

# 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

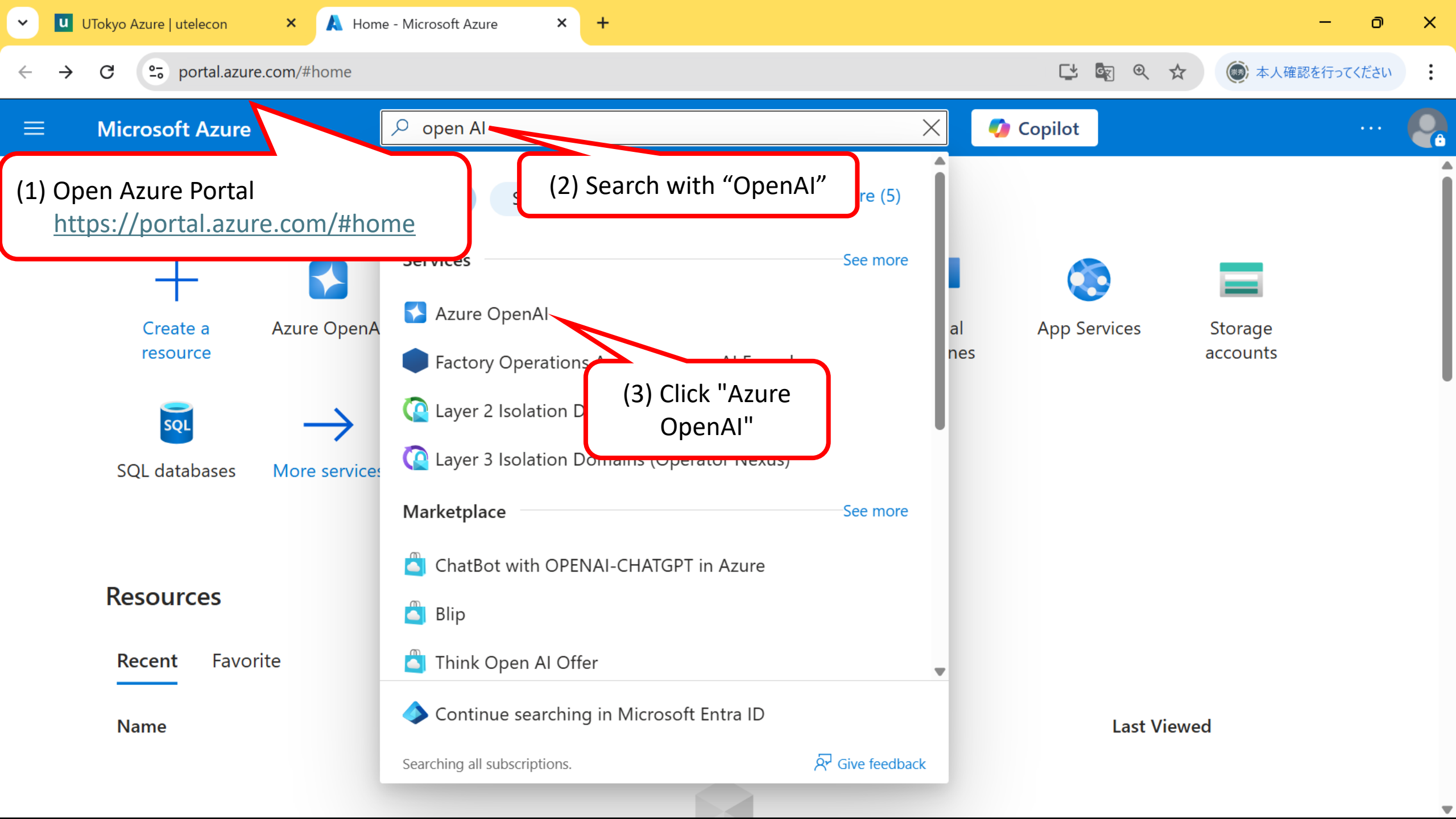
This slide is created for the April 18, 2025 UTokyo Azure Usage Briefing.

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. How to 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



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

(2) Search with "OpenAI"

(3) Click "Azure OpenAI"

U Tokyo Azure | utelecon

Azure AI services - Microsoft Az

← → ↺

https://portal.azure.com/#view/Microsoft\_Azure\_ProjectOxford/CognitiveServicesHub/~ /OpenAI

Google レンズ 📄 🗨 🔍 ☆

👤 本人確認を行ってください

⋮

☰

Microsoft Azure

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

Copilot

📧 🔔 ⚙️ ? 🗨

2573969490@utac.u-to...  
THE UNIVERSITY OF TOKYO (UNI...

👤

Home > Azure AI services

 Azure AI services | Azure OpenAI

📌 ⋮

✕

🔍 Search

✕

⏪

+

 Create

🔧 Manage deleted resources

⚙️ Manage view ▾

🔄 Refresh

⬇️ Export to CSV

🔗 Open query

🏷️ Assign tags

🗑️ Delete

Filter for any field...

Subscription equals all

Type equals all

Resource group equals all ✕

+

 Add filter

▾ More (1)

Showing 0 to 0 of 0 records.

No grouping ▾

☰ List view ▾

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

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

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 to code. Call Centers, Virtual Assistants, Accessibility, Assistants API, Fine Tuning capabilities and experiences. The service can be scaled through Azure OpenAI Service.

A resource group is a container for a collection of resources for an Azure solution.

[Learn more](#)

## Project Details

Subscription \*

Resource group \*

**You will be asked to create a resource group.**

Name \*

openai\_test1

OK Cancel

Create new

**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 subscription.

U Tokyo Azure | utelecon

Create Azure OpenAI - Microsoft

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

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2573969490@utac.u-to...

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI

experiences. The service can be scaled through Standard (t

Learn more

Project Details

Subscription \* ⓘ

1923306880-Yama

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

Previous

Next

Give feedback

Resource group is important for cost management:

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

# Product Availability by Region

With 60+ announced regions, more than any other cloud provider, Azure makes it easy to choose the datacenter and region for your workloads.

© Microsoft 2024. All rights reserved. Release dates, features, and requirements are subject to change prior to final announcement. MICROSOFT MAKES NO WARRANTIES, EXPRESSED, IMPLIED, OR STATUTORY, AS TO THE INFORMATION ON THIS PAGE.

\* Reserved Access Regions: Certain regions are access restricted to support specific customer scenarios, for example in-country disaster recovery. See [Business continuity and disaster recovery \(BCDR\)](#).

- [This page](#) shows resource availabilities in each region.
- SKU stands for Stock Keeping Unit.

Select a Geography

All

● Generally Available

■ In Public Preview

| Geographies                |                               | Japan      |            | Korea         |             |
|----------------------------|-------------------------------|------------|------------|---------------|-------------|
| Products                   | Product SKU                   | Japan East | Japan West | Korea Central | Korea South |
|                            | Language Understanding (LUIS) | ●          | ●          | ●             |             |
|                            | Video Indexer                 | ●          | ●          | ●             |             |
|                            | Azure AI Metrics Advisor      | ●          |            | ●             |             |
|                            | Azure OpenAI Service          | ●          |            | ●             |             |
|                            | Speaker recognition           |            |            |               |             |
|                            | Azure AI Speech               | ●          | ●          | ●             |             |
|                            | Language service              | ●          |            |               |             |
|                            | Azure Video Analyzer          |            |            |               |             |
|                            | Azure AI Video Indexer        | ●          | ●          | ●             |             |
| Azure API Center           |                               |            |            |               |             |
| Azure Active Directory B2C |                               |            |            |               |             |
| Azure Advisor              |                               |            |            |               |             |

“Japan West” doesn't have resources for OpenAI Service

Home > Azure AI services | Azure OpenAI >

# Create Azure OpenAI ...

- Basics
- 2 Network**
- 3 Tags
- 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.

# 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 |

U Tokyo Azure | utelecon

Create Azure OpenAI - Microso

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

Microsoft Azure

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

Copilot

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           | gprr-4o-test-1              |
| Pricing tier   | Standard S0                 |

Previous

Next

Create

Give feedback

UTokyo Azure | utelecon

Microsoft.CognitiveServicesOpenAI-20250405160148

portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai\_test1%2Fproviders%2FMicrosoft.CognitiveServices%2Fdeployments%2FMicrosoft.CognitiveServicesOpenAI-20250405160148

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2

2573969490@utac.u-to...

