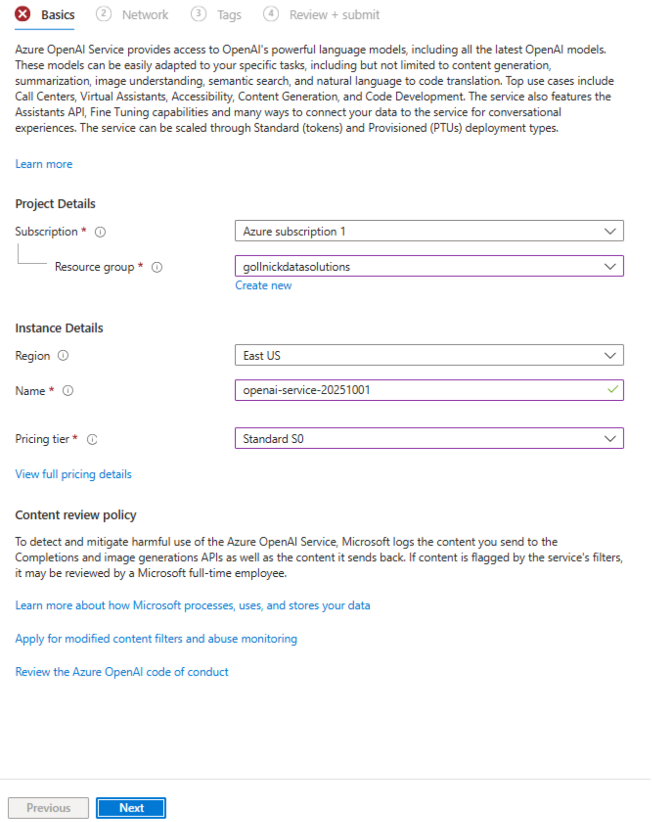
OpenAI Services

gollnickdata.de

OpenAI Service

OpenAI Service

Create new OpenAI service



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 content generation, summarization, image understanding, semantic search, and natural language to code translation. Top use cases include Call Centers, Virtual Assistants, Accessibility, Content Generation, and Code Development. The service also features the Assistants API, Fine Tuning capabilities and many ways to connect your data to the service for conversational experiences. The service can be scaled through Standard (tokens) and Provisioned (PTUs) deployment types.

Learn more

Project Details

Subscription * ⓘ: Azure subscription 1

Resource group * ⓘ: gollnickdatasolutions

Create new

Instance Details

Region ⓘ: East US

Name * ⓘ: openai-service-20251001 ✓

Pricing tier * ⓘ: Standard S0

View full pricing details

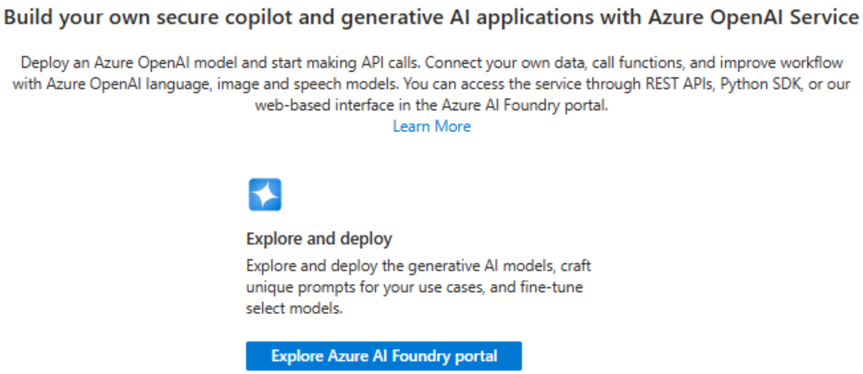
Content review policy

To detect and mitigate harmful use of the Azure OpenAI Service, Microsoft logs the content you send to the Completions and image generations APIs as well as the content it sends back. If content is flagged by the service's filters, it may be reviewed by a Microsoft full-time employee.

Learn more about how Microsoft processes, uses, and stores your data

Apply for modified content filters and abuse monitoring

Review the Azure OpenAI code of conduct



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.

