Introduction to Azure OpenAI

What is Azure OpenAI Service?

What is Azure OpenAI Service?

Overview

- Fully managed offering from Microsoft
- Integrates OpenAI’s advanced AI models into Azure cloud platform
- Enables applications like NLP, computer vision, and more

What is Azure OpenAI Service?

💖 Partnership & Features

- Developed in partnership with OpenAI
- Enterprise-grade features, robust security, data privacy, and regulatory compliance
- Incorporates responsible AI practices

What is Azure OpenAI Service?

🔍 Use Cases

- Content generation, summarization, image understanding, vision, audio
- Semantic search and natural language to code translation
- … and many more.

What is Azure OpenAI Service?

🔑 Access Options

- REST APIs
- Programming Languages / SDK
- Azure AI Foundry

Key Features & Concepts

1. Available Models
2. Custom Model Training with Fine-tuning
3. Capabilities and APIs
4. Supported Programming Languages & SDKs
5. Quotas and Limits
6. Cost Management and Pricing
7. ... and there are many more.

Available Models

Available Models

- o-series models (o3-mini, o1 & o1-mini)
- GPT-4o & GPT-4o mini & GPT-4 Turbo
- GPT-4o audio
- GPT-4
- GPT-3.5
- Embeddings
- DALL-E
- Whisper
- Text to speech (Preview)

o-series models

Reasoning models with advanced problem-solving and increased focus and capability.

GPT-4o & GPT-4o mini & GPT-4 Turbo

The latest most capable Azure OpenAI models with multimodal versions, which can accept both text and images as input.

GPT-4o audio

GPT-4o audio models that support either low-latency, "speech in, speech out" conversational interactions or audio generation.

GPT-4

A set of models that improve on GPT-3.5

and can understand and

generate natural language and code.

GPT-3.5

A set of models that improve on GPT-3

and can understand and

generate natural language and code.

Embeddings

A set of models that can convert text into numerical vector form to facilitate text similarity.

DALL-E

A series of models that can generate original images from natural language.

Whisper

A series of models in preview that can transcribe and translate speech to text.

Text to speech (Preview)

A series of models in preview that can synthesize text to speech.

Important Fact

Retirement:

When a model is retired, it's no longer available for use, and deployments of retired models will return errors.

Deprecation:

A deprecated model is not available to new customers but remains accessible for existing users until it's retired.

Custom Model Training with Fine-tuning

Custom Model Training with Fine-tuning

🔧 Overview of Fine-tuning

- Fine-tune models based on specific requirements
- Tailor model performance for unique use cases (e.g., specialized data, custom output)

Custom Model Training with Fine-tuning

🧠 What is Fine-tuning?

- Retrain a pre-trained model using our own dataset
- Customizes behavior for specific tasks or edge cases (e.g., specific language style or complex tasks)

Custom Model Training with Fine-tuning

💡 Effective Fine-tuning Tips

- Clear use case: Improve output style, handle complex tasks
- Try alternatives: Use prompt engineering or Retrieval Augmented Generation (RAG) first
- Prepare quality data: Ensure dataset is large, high-quality, and properly formatted
- Set success criteria: Use validation data and user testing for measurable outcomes

Custom Model Training with Fine-tuning

⚙️ Key Considerations

- Requires good data, understanding of model limitations, and success measurement methods
- … etc.

Capabilities and APIs

Capabilities and APIs

- Text Generation
- Vision
- Image Generation
- Text to Speech
- Embeddings
- Reasoning
- Function Calling
- Vector Stores
- Chat
- Assistants

... and there are many more.

Text Generation

This capability allows the model to generate human-like text based on a given prompt, useful for writing, summarizing, or dialogue creation.

What is Azure OpenAI

Azure OpenAI refers to the integration of OpenAI's powerful language models and artificial intelligence capabilities into Microsoft Azure, which is Microsoft's cloud computing platform. This partnership allows businesses and developers to access OpenAI's advanced models, such as GPT-3 and subsequent versions, through Azure's infrastructure and services.

Key features of Azure OpenAI include:

1. **Access to Advanced AI Models:** Users can leverage OpenAI's state-of-the-art language models to generate text, summarize information, engage in conversational AI, and more.
2. **Scalability and Reliability:** By hosting OpenAI's models on Azure, users benefit from the platform's scalability, security, and global reach.
3. **Integration with Azure Services:** Azure OpenAI can be integrated with other Azure services, enabling users to build sophisticated applications that leverage AI for various use cases, such as chatbots, content generation, data analysis, and more.
4. **Compliance and Security:** Microsoft emphasizes compliance and security, providing features that help organizations meet regulatory requirements while utilizing AI technologies.

