPT-3 and can understand and generate natural language and code.
Embeddings	A set of models that can convert text into numerical vector form to facilitate text similarity.
DALL-E	A series of models that can generate original images from natural language.
Whisper	A series of models in preview that can transcribe and translate speech to text.
Text to speech (Preview)	A series of models in preview that can synthesize text to speech.

A commonly used model is “GPT-4o.”

Registration is required for the O1 system.

UTokyo Azure | utelecon | Feedback - Microsoft Azure | gpt-4o-mini-realtime-preview

ai.azure.com/resource/models/gpt-4o-mini-realtime-preview/version/2024-12-17/registry/azure-openai?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resource... ☆ 本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service / gprt-4o-test-1 / Models / gpt-4o-mini-realtime-preview All resources 5 gear smiley gprt-4o-test-1 (japaneast, S0) 8

gpt-4o-mini-realtime-preview PREVIEW

Deploy Fine-tune

Details Existing deployments License

The GPT-4o-mini-realtime-preview model introduces a smaller, lower cost model to power realtime speech applications. Like GPT-4o-realtime-preview, GPT-4o-mini-realtime-preview provides a richer and more engaging user experience, at a fraction of the cost.

The introduction of GPT-4o-mini-realtime-preview opens numerous possibilities for businesses in various sectors:

Enhanced customer service: By integrating audio inputs, GPT-4o-mini-realtime-preview enables more dynamic and comprehensive customer support interactions.

Content innovation: Use GPT-4o-mini-realtime-preview's generative capabilities to create engaging and See more

Model Versions

Learn more about regional availability Japan East

Model ID	Availability	Lifecycle	Max request	Retirement Date
2024-12-17	Not available	Preview Not suitable for	Input: 128000	2025年12月17日(水)

Model ID Reference this model ID when deploying the model in code

az ml://registries/azure-openai/models/gpt-4o-mini-realtime-preview/versions/2024-12-17

Quick facts OpenAI

gpt-4o-mini-realtime-preview audio-generation

Last trained October 2023

Models help

Relevant resources

Azure OpenAI Service models

Azure OpenAI Service model deprecations and retirements

Not all models are available in every region.

School | gpt-4o-test-1-1 - Microsoft Azure | gpt-4o - Azure OpenAI Service +

https://ai.azure.com/resource/models/gpt-4o/version/2024-11-20/registry/azure-openai?wsid=/subscriptions/a3b0857d-8c96-43a1-b2ee-d787b83bbe5e/resourceGroups/openai_test1/p... A ⚡ ⚡ ... ⚡

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-1 / Models / gpt-4o Docs All resources 5 🚨 🚧 😊 gpt-4o-test-1-1 (japaneast, S0) ⚡ Help

gpt-4o

Deploy Fine-tune

Details Existing deployments

gpt-4o offers a shift in how AI models process data, providing a richer, more engaging user experience.

Matching the intelligence of gpt-4, GPT-4o exhibits the highest vision and text processing performance.

gpt-4o is engineered for speed and efficiency, enabling cost savings and performance.

The introduction of gpt-4o opens up new possibilities for AI integration.

See more

Model Versions

Learn more about regional availability

Japan East

Model ID	Available regions
2024-05-13	Global

Data, media and languages

Property	Description
Supported data types	Inputs Outputs

Quick facts

gpt-4o chat-completion

Last trained October 2023

Pricing See pricing

Model ID Reference this model ID when deploying the model in code

In this case, processing will take place in "Japan East."

Deployment name * gpt-4o-test

Deployment type Global Standard

Deployment details

Model version 2024-11-20

Capacity 250K tokens per minute (TPM)

Content safety DefaultV2

AI resource gpt-4o-test-1-1

Resource location Japan East

Once a new default is selected under the policy

Deploy Cancel

As of February 2025, accessing the GPT-o1 series requires submitting an access request.

The screenshot shows the Azure AI Foundry interface for the o1 model. A red circle highlights the 'アクセスの要求' (Access Request) link in the deployment modal. A large blue arrow points from this modal to the Microsoft registration page.

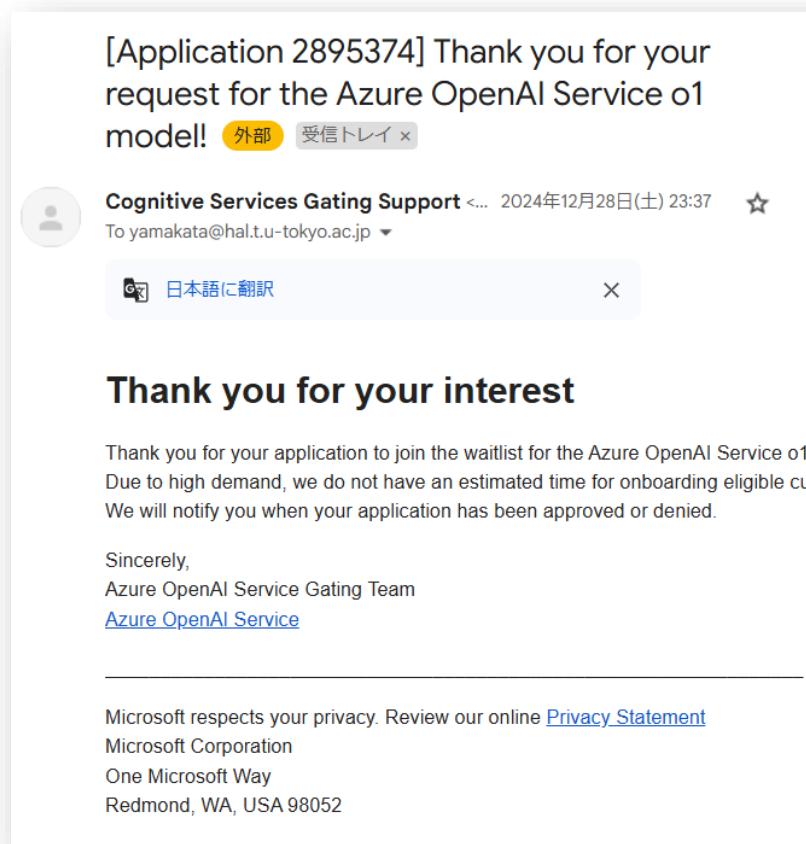
Fill out the form to register

The registration page title is 'Request Access to the Azure OpenAI Service o* family of models'. It includes instructions to read the form carefully, information about registration requirements, and a list of possible causes for denial. A red box highlights the 'Your First Name *' input field.

Approved after a while

In my case...

- o1-preview and o1-mini are approved immediately.
- o1 is approved 15 days after the request.



After 30 minutes



After 18 days



学校

gpt-4o-test-1-2 - Microsoft Azure x gpt-4o-test - Azure OpenAI Service x +

https://ai.azure.com/resource/deployments/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai_test1%2Fproviders%2FMicrosoft.CognitiveSer... A ⚡ ⚡ ... ⚡

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Deployments / gpt-4o-test Docs All resources 5 🔍 ⚡ ⚡ ⚡ gpt-4o-test-1-2 (japaneast, S0) ⚡

← gpt-4o-test

Home Get started Model catalog Playgrounds Chat Assistants PREVIEW Real-time audio Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions Batch jobs Metrics Shared resources Deployments Quota Safety + security Data files Vector stores PREVIEW

Details Metrics Risks & Safety Consume

Open in playground Edit Delete

Endpoint

Target URI: https://gpt-4o-test-1-2.openai.azure.com/openai/deployments/gpt-4o-test

Key: [REDACTED]

Deployment info

Name: gpt-4o-test	Provisioning state: Succeeded
Deployment type: Global Standard	Created on: 2025-04-08T06:38:31.2586957Z
Created by: 1923306880@utac.u-tokyo.ac.jp	Modified on: Apr 8, 2025 3:38 PM
Modified by: 1923306880@utac.u-tokyo.ac.jp	Version update policy: Once a new default version is available
Rate limit (Tokens per minute): 250,000	Rate limit (Requests per minute): 1,500
Model name: gpt-4o	Model version: 2024-11-20
Life cycle status: GenerallyAvailable	Date created: Dec 3, 2024 9:00 AM

Language: Python

SDK: Azure OpenAI SDK

Authentication type: Key Authentication

Get Started

Below are example code snippets for a few use cases. For additional information about Azure OpenAI SDK, see full documentation samples.

1. Authentication using API Key

For OpenAI API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

To create a client with the OpenAI SDK using an API key, initialize the client by passing your API key to the SDK's configuration to authenticate and interact with OpenAI's services seamlessly:

```
import os
from openai import AzureOpenAI

client = AzureOpenAI(
    api_version="2024-12-01-preview",
    endpoint="https://gpt-4o-test-1-2.openai.azure.com/",
    credential=AzureKeyCredential("<API_KEY>")
)
```

2. Install dependencies

Install the Azure Open AI SDK using pip (Requires: Python >=3.8):

```
pip install openai
```

3. Run chat sessions
using the “Chat Playground”

学校

gpt-4o-test-1-2 - Microsoft Azure x gpt-4o-test - Azure OpenAI Service x +

https://ai.azure.com/resource/deployments/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai_test1%2Fproviders%2FMicrosoft.CognitiveSer... A ⚡ ⚡ ... ⚡

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Deployments / gpt-4o-test Docs All resources 5 🔍 ⚡ ⚡ gpt-4o-test-1-2 (japaneast, S0) ⚡

← gpt-4o-test

Home Get started Model catalog Playgrounds Chat Assistants PREVIEW Real-time audio Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions Batch jobs Metrics Shared resources Deployments Quota Safety + security Data files Vector stores PREVIEW

Details Metrics Risks & Safety Consume Open in playground Edit Delete

Endpoint

Target URI: https://gpt-4o-test-1-2.openai.azure.com/openai/deployments/gpt-4o-test
Key: [REDACTED]

Deployment info

Name: gpt-4o-test	Provisioning state: Succeeded
Deployment type: Global Standard	Created on: 2025-04-08T06:38:31.2586957Z
Created by: 1923306880@utac.u-tokyo.ac.jp	Modified on: Apr 8, 2025 3:38 PM
Modified by: 1923306880@utac.u-tokyo.ac.jp	Version update policy: Once a new default version is available
Rate limit (Tokens per minute): 250,000	Rate limit (Requests per minute): 1,500
Model name: gpt-4o	Model version: 2024-11-20
Life cycle status: GenerallyAvailable	Date created: Dec 3, 2024 9:00 AM

Language: Python SDK: Azure OpenAI SDK Authentication type: Key Authentication

Get Started

Below are example code snippets for a few use cases. For additional information about Azure OpenAI SDK, see full documentation samples.

