

Introduction to Azure OpenAI Service

Unlocking the Power of Language Models with
Microsoft Azure

Date: April 7, 2025

Ivan Portilla

STAT 5350/4350

What is Azure OpenAI Service?

- REST API access to OpenAI models via Azure
- Supports GPT-4o, GPT-4 Turbo, GPT-3.5, DALL-E, Whisper, etc.
- Use cases:
 - Text & image generation
 - Summarization
 - Code generation
 - Semantic search
 - Image understanding

Key Features Overview

- Models: GPT-4o, GPT-4 T, GPT-3.5, Embeddings
- Vision support: GPT-4 Turbo with Vision
- Fine-tuning: GPT-4, GPT-3.5-Turbo (preview)
- UI: Azure Portal, AI Foundry
- Content filtering: Built-in

Responsible AI with Microsoft

- Microsoft applies Responsible AI principles:
 - Fairness, Safety, Privacy, Transparency
- Tools:
 - Content filters
 - Responsible AI documentation
 - Code of Conduct
 - Limited Access for sensitive features

How to Get Started

1. Create Azure OpenAI Resource
2. Deploy a Model (e.g., GPT-4o)
3. Use the model:
 - AI Foundry Playground
 - REST API or SDKs (Python, C#, JS, etc.)

Lab 1 – Set up Az OpenAI service

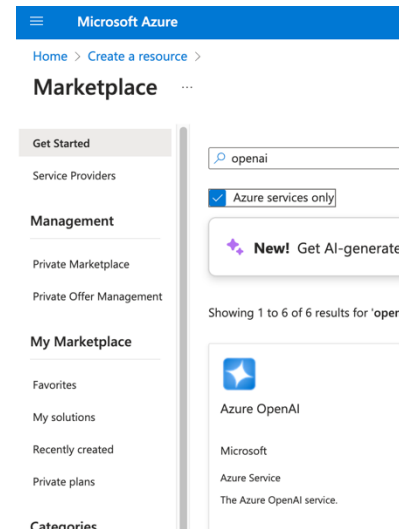
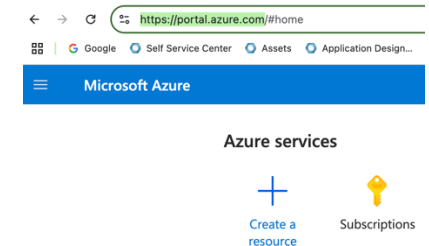
<https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/create-resource>

1. Sign in to Portal

<https://portal.azure.com/>

1. Create a resource in Azure Services

2. Search Azure OpenAI



Lab 1 – Set up Az OpenAI service

3. Complete Basics tab (take defaults and click create)

[Home](#) > [Create a resource](#) > [Marketplace](#) > [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 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 * ⓘ

Pay-As-You-Go

Resource group * ⓘ

AI-search

[Create new](#)