THE UNIVERSITY OF TOKYO (UNI...

Home >

Deployment

Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview

Deployment in progress...

Deployment to resource group 'openai\_test1' is in progress.

Search

DeleteCancelRedeployDownloadRefresh

Overview

Inputs

Outputs

Template

Deployment is in progress

Deployment name : Microsoft.CognitiveService... Start time : 4/5/2025, 4:08:49 PM

Subscription : 1923306880-YamakataTrai... Correlation ID : 02dab288-1f64-4c47-9722...

Resource group : openai\_test1

Deployment details

|  | Resource       | Type                            | Status  |  |
|--|----------------|---------------------------------|---------|--|
|  | gpvt-4o-test-1 | Microsoft.CognitiveServices/acc | Created |  |

Give feedback

Tell us about your experience with deployment

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 >

U Tokyo Azure | utelecon

Microsoft.CognitiveServicesOpenAI-20250405160148

portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGrou...

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2573969490@utac.u-to...  
THE UNIVERSITY OF TOKYO (UNI...

Home >

Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name : Microsoft.CognitiveService...

Subscription : 1923306880-YamakataTrai...

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

UTokyo Azure | utelecon

openai\_test1 - Microsoft Azure

portal.azure.com/#@univtokyo.onmicrosoft.com/resource/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai\_test1/overview

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2

2573969490@utac.u-to...  
THE UNIVERSITY OF TOKYO (UNI...

Home > Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview >

openai\_test1  
Resource group

Search

CreateManage viewDelete resource groupRefreshExport to CSVOpen queryAssign tags

Overview

Activity logAccess control (IAM)TagsResource visualizerEventsSettingsCost ManagementMonitoringAutomationHelp

ResourcesRecommendations

Filter for any field...Type equals allLocation equals allAdd filter

Showing 1 to 1 of 1 records. Show hidden typesNo grouping

| Name           | Type         | Location   |
|----------------|--------------|------------|
| gprt-4o-test-1 | Azure OpenAI | Japan East |

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

U Tokyo Azure | utelecon

gprt-4o-test-1 - Microsoft Azure

portal.azure.com/#@univtokyo.onmicrosoft.com/resource/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai\_test1/providers/Micro... 本人確認を行ってください

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2573969490@utac.u-to...  
THE UNIVERSITY OF TOKYO (UNI...

Home > Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview > openai\_test1 >

gprt-4o-test-1

Azure OpenAI

Go to Azure AI Foundry portal Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Resource Management

Security

Monitoring

Automation

Help

Essentials

Resource group (move)  
openai\_test1

Status  
Active

Location  
Japan East

Subscription (move)  
1923306880-YamakataTraining

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

Tags (edit)  
Add tags

API Kind  
OpenAI

Pricing tier  
Standard

Endpoints  
Click here to view endpoints

Manage keys  
Click here to manage keys

JSON View

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

U Tokyo Azure | utelecon

gprt-4o-test-1 - Microsoft Azure

portal.azure.com/#@univtokyo.onmicrosoft.com/resource/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai\_test1/providers/Micro... 本人確認を行ってください

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2573969490@utac.u-to...  
THE UNIVERSITY OF TOKYO (UNI...

Home > Microsoft.CognitiveServicesOpenAI-20250405160148 | Overview > openai\_test1 >

gprt-4o-test-1

Azure OpenAI

Search

Go to Azure AI Foundry portal

Delete

Add tags

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Resource Management

Security

Monitoring

Automation

Help

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 REST APIs, Python SDK, or our web-based interface in the Azure AI Foundry portal.  
[Learn More](#)

Explore and deploy

Explore and deploy the generative AI models, craft unique prompts for your use cases, and fine-tune select models.

Explore Azure AI Foundry portal

U Tokyo Azure | utelecon

gprt-4o-test-1 - 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 | Azure OpenAI Service

gprt-4o-test-1

Chat playground

All resources5⚙️😊gprt-4o-test-1(japaneast, S0)

Home

Get started

Model catalog

Playgrounds

Chat

Assistants

PREVIEW

Language

PREVIEW

Real-time audio

PREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI Service

PREVIEW

Evaluation

Stored completions

PREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Chat playground

</> View code

📦 Deploy

➞ Import

➞ Export

🧩 Prompt samples

🗨️ Filters feedback

Setup

Hide

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)

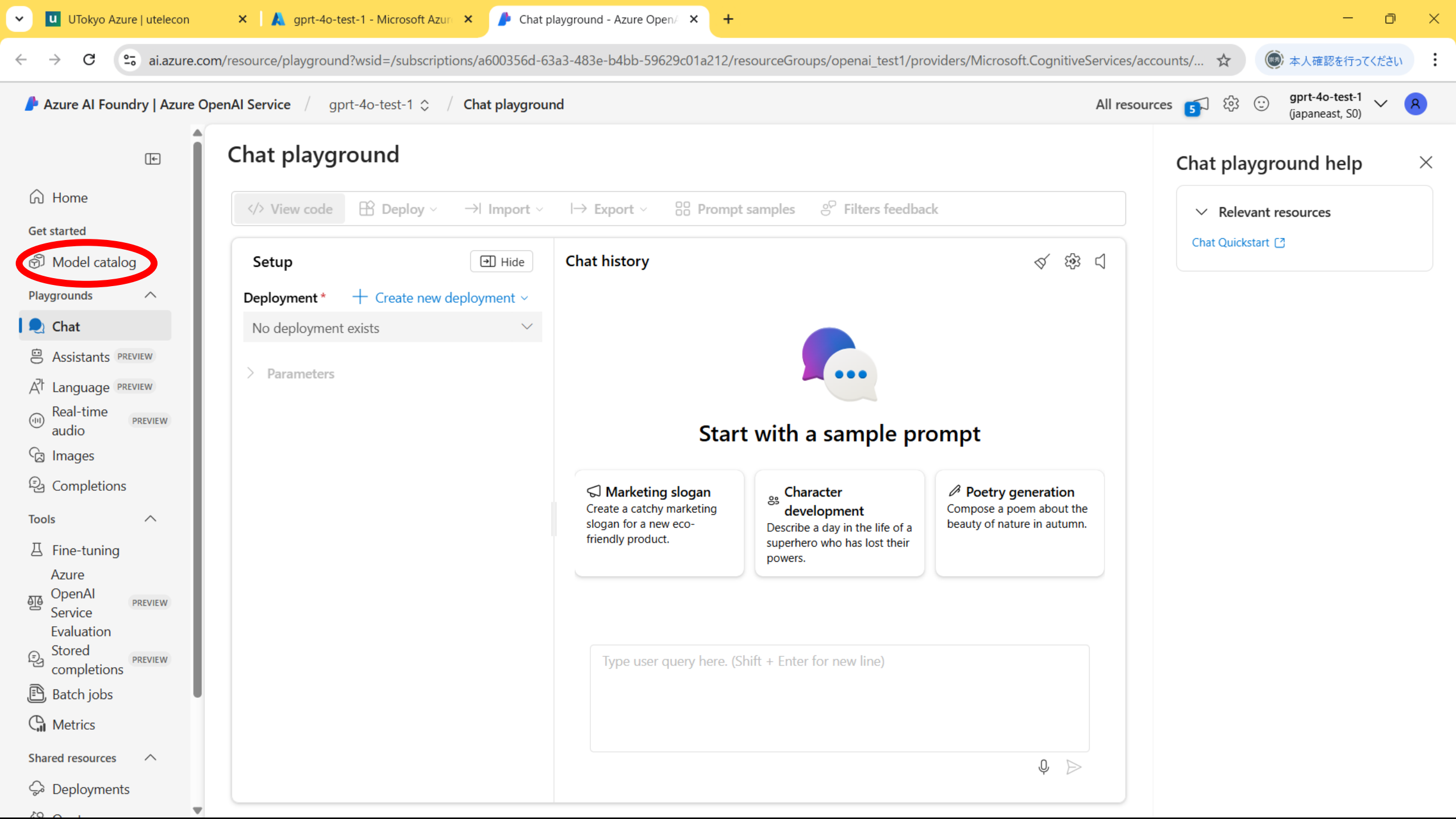
🗣️➞

Chat playground help

Relevant resources

[Chat Quickstart](#)

2. Select a model  
from the model catalog



U Tokyo 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 / gprt-4o-test-1 / Model catalog

All resources 5 gprt-4o-test-1 (japaneast, S0)

Home

Get started

Model catalog

Playgrounds

Chat

Assistants PREVIEW

Language PREVIEW

Real-time audio PREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI Service PREVIEW

Evaluation

Stored completions PREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Find the right model to build your custom AI solution

Announcements

Responses API with CUA

Create your agent with computer-use-preview

Check out modelRead blog

Introducing GPT-4.5 Preview

The latest GPT model that excels at diverse text and image tasks

Check out modelRead blog

4o Mini Audio is here!