1. Authentication using API Key

For OpenAI API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

To create a client with the OpenAI SDK using an API key, initialize the client by passing your API key to the SDK's configuration to authenticate and interact with OpenAI's services seamlessly:

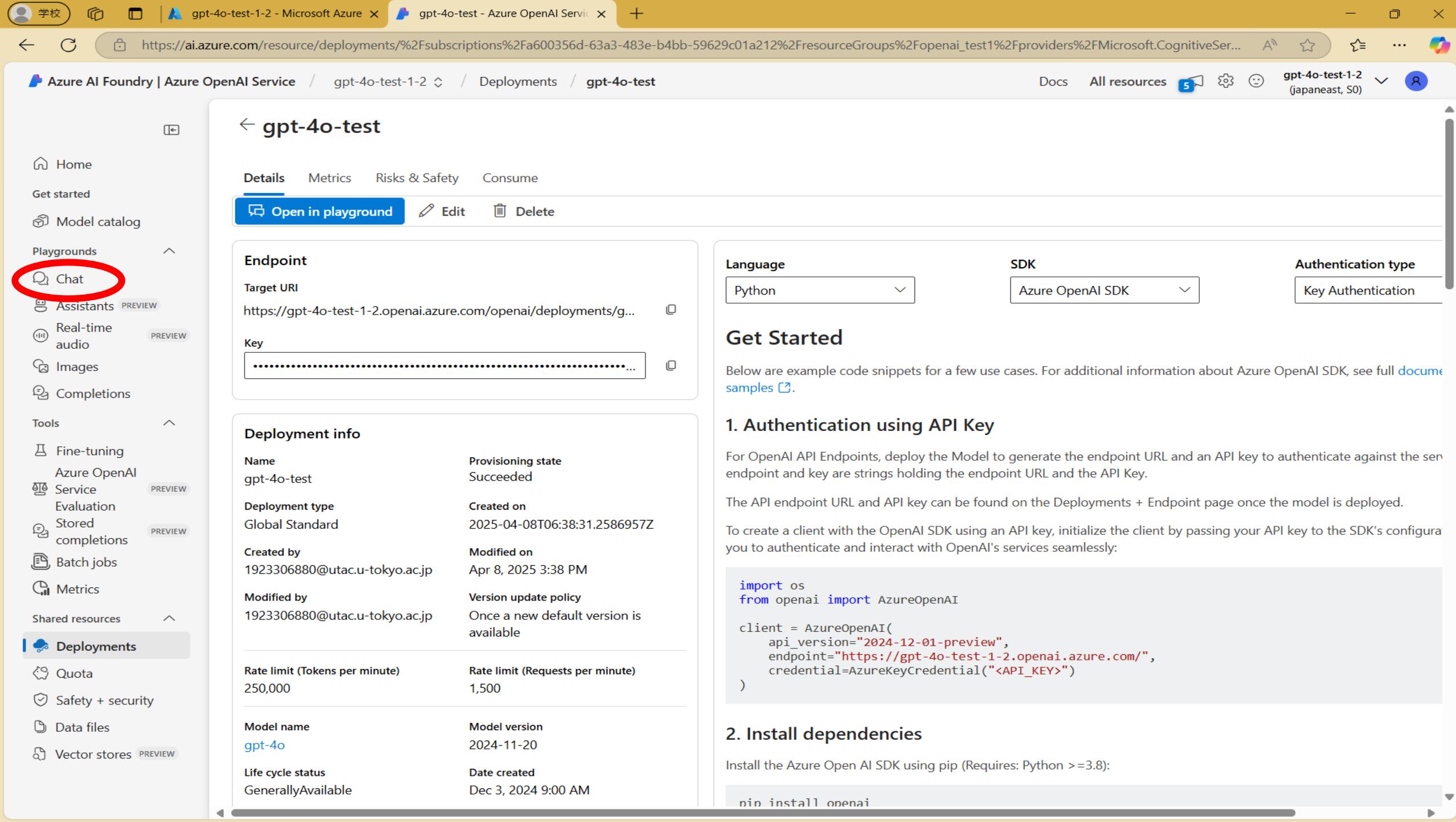
```
import os
from openai import AzureOpenAI

client = AzureOpenAI(
    api_version="2024-12-01-preview",
    endpoint="https://gpt-4o-test-1-2.openai.azure.com/",
    credential=AzureKeyCredential("<API_KEY>")
)
```

2. Install dependencies

Install the Azure Open AI SDK using pip (Requires: Python >=3.8):

```
pip install openai
```



学校

gpt-4o-test-1-2 - Microsoft Azure Chat playground - Azure OpenAI

https://ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/...

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground

Docs All resources 5 🌐⚙️😊 gpt-4o-test-1-2 (japaneast, S0)

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment
gpt-4o-test (version:2024-11-20)

Give the model instructions and context ⓘ

Apply changes Generate prompt

+ Add section

> Add your data
> Parameters

Hide

Chat history

Response format Text

Start with a sample prompt

Historical fiction Write a scene set in ancient Rome, focusing on the daily life of a common citizen.

Dialogue creation Create a conversation between two characters who are meeting for the first time in a mysterious place.

Poetry generation Compose a poem about the beauty of nature in autumn.

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

0/128000 tokens to be sent

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground

Docs All resources 5 gear smiley gpt-4o-test-1-2 (japaneast, S0) Help

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment

gpt-4o-test (version:2024-11-20)

Give the model instructions and context ⓘ

You are an elementary school teacher. Please answer questions from upper-grade elementary school students in a way that is easy for them to understand.

Apply changes Generate prompt

+ Add section

Add your data Parameters

Provide context about the AI: Explain the role or position of the AI model—what it is intended to do, and in what kind of situation or environment it is being used.

Start with a sample prompt

Historical fiction Write a scene set in ancient Rome, focusing on the daily life of a common citizen.

Dialogue creation Create a conversation between two characters who are meeting for the first

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

27/128000 tokens to be sent

The screenshot shows the Azure Chat playground interface. On the left, a sidebar lists various tools and services. The main area is titled 'Chat playground' and contains a 'Setup' section for a deployment named 'gpt-4o-test'. A red box highlights a callout for providing context about the AI's role and environment. Another red box highlights the 'Apply changes' button. To the right, there's a 'Start with a sample prompt' section featuring two examples: 'Historical fiction' and 'Dialogue creation', each with a brief description. At the bottom, there's a large input field for user queries and a status bar indicating token usage.

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground

Docs All resources 5 gear smiley gpt-4o-test-1-2 (japaneast, S0)

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup Chat history Hide

Deployment * + Create new deployment

gpt-4o-test (version:2024-11-15)

Please select the 'Apply changes' button to start a new chat session.

Give the model instructions and set its personality.

You are an elementary school teacher. Your job is to answer questions from upper elementary school students. Your answers should be easy for them to understand.

Do not show this again

Continue Cancel

Add section

Add your data Parameters

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

0/128000 tokens to be sent

Help

Update system message?

Updating the system message will start a new chat session. Previous messages won't be included in new API requests.

Historical fiction

A scene set in ancient Rome, focusing on the daily life of a common

Dialogue creation

Create a conversation between two characters who are meeting for the first

Chat with a sample prompt

UTokyo Azure | utelecon | Feedback - Microsoft Azure | Chat playground - Azure OpenAI

ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/... 本人確認を行ってください

Azure AI Foundry / 25739-m93vv0k1-eastus2 / Chat playground All resources 5 25739-m93vv0k1-eastus2 (eastus2, S0) 8

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment gpt-4o-1 (version:2024-11-20)

Give the model instructions and context

You are an elementary school teacher. Please answer questions from upper-grade elementary school students in a way that is easy for them to understand.

Apply changes Generate prompt

+ Add section

Add your data Parameters

Chat history

Response format Text

Start with a sample prompt

Travel guide Write a travel guide for an imaginary city with unique attractions and culture.

Poetry generation Compose a poem about the beauty of nature in autumn.

Creative storytelling Write a short story about a time traveler who accidentally changes a major historical event.

Please tell me how large language models work.

36/128000 tokens to be sent

Chat playground help Relevant resources Chat Quickstart

School | gpt-4o-test-1-2 - Microsoft Azure | Chat playground - Azure OpenAI

https://ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/acc...

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground Docs All resources 5 gear smiley gpt-4o-test-1-2 (japaneast, S0) Help

Chat playground

Home Get started Model catalog Playgrounds Chat Assistants PREVIEW Real-time audio PREVIEW Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions PREVIEW Batch jobs Metrics Shared resources Deployments

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment* + Create new deployment
gpt-4o-test (version:2024-11-20)

Give the model instructions and context *i*

You are an elementary school teacher. Please answer questions from upper-grade elementary school students in a way that is easy for them to understand.

Apply changes Generate prompt

+ Add section

> Add your data
> Parameters

Chat history

Please tell me how large language models work.

Of course! Let's break it down into simple steps:

A large language model, like the one you're asking about, is a type of computer program that uses math to understand and create text, like words and sentences. It's a little like a super-smart robot brain that's really good at reading and writing.