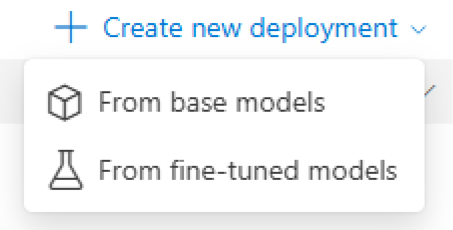
Learn More

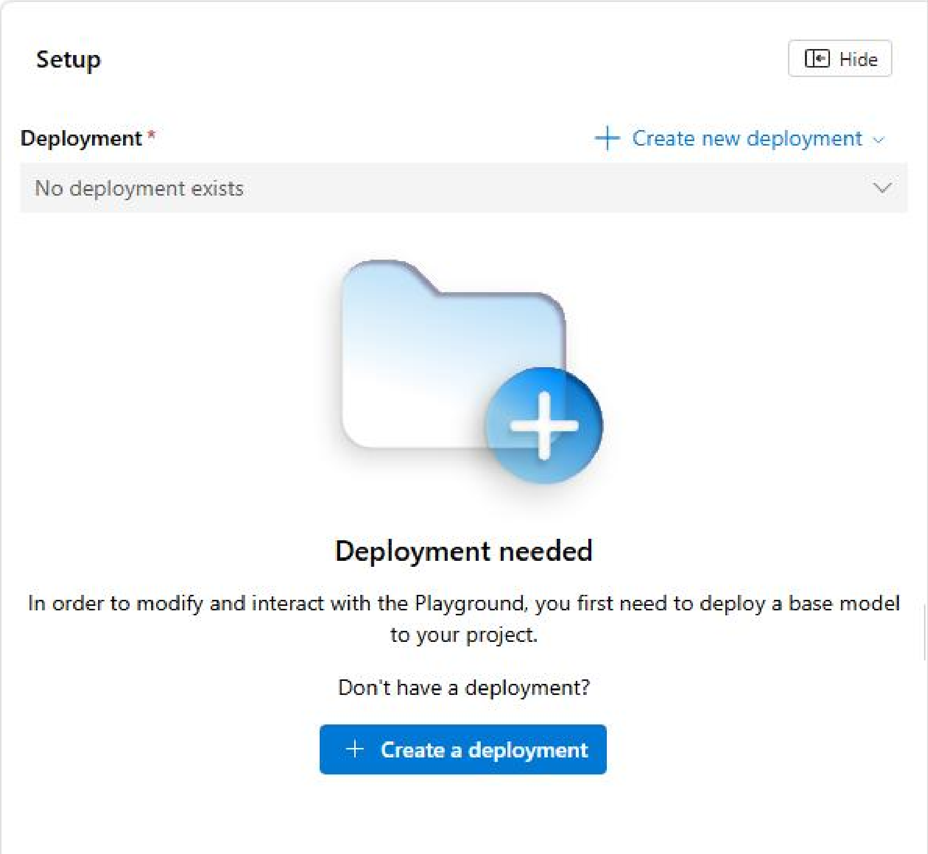
Explore and deploy

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

OpenAI Service

Create Deployment





Deployment *

No deployment exists

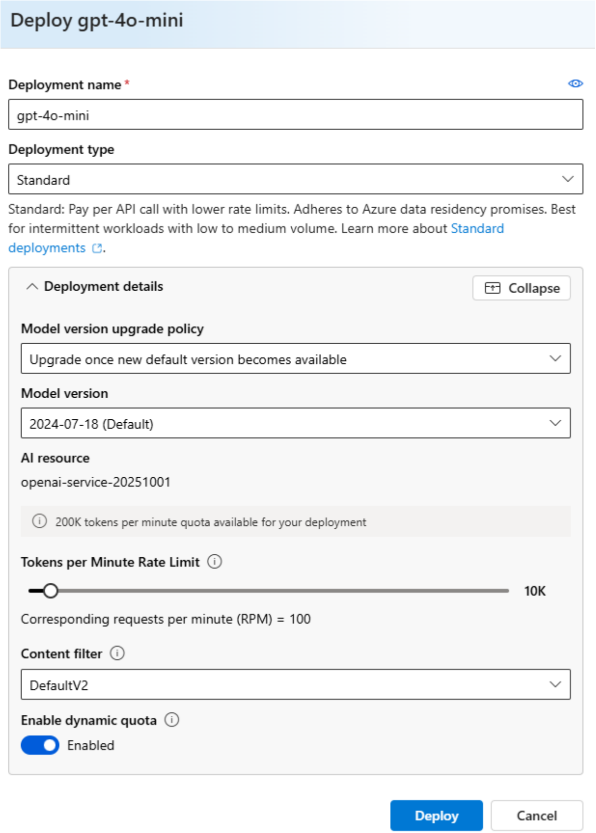
Deployment needed

In order to modify and interact with the Playground, you first need to deploy a base model to your project.

Don't have a deployment?

OpenAI Service

Create Deployment



Deploy gpt-4o-mini

Deployment name *

gpt-4o-mini

Deployment type

Standard

Standard: Pay per API call with lower rate limits. Adheres to Azure data residency promises. Best for intermittent workloads with low to medium volume. Learn more about Standard deployments.

Deployment details

Model version upgrade policy

Upgrade once new default version becomes available

Model version

2024-07-18 (Default)

AI resource

openai-service-20251001

ⓘ 200K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit ⓘ

10K

Corresponding requests per minute (RPM) = 100

Content filter ⓘ

DefaultV2

Enable dynamic quota ⓘ

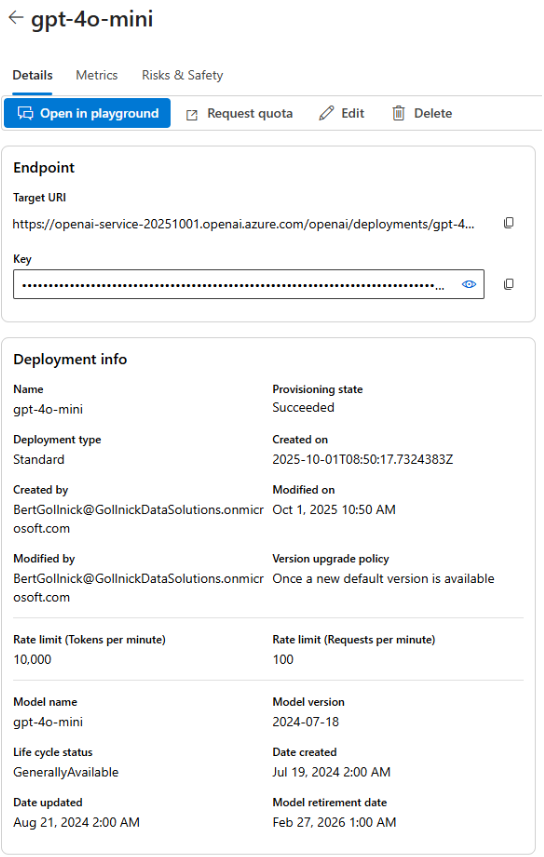
Enabled

gollnickdata.de

Python SDK

Python SDK

Create Deployment



← gpt-4o-mini

Endpoint

Target URI

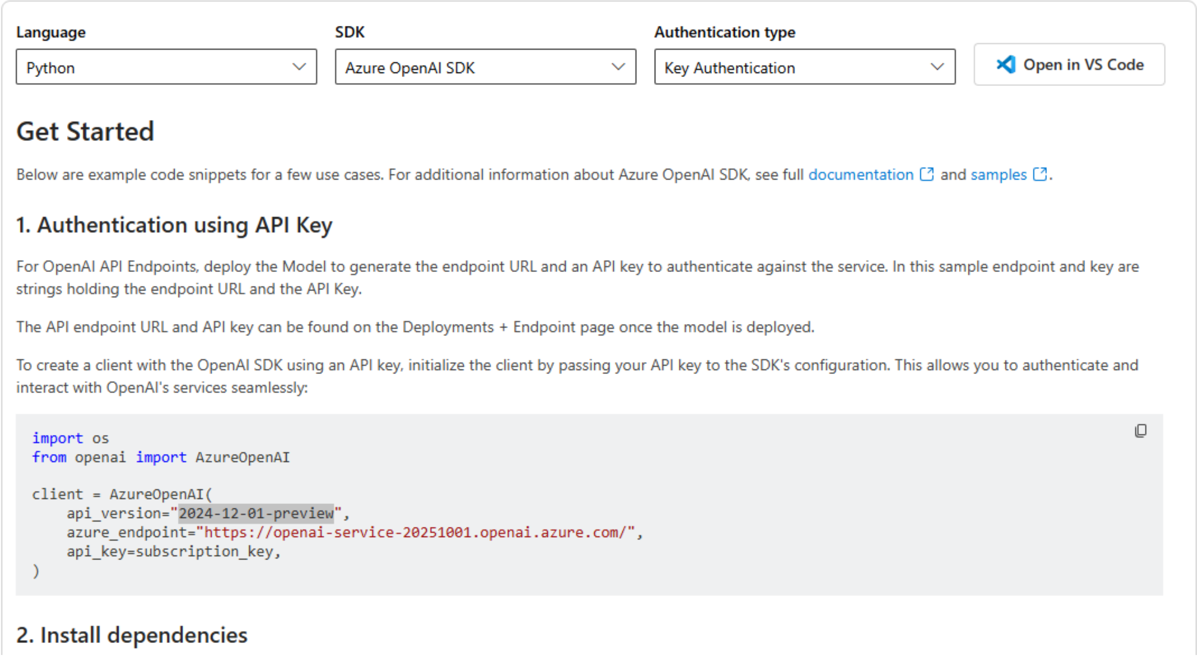
https://openai-service-20251001.openai.azure.com/openai/deployments/gpt-4...

Key

••...

Deployment info

Name gpt-4o-mini	**Provisioning state** Succeeded
Deployment type Standard	**Created on** 2025-10-01T08:50:17.7324383Z
Created by BertGollnick@GollnickDataSolutions.onmicrosoft.com	**Modified on** Oct 1, 2025 10:50 AM
Modified by BertGollnick@GollnickDataSolutions.onmicrosoft.com	**Version upgrade policy** Once a new default version is available
Rate limit (Tokens per minute) 10,000	**Rate limit (Requests per minute)** 100
Model name gpt-4o-mini	**Model version** 2024-07-18
Life cycle status GenerallyAvailable	**Date created** Jul 19, 2024 2:00 AM
Date updated Aug 21, 2024 2:00 AM	**Model retirement date** Feb 27, 2026 1:00 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 and samples.

