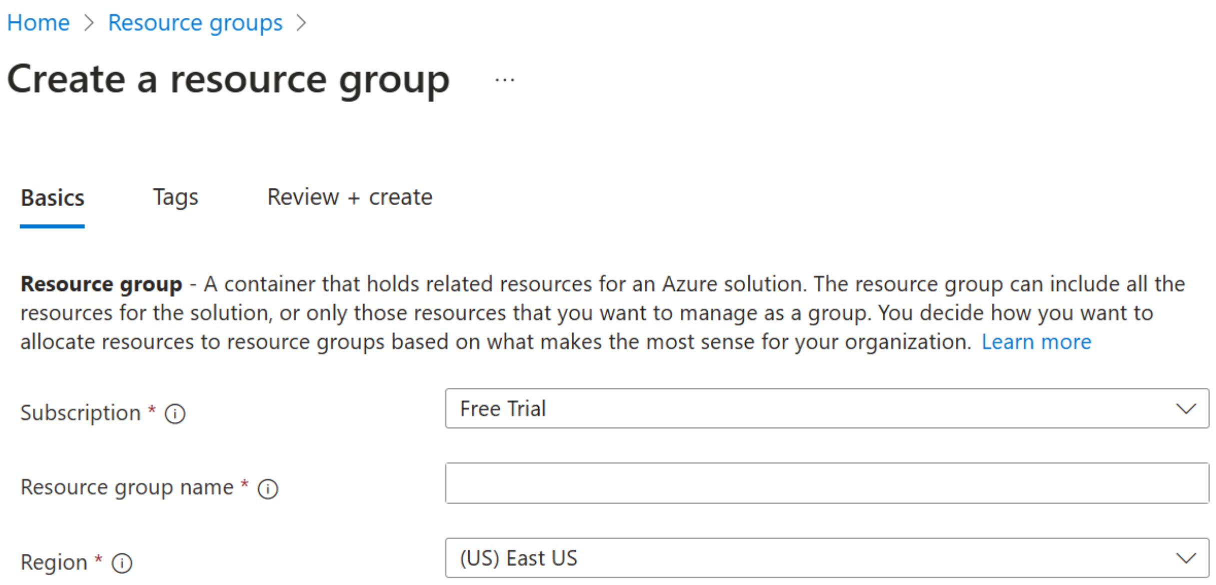
STEPS TO SETUP APIM

STEP 1: Create a Resource Group and select the subscription



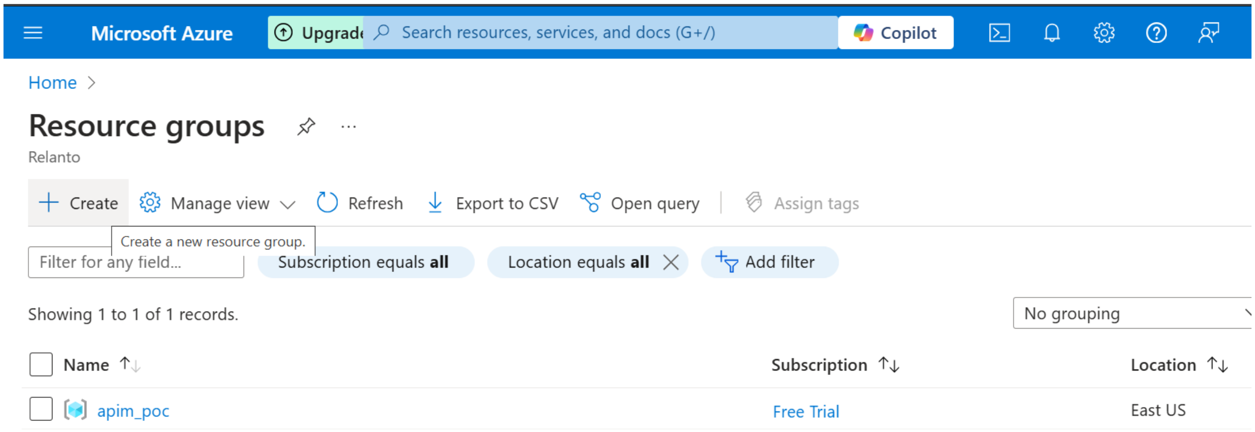
Create a resource group

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. Learn more

Subscription * : Free Trial

Resource group name * :

Region * : (US) East US



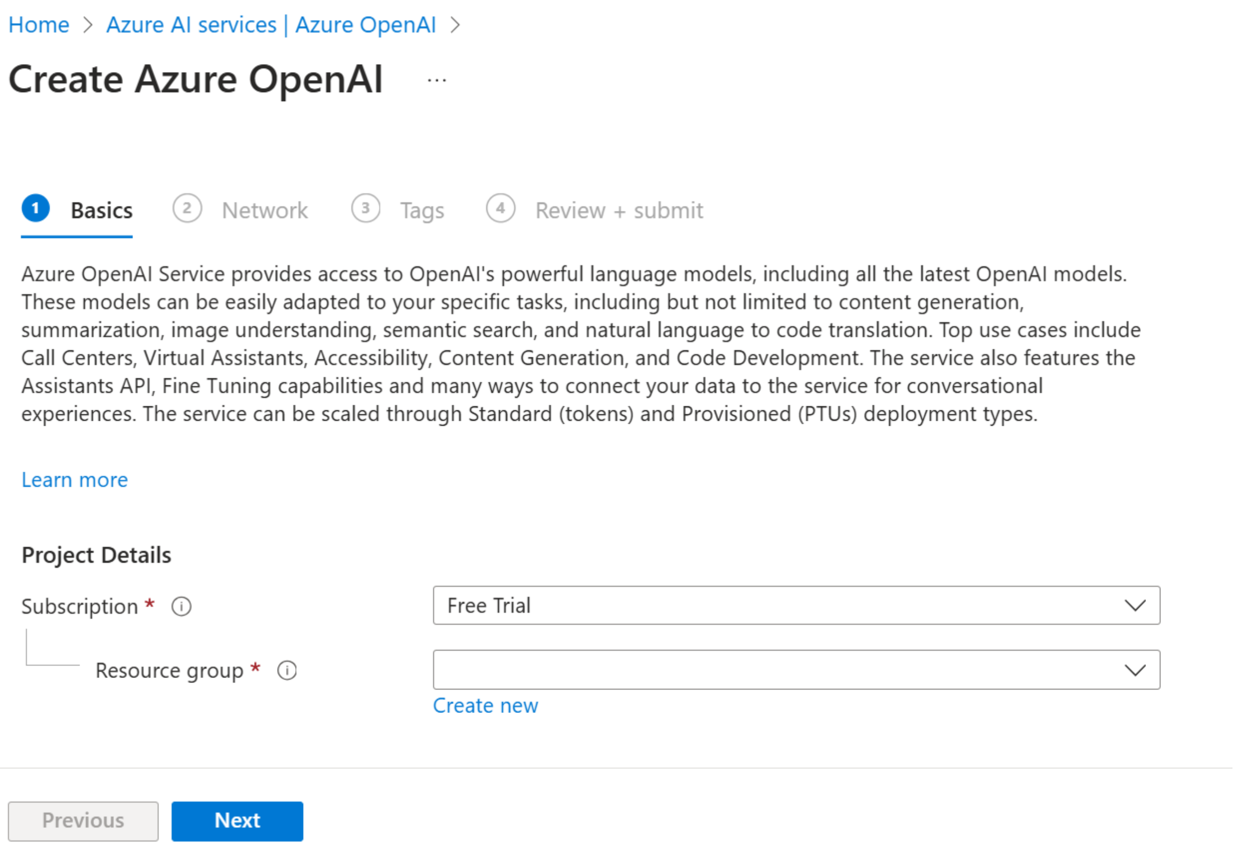
Resource groups

Relanto

Showing 1 to 1 of 1 records.