Advanced audio at a fraction of the cost

Check out modelsRead blog

Inference tasks

Fine-tuning tasks

Search

Models 27

gpt-4.5-preview

Chat completion

o3-mini

Chat completion

computer-use-preview

Responses

gpt-4o-mini-audio-preview

Audio generation

gpt-4o-mini-realtime-preview

Audio generation

o1

Chat completion

o1-mini

Chat completion

gpt-4o

Chat completion

gpt-4o-mini

Chat completion

gpt-4o-audio-preview

gpt-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

## 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](#)

### Quick facts

OpenAI

 **gpt-4o-mini-realtime-preview**

audio-generation

Last trained  
October 2023

## Model Versions

[Learn more about regional availability](#)

Japan East

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

Actually, there are several versions of **GPT-4o**.

### Model ID

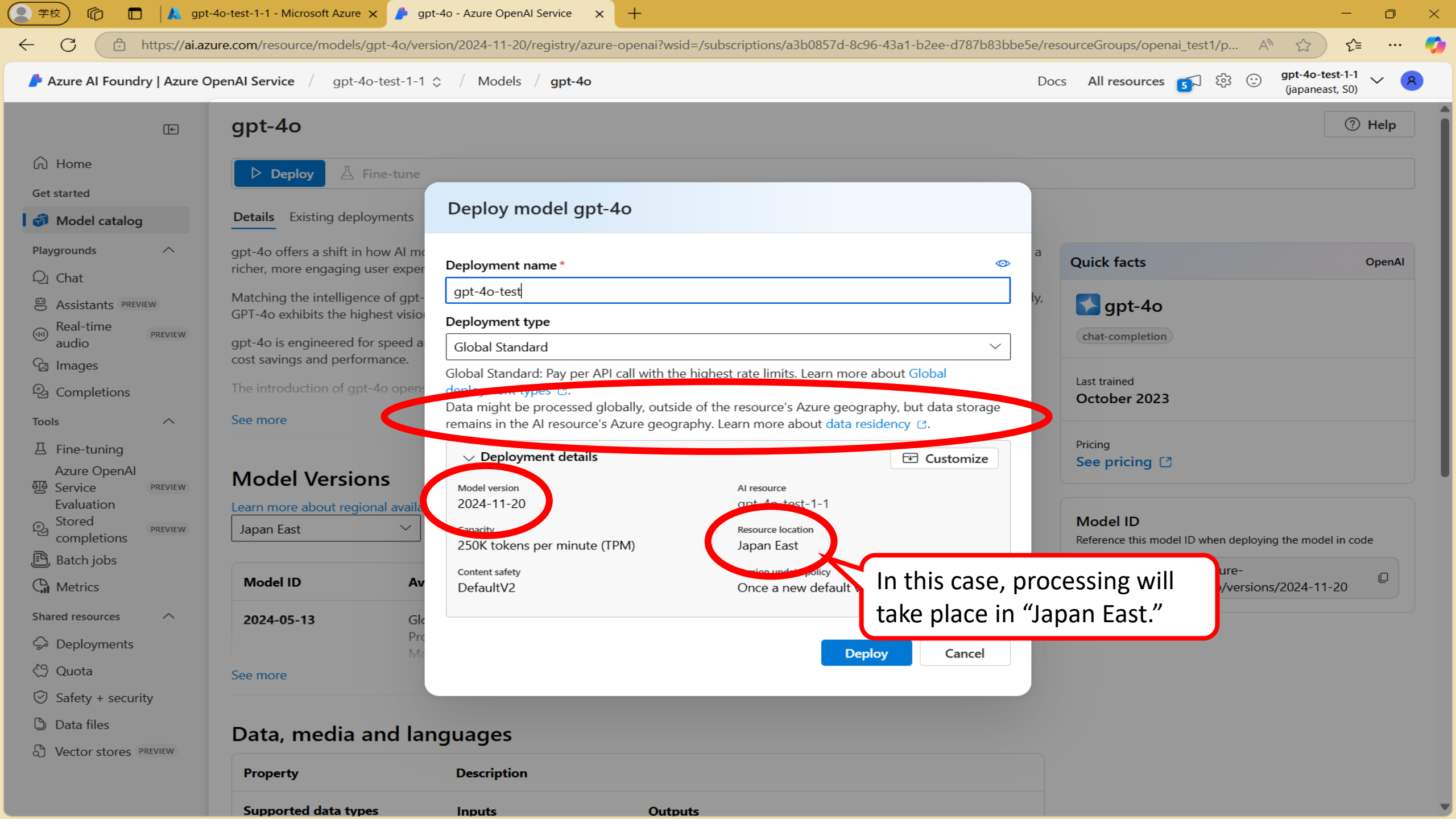
Reference this model ID when deploying the model in code

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

## Models help

### Relevant resources

[Azure OpenAI Service models](#)[Azure OpenAI Service model deprecations and retirements](#)



# How about gpt-o1? (as of February 2025) [[Azure documentation](#)]

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

Registration is required  
for the O1 system.

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

The image shows the Azure AI Foundry console interface for the o1 model. A red circle highlights the 'デプロイ' (Deploy) button. A blue arrow points from this button to a separate screenshot of the 'Request Access to the Azure OpenAI Service' form.

**モデル o1 をデプロイする**

❗ このモデルを使用するには登録が必要です  
アクセスは、Microsoft の資格条件に従って付与されます。現時点では、アクセスの資格がない可能性があります。既にアクセスを要求している場合は、申請を確認していますので、しばらくお待ちください。  
[アクセスの要求](#)

**Fill out the form to register**

**Request Access to the Azure OpenAI Service**  
o\* family of models

\* 必須

Please read all instructions carefully and complete form as instructed

The Azure OpenAI Service o\* family of models require registration. Due to high demand, customers may remain on the waitlist until space is available.

Other models (listed [here](#)) do not require registration. [Learn more about limited access to Azure OpenAI Service here.](#)

Please visit the [Overview of Responsible AI Practices](#) and [Azure OpenAI's transparency note](#), which provides information and guidelines for responsible use of the service as well as system limitations that may be applicable to your scenario.

**Possible causes for a denied application:**

1) You submitted an invalid subscription2) You submitted with a personal email (Example: [@gmail.com](#), [@yahoo.com](#), [@hotmail.com](#), etc.)

**Updates about your request:** Please add [maccount@microsoft.com](#) and [csqgate@microsoft.com](#) to your Safe Senders list to ensure you receive all correspondence pertaining to your registration, including the verification email from Microsoft Vetting Service.