1. Authentication using API Key

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

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. This allows you to authenticate and interact with OpenAI's services seamlessly:

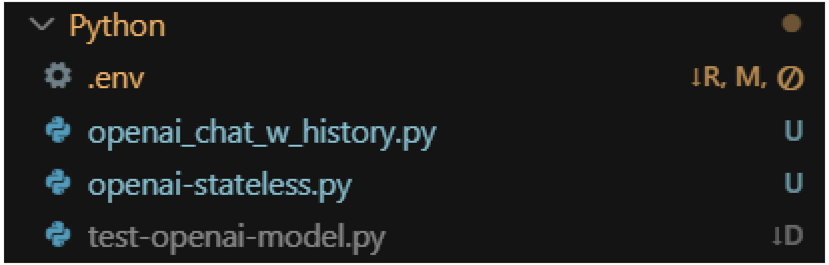
```python
import os
from openai import AzureOpenAI

client = AzureOpenAI(
    api_version="2024-12-01-preview",
    azure_endpoint="https://openai-service-20251001.openai.azure.com/",
    api_key=subscription_key,
)
```

2. Install dependencies

gollnickdata.de

Python SDK

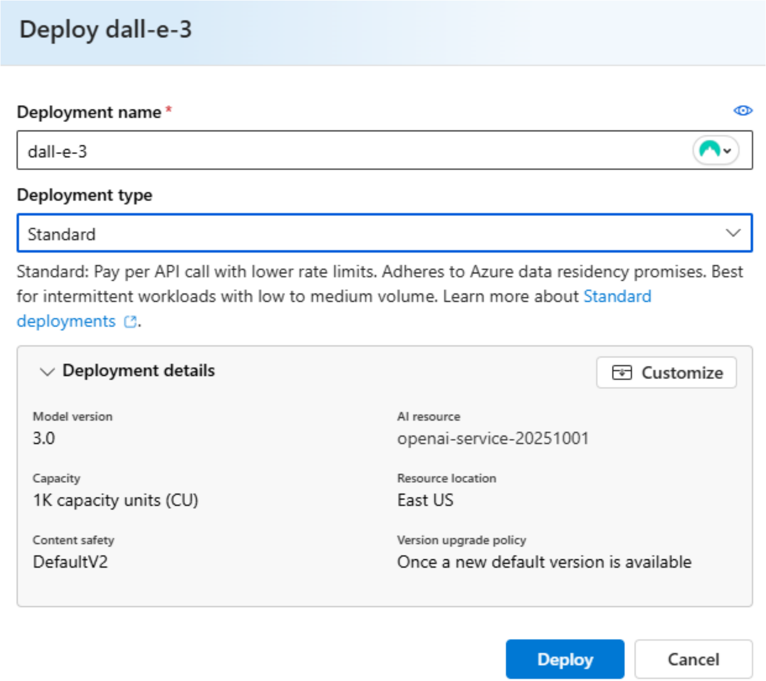
Use Deployment via Python



```
Python                         ●
  .env                  ↓R, M, ⊘
  openai_chat_w_history.py     U
  openai-stateless.py          U
  test-openai-model.py        ↓D
```

Python SDK

Image Generation



Deploy dall-e-3

Deployment name *

dall-e-3

Deployment type

Standard

Standard: Pay per API call with lower rate limits. Adheres to Azure data residency promises. Best for intermittent workloads with low to medium volume. Learn more about Standard deployments.

Deployment details

Customize

Model version
3.0

AI resource
openai-service-20251001

Capacity
1K capacity units (CU)

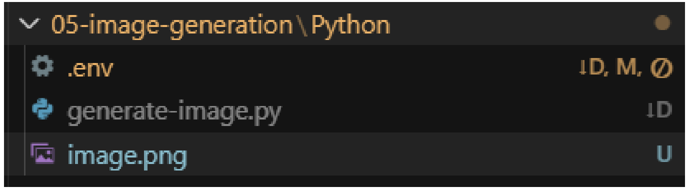
Resource location
East US

Content safety
DefaultV2

Version upgrade policy
Once a new default version is available

Deploy

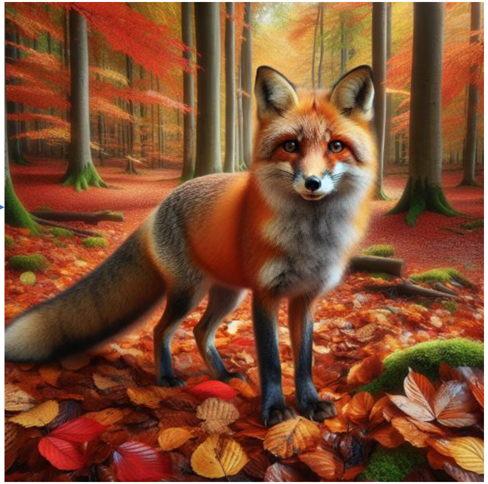
Cancel



05-image-generation \ Python

- .env ↓D, M, ⊘
- generate-image.py ↓D
- image.png U

```
prompt = "A
photograph of a
red fox in an
autumn forest"
```



gollnickdata.de

Deployment as WebApp

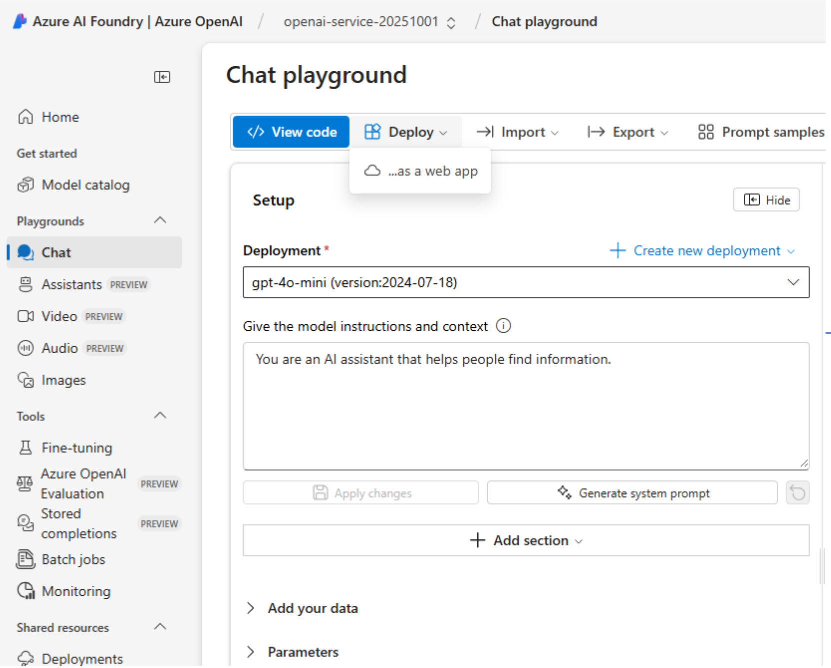
Deployment as WebApp

What is a web app?

- Platform as a service – MS manages the OS, server software, and infrastructure
- Supports multiple languages (e.g. .NET, Java, Python, …)
- Allows for automatic scaling, integration of other services, …

Deployment as WebApp

Deployment



Chat playground

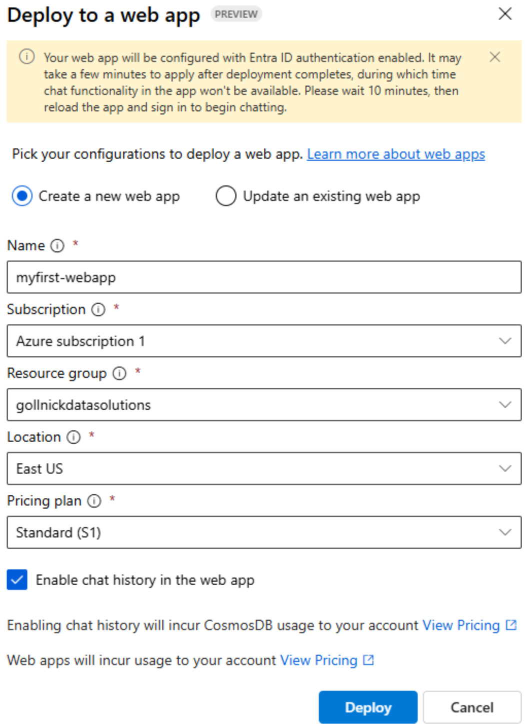
Setup

Deployment *

gpt-4o-mini (version:2024-07-18)

Give the model instructions and context

You are an AI assistant that helps people find information.