Name	Subscription	Location
apim_poc	Free Trial	East US

Step 2: Create a Service in Azure Ai services (OpenAI)



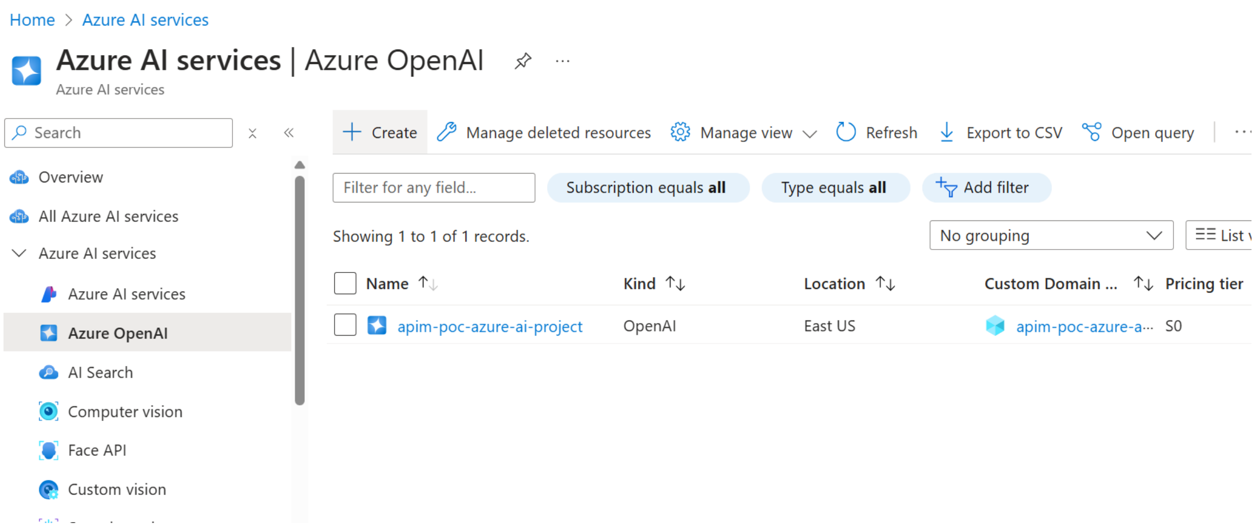
Create Azure OpenAI

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.

Project Details

Subscription * Free Trial

Resource group *



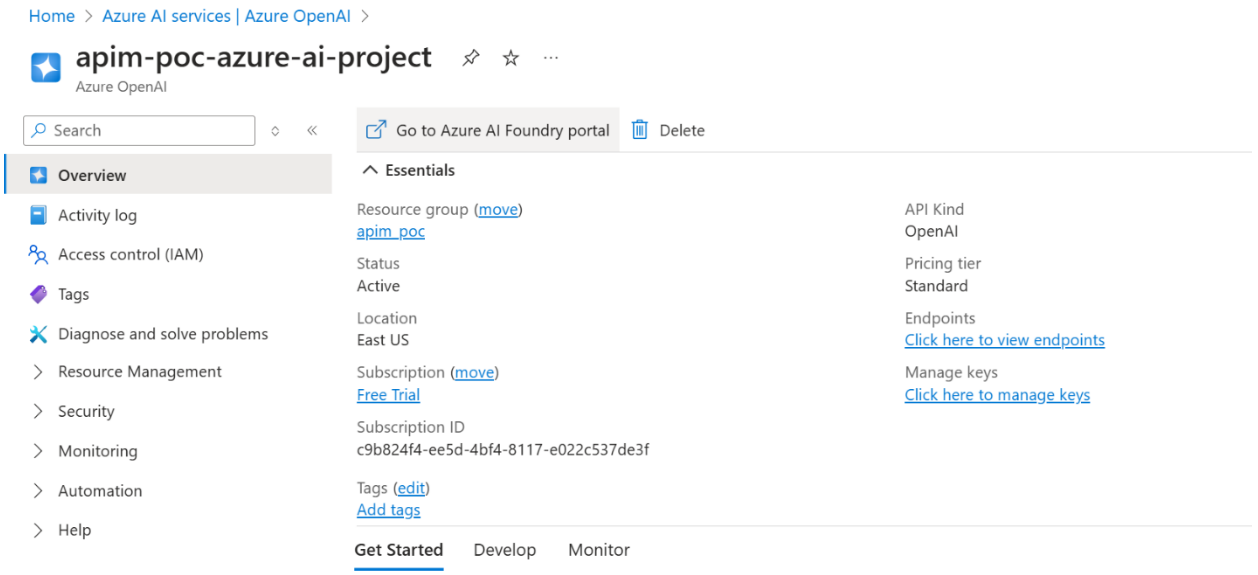
Azure AI services | Azure OpenAI

Showing 1 to 1 of 1 records.

Name	Kind	Location	Custom Domain ...	Pricing tier
apim-poc-azure-ai-project	OpenAI	East US	apim-poc-azure-a...	S0

Step 3: Deploy models inside the created Azure open AI service

1. Open the Azure open AI service that you created.
2. Go to the endpoint section and Copy the Service Endpoint.
3. Click on “Go to AI Foundry portal”