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

279/128000 tokens to be sent

Vision

Enables the model to analyze and understand visual data, making it possible to perform tasks like image classification and object recognition.

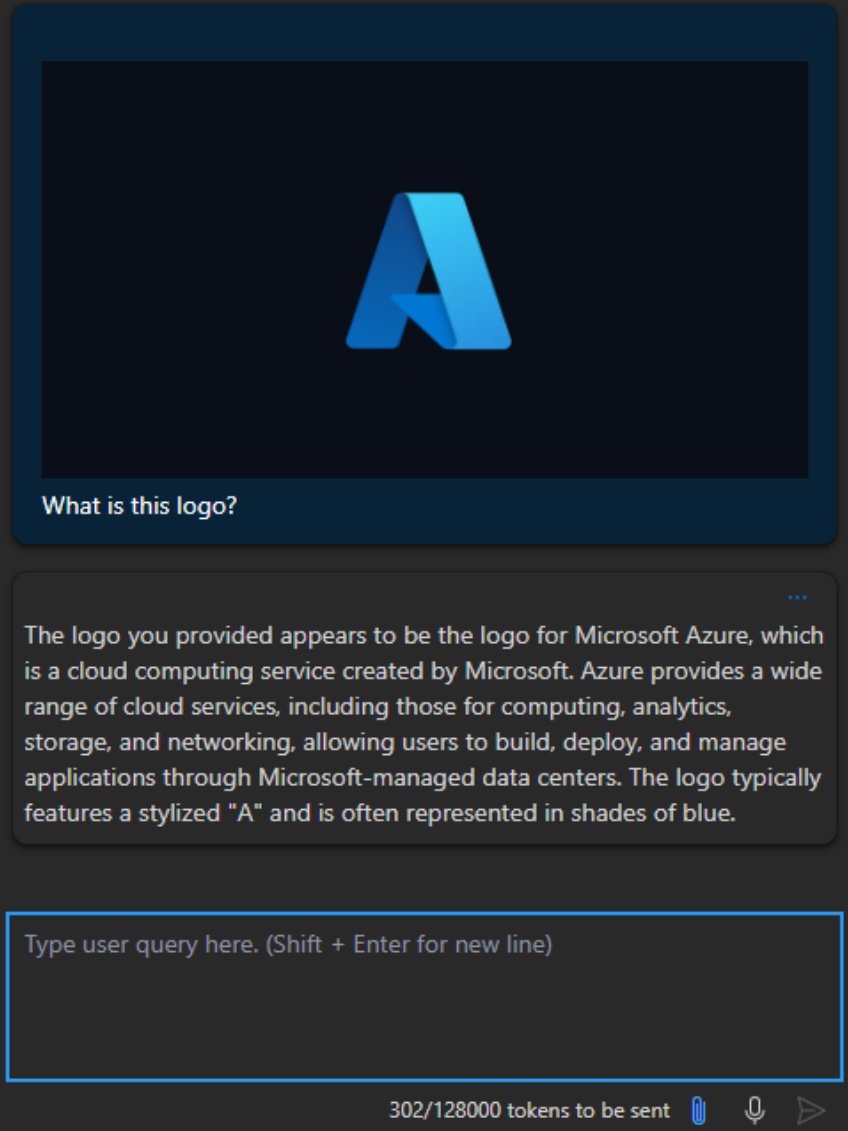
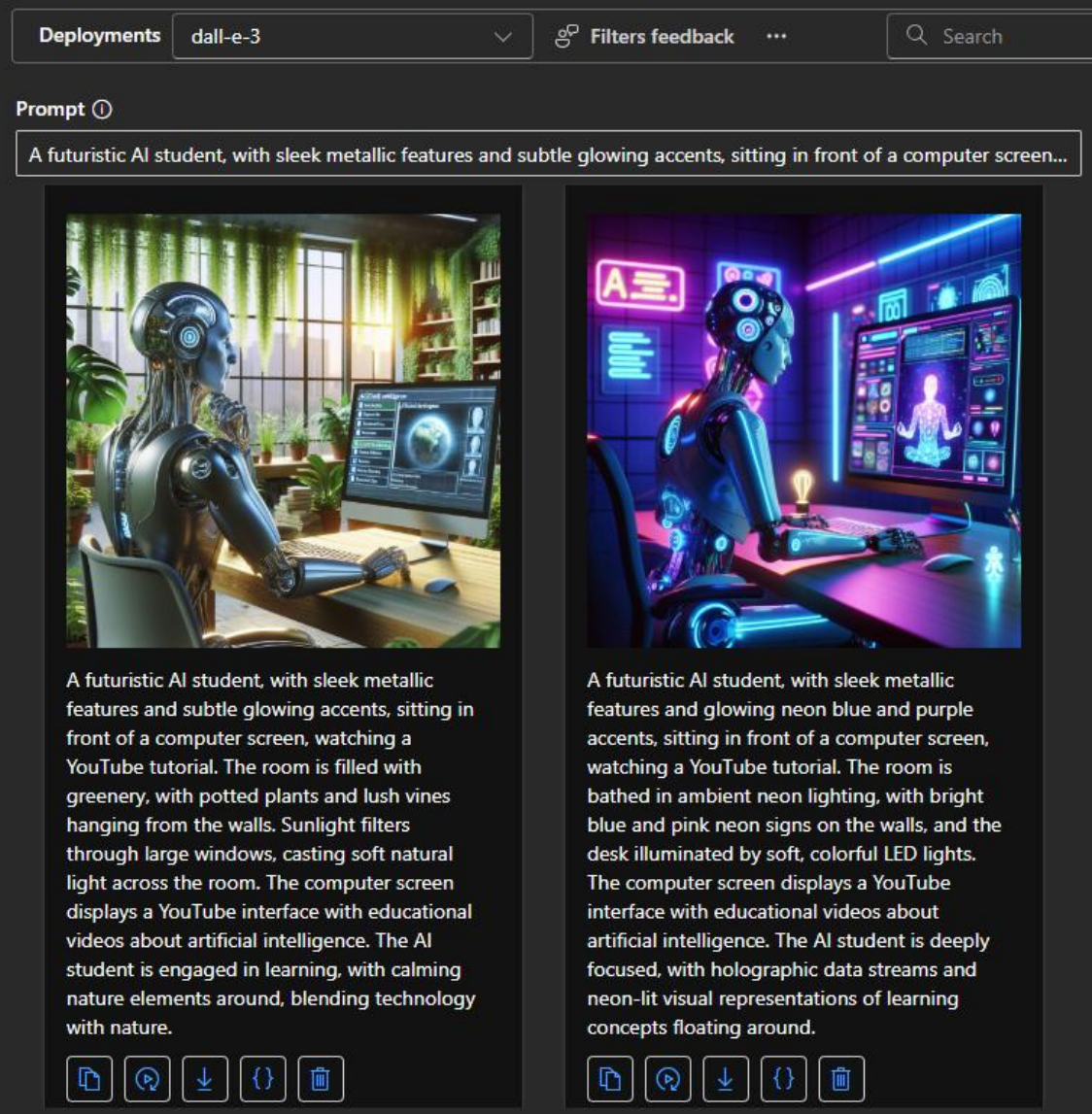