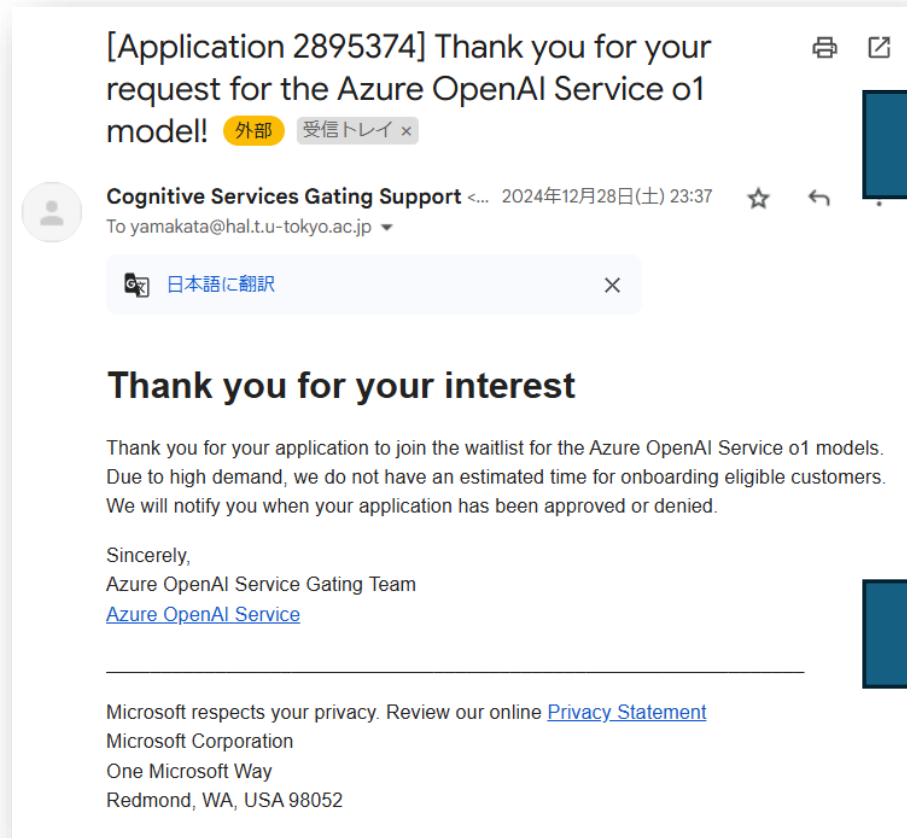
1  
Your First Name \*

回答を入力してください

# Approved after a while

In my case...

- o1-preview and o1-mini are approved immediately.
- o1 is approved 18 days after the request.



After 30 minutes



After 18 days



学校

gpt-4o-test-1-2 - Microsoft Azure

gpt-4o-test - Azure OpenAI Service

https://ai.azure.com/resource/deployments/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai\_test1%2Fproviders%2FMicrosoft.CognitiveSer...

DocsAll resourcesgpt-4o-test-1-2 (japaneast, S0)

HomeGet startedModel catalogPlaygroundsChatAssistantsPREVIEWReal-time audioPREVIEWImagesCompletionsToolsFine-tuningAzure OpenAI ServicePREVIEWEvaluationStored completionsPREVIEWBatch jobsMetricsShared resourcesDeploymentsQuotaSafety + securityData filesVector storesPREVIEW

←gpt-4o-test

DetailsMetricsRisks & SafetyConsume

Open in playgroundEditDelete

Endpoint

Target URIhttps://gpt-4o-test-1-2.openai.azure.com/openai/deployments/g...Key.....

Deployment info

|                                          |                                                              |
|------------------------------------------|--------------------------------------------------------------|
| Namegpt-4o-test                          | Provisioning stateSucceeded                                  |
| Deployment typeGlobal Standard           | Created on2025-04-08T06:38:31.2586957Z                       |
| Created by1923306880@utac.u-tokyo.ac.jp  | Modified onApr 8, 2025 3:38 PM                               |
| Modified by1923306880@utac.u-tokyo.ac.jp | Version update policyOnce a new default version is available |
| Rate limit (Tokens per minute)250,000    | Rate limit (Requests per minute)1,500                        |
| Model namegpt-4o                         | Model version2024-11-20                                      |
| Life cycle statusGenerallyAvailable      | Date createdDec 3, 2024 9:00 AM                              |

LanguagePythonSDKAzure OpenAI SDKAuthentication typeKey Authentication

Get Started

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

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 you 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

gpt-4o-test - Azure OpenAI Service

https://ai.azure.com/resource/deployments/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai\_test1%2Fproviders%2FMicrosoft.CognitiveSer...

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

Docs All resources

gpt-4o-test-1-2 (japaneast, S0)

Home

Get started

Model catalog

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

← gpt-4o-test

Details Metrics Risks & Safety Consume

Open in playground Edit Delete

Endpoint

Target URI

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

Key

.....

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](#).

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 you 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

gpt-4o-test - Azure OpenAI Service

+

https://ai.azure.com/resource/deployments/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceGroups%2Fopenai\_test1%2Fproviders%2FMicrosoft.CognitiveSer...

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

Home

Get started

Model catalog

Playgrounds

Chat

Assistants PREVIEW

Real-time audio PREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI Service PREVIEW

Evaluation PREVIEW

Stored completions PREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Quota

Safety + security

Data files

Vector stores PREVIEW

← gpt-4o-test

Details Metrics Risks & Safety Consume

Open in playground Edit Delete

Endpoint

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

Key  
.....