Resource group (move)
apim_poc

Status
Active

Location
East US

Subscription (move)
Free Trial

Subscription ID
c9b824f4-ee5d-4bf4-8117-e022c537de3f

Tags (edit)
Add tags

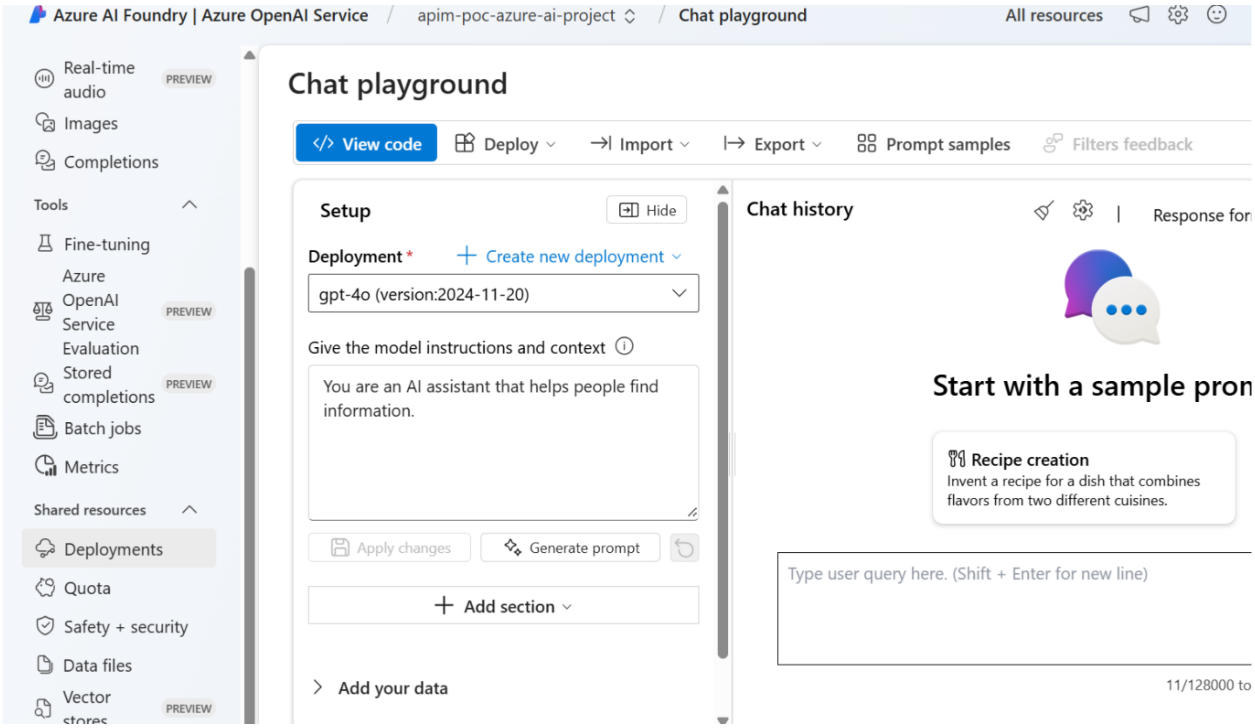
API Kind
OpenAI

Pricing tier
Standard

Endpoints
Click here to view endpoints

Manage keys
Click here to manage keys

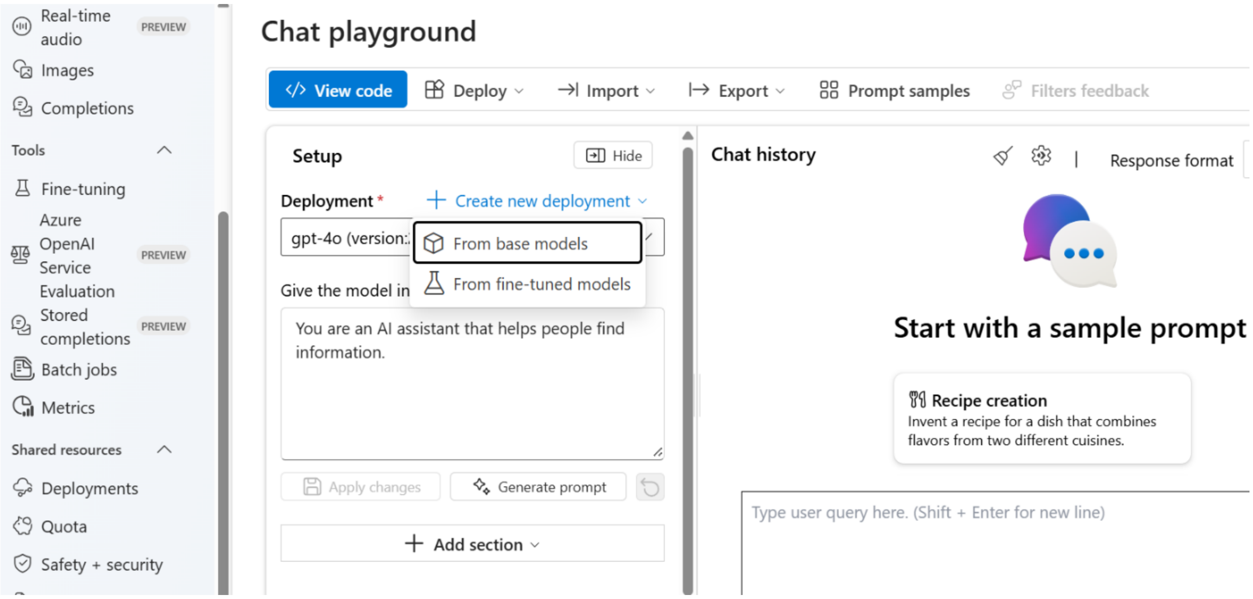
4. Go to the Deployments Section



Deployment *
gpt-4o (version:2024-11-20)

Give the model instructions and context
You are an AI assistant that helps people find information.

5. Create a new Deployment Model and Deploy the model you need in that section by selecting the model and version.



Deployment: gpt-4o (version:

Give the model in

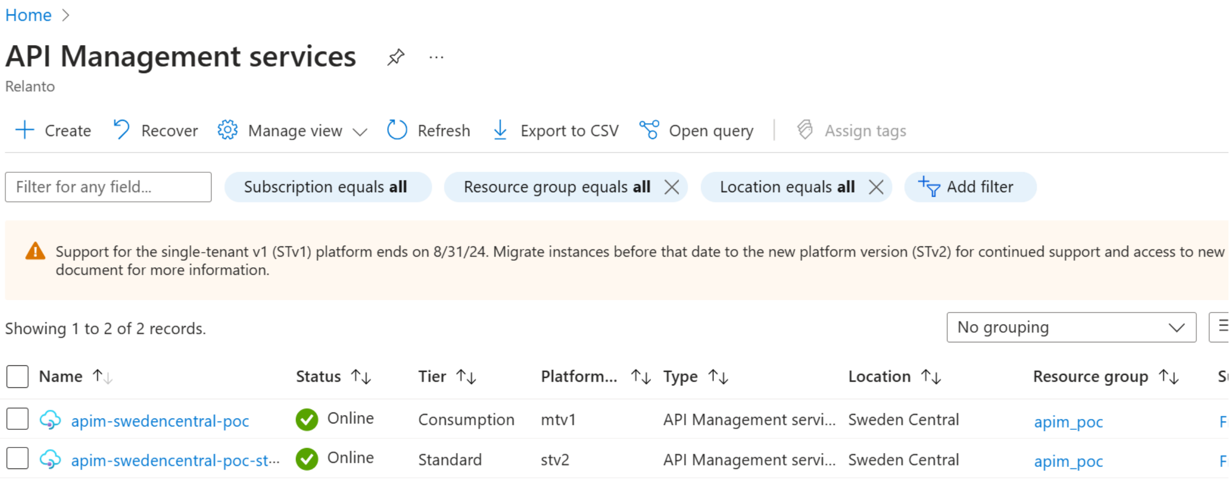
You are an AI assistant that helps people find information.

Recipe creation: Invent a recipe for a dish that combines flavors from two different cuisines.

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

Step 4: Create an APIM:

1. Go to **API Management services** and create APIM**:**



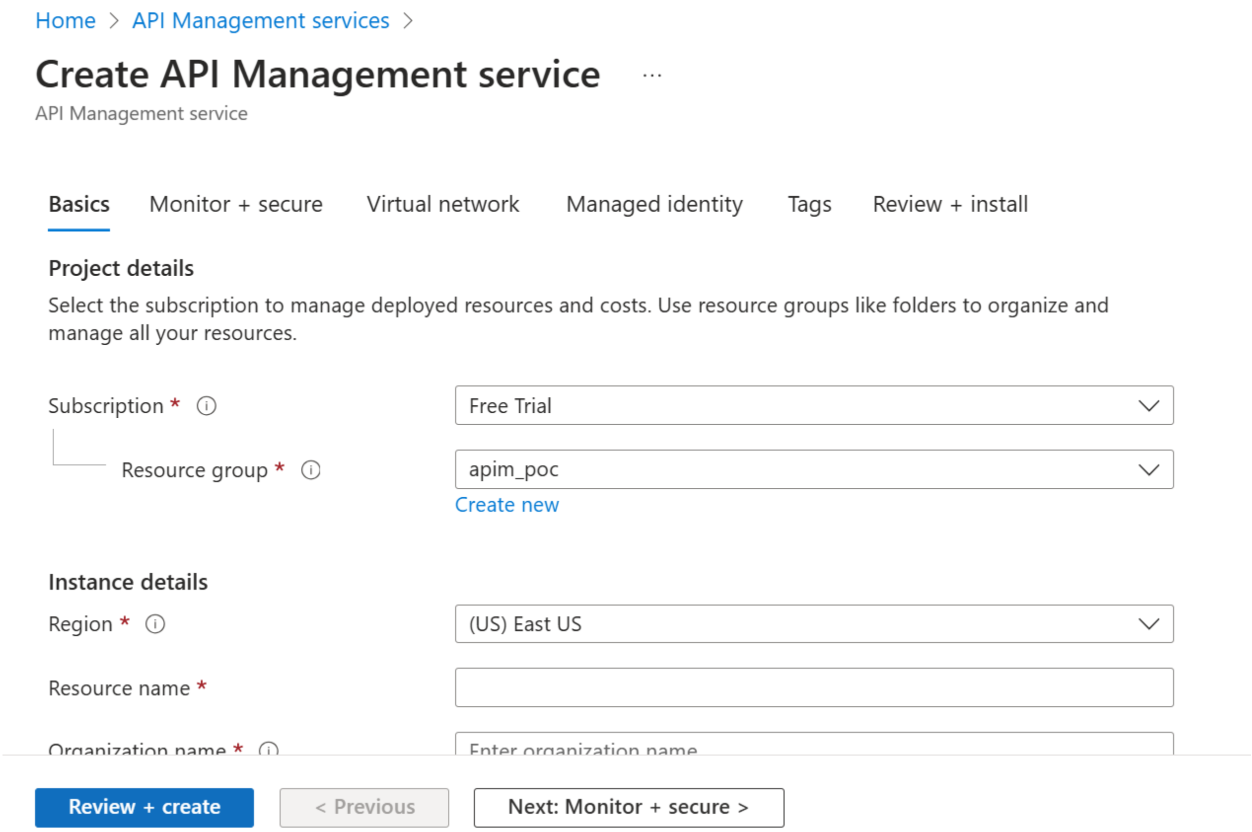
API Management services

Relanto

Support for the single-tenant v1 (STv1) platform ends on 8/31/24. Migrate instances before that date to the new platform version (STv2) for continued support and access to new document for more information.