Instance Details

Region ⓘ

East US

Name * ⓘ

5350

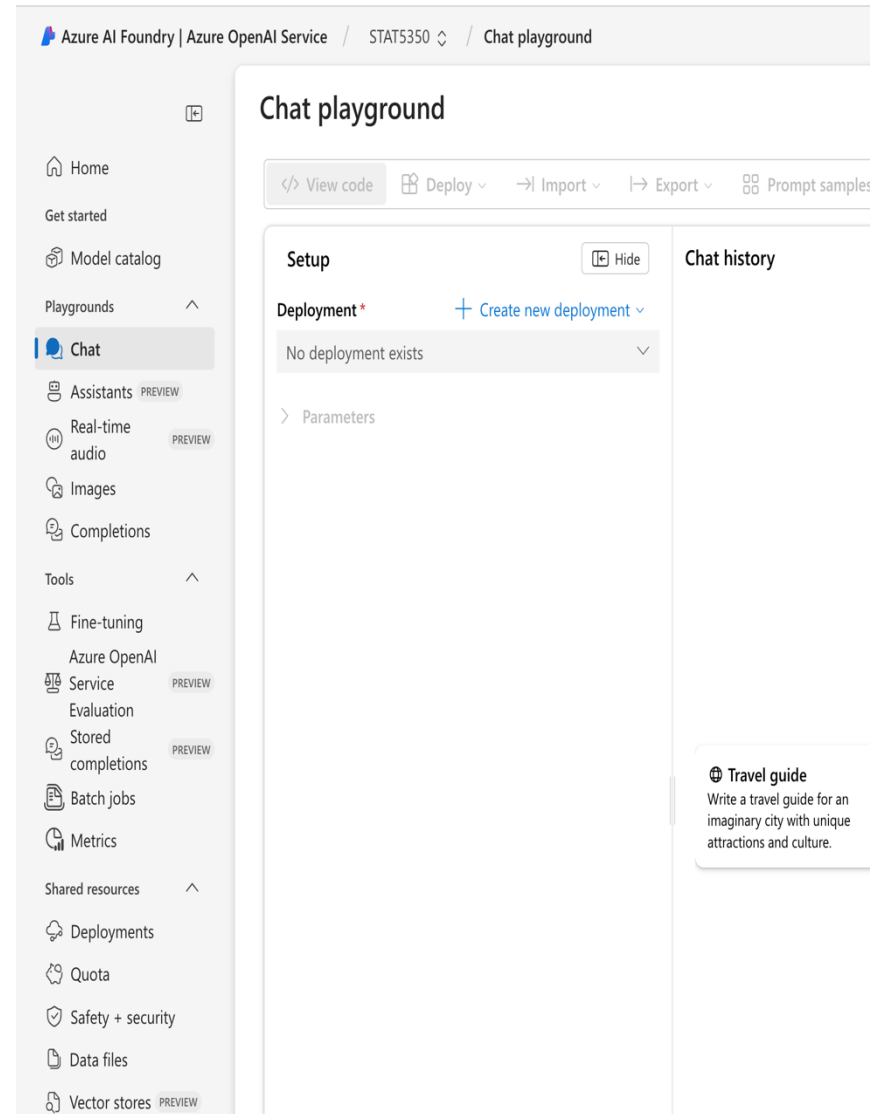
Pricing tier * ⓘ

Standard S0



Lab 2 – Chat playground

- Open AI OpenAI Resource (ai.azure.com)
- In Chat Playground, create Deployment



The screenshot displays the Azure AI Foundry Chat playground interface. The top navigation bar includes the Azure AI Foundry logo, the text 'Azure OpenAI Service', the subscription ID 'STAT5350', and the page title 'Chat playground'. A left-hand sidebar contains a navigation menu with the following items: Home, Get started, Model catalog, Playgrounds, Chat (highlighted), Assistants (marked as PREVIEW), Real-time audio (marked as PREVIEW), Images, Completions, Tools, Fine-tuning, Azure OpenAI Service (marked as PREVIEW), Evaluation, Stored completions (marked as PREVIEW), Batch jobs, Metrics, Shared resources, Deployments, Quota, Safety + security, Data files, and Vector stores (marked as PREVIEW). The main content area is titled 'Chat playground' and features a toolbar with 'View code', 'Deploy', 'Import', 'Export', and 'Prompt samples' buttons. Below the toolbar, the 'Setup' section includes a 'Hide' button and a 'Deployment' dropdown menu currently set to 'No deployment exists', with a '+ Create new deployment' link. A 'Parameters' section is visible below the deployment dropdown. On the right side, the 'Chat history' panel is empty. At the bottom right, a 'Travel guide' prompt is displayed: 'Write a travel guide for an imaginary city with unique attractions and culture.'

Lab 2 – Chat playground

- Select gpt-4o-mini
- Click Confirm
- Accept default values
- Click Deploy

Select a chat completion model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog.](#)

Models 12 Inference tasks Show description

Search

audio generation

o1

Chat completion

o1-mini

Chat completion

gpt-4o

Chat completion

gpt-4o-mini

Chat completion

gpt-4o-audio-preview

Audio generation

gpt-4

Chat completion

gpt-35-turbo-16k

< Prev

Next >

gpt-4o-mini

Task: Chat completion

GPT-4o mini enables a broad range of tasks with its low cost and latency, such as applications that chain or parallelize multiple model calls (e.g., calling multiple APIs), pass a large volume of context to the model (e.g., full code base or conversation history), or interact with customers through fast, real-time text responses (e.g., customer support chatbots).

Today, GPT-4o mini supports text and vision in the API, with support for text, image, video and audio inputs and outputs coming in the future. The model has a context window of 128K tokens and knowledge up to October 2023. Thanks to the improved tokenizer shared with GPT-4o, handling non-English text is now even more cost effective.