Here's how it works:

1. **Learning From Text:** First, the model is "trained" by reading lots and lots of text — like books, articles, websites, and more. It doesn't actually "read" like we do, but it processes patterns in the text to learn how words and sentences fit together.

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

435/128000 tokens to be sent

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground

Docs All resources 5 gear smiley gpt-4o-test-1-2 (japaneast, S0)

Chat playground

View code Deploy Import Export Prompt samples Filters feedback

Setup

Deployment * + Create new deployment
gpt-4o-test (version:2024-11-20)

Give the model instructions and context

You are an elementary school teacher. Please answer questions from upper-grade elementary school students in a way that is easy for them to understand.

Apply changes Generate prompt

+ Add section

Chat history

Please tell me how large language models work.

Of course! Let's break it down into simple steps:

A large language model, like the one you're asking about, is a type of computer program that uses math to understand and create text, like words and sentences. It's a little like a super-smart robot brain that's really good at reading and writing.

Here's how it works:

1. Learning From Text — like books we do, but it processes sentences fit together

If the output is cut off partway, try adjusting the parameters.

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

435/128000 tokens to be sent

Home Get started Model catalog Playgrounds Chat Assistants Real-time audio Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions Batch jobs Metrics Shared resources Deployments

Help

The screenshot shows the Azure AI Foundry Chat playground interface. On the left sidebar, under the 'Chat' category, the 'Parameters' link is circled in red. In the main chat area, the 'Response format' dropdown is also circled in red. A large red callout box highlights the text 'If the output is cut off partway, try adjusting the parameters.' located in the middle-right section of the interface.

School | gpt-4o-test-1-2 - Microsoft Azure | Chat playground - Azure OpenAI | +

https://ai.azure.com/resource/playground?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/acc...

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Chat playground Docs All resources 5 gear smiley gpt-4o-test-1-2 (japaneast, S0) Help

Chat playground

Home Get started Model catalog Playgrounds Chat Assistants PREVIEW Real-time audio PREVIEW Images Completions Tools Fine-tuning Azure OpenAI Service Evaluation Stored completions PREVIEW Batch jobs Metrics Shared resources Deployments

View code Deploy Import Export Prompt samples Filters feedback

Add your data

Parameters

Past messages included 10

Max response 2400 (circled)

Temperature 0.7

Top P 0.95

Stop sequence

Frequency penalty 0

Presence penalty 0

Chat history

Please tell me how large language models work.

Of course! Let's break it down into simple steps:

A large language model, like the one you're asking about, is a type of computer program that uses math to understand and create text, like words and sentences. It's a little like a super-smart robot brain that's really good at reading and writing.

Here's how it works:

- 1. Learning From Text:** First, the model is "trained" by reading lots and lots of text — like books, articles, websites, and more. It doesn't actually "read" like we do, but it processes patterns in the text to learn how words and sentences fit together.

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

435/128000 tokens to be sent

4. Use the service via Web API from your code

How can I send a prompt to a model in the OpenAI Service and receive the result via API from a Google Colaboratory notebook?

Azure AI Foundry | Azure OpenAI Service / gpt-4o-test-1-2 / Deployments / gpt-4o-test Docs All resources 5 gear settings smiley gpt-4o-test-1-2 (japaneast, S0) ?

Playgrounds ^

Chat

Assistants PREVIEW

Real-time audio PREVIEW

Images

Completions

Tools ^

Fine-tuning

Azure OpenAI Service PREVIEW

Evaluation

Stored completions PREVIEW

Batch jobs

Metrics

Shared resources ^

Deployments

Quota

Safety + security

Data files

Vector stores PREVIEW

available

client = AzureOpenAI(

api_version="2024-12-01-preview",

endpoint="https://gpt-4o-test-1-2.openai.azure.com/"

← gpt-4o-test

Details Metrics

Open in playground

Endpoint

Target URI

https://gpt-4o-test-1-2.openai.azure.com/openai/deployments/g...

Key

.....

Deployment info

Name: gpt-4o-test Status: succeeded

Deployment type: Global Standard Created on: 2025-04-09T06:29:31Z

Language: Python SDK: Azure OpenAI SDK

Get Started

Below are example code snippets for a few use cases. For additional information, see the samples.

1. Authentication using API Key

For OpenAI API Endpoints, deploy the Model to generate the endpoint URL and the API Key. The API endpoint URL and API key can be found on the Deployments + E

Copy it with and set it as the “Target URI” in your code.

Copy this with and set as the “API_KEY” in your code.

Caution: If your endpoint information is exposed, anyone can use your resource, and you will be responsible for the charges. Never share or publish this information!

Sample Code for Google Colaboratory

- Please log in with your **ECCS account** and open this in **Colab**.
- If you want to save your own copy, click “**Copy to Drive**” to save it to your Google Drive.
- Then, edit the following section under “**2. Set Parameters**”.

```
import os

# Replace [YOUR TARGET URI] with your actual endpoint
os.environ["AZURE_OPENAI_ENDPOINT"] = "[YOUR TARGET URI]"

# Replace [YOUR KEY] with your actual API key
os.environ["AZURE_OPENAI_KEY"] = "[YOUR KEY]"
```

Rewrite to “Target URI”

Rewrite to Key

```
from openai import AzureOpenAI

client = AzureOpenAI(
    azure_endpoint = os.getenv("AZURE_OPENAI_ENDPOINT"),
    api_key = os.getenv("AZURE_OPENAI_KEY"),
    api_version="2024-12-01-preview",
)
```

```
response = client.chat.completions.create(
    model=os.getenv("AZURE_OPENAI_MODEL"),
    messages=[
        {"role": "system", "content": "You are an Azure expert. Ple
        {"role": "user", "content": "I would like to access a GPT-4
    ],
    max_tokens=800,
    temperature=0.7,
    top_p=0.95,
    frequency_penalty=0,
    presence_penalty=0,
    stop=None,
    stream=False
)
```

Set the context and user query.

```
▶ print(response.choices[0].message.content)
```

→ Accessing a GPT-4 model deployed on Azure OpenAI Service from a notebook running on Google Colab

Step 1: Set Up Your Azure OpenAI Service

1. **Create an Azure OpenAI Service**:

- Log in to the [Azure Portal](<https://portal.azure.com/>).
- Navigate to the Azure OpenAI Service and create a resource.
- Deploy the GPT-4 model in your Azure OpenAI resource. (Ensure you've deployed the desired model version.)

2. **Get Endpoint and API Key**:

- After deploying the model, go to the resource in Azure.
- Note the **endpoint** URL and retrieve the **API key** from the "Keys and Endpoint" section.

Step 2: Set Up Google Colab

1. **Open a New Notebook**:

- Go to [Google Colab](<https://colab.research.google.com/>) and create a new notebook.

2. **Install Required Libraries**:

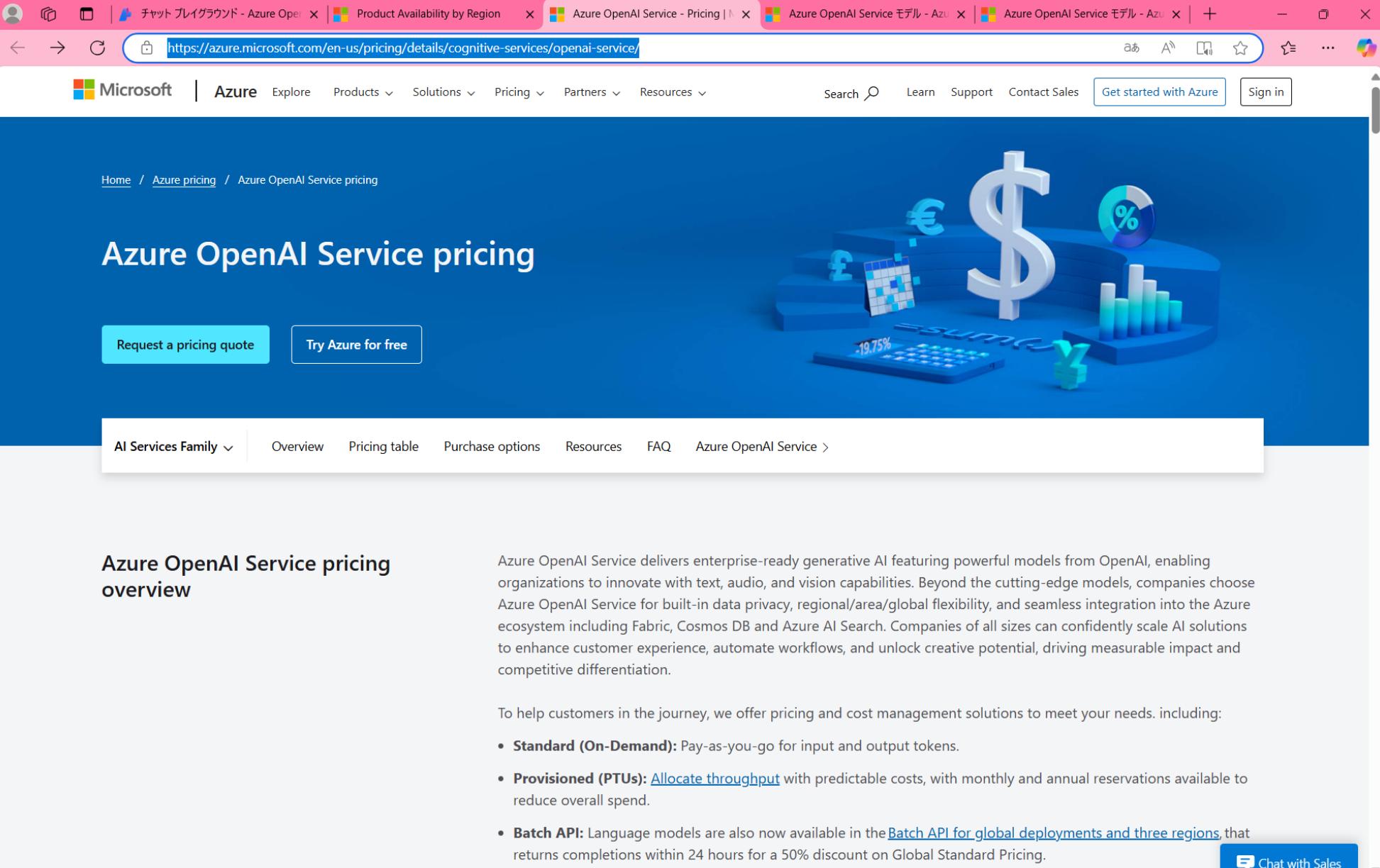
- Install the `requests` library (if not already installed) for making HTTP requests to the Azure OpenAI endpoint.
```python  
!pip install requests  
```