Image Generation

Allows the model to create original images from textual descriptions, ideal for creative tasks like designing visuals or generating concept art.



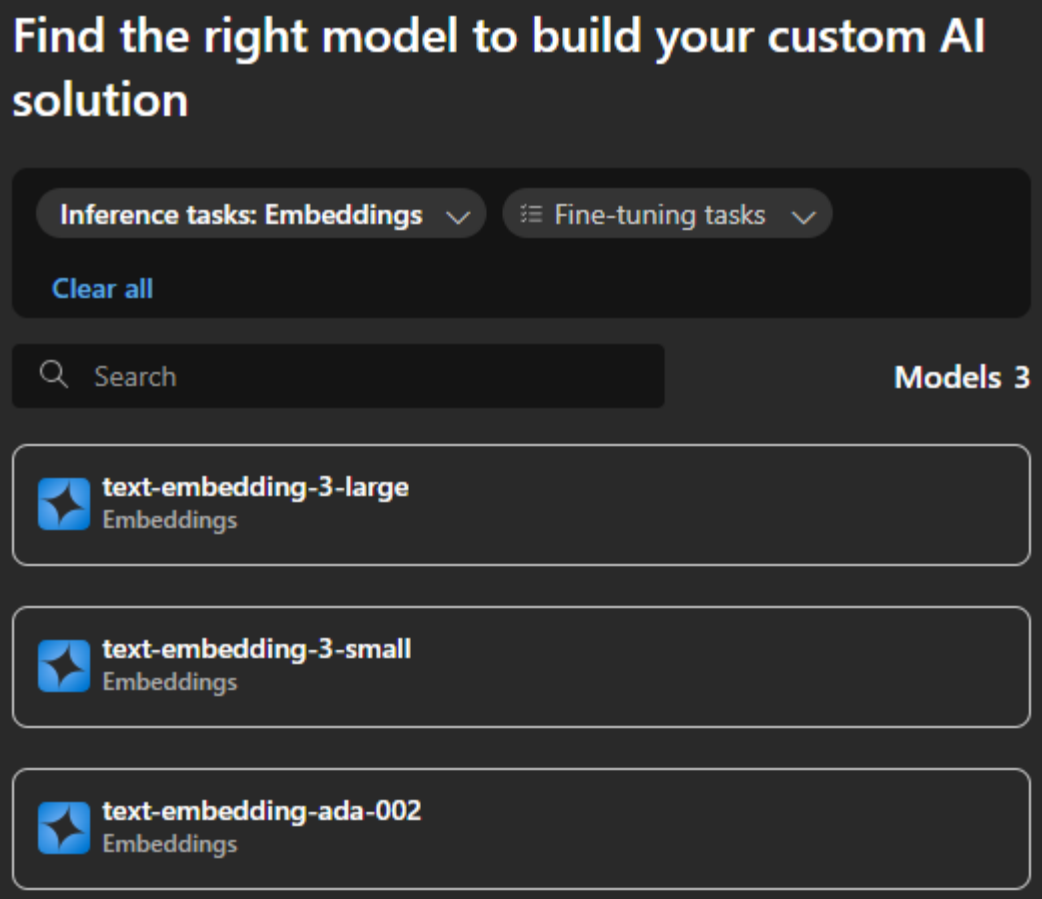
Text to Speech

Converts written text into spoken words, enhancing accessibility and user experience in applications like voice assistants.

```
function getClient(): AzureOpenAI {
  return new AzureOpenAI({
    endpoint,
    azureADTokenProvider,
    apiVersion,
    deployment: deploymentName,
  });
}

async function generateAudioStream(
  client: AzureOpenAI,
  params: SpeechCreateParams
): Promise<NodeJS.ReadableStream> {
  const response = await client.audio.speech.create(params);
  if (response.ok) return response.body;
  throw new Error(`Failed to generate audio stream: ${response.statusText}`);
}
export async function main() {
  console.log("== Text to Speech Sample ==");

  const client = getClient();
  const streamToRead = await generateAudioStream(client, {
    model: deploymentName,
    voice: "alloy",
    input: "the quick brown chicken jumped over the lazy dogs",
  });

  console.log(`Streaming response to ${speechFilePath}`);
  await writeFile(speechFilePath, streamToRead);
  console.log("Finished streaming");
}

main().catch((err) => {
  console.error("The sample encountered an error:", err);
});
```