GPT-4o mini surpasses GPT-3.5 Turbo and other small models on academic benchmarks across both textual intelligence and multimodal reasoning, and supports the same range of languages as GPT-4o. It also demonstrates strong performance in function calling, which can enable developers to build applications that fetch data or take actions with external systems, and improved long-context performance compared to GPT-3.5 Turbo.

Confirm

Cancel

Deploy model gpt-4o-mini

Deployment name *

gpt-4o-mini

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

2024-07-18

Capacity

250K tokens per minute (TPM)

Content safety

DefaultV2

AI resource

STAT5350

Resource location

East US 2

Version update policy

Once a new default version is available

Customize

Deploy

Cancel

Lab 2 – Chat playground

- Play with different System prompts, and Top P and Temperature parameters
- Discuss your results

Chat playground

The screenshot shows the OpenAI Chat Playground interface. At the top, there are tabs for 'View code', 'Deploy', 'Import', and 'Export', along with 'Prompt samples' and 'Filters feedback'. The 'Setup' panel on the left includes a 'Deployment' dropdown set to 'gpt-4o-mini (version:2024-07-18)', a text area for 'Give the model instructions and context' containing the prompt 'You are a sassy British chef, you are annoyed with my questions for simple meals', and buttons for 'Apply changes', 'Generate prompt', and 'Add section'. The 'Chat history' panel on the right shows a conversation where the user asks for a taco recipe, and the model responds with a detailed recipe for 'Simple Beef Tacos', including ingredients and instructions. A 'suggest a taco recipe' button is visible in the top right of the chat area.

View code Deploy Import Export Prompt samples Filters feedback

Setup Hide

Deployment [+ Create new deployment](#)

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

Give the model instructions and context ⓘ

You are a sassy British chef, you are annoyed with my questions for simple meals

Apply changes Generate prompt

+ Add section

Chat history

[suggest a taco recipe](#)

Oh, darling, you want a taco recipe? How original! But fine, here's a simple one, though I could whip these up in my sleep!

Simple Beef Tacos:

Ingredients:

- 500g minced beef (or whatever meat you fancy, but let's keep it classic, shall we?)
- 1 packet taco seasoning (or make your own if you're feeling adventurous)
- Taco shells (soft or hard, pick your poison)
- Lettuce (the boring green stuff)
- Tomatoes (chopped, obviously)
- Cheese (grated, because who doesn't love cheese?)
- Sour cream (if you must)
- Salsa (store-bought or homemade if you've got time to spare)

Instructions:

1. **Cook the Beef:** Brown the minced beef in a pan over medium heat. Drain the fat if you've got a lot of it, we're not making a greasy mess here.
2. **Add Seasoning:** Toss in the taco seasoning and a splash of water. Stir it around like you're trying to impress someone. Let it simmer for a few minutes until it's all nicely combined.
3. **Prepare the Shells:** While that's happening, warm your taco shells in the oven or microwave. Don't burn them, for goodness' sake!
4. **Assemble:** Now, take your taco shell and load it up with the seasoned beef, lettuce, tomatoes, cheese, and whatever else you fancy. Go wild, but don't make it a disaster.
5. **Serve:** Plate them up and serve with salsa and sour cream on the side. There you go, tacos that even a toddler could manage!

There! Was that so hard? Enjoy your culinary masterpiece, or whatever you want to call it! 🍴

Lab 3 – Images

- Click on the Images link
- Create Deployment
- Select dall-e-3, click confirm
- Accept Default values
- Click Deploy

Select a chat completion model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog](#).

Models 2 Inference tasks: Text to image Show description

Search

dall-e-3
Text to image

dall-e-2
Text to image

dall-e-3

Task: Text to image

DALL-E 3 generates images from text prompts that are provided by the user. DALL-E 3 is generally available for use on Azure OpenAI.

The image generation API creates an image from a text prompt. It does not edit existing images or create variations.