5. Other Topics

5. 1 Pricing

Estimating the Cost

Azure OpenAI Service pricing (see here)



The screenshot shows the Azure OpenAI Service pricing page. At the top, there's a navigation bar with links for Microsoft, Azure, Explore, Products, Solutions, Pricing, Partners, Resources, Search, Learn, Support, Contact Sales, Get started with Azure, and Sign in. The main heading is "Azure OpenAI Service pricing". Below it are two buttons: "Request a pricing quote" and "Try Azure for free". To the right is a large graphic featuring a large dollar sign, a bar chart, a pie chart, and various currency symbols like euro and yen. Below the graphic is a menu bar with "AI Services Family", "Overview", "Pricing table", "Purchase options", "Resources", "FAQ", and "Azure OpenAI Service >". The main content area has a section titled "Azure OpenAI Service pricing overview" with a paragraph about the service's capabilities and a list of pricing options.

Azure OpenAI Service pricing overview

Azure OpenAI Service delivers enterprise-ready generative AI featuring powerful models from OpenAI, enabling organizations to innovate with text, audio, and vision capabilities. Beyond the cutting-edge models, companies choose Azure OpenAI Service for built-in data privacy, regional/area/global flexibility, and seamless integration into the Azure ecosystem including Fabric, Cosmos DB and Azure AI Search. Companies of all sizes can confidently scale AI solutions to enhance customer experience, automate workflows, and unlock creative potential, driving measurable impact and competitive differentiation.

To help customers in the journey, we offer pricing and cost management solutions to meet your needs. including:

- **Standard (On-Demand):** Pay-as-you-go for input and output tokens.
- **Provisioned (PTUs):** [Allocate throughput](#) with predictable costs, with monthly and annual reservations available to reduce overall spend.
- **Batch API:** Language models are also now available in the [Batch API for global deployments and three regions](#), that returns completions within 24 hours for a 50% discount on Global Standard Pricing.

[Chat with Sales](#)

OpenAI GPT Pricing

GPT-4o

GPT-4o is the most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4o-2024-1120 Global	Input: ¥393.88751 Cached Input: ¥196.9438 Output: ¥1,575.5501	Input: ¥196.94376 Output: ¥787.7751

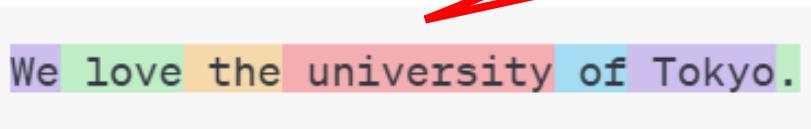
o1

o1 is the new reasoning model series for complex tasks. The model has 200K context and an October 2023 knowledge cutoff.

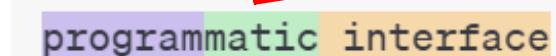
Model	Pricing (1M Tokens)
o1 1217 Global	Input: ¥2,363.3251 Cached Input: ¥1,181.6626 Output: ¥9,453.300060

OpenAI GPT Pricing

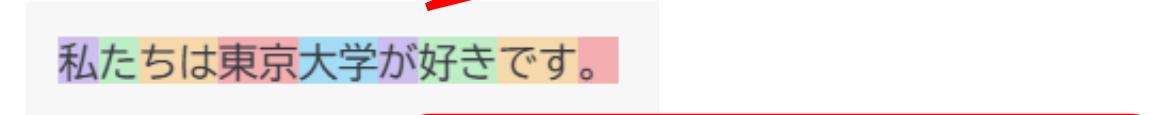
- The cost is determined by the number of tokens in both the **input** and **output**.
- To check the number of tokens in OpenAI, you can use the [web-based tokenizer](#).
- In English, approximately **100 tokens ≈ 75 words**, but **low-frequency words** may require more tokens.
- Since Japanese is considered a low-frequency language in the tokenizer's vocabulary, it generally results in **significantly more tokens** than English for the same amount of text.

Tokens	Characters
7	32
 The text "We love the university of Tokyo." is shown with each word colored differently: We (purple), love (green), the (yellow), university (pink), of (light blue), and Tokyo (teal). This visual representation serves as a reference for character-to-token mapping.	

High-frequency English words:
1 token ≈ 1 word

Tokens	Characters
3	2
 The text "programmatic interface" is shown with each word colored differently: programmatic (purple), interface (orange). Red arrows point from the "Tokens" and "Characters" values to this example.	

Low-frequency English words:
Sometimes 2 tokens ≈ 1 word,
or even more

Tokens	Characters
9	14
 The text "私たちは東京大学が好きです。" is shown with each character colored differently: 私 (purple), た (pink), は (light blue), 東 (yellow), 京 (teal), 大 (orange), 学 (light blue), が (pink), 好 (purple), す (pink), で (light blue), す (pink). Red arrows point from the "Tokens" and "Characters" values to this example.	

Japanese characters:
Often 1 token ≈ 1 character

Tokens	Characters
5	4
 The text "深??学習" is shown with each character colored differently: 深 (purple), ? (pink), ? (pink), 学 (light blue), 學 (teal), 習 (orange). Red arrows point from the "Tokens" and "Characters" values to this example.	

In some cases:
2 tokens ≈ 1 character,
or 3 tokens ≈ 1 character
Example: The Japanese character
「層」 requires 2 tokens.

You can check the number of tokens under "Deployments > [Name] > Metrics".

The screenshot shows the Azure OpenAI Service interface. On the left, a sidebar lists categories like Playgrounds, Assistants, Real-time audio, Images, Completions, Tools, Fine-tuning, Azure OpenAI Service, Evaluation, Stored completions, Batch jobs, and Metrics. The 'Metrics' item in the Tools section is circled in red. The main content area shows a deployment named 'gpt-4o-test'. Below the deployment name, there are tabs for 'Detail', 'Metrics' (which is selected and highlighted with a blue underline), 'Risks & Safety', and 'Consume'. A callout box points to the 'Metrics' tab with the text: 'You can check the number of tokens under "Deployments > [Name] > Metrics".' The 'Metrics' section displays four key metrics: 'Total requests' (2), 'Total token count' (1.33K), 'Completion token count' (1.21K), and 'Prompt token count' (122). At the bottom, there is a chart titled 'Token usage' showing individual token counts.

Playgrounds

Assistants PREVIEW

Real-time audio PREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI Service PREVIEW

Evaluation

Stored completions PREVIEW

Batch jobs

Metrics

Deployments

Quota

Safety + security

Data files

Vector stores PREVIEW

← gpt-4o-test

Detail Metrics Risks & Safety Consume

Open in Azure Monitor 04/01/2025 - 04/08/2025 Last day 7D 1M 3M 1Y

2 Total requests

1.33K Total token count

1.21K Completion token count

122 Prompt token count

Token usage

tokens
888
814
740
666
592
518
444
370

5.2 Token Limit per Minute

Azure OpenAI Service has quota limits

UTokyo Azure | utelecon x | A gppt-4o-test-1 - Microsoft Azure x Model quota - Azure OpenAI Service x + ai.azure.com/resource/quota?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai_test1/providers/Microsoft.CognitiveServices/accounts/gppt-4o-test-1 Model quota

本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service / gppt-4o-test-1 / Model quota All resources 5 gear smiley gppt-4o-test-1 (japaneast, S0) 8 Help

Model catalog

Playgrounds

Chat

Assistants PREVIEW

Language PREVIEW

Real-time audio PREVIEW

Images

Completions

Fine-tuning

Azure OpenAI Service PREVIEW

Evaluation

Stored completions PREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Quota

Safety + security

Data files

Vector stores PREVIEW

Monitor and track your quota usage

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

1923306880-YamakataTraining

Azure OpenAI Service standard + batch Azure OpenAI Service provisioned (classic) Azure OpenAI Service provisioned

Request quota Refresh Reset view Group by Quota type, Model & Region

Deployment ↑	Model name	Version	Quota ty...	Region	Resource	Quota
> GlobalStandard						
> Standard						

Search Filter Show all quota Columns

Quota with lowest availability

Total Quota Quota (Model + Region + Quota type)

● Available quota ● Assigned quota

Azure OpenAI Service model quota is managed at the subscription level, and is given per region per model type. All deployments of the same type running one or multiple Azure AI resource in the same region share a single quota limit.

[Home](#)[Get started](#)[Model catalog](#)[Playgrounds](#)[Chat](#)[Assistants PREVIEW](#)[Real-time audio PREVIEW](#)[Images](#)[Completions](#)[Tools](#)[Fine-tuning](#)[Azure OpenAI Service Evaluation PREVIEW](#)[Stored completions PREVIEW](#)[Batch jobs](#)[Metrics](#)[Shared resources](#)[Deployments](#)[Quota](#)[Safety + security](#)[Data files](#)[Vector stores PREVIEW](#)

Monitor and track your quota usage

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

1923306880-YamakataTraining

Azure OpenAI Service standard + batch

Azure OpenAI Service provisioned (classic)

Azure OpenAI Service provisioned

[Request quota](#)[Refresh](#)[Reset view](#)

Group by

Quota type, Model & Region

 Search

Deployment ↑

Model n...

Vers...

Quota...

i

gpt-4o

GlobalStan...