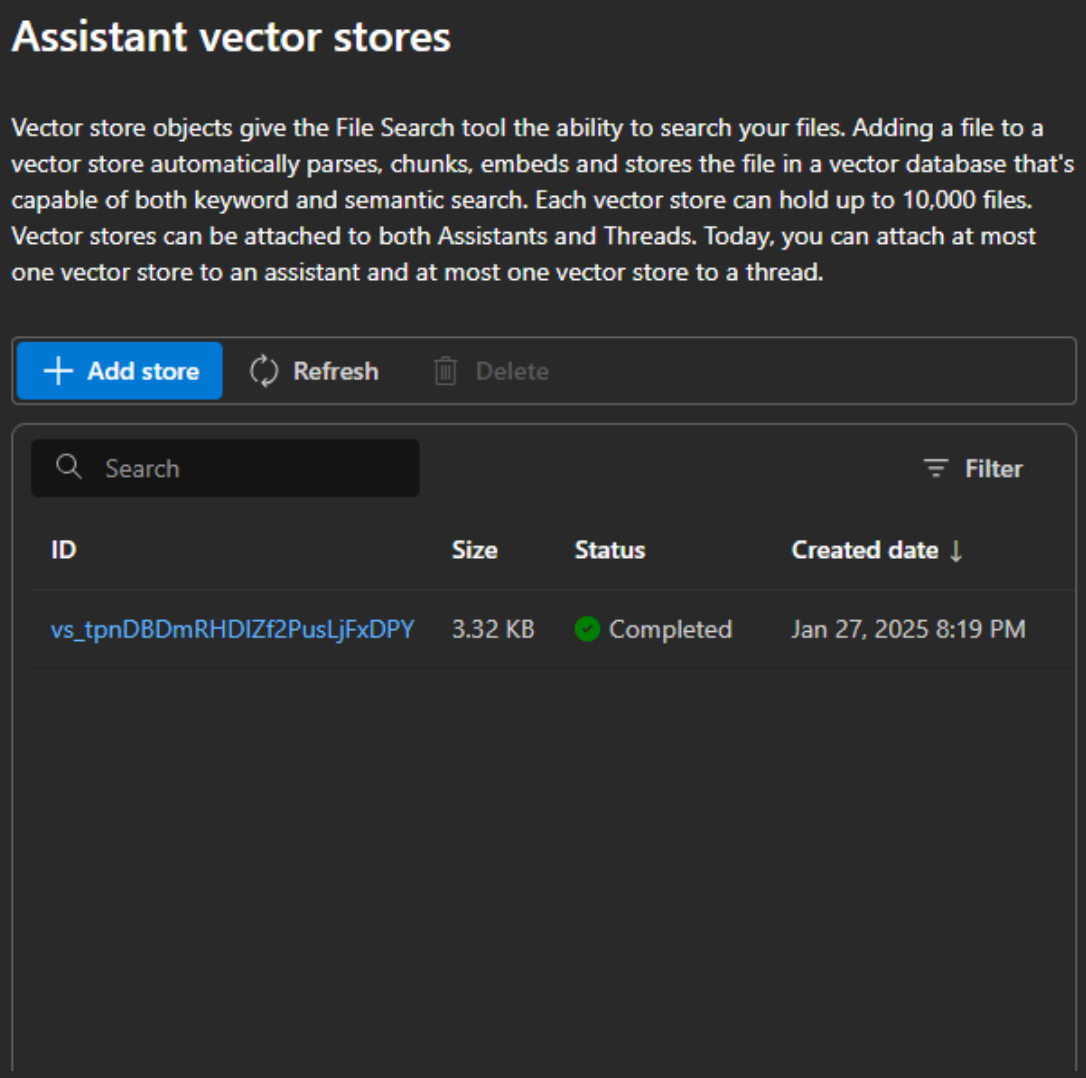
Embeddings

Transforms text into numerical vectors, which helps in tasks like semantic search, text similarity, and improving data analysis accuracy.