Showing 1 to 2 of 2 records.

Name	Status	Tier	Platform...	Type	Location	Resource group	S
apim-swedencentral-poc	Online	Consumption	mtv1	API Management servi...	Sweden Central	apim_poc	F
apim-swedencentral-poc-st...	Online	Standard	stv2	API Management servi...	Sweden Central	apim_poc	F



Home > API Management services >

Create API Management service ...

API Management service

Basics | Monitor + secure | Virtual network | Managed identity | Tags | Review + install

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

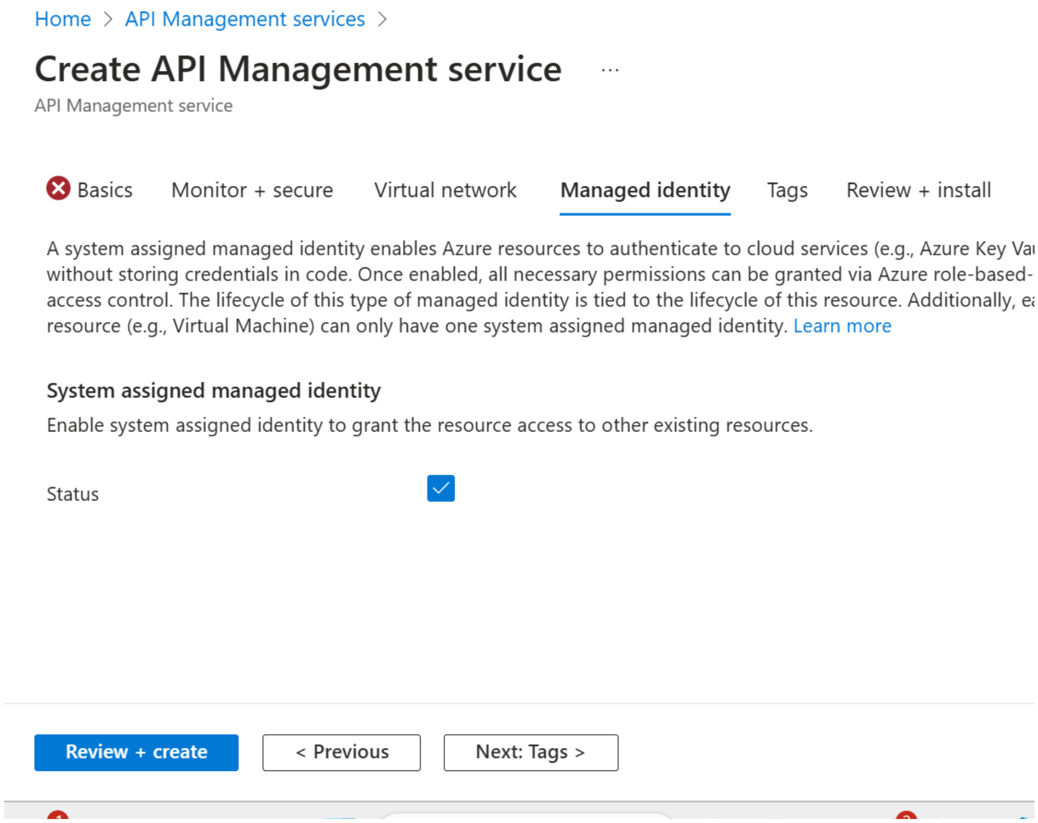
Subscription * ⓘ	Free Trial
Resource group * ⓘ	apim_poc Create new

Instance details

Region * ⓘ	(US) East US
Resource name *	
Organization name * ⓘ	Enter organization name

Note: 1. Use standard tier to meet the requirements

2. Enable status of Managed identity section



Home > API Management services >

Create API Management service ...

API Management service

☒ Basics | Monitor + secure | Virtual network | Managed identity | Tags | Review + install

A system assigned managed identity enables Azure resources to authenticate to cloud services (e.g., Azure Key Vau without storing credentials in code. Once enabled, all necessary permissions can be granted via Azure role-based-access control. The lifecycle of this type of managed identity is tied to the lifecycle of this resource. Additionally, ea resource (e.g., Virtual Machine) can only have one system assigned managed identity. Learn more

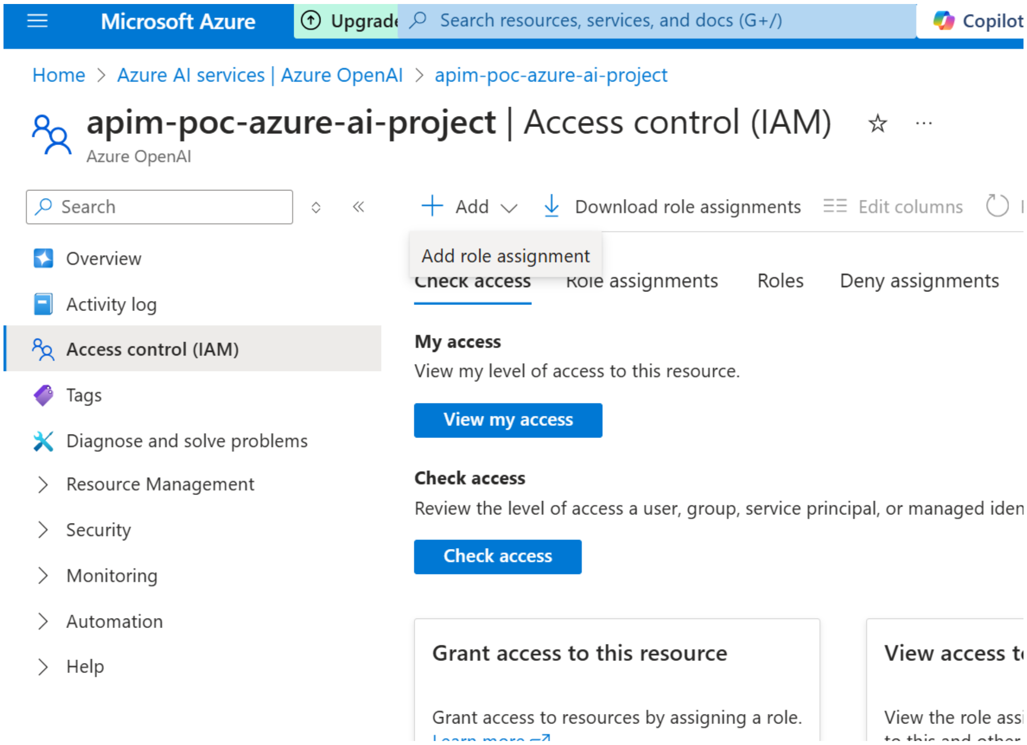
System assigned managed identity

Enable system assigned identity to grant the resource access to other existing resources.

Status ☑

Step 5: Authenticating Azure Openai service with APIM using Managed identity

1. Now after creating APIM go to the Azure openAI service's Access control and click on Add role assignment



My access

View my level of access to this resource.