Deploy to a web app PREVIEW

Your web app will be configured with Entra ID authentication enabled. It may take a few minutes to apply after deployment completes, during which time chat functionality in the app won't be available. Please wait 10 minutes, then reload the app and sign in to begin chatting.

Pick your configurations to deploy a web app. Learn more about web apps

● Create a new web app ○ Update an existing web app

Name *

myfirst-webapp

Subscription *

Azure subscription 1

Resource group *

gollnickdatasolutions

Location *

East US

Pricing plan *

Standard (S1)

☑ Enable chat history in the web app

Enabling chat history will incur CosmosDB usage to your account View Pricing

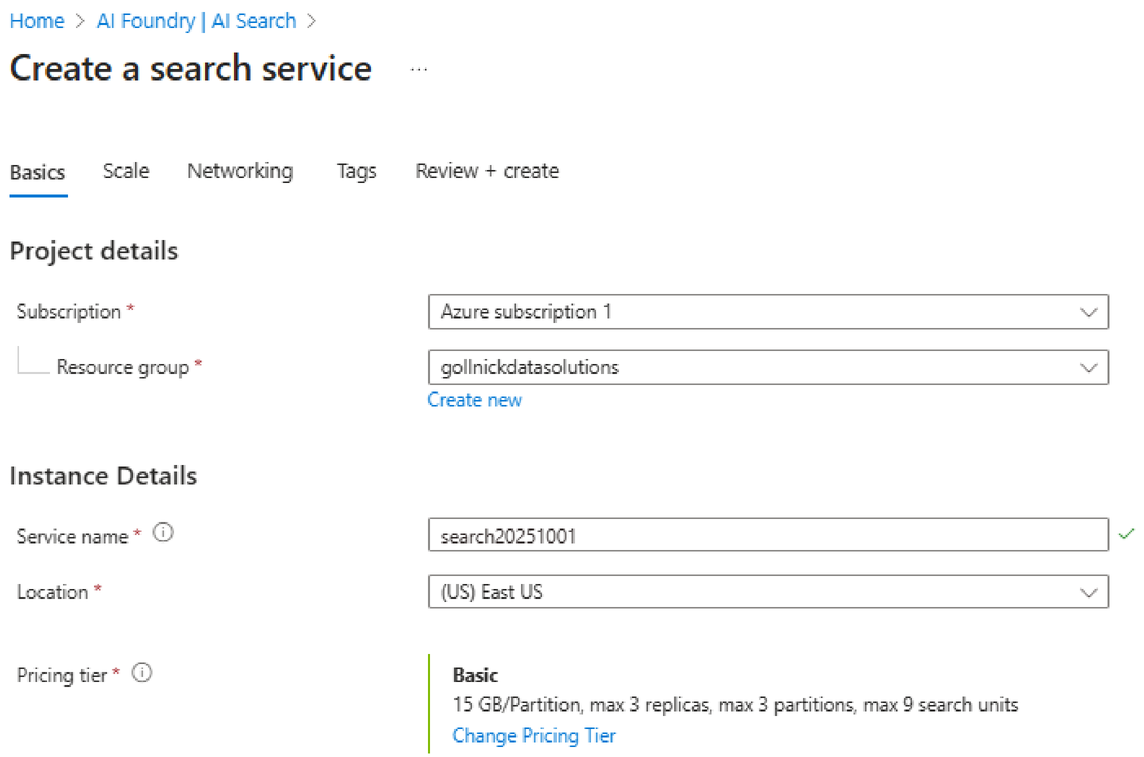
Web apps will incur usage to your account View Pricing

gollnickdata.de

RAG service

RAG service

Create AI Search



Create a search service ···

Project details

Subscription *	Azure subscription 1
Resource group *	gollnickdatasolutions Create new

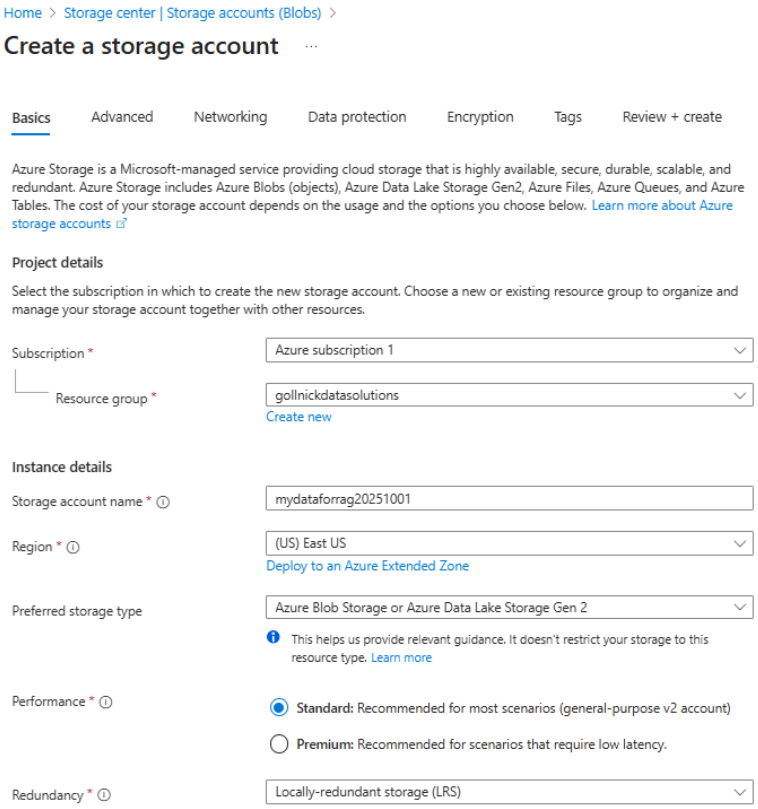
Instance Details

Service name * ⓘ	search20251001 ✓
Location *	(US) East US
Pricing tier * ⓘ	**Basic** 15 GB/Partition, max 3 replicas, max 3 partitions, max 9 search units Change Pricing Tier

gollnickdata.de

RAG service

Create new storage account



Home > Storage center | Storage accounts (Blobs) >

Create a storage account ...