East US 2

gpt-4o

GlobalStan...

Japan East

gpt-4o

GlobalStan...

gpt-4o-test

gpt-4o

202...

GlobalStan...

Japan ...

gpt-4o-test-1-2

250K TPM



gpt-4o-mini-real

GlobalStan...

Azure OpenAI Service model quota is managed at the subscription level, and is given per region per model type. All deployments of the same type running one or multiple Azure AI resource in the same region share a single quota limit.

Model availability across regions and

- K stands for 1,000, and TPM means Tokens per Minute, i.e., the number of tokens allowed per minute.
- By default, there is a limit of 250,000 tokens per minute.
- Quota varies depending on the model and region ([see reference](#)).

Home

Get started

Model catalog

Playgrounds

Chat

Assistants PREVIEW

Real-time audio PREVIEW

Images

Completions

Fine-tuning

Azure OpenAI Service Evaluation PREVIEW

Stored completion

Batch jobs

Metrics

Shared resources

Deployments

Quota

Safety + security

Data files

Vector stores PREVIEW

Monitor and track your quota usage

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  

Azure OpenAI Service standard + batch Azure OpenAI Service provisioned (classic) Azure OpenAI Service provisioned

Edit quota allocated to this deployment

Capacity Tokens Per Minute (thousands) - gpt-4o - GlobalStandard  1024K/30000K

 Confirm 

Drag the slider to increase it to the desired size.

202... GlobalStan... Japan ... gpt-4o-test-1-2 250K TPM 

gpt-4o-mini-real GlobalStan...
East US 2 gpt-4o... GlobalStan...
Standard 6 of 6 Requests Per

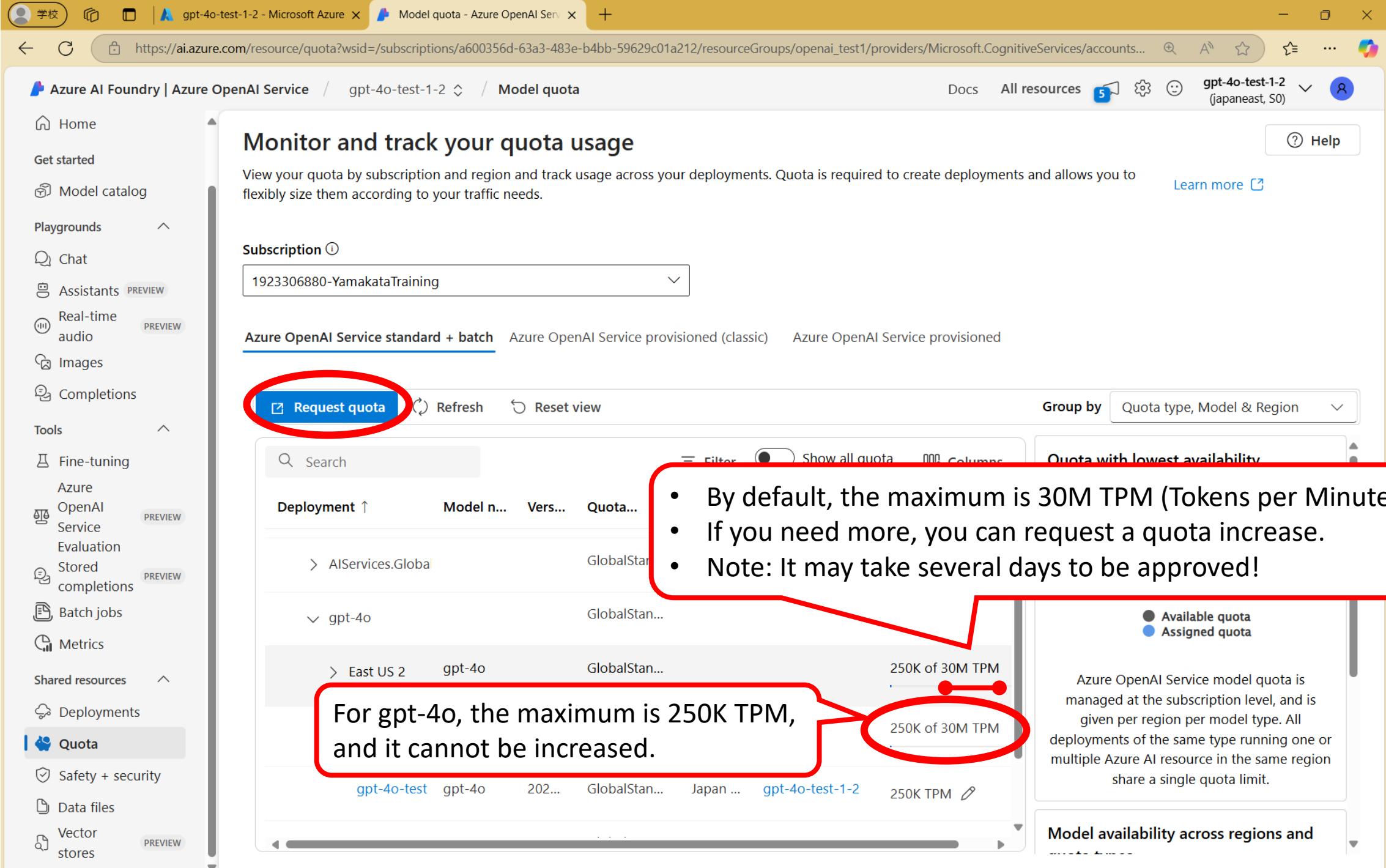
Quota with lowest availability

Total Quota 16M 0 Quota (Model + Region + Quota type)

Available quota  Assigned quota 

Azure OpenAI Service model quota is managed at the subscription level, and is given per region per model type. All deployments of the same type running one or multiple Azure AI resource in the same region share a single quota limit.

Model availability across regions and



Request a quota increase

- Quota Increase Request Form
- Provide your **full name, email address, organization name, and the type and size** of the quota you are requesting.
- Include a **justification** explaining why you are requesting a quota increase. There is a manual review process.
- it may take **several days for approval**, so be sure to submit your request early.

The screenshot shows a Microsoft Edge browser window with a pink header bar. The address bar contains the URL: <https://customervoice.microsoft.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180B...>. The main content area has a black header with the Microsoft Azure logo. Below the header, the title "Azure OpenAI Service: Request for Quota Increase" is displayed. A descriptive text block explains that the form is used to submit requests for quota increases in quota for standard deployment type. It states that requests are being accepted and will be filled in the order received, with priority given to customers who generate traffic that consumes existing quota allocation. It also notes that requests may be denied if conditions are not met. A red asterisk followed by "必須" (Required) is shown next to a section of text. Below this, instructions advise reading the form carefully and answering questions completely before submission. A note at the bottom states that the form is for external users and not Microsoft personnel, with a link to more information. The form includes a field for entering a first name.

Microsoft Azure

Azure OpenAI Service: Request for Quota Increase

This form is used to submit requests for quota increases in quota for the standard deployment type. Quota increase requests are being accepted and will be filled in the order they are received. Priority will be given to customers who generate traffic that consumes the existing quota allocation, and your request may be denied if this condition is not met.

* 必須

Read the instructions carefully and answer each question completely before submitting the request

Use this form to request an increase due to your forecasted usage for Azure OpenAI Service. Microsoft will use the information you provide to assess your usage volume and patterns, allowing us to allocate the necessary GPU capacity to support your work. We will make every effort to accommodate your request; however, allocation is based on our current capacity and future deployments, and is subject to availability.

Please Note: For Microsoft personnel, do not fill out this form. Find more information [here](#).

1. First Name *

Enter your first name.

5.3 Azure OpenAI Service Data, Privacy, and Security

Will the data be used for Azure or OpenAI?

Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn

Your prompts (inputs) and completions (outputs), your embeddings, and your training data:

- are NOT available to other customers.
- are NOT available to OpenAI.
- are NOT used to improve OpenAI models.
- are NOT used to train, retrain, or improve Azure OpenAI Service foundation models.
- are NOT used to improve any Microsoft or 3rd party products or services without your permission or instruction.
- Your fine-tuned Azure OpenAI models are available exclusively for your use.

The Azure OpenAI Service is operated by Microsoft as an Azure service; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g. ChatGPT, or the OpenAI API).

Purpose

- Two Options for Using Pretrained Models
 - **Azure OpenAI Service**: For immediate access to OpenAI models
 - **Azure AI Foundry**: For access to various models including OpenAI

- **Roadmap for Azure OpenAI Service**

1. Create a resource group
2. Select a model from the model catalog (e.g., choose “gpt-4o”)
3. Run chat sessions using the “Chat Playground”
4. Use the service via Web API from your code

Additional Topics

- 5.1 Pricing
- 5.2 Token limits per minute
- 5.3 Data, privacy, and security

- **Using Other Pretrained Models (via Azure AI Foundry)**
 - Example: DeepSeek R1

Get started with DeepSeek

If you want to use a model other than OpenAI, use AI Foundry

List of models as of February 2025

See [here](#) for the latest version

company	model	type
AI21 Labs	AI21-Jamba-1.5-Mini, AI21-Jamba-1.5-Large	chat-completion
Azure OpenAI	o3-mini, o1, o1-preview, o1-mini, gpt-4o, gpt-4o-mini	chat-completion
Azure OpenAI	gpt-4o-realtime-preview	real-time
Azure OpenAI	text-embedding-3-large, text-embedding-3-small	embeddings
Cohere	Cohere-embed-v3-english, Cohere-embed-v3-multilingual	embeddings
Cohere	Cohere-command-r-plus-08-2024, Cohere-command-r-08-2024, Cohere-command-r-plus, Cohere-command-r	chat-completion
Core42	jais-30b-chat	Open Source Version
DeepSeek	DeepSeek-R1	chat-completion
Meta	Llama-3.3-70B-Instruct, Llama-3.2-11B-Vision-Instruct, Llama-3.2-90B-Vision-Instruct, Meta-Llama-3.1-405B-Instruct, Meta-Llama-3.1-8B-Instruct, Meta-Llama-3.1-70B-Instruct, Meta-Llama-3-8B-Instruct, Meta-Llama-3-70B-Instruct	chat-completion
Microsoft	Phi-3-mini-128k-instruct, Phi-3-mini-4k-instruct, Phi-3-small-8k-instruct, Phi-3-medium-128k-instruct, Phi-3-medium-4k-instruct, Phi-3.5-vision-instruct, Phi-3.5-MoE-instruct, Phi-3-small-128k-instruct, Phi-3.5-mini-instruct, Phi-4	chat-completion
Mistral AI	Mistral-3B, Mistral-large, Mistral-small, Mistral-Nemo, Mistral-large-2407, Mistral-Large-2411, Codestral-2501	chat-completion
NTT Data	Tsuzumi-7b	chat-completion