Check access

Review the level of access a user, group, service principal, or managed iden

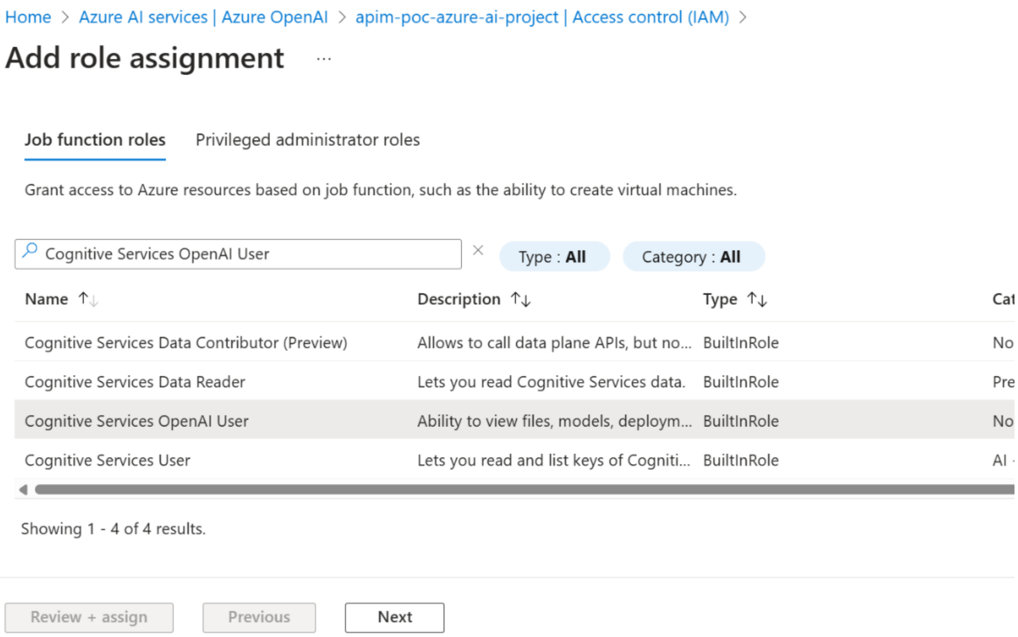
Grant access to this resource

Grant access to resources by assigning a role.

View access t

View the role ass

2. Add Cognitive Services OpenAI User(Built in) role:


Home > Azure AI services | Azure OpenAI > apim-poc-azure-ai-project | Access control (IAM) >
Add role assignment
Job function roles
Privileged administrator roles
Type : All
Category : All
Review + assign
Previous
Next

Grant access to Azure resources based on job function, such as the ability to create virtual machines.

Cognitive Services OpenAI User

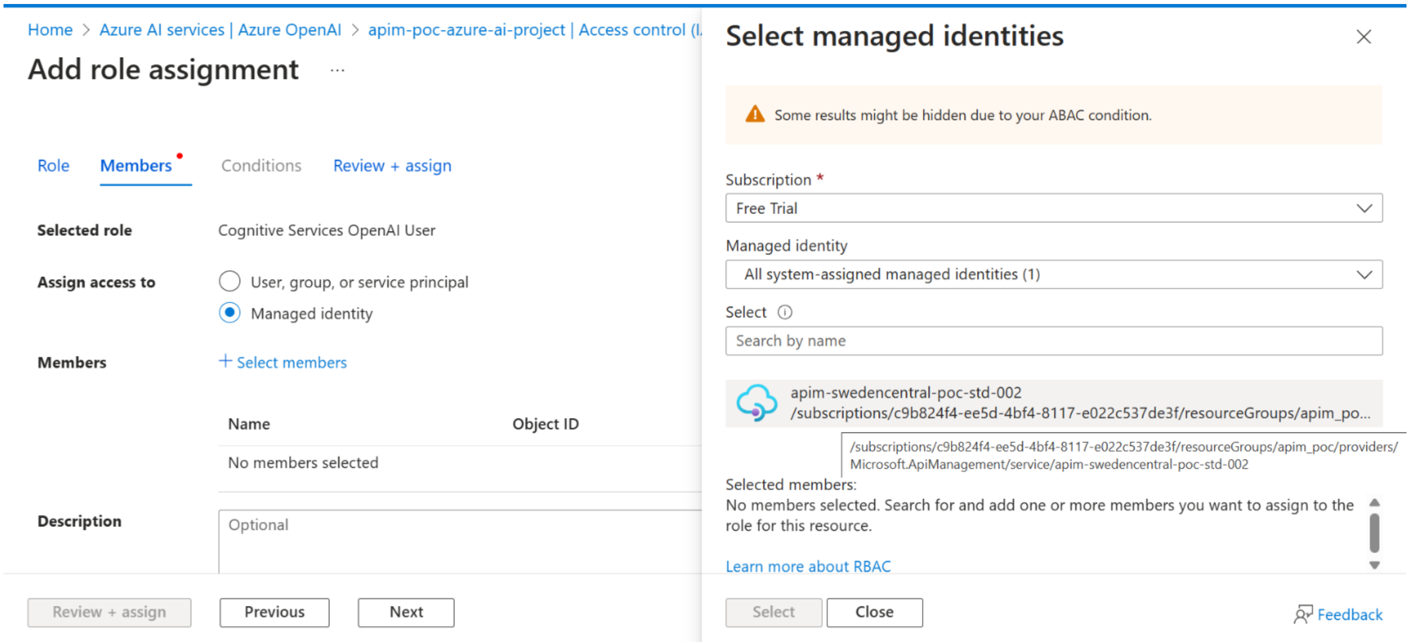
Name	Description	Type	Cat
Cognitive Services Data Contributor (Preview)	Allows to call data plane APIs, but no...	BuiltInRole	No
Cognitive Services Data Reader	Lets you read Cognitive Services data.	BuiltInRole	Pre
Cognitive Services OpenAI User	Ability to view files, models, deploym...	BuiltInRole	No
Cognitive Services User	Lets you read and list keys of Cogniti...	BuiltInRole	AI

Showing 1 - 4 of 4 results.

Note:

(1) Assign access to Managed identity

(2) Select the APIM according to the image below



Home > Azure AI services | Azure OpenAI > apim-poc-azure-ai-project | Access control (I

Add role assignment ···

Role | Members | Conditions | Review + assign

Selected role Cognitive Services OpenAI User

Assign access to ○ User, group, or service principal ● Managed identity

Members + Select members

Name	Object ID
No members selected	

Description Optional

Review + assign | Previous | Next

Select managed identities ×

Some results might be hidden due to your ABAC condition.

Subscription * Free Trial

Managed identity: All system-assigned managed identities (1)

Select ⓘ Search by name

apim-swedencentral-poc-std-002
/subscriptions/c9b824f4-ee5d-4bf4-8117-e022c537de3f/resourceGroups/apim_po...

/subscriptions/c9b824f4-ee5d-4bf4-8117-e022c537de3f/resourceGroups/apim_poc/providers/Microsoft.ApiManagement/service/apim-swedencentral-poc-std-002

Selected members:
No members selected. Search for and add one or more members you want to assign to the role for this resource.