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more about Azure storage accounts

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *: Azure subscription 1

Resource group *: gollnickdatasolutions

Create new

Instance details

Storage account name *: mydataforrag20251001

Region *: (US) East US

Deploy to an Azure Extended Zone

Preferred storage type: Azure Blob Storage or Azure Data Lake Storage Gen 2

This helps us provide relevant guidance. It doesn't restrict your storage to this resource type. Learn more

Performance *:

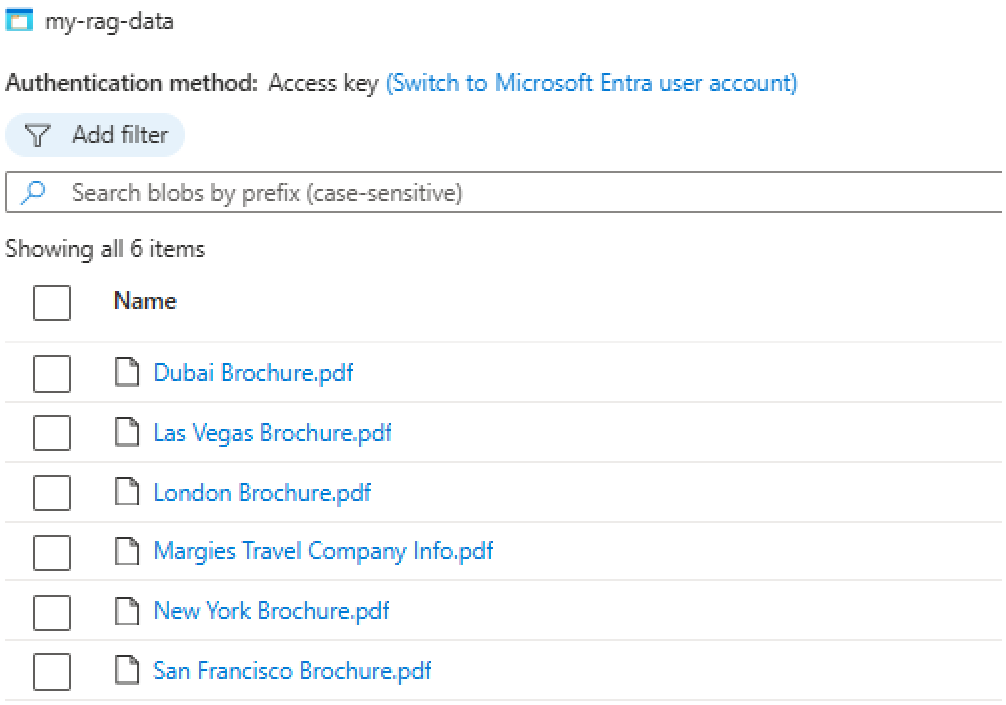
● **Standard:** Recommended for most scenarios (general-purpose v2 account)

○ **Premium:** Recommended for scenarios that require low latency.

Redundancy *: Locally-redundant storage (LRS)

RAG service

Create Storage Container and Data Upload



my-rag-data

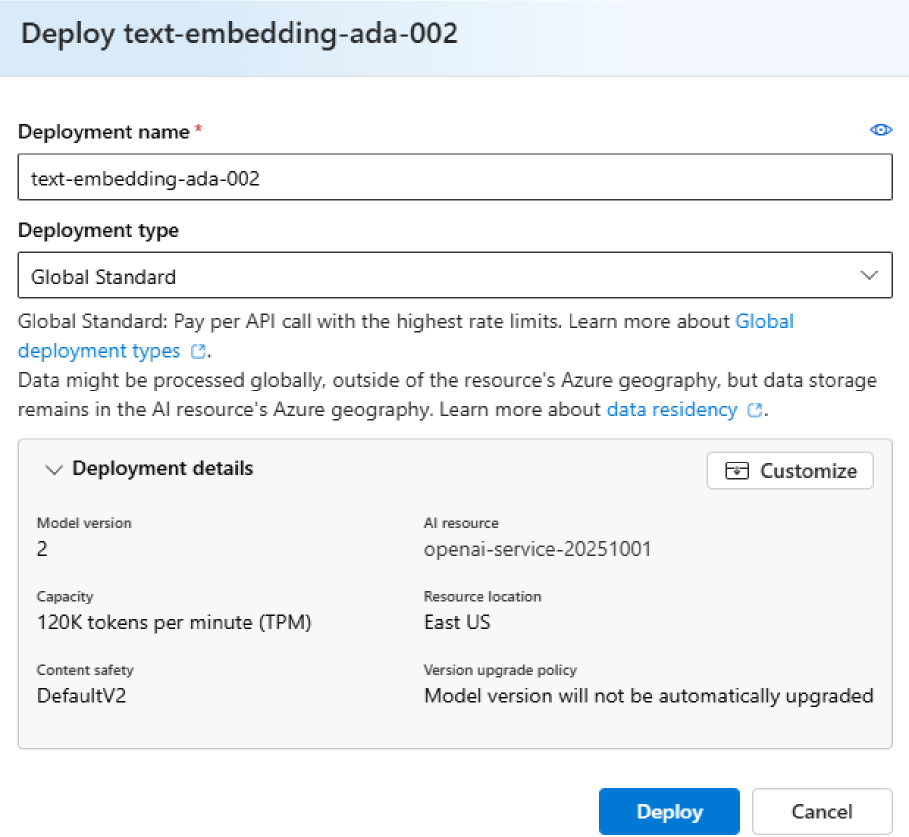
Authentication method: Access key (Switch to Microsoft Entra user account)

Showing all 6 items

□	Name
□	Dubai Brochure.pdf
□	Las Vegas Brochure.pdf
□	London Brochure.pdf
□	Margies Travel Company Info.pdf
□	New York Brochure.pdf
□	San Francisco Brochure.pdf

RAG service

Depoy Embedding Model



Deploy text-embedding-ada-002

Deployment name *

text-embedding-ada-002

Deployment type

Global Standard

Global Standard: Pay per API call with the highest rate limits. Learn more about Global deployment types.

Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about data residency.

Deployment details

Model version
2

AI resource
openai-service-20251001

Capacity
120K tokens per minute (TPM)

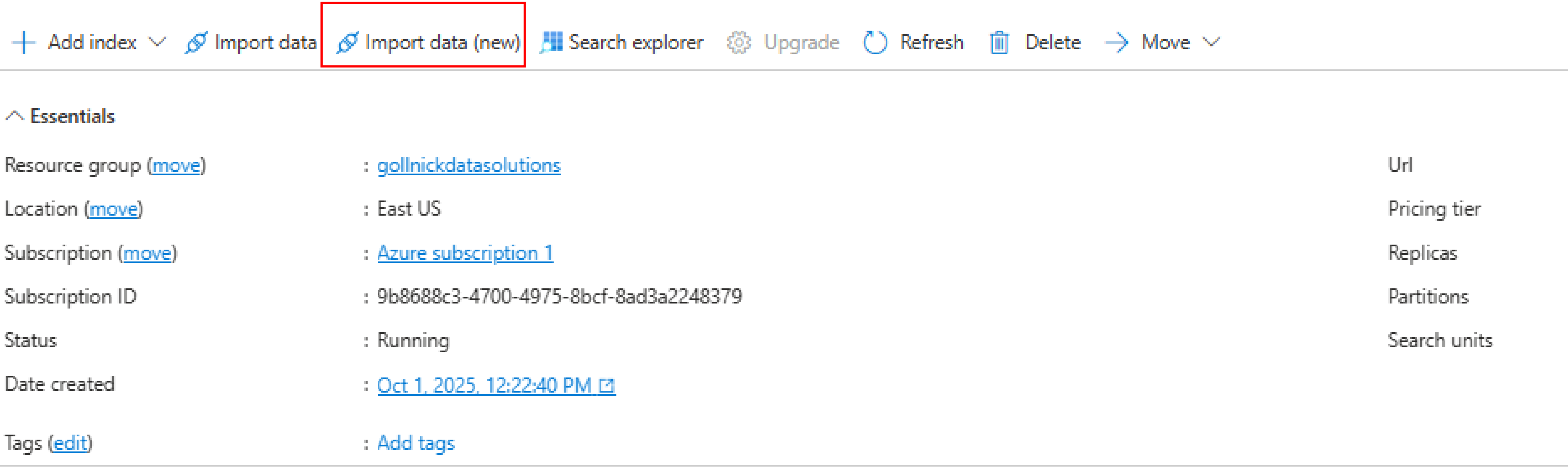
Resource location
East US

Content safety
DefaultV2

Version upgrade policy
Model version will not be automatically upgraded

RAG service

Import and Vectorize Data



Resource group (move)	: gollnickdatasolutions
Location (move)	: East US
Subscription (move)	: Azure subscription 1
Subscription ID	: 9b8688c3-4700-4975-8bcf-8ad3a2248379
Status	: Running
Date created	: Oct 1, 2025, 12:22:40 PM
Tags (edit)	: Add tags