1. Create a Hub and a Project



音声で読み上げる

音声オプション

Microsoft Azure

AI founda

Copilot

1923306880@utac.u-to...
THE UNIVERSITY OF TOKYO

Azure services

Create a resource

More services

Services (31)

Services

- Azure AI Foundry
- Factory Operations Agent in Azure AI Foundry
- Azure Data Boundaries
- AI Search

Documentation

- Deploy models in Azure AI Foundry portal - Azure AI Foundry
- Create an Azure AI Foundry project in Azure AI Foundry portal - Azure AI F...
- Microsoft Azure AI Fundamentals: Document Intelligence and Knowledge ...

See more

Resources

Recent

Name

- Alfoundarytest
- gpt-4o-test-1-2
- openai_test1
- 1923306880-YamakataLab
- gpt-4o-test-1-1
- 1923306880-YamakataTraining
- openai_test1

Resource group

Subscription

Azure OpenAI

Subscription

Resource group

Last Viewed

- a few seconds ago
- an hour ago
- an hour ago
- an hour ago
- an hour ago
- 2 hours ago

Select "Azure AI Foundry"

Cost alerts

Quotas

Users

Resource groups

音声で読み上げる



Microsoft Azure

Search resources, services, and docs (G+/-)

Copilot

1923306880@utac.u-to...
THE UNIVERSITY OF TOKYO

Home >

Azure AI Foundry

The University of Tokyo

+ Create Manage view Open query Assign tags

Resource group equals all Location equals all Add filter

No grouping List view

Start by creating a Hub

Resource group ↑	Type ↑	Location ↑	Subscription ↑↓	...
openai_test1	Azure AI services	East US 2	1923306880-YamakataLab	...
rg-1923306880-1245_ai	Azure AI hub	East US 2	1923306880-YamakataLab	...
openai_test1	Azure AI services	East US 2	1923306880-YamakataTraining	...
openai_test1	Azure AI project	East US	1923306880-YamakataTraining	...
AI-Foundry-hub0	Azure AI project	East US 2	1923306880-YamakataLab	...
AI-Foundry-hub0	Azure AI hub	East US 2	1923306880-YamakataLab	...
openai_test1	Azure AI hub	East US	1923306880-YamakataTraining	...
openai_test1	Azure AI services	East US	1923306880-YamakataTraining	...
AI-Foundry-hub0	Azure AI services	East US 2	1923306880-YamakataLab	...

< Previous

Page 1 of 1

Next >

Give feedback

音声で読み上げる 音声オプション

Microsoft Azure Search resources, services, and docs (G+/-) Copilot 9 🔍 🚧 🌐 1923306880@utac.u-to... THE UNIVERSITY OF TOKYO

Home > Azure AI Foundry >

Azure AI hub

Create an Azure AI hub resource

Basics Storage Inbound Access Outbound Access Encryption Identity Tags Review + create

Organization

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources. An AI hub is a collaboration environment for a team to share project work, model endpoints, compute, (data) connections, security settings, govern usage.

Subscription * ① 1923306880-YamakataTraining

Resource group * ① (New) AI-Foundry-hub0 Create new

Region * ① East US 2

Resource details

Name * ① AI-Foundry-test0 ✓

Friendly name ① AI Foundry test0

Default project resource group ①

Same as hub resource group

Review + create (circled in red)

< Previous Next : Storage

The interface is largely the same as Azure OpenAI Service.

[Home](#) > [Azure AI Foundry](#) >

Azure AI hub

Create an Azure AI hub resource

✓ Validation passed[Basics](#) [Storage](#) [Inbound Access](#) [Outbound Access](#) [Encryption](#) [Identity](#) [Tags](#) [Review + create](#)

Basics

Subscription	1923306880-YamakataTraining
Resource group	(New) AI-Foundry-hub0
Name	AI-Foundry-test0
Default project resource group	(New) AI-Foundry-hub0

Resources

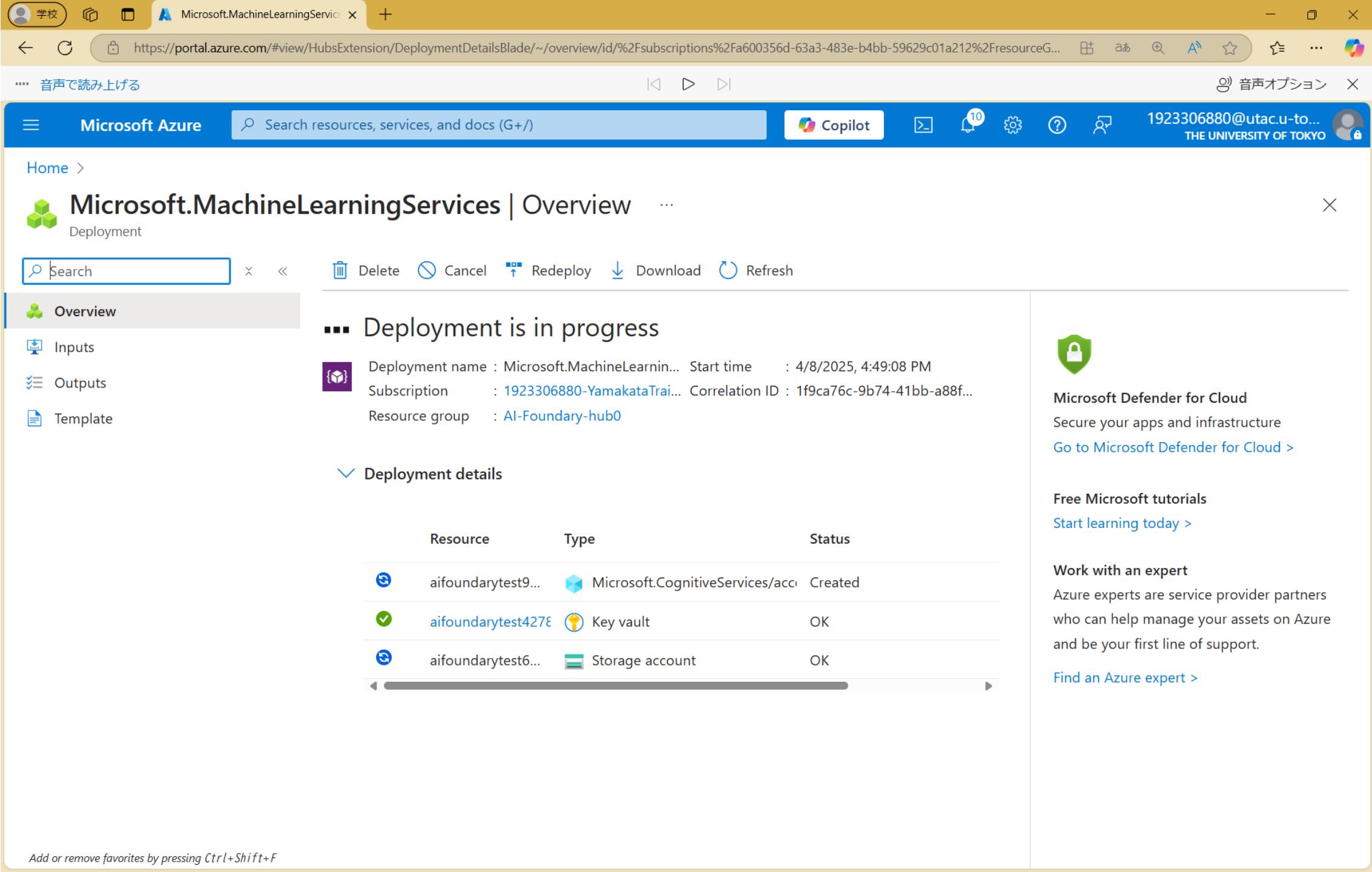
Region	East US 2
AI Services	(new) aifoundarytest9323385592
Storage account	(new) aifoundarytest6713422946
Key vault	(new) aifoundarytest4278903108
Application insights	None
Container registry	None

Create

< Previous

Next >

Download a template for automation





音声で読み上げる 音声オプション

Microsoft Azure Search resources, services, and docs (G+/-) Copilot 10 🔍 ⚙️ 🌐 🔍 1923306880@utac.u-to... THE UNIVERSITY OF TOKYO

Home > Microsoft.MachineLearningServices | Overview Deployment

Search X ↪ Delete Cancel Redeploy Download Refresh

Deployment succeeded Deployment 'Microsoft.MachineLearningServices' to resource group 'AI-Foundary-hub0' was successful.