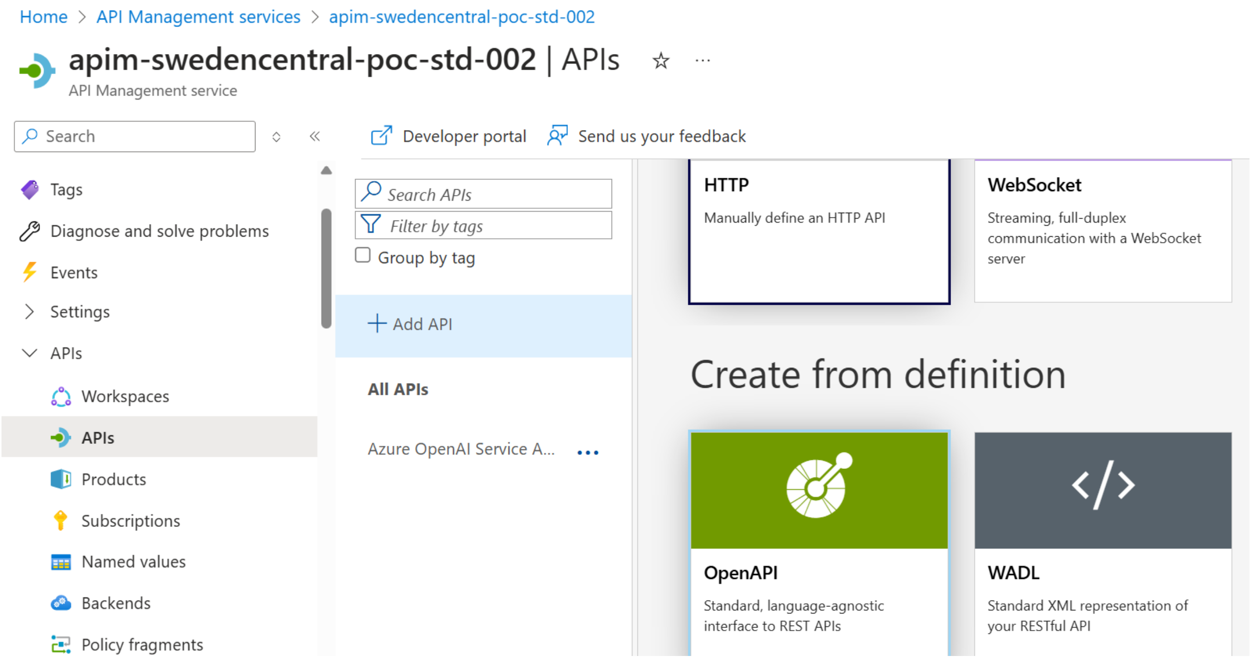
Learn more about RBAC

Select | Close | Feedback

Step 6:Go to API's Section of the APIM and Import OpenAPI(Swagger) documentation for the specific model and version you are using

Download the json file from below link and upload it in the next step(This is for openai gpt-4o model)

Ref: https://github.com/HoussemDellai/ai-course/blob/main/300_apim_genai_openai/openapi-AzureOpenAI-2024-10-21-



Home > API Management services > apim-swedencentral-poc-std-002

apim-swedencentral-poc-std-002 | APIs ☆ ···

API Management service

Search | Developer portal | Send us your feedback

Tags
Diagnose and solve problems
Events
Settings
APIs
Workspaces
APIs
Products
Subscriptions
Named values
Backends
Policy fragments

Search APIs
Filter by tags
□ Group by tag

+ Add API

All APIs

Azure OpenAI Service A... ···

HTTP Manually define an HTTP API

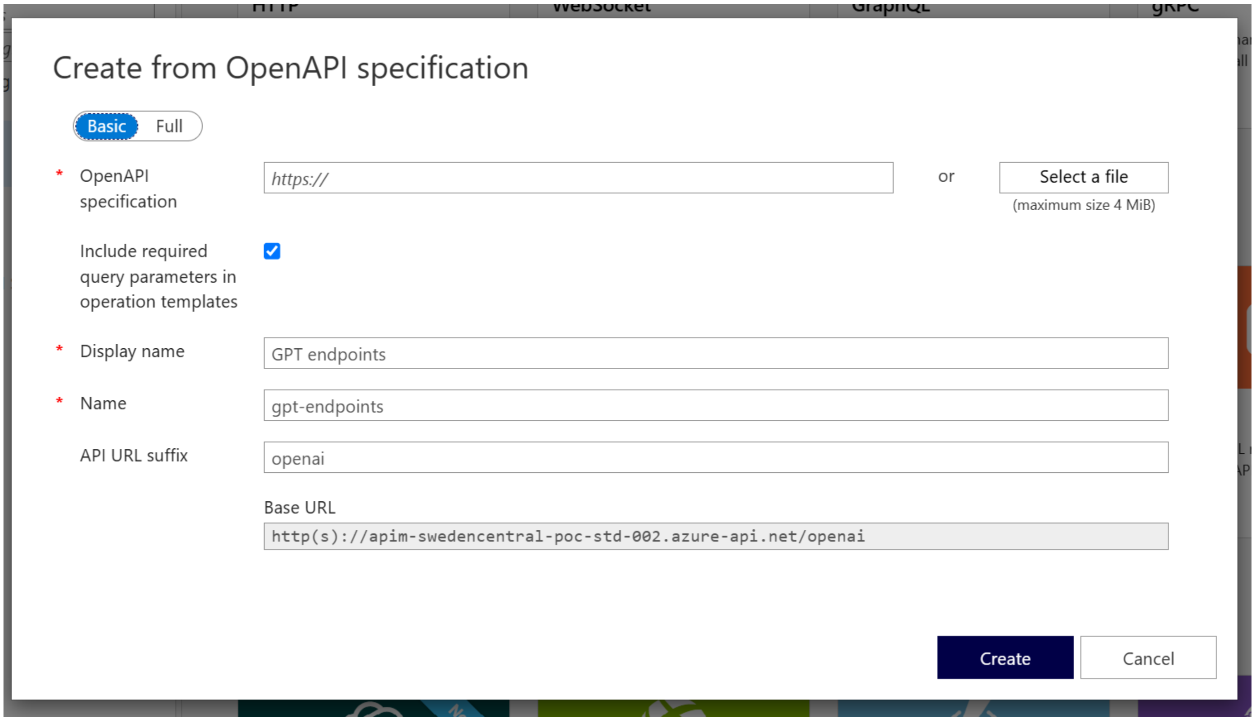
WebSocket Streaming, full-duplex communication with a WebSocket server

Create from definition

OpenAPI Standard, language-agnostic interface to REST APIs

WADL Standard XML representation of your RESTful API

Note: Add a suffix for your APIM URL endpoint.



Create from OpenAPI specification

Basic | Full

* OpenAPI specification: https:// or Select a file (maximum size 4 MiB)

Include required query parameters in operation templates: ☑

* Display name: GPT endpoints

* Name: gpt-endpoints

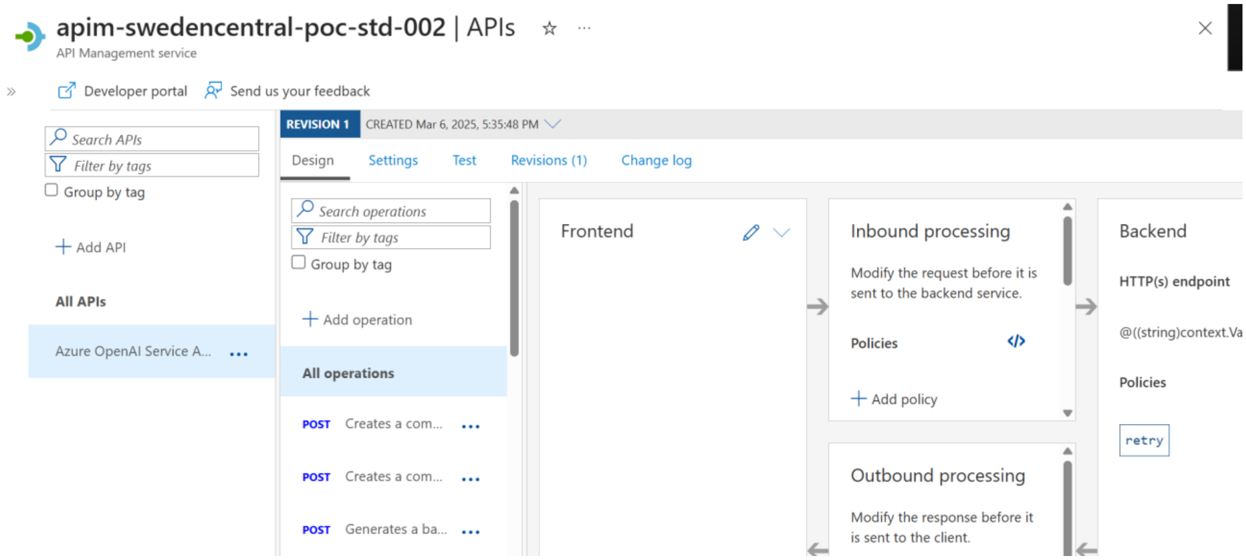
API URL suffix: openai

Base URL
`http(s)://apim-swedencentral-poc-std-002.azure-api.net/openai`

Create | Cancel

Step 7: Configuring APIM's API section:

1. Go to the API which was just imported and go to the design section



apim-swedencentral-poc-std-002 | APIs

API Management service

Developer portal | Send us your feedback

Search APIs

Filter by tags

□ Group by tag

+ Add API

All APIs

Azure OpenAI Service A...