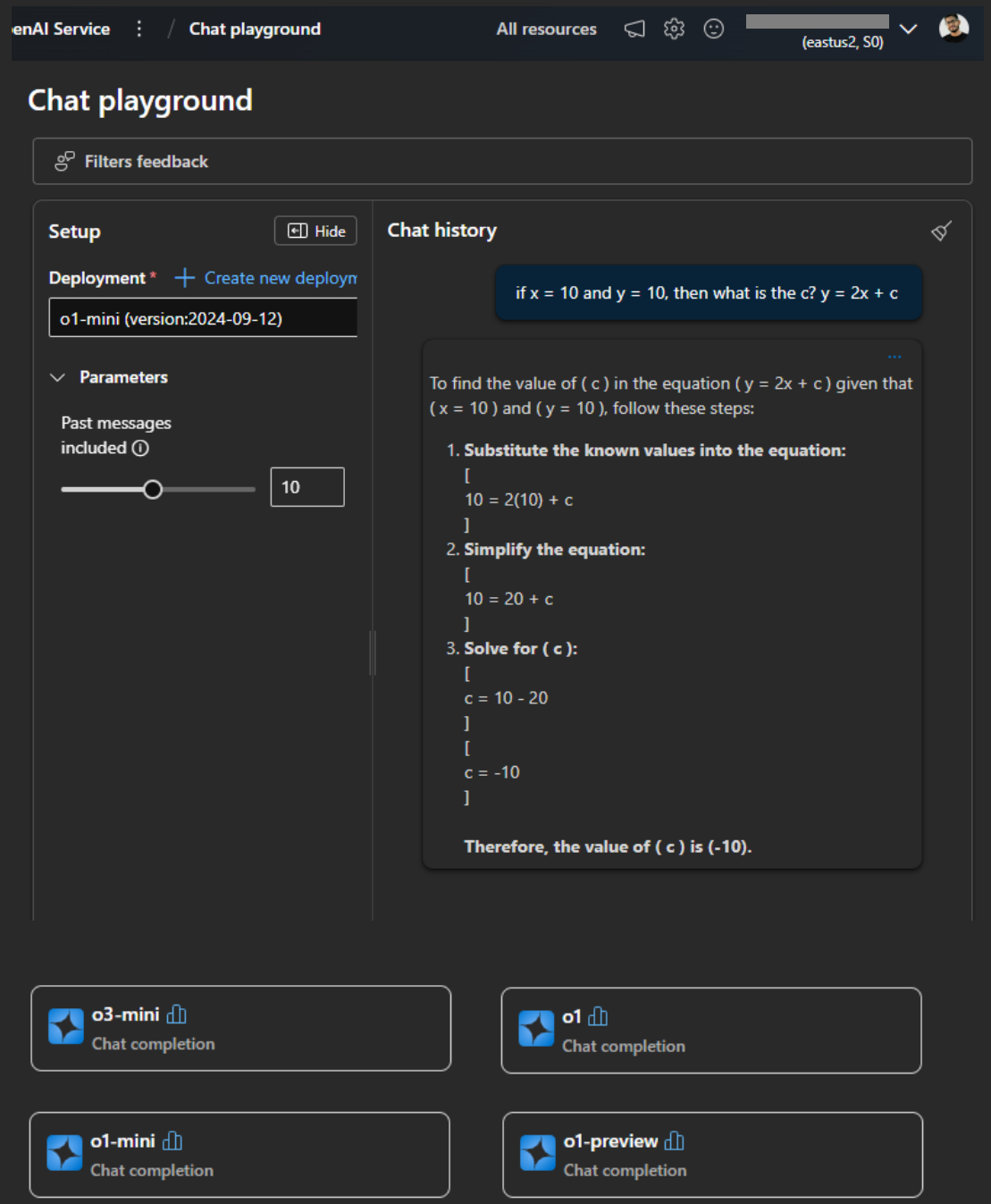
Vector Stores

Stores and organizes embeddings in a way that makes it easier to perform complex searches, retrieve specific information, and manage large datasets for enhanced AI interaction.



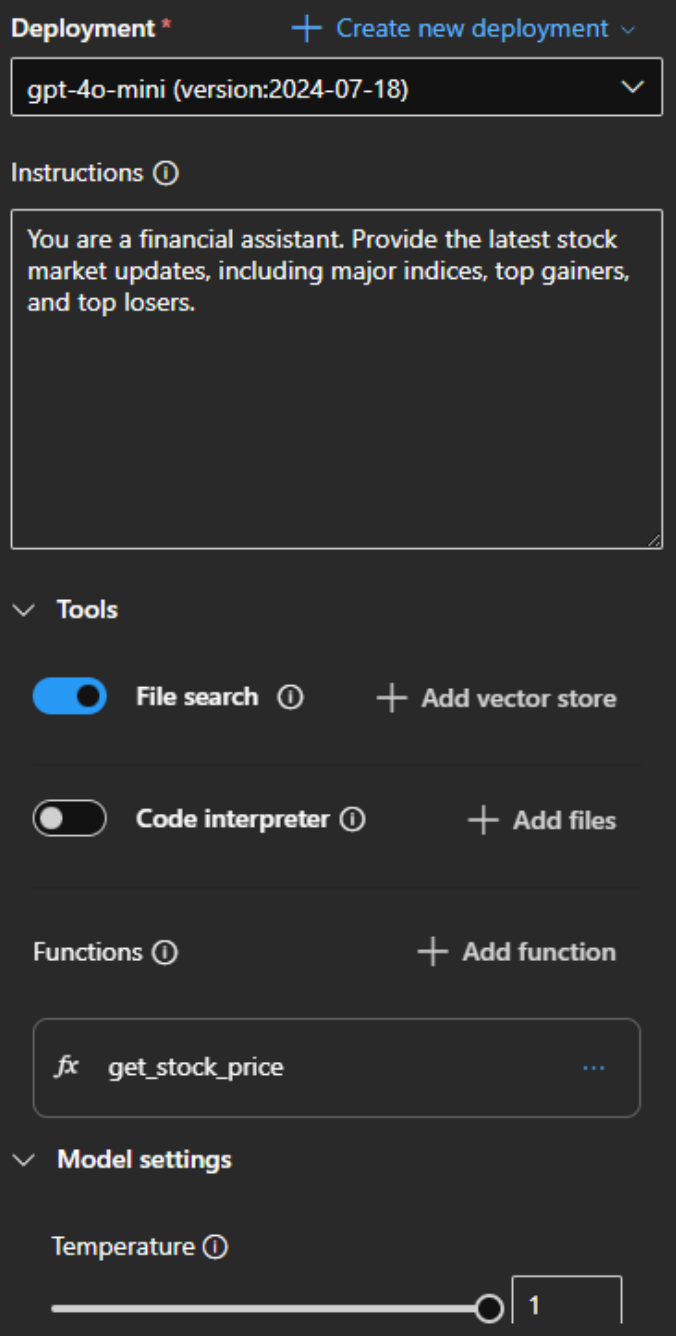
Reasoning

Supports logical problem-solving and decision-making tasks, making it suitable for applications in fields like science, engineering, and coding.