Go to resource Go to resource group

Overview

Inputs Outputs Template

✓ Your deployment is complete

Deployment name : Microsoft.MachineLearnin... Start time : [redacted]
Subscription : 1923306880-YamakataTrai... Correlation ID : 1f9ca76c-9b74-41bb-a88f...
Resource group : AI-Foundary-hub0

> Deployment details

Next steps

Go to resource

Cost management Get notified to stay within your budget and prevent unexpected charges on your bill. Set up cost alerts >

Microsoft Defender for Cloud Secure your apps and infrastructure Go to Microsoft Defender for Cloud >

Free Microsoft tutorials Start learning today >

Work with an expert Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support. Find an Azure expert >



音声で読み上げる 音声オプション ×

Microsoft Azure Search resources, services, and docs (G+/-) Copilot 10 🔍 ⚙️ 🌐 🌐 1923306880@utac.u-to... THE UNIVERSITY OF TOKYO

Home > Microsoft.MachineLearningServices | Overview >

AI-Foundry-test0

Azure AI hub



+ Create project

Download config.json

Delete

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Resource visualizer
- Events
- Settings
- Monitoring
- Automation
- Support + troubleshooting

Essentials

JSON View

Resource group Project resource group (default)

[AI-Foundry-hub0](#) [AI-Foundry-hub0](#)

Location Storage

East US 2 [aifoundarytest6713422946](#)

Subscription Container Registry (edit)

[1923306880-YamakataTraining](#) [...](#)

Subscription ID Application Insights (edit)

a600356d-63a3-483e-b4bb-59629c01a212 [...](#)

Key Vault Provisioning State

[aifoundarytest4278903108](#) Succeeded

Govern the environment for your team in AI Foundry



Your Azure AI hub provides enterprise-grade security, and a collaborative environment to build AI solutions. Centrally audit usage and cost, and set up connections to your company resources that all projects can use. [learn more about the Azure AI Foundry](#)

Launch Azure AI Foundry

School

AI-Foundry-test0 - Microsoft Az... Hub overview - Azure AI Foundry

https://ai.azure.com/managementCenter/hub/overview?tid=f07baf4f-2b70-47d7-9a02-0875caf94c84&wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourcegro... Docs All hubs + projects 5 Hub AI-Foundry-test0 Help

Azure AI Foundry | Management center / Hub overview

AI Foundry test0

Management center

All hubs + projects Quota

Hub (AI-Foundry-test0) Overview

Users Models + endpoints Connected resources Compute

+ New project Refresh Delete project Reset view

Description Add a hub description here

Hub properties

Name	Location
AI-Foundry-test0	eastus2
Subscription	Resource Group
1923306880-YamakataTraining	AI-Foundry-hub0

Manage in Azure Portal Get API endpoints and keys View subscription quota

Delete hub

Create your first project

Projects help you organize your AI assets and save "state" (when a system remembers previous events or interactions) across different tools, data connections, and more.

Learn more about projects in Azure AI Foundry

Users 1 View all →

AI-Foundry-test0 Azure AI Administrator

New user

The screenshot shows the Azure AI Foundry Hub overview page. On the left, there's a sidebar with 'Management center' options like 'All hubs + projects' and 'Quota'. Below that is a tree view for 'Hub (AI-Foundry-test0)' with 'Overview' selected, followed by 'Users', 'Models + endpoints', 'Connected resources', and 'Compute'. In the main content area, there's a header 'AI Foundry test0' with a 'New project' button (which has a red circle around it), 'Refresh', 'Delete project', and 'Reset view' buttons. Below this is a large 'Create your first project' section with a file icon. To the right, there's a 'Description' field with placeholder text 'Add a hub description here'. Under 'Hub properties', it shows the hub's name ('AI-Foundry-test0'), location ('eastus2'), subscription ('1923306880-YamakataTraining'), and resource group ('AI-Foundry-hub0'). There are links to 'Manage in Azure Portal', 'Get API endpoints and keys', and 'View subscription quota'. A prominent red-bordered 'Delete hub' button is located here. At the bottom, there's a 'Users' section showing one user ('AI-Foundry-test0') with the role 'Azure AI Administrator', and a 'New user' button.

School

AI-Foundry-test0 - Microsoft Az... Hub overview - Azure AI Foundry

https://ai.azure.com/managementCenter/hub/overview?tid=f07baf4f-2b70-47d7-9a02-0875caf94c84&wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourcegro... Docs All hubs + projects Hub 5 AI-Foundry-test0 Help

AI Foundry test0

All hubs + projects Quota

Hub (AI-Foundry-test0) Overview Users Models + endpoints Connected resources Compute

Create a project

Projects are easy-to-manage containers for your work—and the key to collaboration, organization, and connecting data and other services.

Current hub: AI-Foundry-test0

Project name *: 1923306880-1774

Create Cancel

connections, and more.

Learn more about projects in Azure AI Foundry

Users 1 View all →

AI-Foundry-test0 Azure AI Administrator

+ New user

The screenshot shows the Azure AI Foundry Management Center Hub overview page. A modal dialog box titled "Create a project" is open in the center. It contains a brief description of what projects are, the current hub set to "AI-Foundry-test0", and a "Project name" input field containing "1923306880-1774". At the bottom of the modal are two buttons: "Create" (highlighted with a red oval) and "Cancel". In the background, the main hub overview shows basic information like location ("eastus2") and resource group ("AI-Foundry-hub0"). The left sidebar includes sections for Management center, All hubs + projects, Quota, and various hub settings like Overview, Users, Models + endpoints, Connected resources, and Compute. A "New project" button is also visible in the top navigation bar.

Azure AI Foundry / 1923306880-1774 / Overview

1923306880-1774

Add a project description (optional)

Endpoints and keys

View all endpoints

API Key

.....

Included capabilities

Azure AI inference

Azure OpenAI Service

Azure AI Services

Use the following endpoint to call your Azure OpenAI Service models:

Azure OpenAI Service endpoint

<https://aifoundarytest9323385592.openai.azure.com/>

[x] API documentation

Project details

Project connection string

eastus2.api.azureml.ms;a600356d-63a3-483...

Subscription

1923306880-YamakataTraining

Subscription ID

a600356d-63a3-483e-b4bb-59629c01a212

Location

eastus2

Manage project settings

- Add users
- View quota
- Connect resources
- Track costs

Open in management center

Nail the basics with these steps

Management center

This part is also largely the same as Azure OpenAI Service.

The screenshot shows the Azure AI Foundry Overview page for project 1923306880-1774. On the left, there's a sidebar with various project management options like Model catalog, Playgrounds, AI Services, and different build and customize sections. The main area has tabs for Overview, Model catalog, Playgrounds, AI Services, and so on. The Overview tab is active. It displays the project ID '1923306880-1774', a placeholder for a project description, and a 'Endpoints and keys' section. This section contains an 'API Key' field (redacted), a list of 'Included capabilities' (Azure AI inference, Azure OpenAI Service, Azure AI Services), and a link to the 'Azure OpenAI Service endpoint' (https://aifoundarytest9323385592.openai.azure.com/). Below this is an 'API documentation' button. To the right is a 'Project details' panel showing project connection string, subscription information, subscription ID, location, and manage project settings. A red callout box highlights the 'Endpoints and keys' section, and a red arrow points from it to the text 'This part is also largely the same as Azure OpenAI Service.' at the top right. At the bottom, there's a note 'Nail the basics with these steps' and a 'Management center' button.

2. Select a model
from the model catalog

[!\[\]\(fcd24c7d945c5138d054baf3d01d90e2_img.jpg\) Overview](#)[!\[\]\(0aa53286f9a450bc93b78d504d842c2f_img.jpg\) Model catalog](#)[Playgrounds](#)[AI Services](#)[Build and customize](#)[Agents PREVIEW](#)[Templates PREVIEW](#)[Fine-tuning](#)[Prompt flow](#)[Assess and improve](#)[Tracing PREVIEW](#)[Evaluation](#)[Safety + security](#)[My assets](#)[Models + endpoints](#)[Data + indexes](#)[Web apps](#)[Management center](#)

Find the right model to build your custom AI solution

[Help](#)

Announcements

Introducing Llama 4 models



Build highly personalized experiences for every use case at a lower cost

[Check out models](#)[Read blog](#)

Responses API with CUA



Create your agent with computer-use-preview

[Check out model](#)[Read blog](#)

News from Mistral AI: Mistral Small 3.1



Enhanced Mistral Small 3 with multimodal capabilities and a 128k context length.

[Check out model](#)

NVIDIA NIM now available!

TCO and performance optimization for your AI applications

[Check out models](#)

Model leaderboards

See what models are performing best in different criteria.

[Browse leaderboards](#)

Quality

- 1 o1
- 2 DeepSeek-R1
- 3 gpt-4.5-preview

Cost

- 1 Minstral-3B
- 2 Mistral-Nemo
- 3 mistral-small-2503

Throughput

- 1 Llama-3.2-1B-Instruct
- 2 o3-mini
- 3 o1-mini

[Collections](#) [Industry](#) [Capabilities](#) [Deployment options](#) [Inference tasks](#) [Fine-tuning tasks](#)[Licenses](#)[Compare models](#)[Search](#)**Models 1900**

Find the right model to build your custom AI solution

[Help](#)[Overview](#)[Model catalog](#)[Playgrounds](#)[AI Services](#)[Build and customize](#)[Agents PREVIEW](#)[Templates PREVIEW](#)[Fine-tuning](#)[Prompt flow](#)[Assess and improve](#)[Tracing PREVIEW](#)[Evaluation](#)[Safety + security](#)[My assets](#)[Models + endpoints](#)[Data + indexes](#)[Web apps](#)[Management center](#)[Quality](#)

1 o1

2 DeepSeek-R1

3 gpt-4.5-preview

[Cost](#)

1 Minstral-3B

2 Mistral-Nemo

3 mistral-small-2503

[Throughput](#)

1 Llama-3.2-1B-Instruct

2 o3-mini

3 o1-mini

[Collections](#) [Industry](#) [Capabilities](#) [Deployment options](#) [Inference tasks](#) [Fine-tuning tasks](#)[Licenses](#)[Search](#)

Models 1900

[gpt-4.5-preview](#)

Chat completion

[o3-mini](#)

Chat completion

[Llama-4-Scout-17B-16E](#)

Chat completion

[Llama-4-Scout-17B-16E-Instr...](#)

Chat completion

[Llama-4-Maverick-17B-128E...](#)

Chat completion

[DeepSeek-V3](#)

Chat completion

[DeepSeek-R1](#)

Chat completion

[Phi-4-mini-instruct](#)

Chat completion

[Phi-4-multimodal-instruct](#)

Chat completion

[Phi-4](#)

Chat completion

[mistral-small-2503](#)

Chat completion, Completions, C...