Deployment info

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

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 you 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
```



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

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 gpt-4o-test-1-2 (japaneast, S0)

# Chat playground Help

[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

## 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

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

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.

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.

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

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 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:2024-11-01) Please select the 'Apply changes' button to save the deployment.

Give the model instructions and context. You can use the 'Generate prompt' button to generate a prompt for the model.

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

Apply changes Generate prompt

+ Add section

## Chat history

Response format Text

### Start with a sample prompt

historical fiction

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

Dialogue creation

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

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

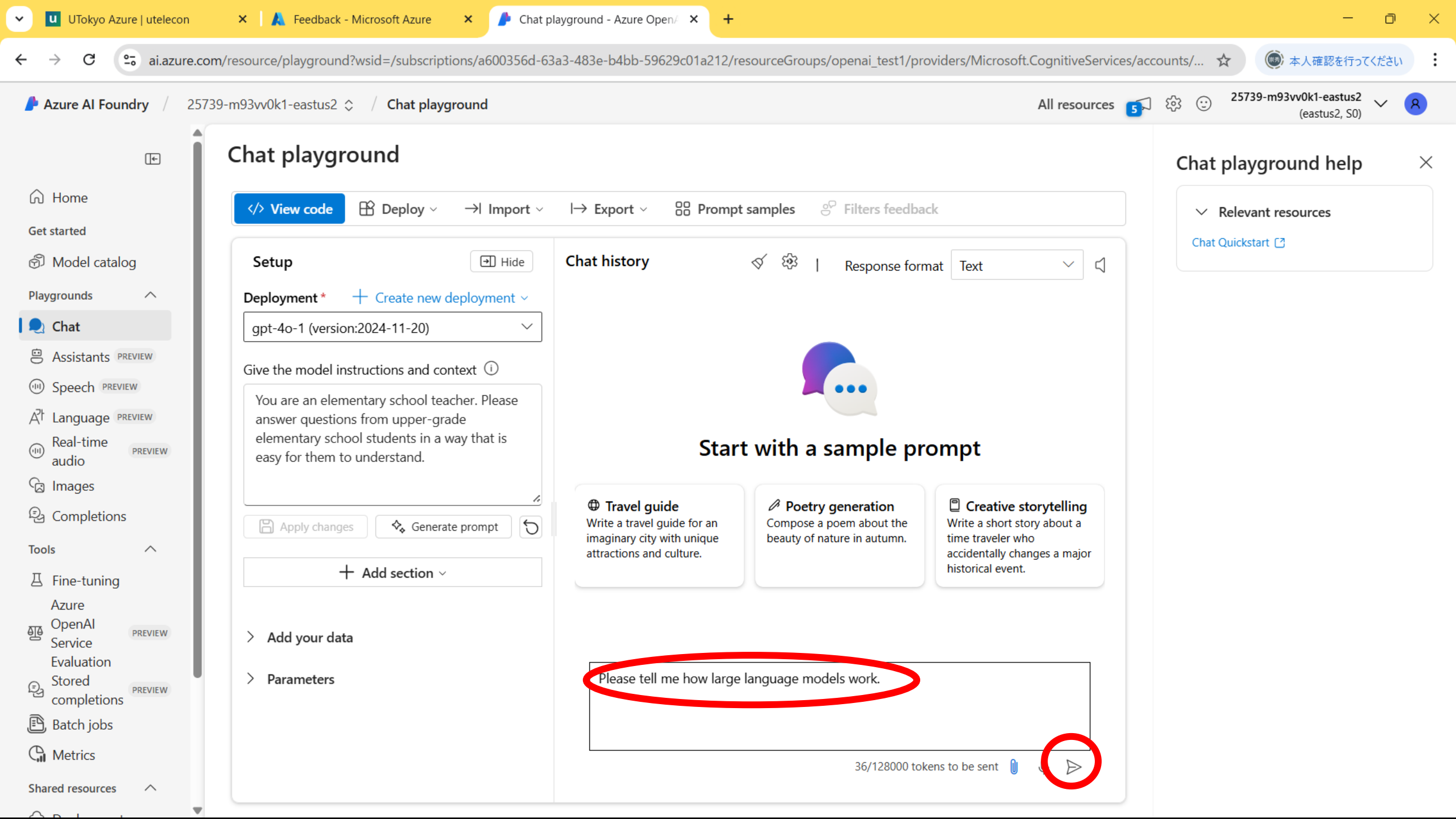
0/128000 tokens to be sent

### Update system message?

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

☐ Do not show this again

**Continue** Cancel



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/acc...🔍🔖🌟🔍...

Azure AI Foundry | Azure OpenAI Service

gpt-4o-test-1-2

Chat playground

Docs

All resources

5

⚙️

😊

gpt-4o-test-1-2  
(japaneast, S0)

▼

👤

🔖

Home

Get started

Model catalog

Playgrounds

Chat

AssistantsPREVIEW

Real-time audioPREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI ServicePREVIEW

Evaluation

Stored completionsPREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Chat playground

🔍 Help

View code

Deploy

Import

Export

Prompt samples

Filters feedback

Setup

Hide

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

Chat history

🗑️⚙️

Response format

Text

🔊

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

🔒🔊🔍

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

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 gpt-4o-test-1-2 (japaneast, S0)

# Chat playground

Help

View code Deploy Import Export Prompt samples Filters feedback

## Setup

Hide

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**

## Chat history

Response format Text

Please tell me how large language models work.

Of course! Let's

A large language

program that us

It's a little like a

Here's how it works:

- 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

If the model stops outputting a response before it's complete, adjust the "parameters."

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/acc...

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

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

HomeGet startedModel catalogPlaygroundsChatAssistants PREVIEWReal-time audio PREVIEWImagesCompletionsToolsFine-tuningAzureOpenAI Service PREVIEWEvaluationStored completions PREVIEWBatch jobsMetricsShared resourcesDeployments

Chat playground

Help

View codeDeployImportExportPrompt samplesFilters feedback

Add your data

Parameters

Past messages included10

Max response2400

Temperature0.7

Top P0.95

Stop sequence

Frequency penalty0

Presence penalty0

Chat history

Response formatText

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. How to use the service via Web API from your code

I want to send a prompt to the model and receive the response via Web API from a Google Colaboratory notebook?

## 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 the “Target URI”

Rewrite to the “Key”

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.Cognit...

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)

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

## gpt-4o-test

Details Metrics

Open in playground

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

Endpoint

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

Key  
..... 

Copy this with  and set as the "API\_KEY" in your code.

Deployment info

|                 |                 |                              |
|-----------------|-----------------|------------------------------|
| Name            | gpt-4o-test     | Succeeded                    |
| Deployment type | Global Standard | Created on                   |
|                 |                 | 2025-04-08T06:38:31.2586957Z |

Language  
Python

SDK  
Azure OpenAI SDK

### Get Started

Below are example code snippets for a few use cases. For additional information, see [samples](#).

#### 1. Authentication using API Key

For OpenAI API Endpoints, deploy the Model to generate the endpoint URL 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 + Endpoints page.

To create a client with the OpenAI SDK using an API key, initialize the client with the endpoint and API key.

```
api_version="2024-12-01-preview",  
endpoint="https://gpt-4o-test-1-2.openai.azure.com/"
```

**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!

```
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)

#### 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 API:

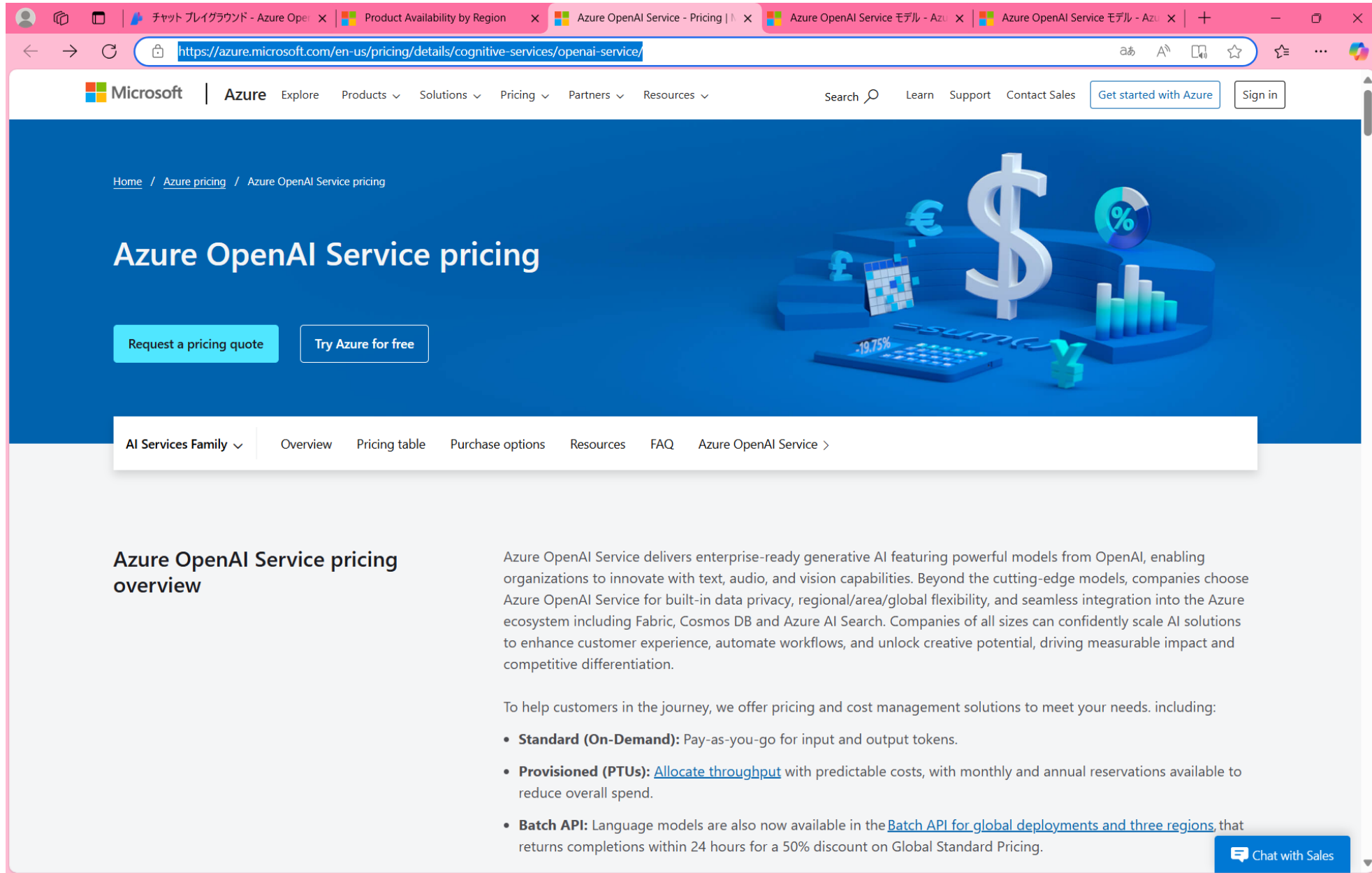
```
```python
!pip install requests
```
```

## 5. Other Topics

# 5. 1 Pricing

Estimating the Cost

# Azure OpenAI Service pricing ([see here](#))

A screenshot of a web browser displaying the Azure OpenAI Service pricing page. The browser's address bar shows the URL 'https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/'. The page features a blue header with the Microsoft logo and navigation links. The main content area has a blue background with a large 3D graphic of financial symbols (dollar, euro, yen, percentage, and a bar chart) and a calculator. Below this, there's a white navigation bar with links like 'AI Services Family', 'Overview', 'Pricing table', 'Purchase options', 'Resources', 'FAQ', and 'Azure OpenAI Service >'. The main text area is divided into two columns: the left column contains the heading 'Azure OpenAI Service pricing overview', and the right column contains a paragraph about the service's capabilities and pricing options, followed by a list of pricing models: Standard (On-Demand), Provisioned (PTUs), and Batch API. A 'Chat with Sales' button is visible in the bottom right corner.