REVISION 1 | CREATED Mar 6, 2025, 5:35:48 PM

Design | Settings | Test | Revisions (1) | Change log

Search operations

Filter by tags

□ Group by tag

+ Add operation

All operations

POST Creates a com...

POST Creates a com...

POST Generates a ba...

Frontend

Inbound processing

Modify the request before it is sent to the backend service.

Policies

+ Add policy

Outbound processing

Modify the response before it is sent to the client.

Backend

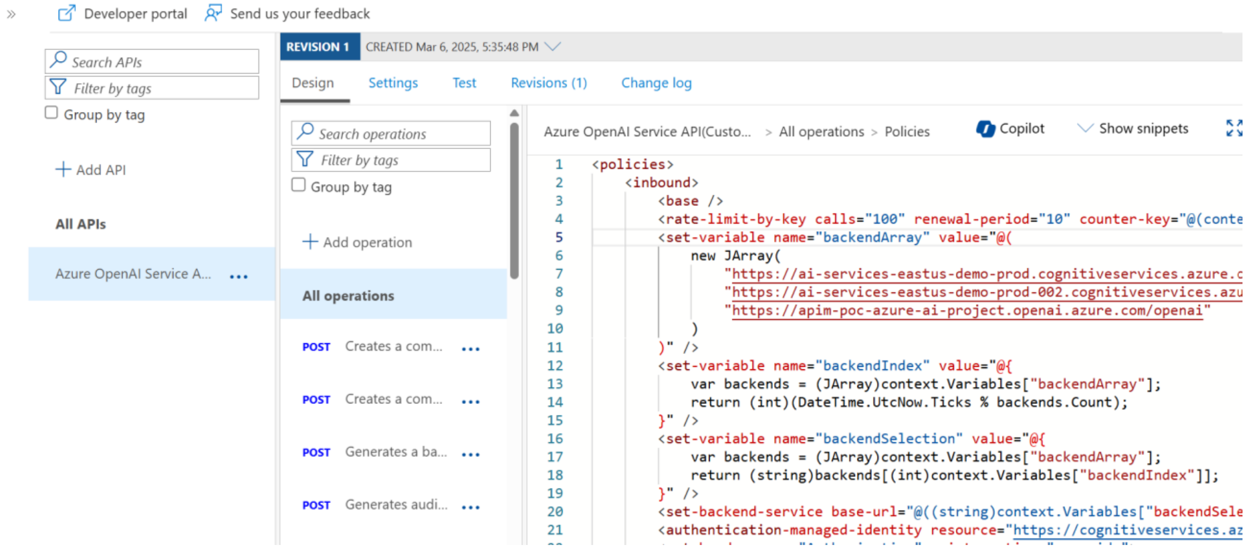
HTTP(s) endpoint

@((string)context.Va

Policies

retry

2. Go to the policy in the inbound processing and add the below policy

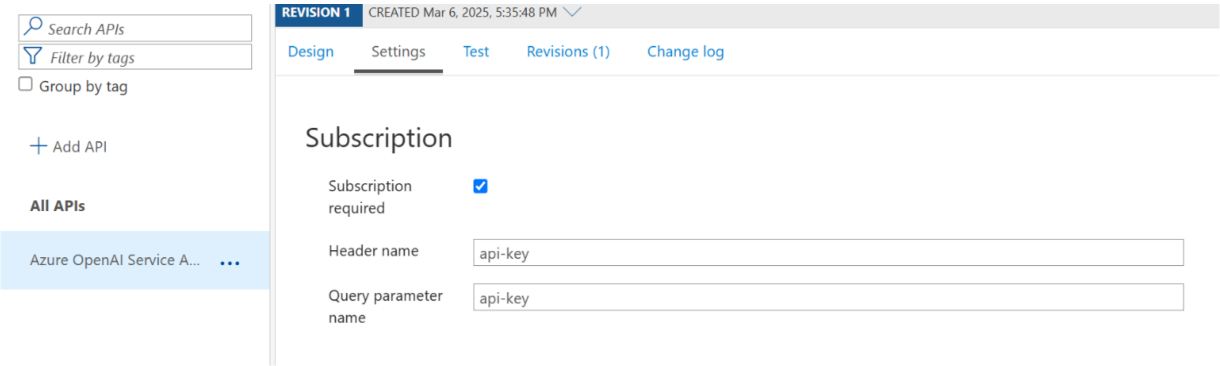


```
<policies>
    <inbound>
        <base />
        <rate-limit-by-key calls="100" renewal-period="10" counter-key="@(conte
        <set-variable name="backendArray" value="@(
            new JArray(
                "https://ai-services-eastus-demo-prod.cognitiveservices.azure.c
                "https://ai-services-eastus-demo-prod-002.cognitiveservices.azu
                "https://apim-poc-azure-ai-project.openai.azure.com/openai"
            )
        )" />
        <set-variable name="backendIndex" value="@{
            var backends = (JArray)context.Variables["backendArray"];
            return (int)(DateTime.UtcNow.Ticks % backends.Count);
        }" />
        <set-variable name="backendSelection" value="@{
            var backends = (JArray)context.Variables["backendArray"];
            return (string)backends[(int)context.Variables["backendIndex"]];
        }" />
        <set-backend-service base-url="@((string)context.Variables["backendSele
        <authentication-managed-identity resource="https://cognitiveservices.az
```

Policy used (Replace your backend URL’s in the JArray section):

```
<policies>
  <inbound>
    <base />
    <rate-limit-by-key calls="100" renewal-period="10" counter-key="@(context.Request.IpAddress)" />
    <set-variable name="backendArray" value="@(
      new JArray(
        "https://ai-services-eastus-demo-prod.cognitiveservices.azure.com/openai",
        "https://ai-services-eastus-demo-prod-002.cognitiveservices.azure.com/openai",
        "https://apim-poc-azure-ai-project.openai.azure.com/openai"
      )
    )" />
    <set-variable name="backendIndex" value="@{
      var backends = (JArray)context.Variables["backendArray"];
      return (int)(DateTime.UtcNow.Ticks % backends.Count);
    }" />
    <set-variable name="backendSelection" value="@{
      var backends = (JArray)context.Variables["backendArray"];
      return (string)backends[(int)context.Variables["backendIndex"]];
    }" />
    <set-backend-service base-url="@((string)context.Variables["backendSelection"])" />
    <authentication-managed-identity resource="https://cognitiveservices.azure.com" output-token-variable-name="msi-access-token" ignore-error="false" />
    <set-header name="Authorization" exists-action="override">
      <value>@("Bearer " + (string)context.Variables["msi-access-token"])</value>
    </set-header>
  </inbound>
  <backend>
    <retry condition="@(context.Response == null || context.Response.StatusCode == 500 || context.Response.StatusCode == 429 || context.Response.StatusCode != 200)" count="20" interval="1">
      <forward-request timeout="@((100 + (context.Variables.ContainsKey(&quot;retryCount&quot;) ? (int)context.Variables[&quot;retryCount&quot;] * 100 : 0)))" buffer-request-body="true" />
      <set-variable name="backendIndex" value="@{
        var backends = (JArray)context.Variables["backendArray"];
        return ((int)context.Variables["backendIndex"] + 1) % backends.Count;
      }" />
      <set-variable name="backendSelection" value="@{
        var backends = (JArray)context.Variables["backendArray"];
        return (string)backends[(int)context.Variables["backendIndex"]];
      }" />
      <set-backend-service base-url="@((string)context.Variables["backendSelection"])" />
    </retry>
  </backend>
  <outbound>
    <base />
    <set-header name="X-Selected-Backend" exists-action="override">
      <value>@((string)context.Variables["backendSelection"])</value>
    </set-header>
  </outbound>
  <on-error>
    <base />
  </on-error>
</policies>
```