Url

Pricing tier

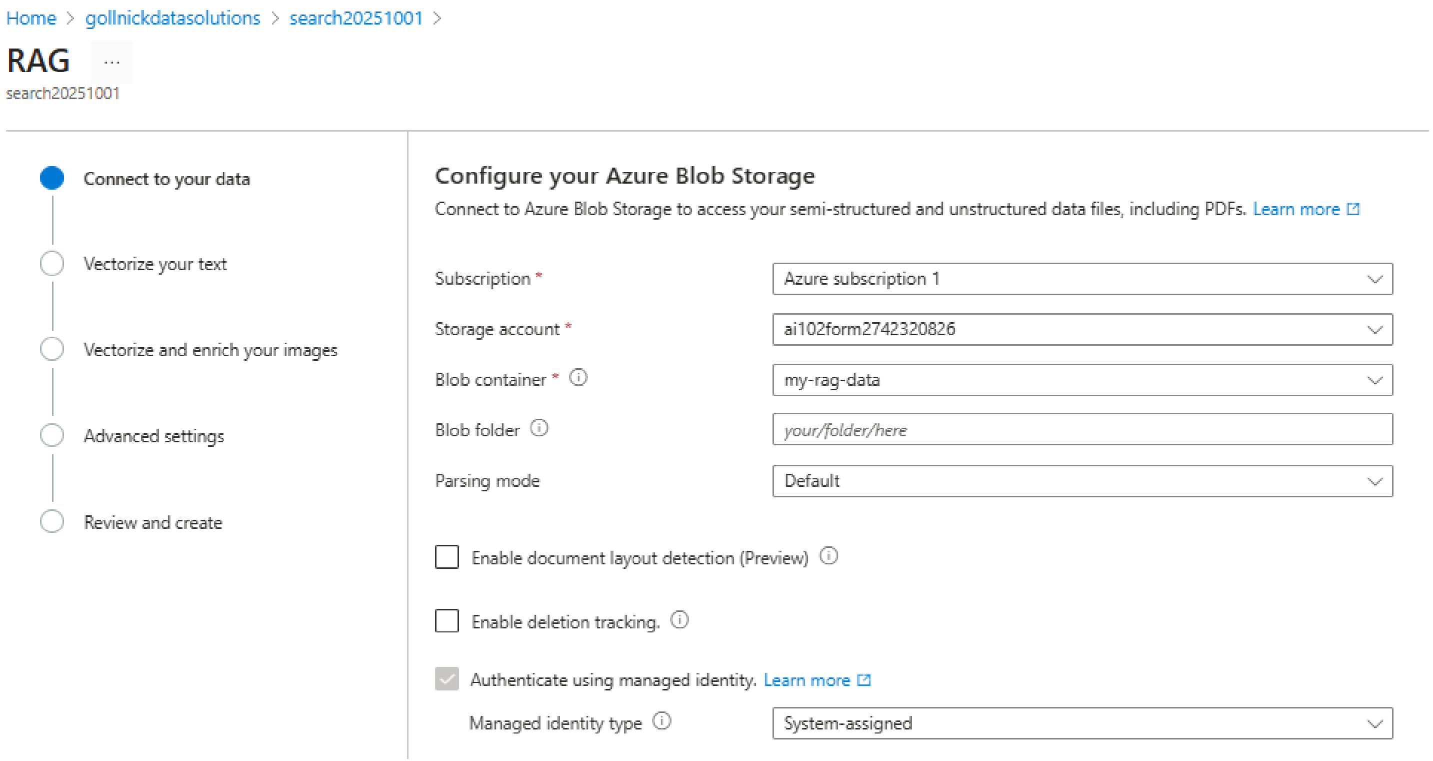
Replicas

Partitions

Search units

RAG service

Import and Vectorize Data



Home > gollnickdatasolutions > search20251001 >

RAG ...

search20251001

Configure your Azure Blob Storage

Connect to Azure Blob Storage to access your semi-structured and unstructured data files, including PDFs. Learn more

Subscription *	Azure subscription 1
Storage account *	ai102form2742320826
Blob container *	my-rag-data
Blob folder	*your/folder/here*
Parsing mode	Default

☐ Enable document layout detection (Preview)

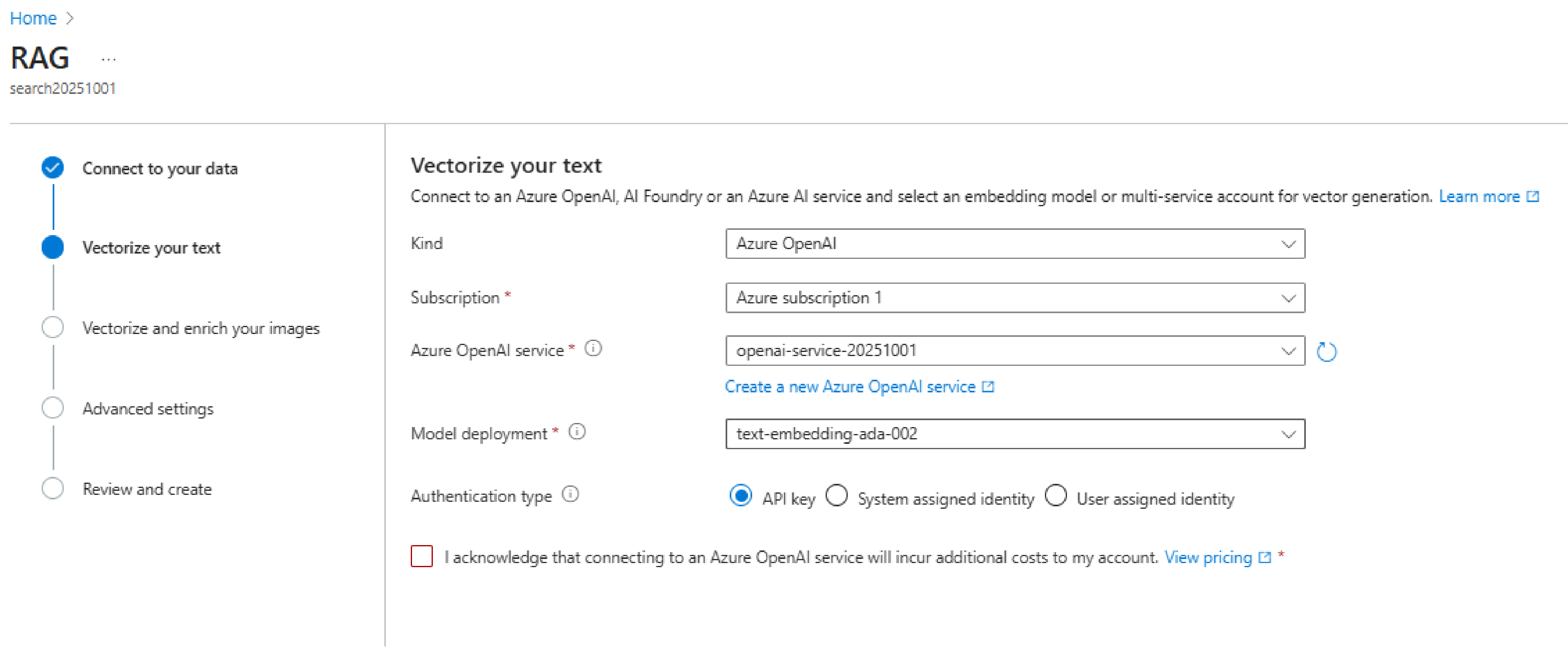
☐ Enable deletion tracking.