Home / [Azure pricing](#) / Azure OpenAI Service pricing

## Azure OpenAI Service pricing

[Request a pricing quote](#) [Try Azure for free](#)

[AI Services Family](#) [Overview](#) [Pricing table](#) [Purchase options](#) [Resources](#) [FAQ](#) [Azure OpenAI Service >](#)

### 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<br>Cached Input: ¥196.9438<br>Output: ¥1,575.5501 | Input: ¥196.94376<br>Output: ¥787.7751 |

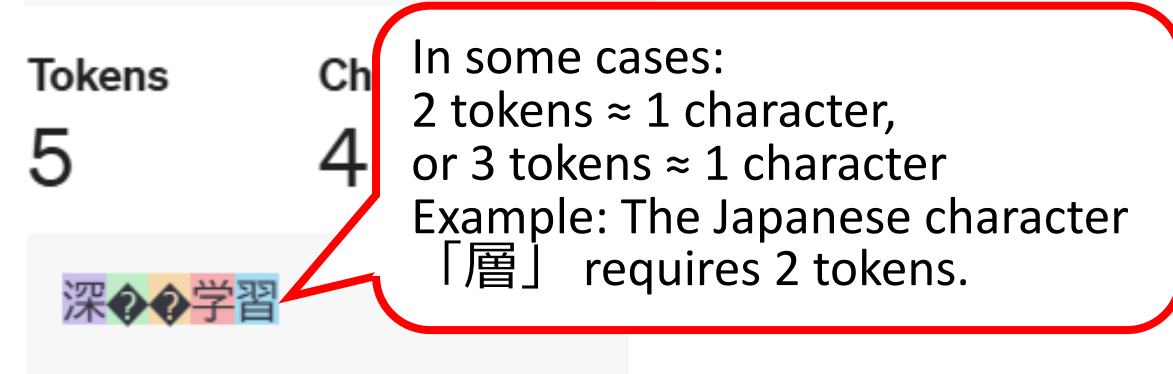
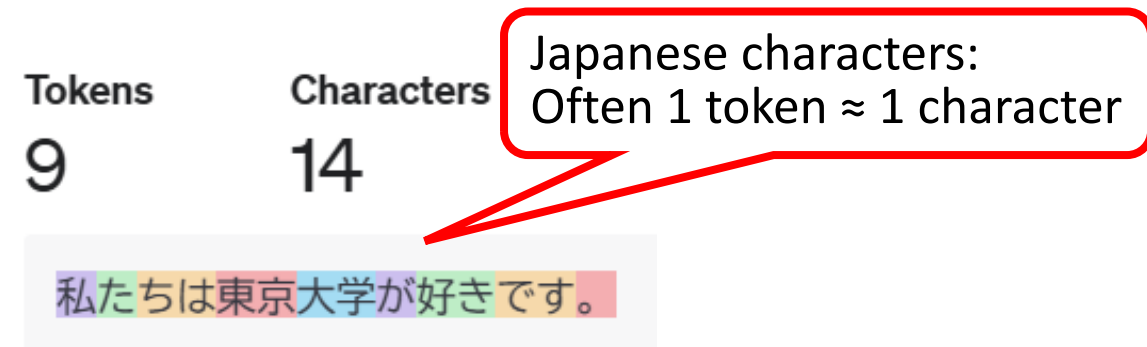
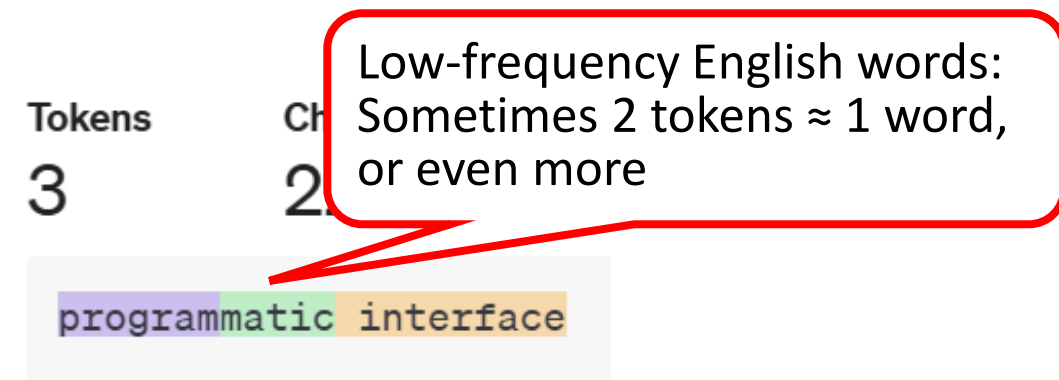
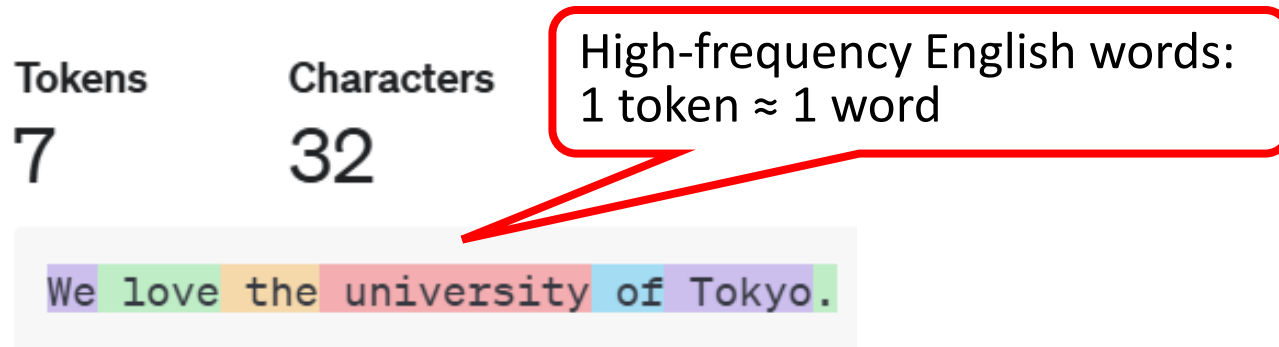
## o1

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<br>Cached Input: ¥1,181.6626<br>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  $\approx$  75 words**, but **low-frequency words** may require more tokens.
- Since Japanese is recognized a low-frequency language in the tokenizer's vocabulary, it generally results in **significantly more tokens** than English for the same amount of text.



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



Total requests

Total token count

Prompt token count

Completion token count

### Token usage

| Series         | Point 1 | Point 2 |
|----------------|---------|---------|
| Blue Circles   | 444     | 888     |
| Orange Squares | 370     | 740     |

## 5.2 Token Limit per Minute

Azure OpenAI Service has quota limits

U Tokyo Azure | utelecon

gprt-4o-test-1 - Microsoft Azure

Model quota - Azure OpenAI Service

ai.azure.com/resource/quota?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/openai\_test1/providers/Microsoft.CognitiveServices/accounts/gprt-...

本人確認を行ってください

Azure AI Foundry | Azure OpenAI Service

gprt-4o-test-1

Model quota

All resources5

gprt-4o-test-1(japaneast, S0)

Help

Model catalog

Playgrounds

Chat

AssistantsPREVIEW

LanguagePREVIEW

Real-time audioPREVIEW

Images

Completions

Tools

Fine-tuning

Azure OpenAI ServicePREVIEW

Evaluation

Stored completionsPREVIEW

Batch jobs

Metrics

Shared resources

Deployments

Quota

Safety + security

Data files

Vector storesPREVIEW

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 byQuota type, Model & Region

Search

Filter

Show all quota

Columns

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

Quota with lowest availability

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.

- **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](#)).