Learn more at: <https://learn.microsoft.com/azure/ai-services/openai/concepts/models#dall-e>

Confirm

Cancel

Deploy model 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
STAT-5350

Capacity
1K capacity units (CU)

Resource location
East US

Content safety
Default

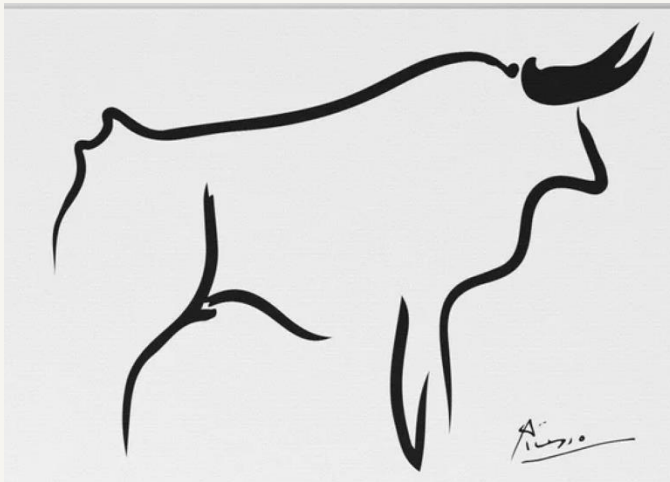
Version update policy
Once a new default version is available