☑ Authenticate using managed identity. Learn more

Managed identity type: System-assigned

gollnickdata.de

RAG service



Home >

RAG ...

search20251001

Vectorize your text

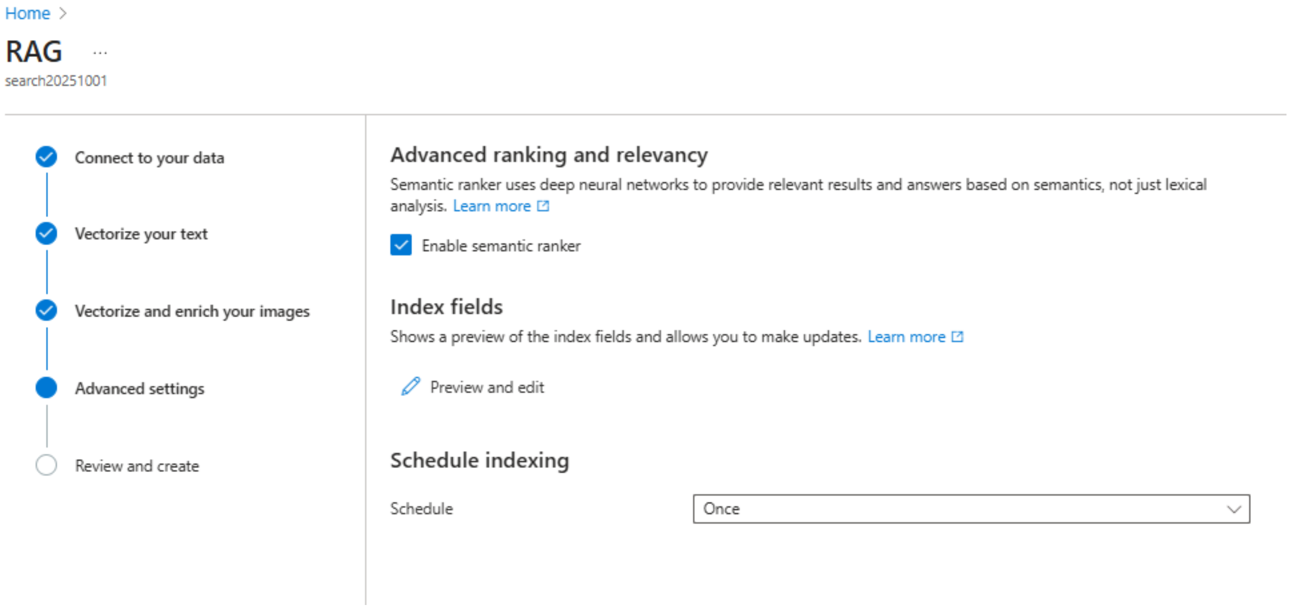
Connect to an Azure OpenAI, AI Foundry or an Azure AI service and select an embedding model or multi-service account for vector generation. Learn more

Kind	Azure OpenAI
Subscription *	Azure subscription 1
Azure OpenAI service *	openai-service-20251001 Create a new Azure OpenAI service
Model deployment *	text-embedding-ada-002
Authentication type	● API key ○ System assigned identity ○ User assigned identity

□ I acknowledge that connecting to an Azure OpenAI service will incur additional costs to my account. View pricing *

RAG service

Import and Vectorize Data



RAG ...

search20251001

- ✓ Connect to your data
- ✓ Vectorize your text
- ✓ Vectorize and enrich your images
- ● Advanced settings
- ○ Review and create

Advanced ranking and relevancy

Semantic ranker uses deep neural networks to provide relevant results and answers based on semantics, not just lexical analysis. Learn more

☑ Enable semantic ranker

Index fields

Shows a preview of the index fields and allows you to make updates. Learn more

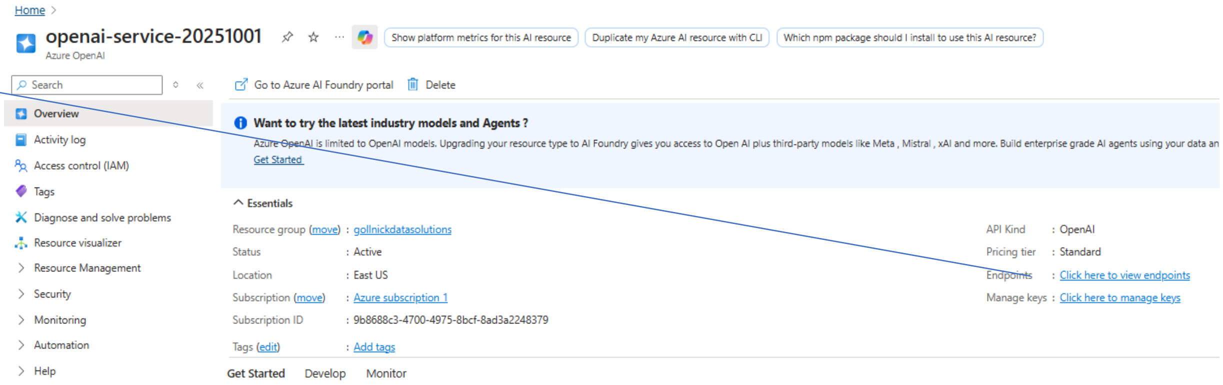
Schedule indexing

Schedule: Once

RAG service

Use Vector Search via Python

```
AZURE_OAI_ENDPOINT=

AZURE_OAI_KEY=

AZURE_OAI_DEPLOYMENT=

AZURE_SEARCH_ENDPOINT=

AZURE_SEARCH_KEY=

AZURE_SEARCH_INDEX=
```


Home >
openai-service-20251001
Azure OpenAI
Show platform metrics for this AI resource
Duplicate my Azure AI resource with CLI
Which npm package should I install to use this AI resource?
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Resource visualizer
Resource Management
Security
Monitoring
Automation
Help
Go to Azure AI Foundry portal
Delete
Get Started
Develop
Monitor

Want to try the latest industry models and Agents ?

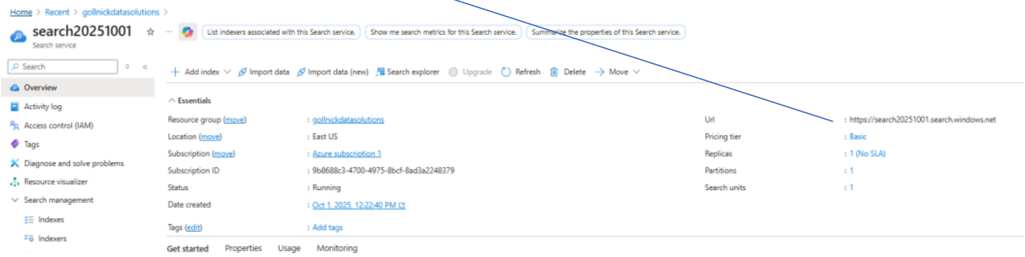
Azure OpenAI is limited to OpenAI models. Upgrading your resource type to AI Foundry gives you access to Open AI plus third-party models like Meta , Mistral , xAI and more. Build enterprise grade AI agents using your data an