Azure OpenAI Service provisioned

Group by Quota type, Model & Region

Search

Filter

Show all quota

Columns

Deployment ↑

Model n...

Ver...

Region

Resource

Quota allocation

gpt-4o

GlobalSt...

> East US 2

gpt-4o

GlobalStan...

250K of 30M TPM

> Japan East

gpt-4o

GlobalStan...

250K of 30M TPM

[gpt-4o-test](#)

gpt-4o

202...

GlobalStan...

Japan ...

[gpt-4o-test-1-2](#)

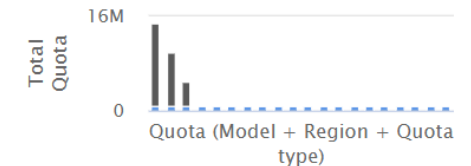
250K TPM



> gpt-4o-mini-real

GlobalStan...

### Quota with lowest availability



● 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

学校 | gpt-4o-test-1-2 - Microsoft Azure | Model quota - Azure OpenAI Service | +

https://ai.azure.com/resource/quota?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 / Model quota Docs All resources 5 gpt-4o-test-1-2 (japaneast, S0)

## 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** ①

1923306880-YamakataTraining

**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 Cancel

**Group by** Quota type, Model & Region

### Quota with lowest availability

Total Quota 16M

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

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

gpt-4o-mini-real GlobalStan...

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

Standard

Drag the slider to increase it to the desired size.

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

https://ai.azure.com/resource/quota?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 / Model quota Docs All resources 5 gpt-4o-test-1-2 (japaneast, S0)

## 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 ①

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 Filter Show all quota Columns Quota with lowest availability

| Deployment ↑        | Model n... | Vers... | Quota...        |
|---------------------|------------|---------|-----------------|
| > AIServices.Global |            |         | GlobalStar...   |
| ▼ gpt-4o            |            |         | GlobalStan...   |
| > East US 2         | gpt-4o     |         | GlobalStan...   |
|                     |            |         | 250K of 30M TPM |
|                     |            |         | 250K of 30M TPM |
| gpt-4o-test         | gpt-4o     | 202...  | GlobalStan...   |
|                     |            |         | Japan ...       |
| gpt-4o-test-1-2     |            |         | 250K TPM        |

● 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

- By default, the maximum is 30M TPM (Tokens per Minute).
- If you need more, you can request a quota increase.
- Note: It may take several days to be approved!

For gpt-4o, the maximum is 250K TPM, and it cannot be increased.

# 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.
- Since it is a manual review process, it may take **several days for approval**, so be sure to **submit your request early**.

The screenshot shows a web browser window with the URL <https://customervoice.microsoft.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180B...>. The page has a dark header with the Microsoft Azure logo and the title "Azure OpenAI Service: Request for Quota Increase". Below the title, a paragraph states: "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." The main content area is white and contains a section titled "\* 必須" (Required). It instructs users to "Read the instructions carefully and answer each question completely before submitting the request" and provides a detailed explanation of the form's purpose: "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." A "Please Note" section advises users to contact Microsoft personnel for more information. The first question is "1. First Name \*" with a prompt to "Enter your 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 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                                                                                                                                                                                                                           | chat-completion |
| 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   | Ministral-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

Microsoft Azure portal interface showing search results for "AI founda".

**Search Bar:** AI founda

**Results:**

- Services (31):**
  - Azure AI Foundry (Highlighted with a red box and text: "Select '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 ...
  - AI-900 - Training

**Recent Resources:**

| Name                        | Type           | Last Viewed       |
|-----------------------------|----------------|-------------------|
| Alfoundarytest              | Resource group | a few seconds ago |
| gpt-4o-test-1-2             | Subscription   | an hour ago       |
| openai_test1                | Resource group | an hour ago       |
| 1923306880-YamakataLab      | Subscription   | an hour ago       |
| gpt-4o-test-1-1             | Azure OpenAI   | an hour ago       |
| 1923306880-YamakataTraining | Subscription   | an hour ago       |
| openai_test1                | Resource group | 2 hours ago       |

学校

Azure AI Foundry - Microsoft Azure

+

https://portal.azure.com/#browse/Microsoft.MachineLearningServices%2Faistudio

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

The University of Tokyo

+

Create

⚙️

Manage view

🔄

Refresh

🔍

Open query

🏷️

Assign tags

Project

Collaborate, organize, and track work to build apps.

Hub

Grouping container for projects. Provides security, connectivity, and compute management.

Resource group equals all

Location equals all

Add filter

No grouping

List view

|                          | Resource group ↑↓     | Type ↑↓           | Location ↑↓ | Subscription ↑↓             |     |
|--------------------------|-----------------------|-------------------|-------------|-----------------------------|-----|
| <input type="checkbox"/> | openai_test1          | Azure AI services | East US 2   | 1923306880-YamakataLab      | ... |
| <input type="checkbox"/> | rg-1923306880-1245_ai | Azure AI hub      | East US 2   | 1923306880-YamakataLab      | ... |
| <input type="checkbox"/> | openai_test1          | Azure AI services | East US 2   | 1923306880-YamakataTraining | ... |
| <input type="checkbox"/> | openai_test1          | Azure AI project  | East US     | 1923306880-YamakataTraining | ... |
| <input type="checkbox"/> | AI-Foundry-hub0       | Azure AI project  | East US 2   | 1923306880-YamakataLab      | ... |
| <input type="checkbox"/> | AI-Foundry-hub0       | Azure AI hub      | East US 2   | 1923306880-YamakataLab      | ... |
| <input type="checkbox"/> | openai_test1          | Azure AI hub      | East US     | 1923306880-YamakataTraining | ... |
| <input type="checkbox"/> | openai_test1          | Azure AI services | East US     | 1923306880-YamakataTraining | ... |
| <input type="checkbox"/> | AI-Foundry-hub0       | Azure AI services | East US 2   | 1923306880-YamakataLab      | ... |

< Previous

Page 1 of 1

Next >

Give feedback

Start by creating a Hub

学校 Azure AI hub - Microsoft Azure

https://portal.azure.com/#view/Microsoft\_Azure\_MLTeamAccounts/CreateAIStudioResourceBlade/\_provisioningContext~//%7B"initialValues"%3A%7B"subscriptionIds"%...

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 > 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 \* ⓘ 1923306880-YamakataTraining

Resource group \* ⓘ (New) AI-Foundry-hub0  
[Create new](#)

Region \* ⓘ East US 2

## Resource details

Name \* ⓘ AI-Foundry-test0 ✓

Friendly name ⓘ AI Foundry test0

Default project resource group ⓘ

Same as hub resource group ☒

**Review + create** < Previous Next : Storage

The interface is largely the same as Azure OpenAI Service.

学校

Azure AI hub - Microsoft Azure

+

←https://portal.azure.com/#view/Microsoft\_Azure\_MLTeamAccounts/CreateAIStudioResourceBlade/\_provisioningContext~//%7B"initialValues"%3A%7B"subscriptionIds"%...→

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

Validation passed

BasicsStorageInbound AccessOutbound AccessEncryptionIdentityTagsReview + create

Basics

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

Home >

# Microsoft.MachineLearningServices | Overview

Deployment

Search Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

## Deployment is in progress

Deployment name : Microsoft.MachineLearnin... Start time : 4/8/2025, 4:49:08 PM  
Subscription : 1923306880-YamakataTrai... Correlation ID : 1f9ca76c-9b74-41bb-a88f...  
Resource group : AI-Foundary-hub0

### Deployment details

| Resource           | Type                            | Status  |
|--------------------|---------------------------------|---------|
| aifoundarytest9... | Microsoft.CognitiveServices/acc | Created |
| aifoundarytest4278 | Key vault                       | OK      |
| aifoundarytest6... | Storage account                 | OK      |



### 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 >](#)

学校

Microsoft.MachineLearningService

+

https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2Fa600356d-63a3-483e-b4bb-59629c01a212%2FresourceG...

🔍 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

Delete

Cancel

Redeploy

Download

Refresh

Overview

Inputs

Outputs

Template

✔ Your deployment is complete

Deployment name : Microsoft.MachineLearnin...

Subscription : 1923306880-YamakataTrai...