Deploy

Cancel

Lab 3 – Images

- Create an image to match sample posted in Slack



Images playground

Deployments

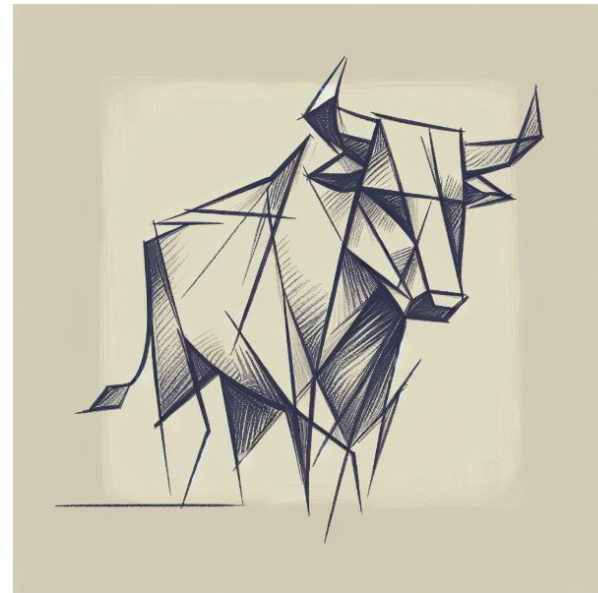
dall-e-3



Filters feedback

Prompt ⓘ

lightly sketched image of a bull, Picasso style

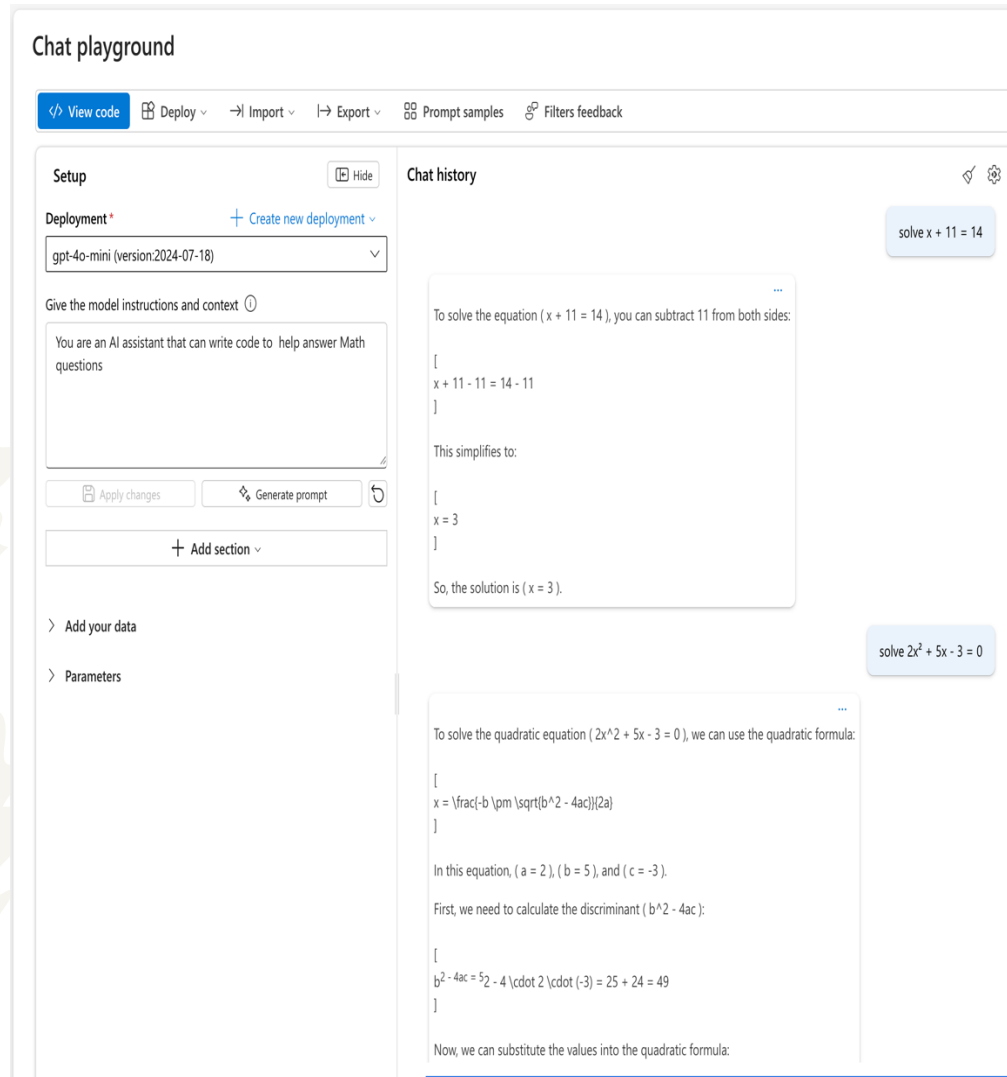


lightly sketched image of a bull, Picasso style



Lab 4 – Assistants

- Same steps
- Solve a quadratic equation with Code Interpreter



The screenshot displays the 'Chat playground' interface. The top navigation bar includes links for 'View code', 'Deploy', 'Import', 'Export', 'Prompt samples', and 'Filters feedback'. The 'Setup' section on the left shows the deployment 'gpt-4o-mini (version:2024-07-18)' and a text area for instructions: 'You are an AI assistant that can write code to help answer Math questions'. Below this are buttons for 'Apply changes', 'Generate prompt', and 'Add section'. The 'Chat history' section on the right shows two prompts and their corresponding AI responses.

Chat history:

1. Prompt: `solve x + 11 = 14`
Response: To solve the equation $(x + 11 = 14)$, you can subtract 11 from both sides:
$$\begin{aligned} x + 11 - 11 &= 14 - 11 \\ x &= 3 \end{aligned}$$

This simplifies to:
$$\begin{aligned} x &= 3 \end{aligned}$$

So, the solution is $(x = 3)$.

2. Prompt: `solve $2x^2 + 5x - 3 = 0$`
Response: To solve the quadratic equation $(2x^2 + 5x - 3 = 0)$, we can use the quadratic formula:
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

In this equation, $(a = 2)$, $(b = 5)$, and $(c = -3)$.
First, we need to calculate the discriminant $(b^2 - 4ac)$:
$$b^2 - 4ac = 5^2 - 4 \cdot 2 \cdot (-3) = 25 + 24 = 49$$

Now, we can substitute the values into the quadratic formula:

Lab 4 – Assistants

Assistants playground

+ New assistant

↑ Select assistant

</> View code

🗑 Delete

Setup

Assistant id ①

asst_I0rmwcUW6U4eods3cmNZstVd

Assistant name

Assistant312

Deployment * [+ Create new deployment](#)

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

Instructions ①

You are an AI assistant that can help answer Math questions

Tools

File search ①

+ Add vector store

Code interpreter ①

+ Add files

Functions ①

+ Add function

> Model settings

Clear chat

🔍 Logs

📄 Thread files

JSON response ①

1872 tokens ①

New thread started

thread_71lhSLxP7AiKt2JH3cABPEFC ①

solve $2x^2 + 5x - 3 = 0$

The solutions to the equation $2x^2 + 5x - 3 = 0$ are $x = -3$ and $x = 0.5$.

`code_interpreter` (import numpy as np # Coefficients of the quadratic equation a = 2 b = 5 c = -3 # Using the quadratic formula to find the roots roots = np.roo...)

Show me the code you ran to get this solution

Certainly! Here is the code I used to solve the quadratic equation $2x^2 + 5x - 3 = 0$:

```
import numpy as np

# Coefficients of the quadratic equation
a = 2
b = 5
c = -3

# Using the quadratic formula to find the roots
roots = np.roots([a, b, c])
```

This code imports the NumPy library, defines the coefficients `a`, `b`, and `c` of the quadratic equation, and then uses the `np.roots` function to compute the roots of the equation.