Get Started

Essentials

Resource group (move) : gollnickdatasolutions
Status : Active
Location : East US
Subscription (move) : Azure subscription 1
Subscription ID : 9b8688c3-4700-4975-8bcf-8ad3a2248379
Tags (edit) : Add tags

API Kind : OpenAI
Pricing tier : Standard
Endpoints : Click here to view endpoints
Manage keys : Click here to manage keys



Essentials

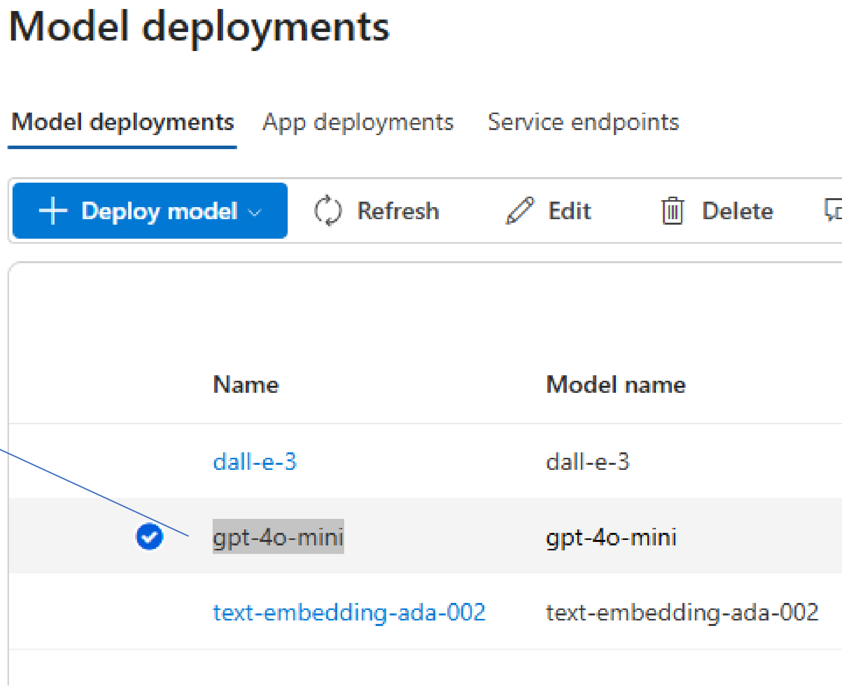
Resource group (move) : gollnickdatasolutions
Location (move) : East US
Subscription (move) : Azure subscription 1
Subscription ID : 9b8688c3-4700-4975-8bcf-8ad3a2248379
Status : Running
Date created : Oct 1, 2025, 12:22:40 PM
Tags (edit) : Add tags

Url : https://search20251001.search.windows.net
Pricing tier : Basic
Replicas : 1 (No SLA)
Partitions : 1
Search units : 1

RAG service

Use Vector Search via Python

```
AZURE_OAI_ENDPOINT=

AZURE_OAI_KEY=

AZURE_OAI_DEPLOYMENT=

AZURE_SEARCH_ENDPOINT=

AZURE_SEARCH_KEY=

AZURE_SEARCH_INDEX=
```



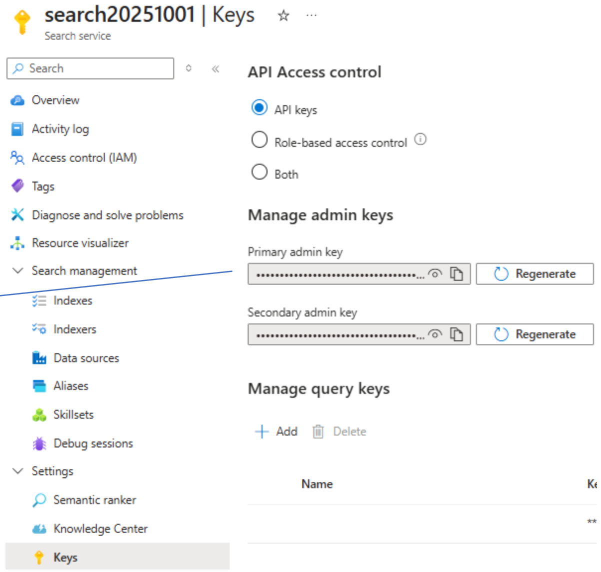
Model deployments

Name	Model name
dall-e-3	dall-e-3
gpt-4o-mini	gpt-4o-mini
text-embedding-ada-002	text-embedding-ada-002

RAG service

Use Vector Search via Python

```
AZURE_OAI_ENDPOINT=

AZURE_OAI_KEY=

AZURE_OAI_DEPLOYMENT=

AZURE_SEARCH_ENDPOINT=

AZURE_SEARCH_KEY=

AZURE_SEARCH_INDEX=
```



search20251001 | Keys ☆ ···
Search service

API Access control

● API keys
○ Role-based access control ⓘ
○ Both

Manage admin keys

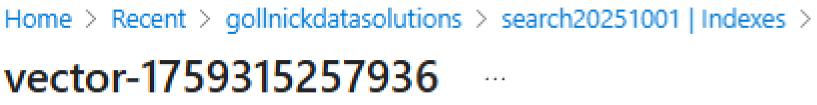
Primary admin key
••••••••••••••••••••••••••••••••••... Regenerate

Secondary admin key
••••••••••••••••••••••••••••••••••... Regenerate

Manage query keys

+ Add Delete

Name	Ke
	**

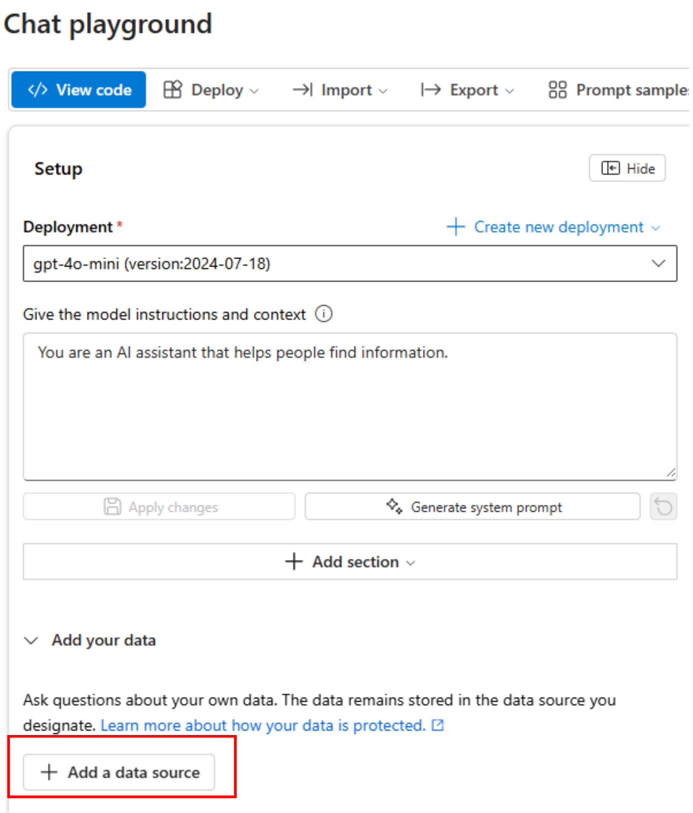


Home > Recent > gollnickdatasolutions > search20251001 | Indexes >

vector-1759315257936 ···

RAG service

Alternatively: run via Azure AI Foundry



Chat playground

Setup

Deployment *

gpt-4o-mini (version:2024-07-18)

Give the model instructions and context

You are an AI assistant that helps people find information.

Add your data

Ask questions about your own data. The data remains stored in the data source you designate. Learn more about how your data is protected.