[gpt-4o-mini-audio-preview](#)

Audio generation

[gpt-4o-mini-realtime-preview](#)

Audio generation

[o1](#)

Chat completion

[o1-mini](#)

Chat completion

[ant-4o](#)[ant-4o-mini](#)[ant-4o-audio-preview](#)[ant-4o-realtime-preview](#)

DeepSeek-R1

[Deploy](#)[Details](#) Benchmarks Existing deployments License[Learn more: \[original model announcement\]](#)

DeepSeek-R1 excels at reasoning tasks using a step-by-step training process, such as language, scientific reasoning, and coding tasks. It features 671B total parameters with 37B active parameters, and 128k context length.

DeepSeek-R1 builds on the progress of earlier reasoning-focused models that improved performance by extending Chain-of-Thought (CoT) reasoning. DeepSeek-R1 takes things further by combining reinforcement learning (RL) with fine-tuning on carefully chosen datasets. It evolved from an earlier version, DeepSeek-R1-Zero, which relied solely on RL and showed strong reasoning skills but had issues like hard-to-read outputs and language inconsistencies. To address these limitations, DeepSeek-R1 incorporates a small amount of cold-start data and follows a refined training pipeline that blends reasoning-oriented RL with supervised fine-

[See more](#)

Data, media and languages

Property	Description	
Supported data types	Inputs	Outputs
	text	text

Property	Description
Supported languages	en, zh

Transparency

Model alignment

Quick facts

 DeepSeek-R1[chat-completion](#)

Last trained

Not available

Benchmarks

0.91 Quality index
AI quality**2.36** USD per 1M token
Estimated cost

Model ID

Reference this model ID when deploying the model in

<azureml://registries/azureml-deepseek/models/DeepSeek-R1/versions/1>

DeepSeek-R1

PREVIEW

Deploy

Details

Benchmarks

Overview Pricing and terms



DeepSeek-R1 is provided by Microsoft as a First Party Consumption Service.

[Learn more about Models as a Service.](#)

Project Name

1923306880-1774

ⓘ Serverless API is available in the following regions: [eastus2, westus3, northcentralus, eastus, southcentralus, westus].

Deployment name *

DeepSeek-R1-kvjgv

Content filter (preview)

 Enabled

ⓘ Content filtering uses default configuration and is billed through Azure AI Content Safety. [Learn more](#)

Deploy

Cancel

Quick facts

DeepSeek-R1

chat-completion

Last trained

Not available

Benchmarks

0.91 Quality index
AI quality2.36 USD per 1M token
Estimated cost

Model ID

Reference this model ID when deploying the model in

azureml://registries/azureml-deepseek/models/DeepSeek-R1/versions/1

Transparency

Model alignment

Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents

PREVIEW

Templates

PREVIEW

Fine-tuning

Prompt flow

Assess and improve

^

Tracing

PREVIEW

Evaluation

Safety + security

My assets

^

Models + endpoints

Data + indexes

Web apps

Management center

DeepSeek-R1 PREVIEW

Deploy

Details

Benchmarks

Existing deployments

License

Learn more: [original r...

DeepSeek-R1 excels at tasks. It features 671B...

DeepSeek-R1 builds on Thought (CoT) reasoning, chosen datasets. It even but had issues like han...

See more

Data, media

Property

Supported data typ...

text

text

Supported languages

en, zh

Transparency

Model alignment

Model SKU	Input Pricing in USD (1K Tokens)	Output Pricing in USD (1K Tokens)
DeepSeek-R1 Global	\$0.00135	\$0.0054
DeepSeek-R1 Regional	\$0.001485	\$0.00594

Overview Pricing and terms

Pricing

[See the pricing details](#)

Legal

License information can be found on the model catalog model details page

Deploy

Cancel

Benchmarks

0.91 Quality index	2.36 USD per 1M tokens
AI quality	Estimated cost

Model ID
Reference this model ID when deploying the model in
`azureml://registries/azureml-deepseek/models/DeepSeek-R1/versions/1`

Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents PREVIEWTemplates PREVIEW

Fine-tuning

Prompt flow

Assess and improve

Tracing PREVIEW

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Management center

DeepSeek-R1-kvjgv

[Overview](#)[Model catalog](#)[Playgrounds](#)[AI Services](#)[Build and customize](#)[Agents PREVIEW](#)[Templates PREVIEW](#)[Fine-tuning](#)[Prompt flow](#)[Assess and improve](#)[Tracing PREVIEW](#)[Evaluation](#)[Safety + security](#)[My assets](#)[Models + endpoints](#)[Data + indexes](#)[Web apps](#)[Management center](#)

Details

[Open in playground](#)[Refresh](#)[Edit](#)[Delete](#)

Endpoint

Target URI

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com/>

Key

.....

Compute type

Consumption

Swagger URI

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com/>

API Routes

Azure AI model inference: Chat Completion

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com/>

deepseek-r1: Chat Completion

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com/>

Deployment info

Name

DeepSeek-R1-kvjgv

Language

Python

SDK

Azure AI Inference...

Authentication type

Key Authentication

Get Started

Below are example code snippets for a few use cases. For additional information about Azure AI Inference SDK, see full [documentation](#) and [samples](#).

1. Authentication using API Key

For Serverless API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service. In this sample endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

Azure inferencing package includes a client for Chat Completion, [ChatCompletionsClient](#). A client can be authenticated using the API key. The code sample creates and authenticates a synchronous ChatCompletionsClient:

```
from azure.ai.inference import ChatCompletionsClient
from azure.core.credentials import AzureKeyCredential

# For Serverless API or Managed Compute endpoints
client = ChatCompletionsClient(
    endpoint="https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com",
    credential=AzureKeyCredential("<API_KEY>")
```

3. Run a chat
from the chat playground

DeepSeek-R1-kvjgv

 Overview Model catalog Playgrounds AI ServicesBuild and customize  Agents  Templates  Fine-tuning Prompt flowAssess and improve  Tracing  Evaluation Safety + securityMy assets  Models + endpoints Data + indexes Web apps Management center

Details

 Open in playground Refresh Edit Delete

Endpoint

Target URI

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com>

Key

.....

Compute type

Consumption

Swagger URI

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com>

API Routes

Azure AI model inference: Chat Completion

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com>

deepseek-r1: Chat Completion

<https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com>

Deployment info

Name

DeepSeek-R1-kvjgv

Language

Python

SDK

Azure AI Inference...

Authentication type

Key Authentication

Get Started

Below are example code snippets for a few use cases. For additional information about Azure AI Inference SDK, see full [documentation](#) and [samples](#).

1. Authentication using API Key

For Serverless API Endpoints, deploy the Model to generate the endpoint URL and an API key to authenticate against the service. In this sample endpoint and key are strings holding the endpoint URL and the API Key.

The API endpoint URL and API key can be found on the Deployments + Endpoint page once the model is deployed.

Azure inferencing package includes a client for Chat Completion, [ChatCompletionsClient](#). A client can be authenticated using the API key. The code sample creates and authenticates a synchronous ChatCompletionsClient:

```
from azure.ai.inference import ChatCompletionsClient  
from azure.core.credentials import AzureKeyCredential  
  
# For Serverless API or Managed Compute endpoints  
client = ChatCompletionsClient(  
    endpoint="https://DeepSeek-R1-kvjgv.eastus2.models.ai.azure.com",  
    credential=AzureKeyCredential("<API_KEY>")  
)
```

← Chat playground

 View code

Setup

 Hide

Deployment *

Create new deployment

DeepSeek-R1-kvjgv

Parameters

Chat session

 Overview Model catalog Playgrounds AI Services

Build and customize ⏹

 Agents PREVIEW Templates PREVIEW Fine-tuning Prompt flow

Assess and improve ⏹

 Tracing PREVIEW Evaluation Safety + security

My assets ⏹

 Models + endpoints Data + indexes Web apps Management center

Start typing here



School | AI-Foundry-test0 - Microsoft Az... | Chat playground - Azure AI Found... | DeepSeek R1: Improved Perform... | +

https://ai.azure.com/playground/chat?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foundry-hub0/providers/Microsoft.MachineLearningSer...

Azure AI Foundry / 1923306880-1774 / Playgrounds / Chat playground

Docs All hubs + projects 5 Project 1923306880-1774 8 Help

← Chat playground

View code

Setup Hide

Deployment * Create new deployment

DeepSeek-R1-kvjgv

Parameters

Chat session

Could you tell me what is good about the University of Tokyo?

Okay, so I need to figure out what's good about the University of Tokyo. Let me start by recalling what I know. The University of Tokyo, often called Todai, is a prestigious university in Japan. I think it's one of the top universities in Asia. But why exactly is it considered good? Well, first, maybe its academic reputation. It's probably highly ranked globally. I remember seeing it in rankings like QS or Times Higher Education. Then, there's research output. Top universities usually have strong research programs. Maybe Todai has notable research facilities or contributions in various fields. They might have produced Nobel laureates or significant scientific breakthroughs. I think Japan has a strong tradition in technology and engineering, so Todai's engineering department might be particularly strong.

Another aspect could be the faculty. Having renowned professors and experts in their fields would attract students. Also, the alumni network. Graduates from Todai probably go on to hold influential positions in government, business, and academia, which can be beneficial for networking.

Campus and facilities might be another point. Maybe the university has a beautiful campus, especially since it's in Tokyo, which is a major city. Access to resources like libraries, labs, and partnerships with industries could be a plus. Also, international collaborations. Many top universities have exchange programs and partnerships with

Management center

Start typing here