3. Specify the header and query parameter name and enable the checkbox(This will be used while making an API call)



Subscription

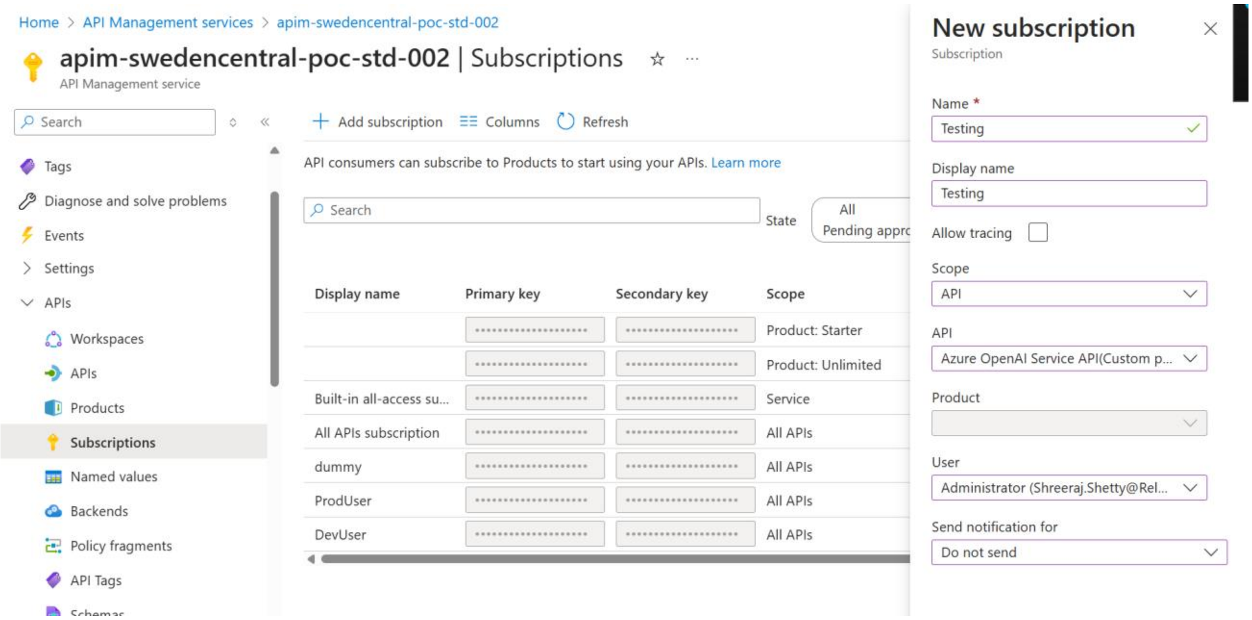
Subscription required ☑

Header name: api-key

Query parameter name: api-key

Step 8: Subscription Key Creation

1. Create a subscription key in the Subscription section of APIs under APIM



API consumers can subscribe to Products to start using your APIs. Learn more

Display name	Primary key	Secondary key	Scope
	••••••••••••••••••••	••••••••••••••••••••	Product: Starter
	••••••••••••••••••••	••••••••••••••••••••	Product: Unlimited
Built-in all-access su...	••••••••••••••••••••	••••••••••••••••••••	Service
All APIs subscription	••••••••••••••••••••	••••••••••••••••••••	All APIs
dummy	••••••••••••••••••••	••••••••••••••••••••	All APIs
ProdUser	••••••••••••••••••••	••••••••••••••••••••	All APIs
DevUser	••••••••••••••••••••	••••••••••••••••••••	All APIs

Name * Testing

Display name Testing

Allow tracing □

Scope API

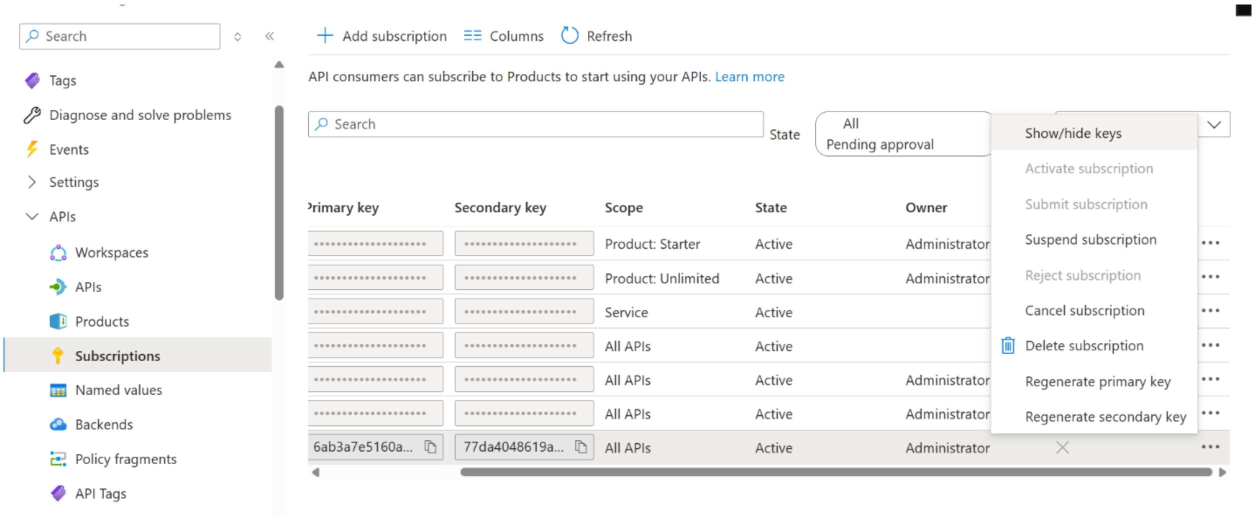
API Azure OpenAI Service API(Custom p...

Product

User Administrator (Shreeraj.Shetty@Rel...

Send notification for Do not send

2. Retrieve the key and copy



API consumers can subscribe to Products to start using your APIs. Learn more

›rimary key	Secondary key	Scope	State	Owner
••••••••••••••••••••	••••••••••••••••••••	Product: Starter	Active	Administrator
••••••••••••••••••••	••••••••••••••••••••	Product: Unlimited	Active	Administrator
••••••••••••••••••••	••••••••••••••••••••	Service	Active	
••••••••••••••••••••	••••••••••••••••••••	All APIs	Active	
••••••••••••••••••••	••••••••••••••••••••	All APIs	Active	Administrator
••••••••••••••••••••	••••••••••••••••••••	All APIs	Active	Administrator
6ab3a7e5160a...	77da4048619a...	All APIs	Active	Administrator

Step 9: Testing

This Apim’s performance can be tested using Apache benchmark docker image

Run the below Docker command:

```
docker run --rm -v <Path to your payload file>:/data jordi/ab -n 10 -c 10 -l -T "application/json" -H "api-key: <your-subscription-key>" -p /data/apim-body.json -v 2 -e /data/apim-errors.csv -g /data/apim-results.tsv https://apim-swedencentral-poc-std-002.azure-api.net/openai/deployments/gpt-4o/chat/completions?api-version=2024-08-01-preview
```

Note:

-n represents the number of requests

-c represents the concurrency

That means in every second c number of requests are sent parallelly.

Step 10: Checking the logs

Go to the Monitoring section of the APIM and query to get the logs.

Query:

1. **To get the Details:**

```
ApiManagementGatewayLogs

| where TimeGenerated > ago(1d)

| where IsRequestSuccess == true

| where ResponseCode == 200

| project TimeGenerated, BackendUrl| top 10 by TimeGenerated desc
```

2. To get the Count of the Requests hit to particular Backend:

```
ApiManagementGatewayLogs
| where TimeGenerated > ago(1d)
| where IsRequestSuccess == true
| where ResponseCode == 200
| project TimeGenerated, BackendUrl
| top 50 by TimeGenerated desc  // Fetch last 30 requests
| summarize HitCount = count() by BackendUrl
| order by HitCount desc
```