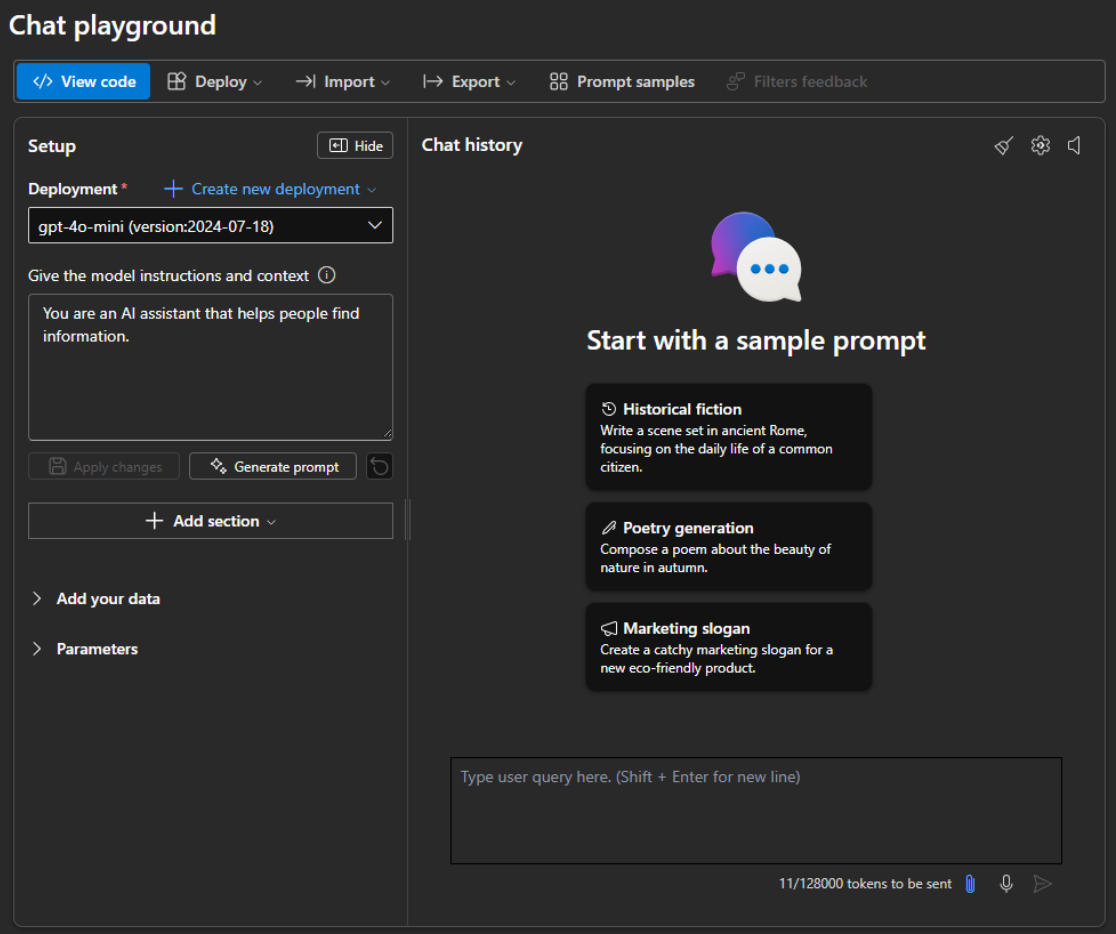
Function Calling

This feature allows the model to trigger specific functions or APIs based on user input, enabling dynamic and automated interactions.