Resource group : AI-Foundary-hub0

> Deployment details

✓ Next steps

Go to resource

✔ Deployment succeeded

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

Go to resource

Go to resource group

💰

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 >

Add or remove favorites by pressing Ctrl+Shift+F

Add or remove favorites by pressing **Ctrl+Shift+F**

Compute

+ New project

 Refresh Delete project [Reset view](#)

[Learn more about projects in Azure AI Foundry](#)  

 Delete hub

AI-Foundary-hub0

1



## Azure AI Administrator

+ New user

## Azure AI Administrator

+ New user

[Learn more about projects in Azure AI Foundry](#)  

学校 AI-Foundary-test0 - Microsoft Az Overview - Azure AI Foundry

https://ai.azure.com/build/overview?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foundary-hub0/providers/Microsoft.MachineLearningServi...

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](#)

**My assets**

- Models + endpoints
- Data + indexes
- Web apps

[Management center](#)

**Nail the basics with these steps**

Step 1 Step 2 Step 3

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

2. Select a model  
from the model catalog

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

## 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 Foundry!**  
TCO and performance optimized for your AI applications  
[Check out models](#)

### Model leaderboards

See what models are performing best in different criteria.

Browse leaderboards

**Quality**

- o1
- DeepSeek-R1
- gpt-4.5-preview

**Cost**

- Ministral-3B
- Mistral-Nemo
- mistral-small-2503

**Throughput**

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

Collections Industry Capabilities Deployment options Inference tasks Fine-tuning tasks Licenses

[Compare models](#)

Search

Models 1900

学校

AI-Foundry-test0 - Microsoft Az...

Model catalog - Azure AI Foundry

+

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

Azure AI Foundry

1923306880-1774

Model catalog

Docs

All hubs + projects

5

Settings

Profile

Project 1923306880-1774

Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents

Templates

Fine-tuning

Prompt flow

Assess and improve

Tracing

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Management center

Find the right model to build your custom AI solution

Help

Quality

Cost

Throughput

1 o1

2 DeepSeek-R1

3 gpt-4.5-preview

1 Ministral-3B

2 Mistral-Nemo

3 mistral-small-2503

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

gpt-4.5-preview

Chat completion

o3-mini

Chat completion

Llama-4-Scout-17B-16E

Chat completion

Llama-4-Scout-17B-16E-Inst...

Chat completion

Llama-4-Maverick-17B-128E...

Chat completion

DeepSeek-V3

Chat completion

DeepSeek-R1

Chat completion

computer-use-preview

Responses

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

学校

AI-Foundry-test0 - Microsoft Az

DeepSeek-R1 - Azure AI Foundry

+

←↻🔒https://ai.azure.com/explore/models/DeepSeek-R1/version/1/registry/azureml-deepseek?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foun...🔍🗒️⭐🔖⋮🌐

Azure AI Foundry

1923306880-1774

Model catalog

DeepSeek-R1

DocsAll hubs + projects5⚙️😊Project1923306880-1774

← DeepSeek-R1PREVIEW?

▶ 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-tuning on curated datasets, resulting in a model that achieves state-of-the-art performance on reasoning benchmarks.  
[See more](#)

Data, media and languages

| Property             | Description                   |
|----------------------|-------------------------------|
| Supported data types | Inputs<br>textOutputs<br>text |
| Supported languages  | en, zh                        |

Transparency

Model alignment

Quick facts

De

🌐 DeepSeek-R1

chat-completion

Last trained

Not available

Benchmarks

0.91Quality indexAI quality

2.36USD per 1M tokEstimated 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

AgentsPREVIEW

TemplatesPREVIEW

Fine-tuning

Prompt flow

Assess and improve

TracingPREVIEW

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

→ Management center

学校

AI-Foundry-test0 - Microsoft Az...

DeepSeek-R1 - Azure AI Foundry

+

←↻🔒https://ai.azure.com/explore/models/DeepSeek-R1/version/1/registry/azureml-deepseek?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foun...🔍🔊🌟🔖⋮🌈

Azure AI Foundry

1923306880-1774

Model catalog

DeepSeek-R1

DocsAll hubs + projects5⚙️😊Project1923306880-1774

Overview

Model catalog

Playgrounds

AI Services

Build and customize

AgentsPREVIEW

TemplatesPREVIEW

Fine-tuning

Prompt flow

Assess and improve

TracingPREVIEW

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Management center

←DeepSeek-R1PREVIEW

Deploy

Details

Benchmarks

Learn more: [original r

DeepSeek-R1 excels a

tasks. It features 671B

DeepSeek-R1 builds o

Thought (CoT) reason

chosen datasets. It ev

but had issues like ha

small amount of cold-

using an curated da

See more

Data, media

Property

Supported data ty

Supported langua

Transparency

Model alignment

Serverless API deployment for DeepSeek-R1

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

0.91Quality indexAI quality

2.36USD per 1M tokEstimated cost

Model ID

Reference this model ID when deploying the model in

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

学校 AI-Foundry-test0 - Microsoft Az DeepSeek-R1 - Azure AI Foundry

https://ai.azure.com/explore/models/DeepSeek-R1/version/1/registry/azureml-deepseek?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foun...

Azure AI Foundry / 1923306880-1774 / Model catalog / DeepSeek-R1 Docs All hubs + projects 5 Project 1923306880-1774

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 repository](#)

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

DeepSeek-R1 builds on the DeepSeek-V2 architecture, which uses a mixture of experts (MoE) design. It has been trained on a large dataset, but had issues like hallucinations and a small amount of cold-start knowledge.

[See more](#)

### Serverless API deployment

Overview **Pricing and terms**

#### Pricing

[See the pricing details](#)

#### Legal

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

[Deploy](#) Cancel

| 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                         |

**Data, media**

**Property**

**Supported data type**

text text

**Supported languages**

en, zh

**Transparency**

**Model alignment**

**Benchmarks**

Last trained **Not available**

**0.91** Quality index AI quality

**2.36** USD per 1M tokens Estimated cost [View details](#)

**Model ID**

Reference this model ID when deploying the model in your application

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

学校

AI-Foundry-test0 - Microsoft Az...

DeepSeek-R1-kvjgv - Azure AI Fo...

DeepSeek R1: Improved Performa...

+

https://ai.azure.com/build/deployments/serverless/DeepSeek-R1-kvjgv?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foundry-hub0/provid...

Azure AI Foundry

1923306880-1774

Models + endpoints

DeepSeek-R1-kvjgv

Docs

All hubs + projects

5

Settings

Feedback

Project

1923306880-1774

Avatar

Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents

Templates

Fine-tuning

Prompt flow

Assess and improve

Tracing

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Management center

DeepSeek-R1-kvjgv

Details

Open in playground

Refresh

Edit

Delete

Endpoint

Target URI

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

Key

.....

Compute type

Consumption

Swagger URI

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

API Routes

Azure AI model inference: Chat Completion

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

deepseek-r1: Chat Completion

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

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

学校

AI-Foundry-test0 - Microsoft Az...

DeepSeek-R1-kvjgv - Azure AI Fo...

DeepSeek R1: Improved Performa...

+

https://ai.azure.com/build/deployments/serverless/DeepSeek-R1-kvjgv?wsid=/subscriptions/a600356d-63a3-483e-b4bb-59629c01a212/resourceGroups/AI-Foundry-hub0/provid...

Azure AI Foundry

1923306880-1774

Models + endpoints

DeepSeek-R1-kvjgv

Docs

All hubs + projects

5

Project

1923306880-1774

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-kvjgv

Details

Open in playground

Refresh

Edit

Delete

Endpoint

Target URI

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

Key

.....

Compute type

Consumption

Swagger URI

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

API Routes ⓘ

Azure AI model inference: Chat Completion

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

deepseek-r1: Chat Completion

https://DeepSeek-R1-kvjgv.eastus2.models.ai.azu...

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>")
)
```

- 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

## Chat playground

Help

[View code](#)

### Setup

Hide

Deployment \* [Create new deployment](#)

DeepSeek-R1-kvjgv

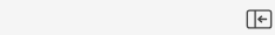
> Parameters

### Chat session



Start typing here





Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents

Templates

Fine-tuning

Prompt flow

Assess and improve

Tracing

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Management center

## Chat playground

Help

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



Start typing here