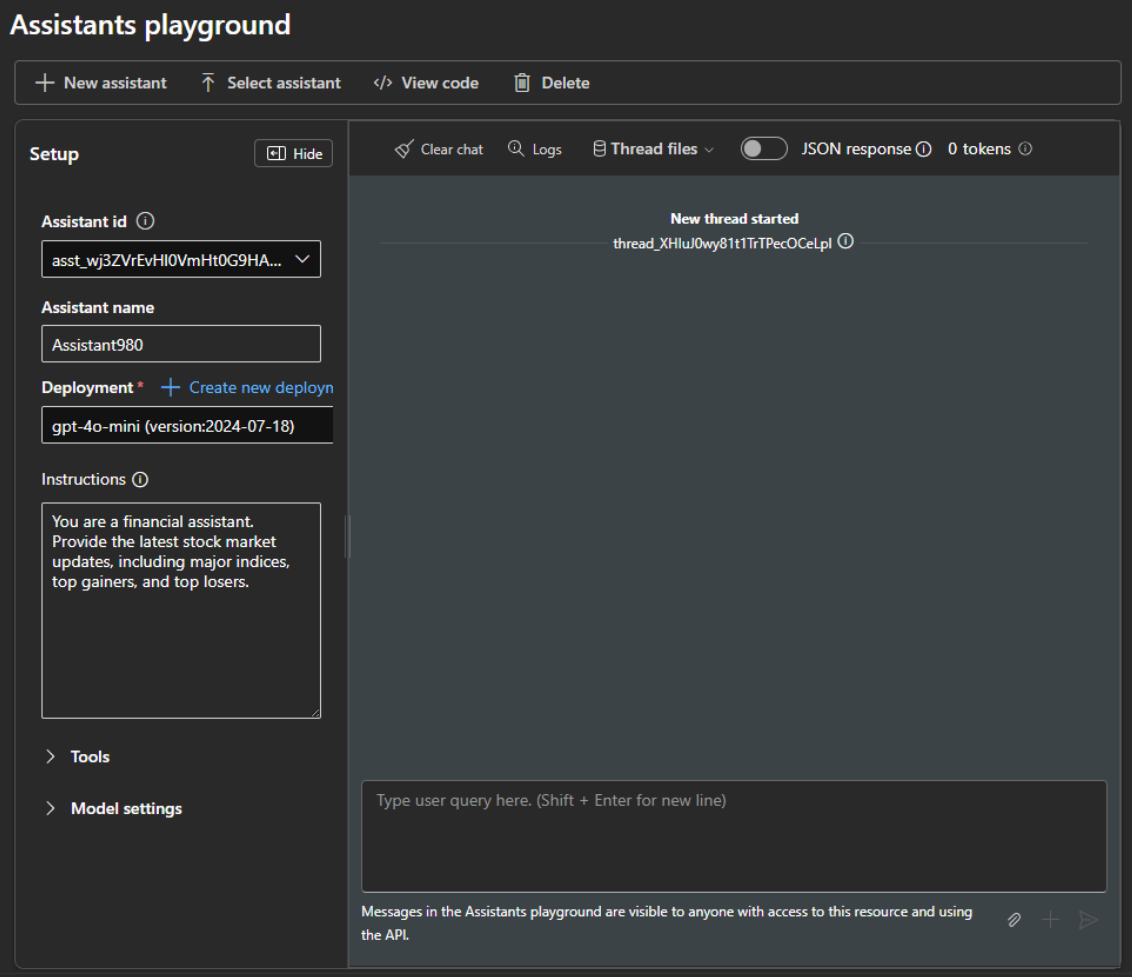
Chat

Powers conversational AI, enabling natural language interactions where the model can respond to queries and hold contextual conversations.



Assistants

Enables the creation of virtual assistants that can handle tasks such as answering questions, managing schedules, and providing recommendations.



Supported Programming Languages & SDKs

Supported Programming Languages & SDKs

Other Languages and Tools

(for broader access)

Other Languages and Tools

PowerShell

Azure CLI

REST API

Quotas and Limits

Quotas and Limits

- Essential to understand to ensure a seamless experience
- Helps prevent service interruptions
- Vary based on subscription tier, region, and API usage specifics
- Quotas and limitations may evolve as Azure OpenAI develops
- We'll revisit these in future sessions when working on projects
- Get a general sense of usage and restrictions to anticipate

Cost Management and Pricing

Cost Management and Pricing

- Understand pricing to prevent unexpected charges.
- Costs vary by model type and usage.
- Costs increase with project scale.
- Use Azure tools for real-time tracking.
- Free tiers available for small projects.
- Monitor usage in the Azure portal.
- Check the Azure pricing page for details.

... and there are many more

… and there are many more

- Security and Compliance
- Data Privacy
- Responsible AI
- Integration with Azure Ecosystem
- Scalability
- Model Monitoring and Analytics
- Model Deployment Options
- Community and Support
- … etc.

Azure OpenAI Community & Support

- Official Documentation
- Microsoft Learn
- Tech Communities, Meetups & Events
- Forums & Tech Groups
- Support and Help Options
- Videos & Tutorials
- Blogs
- … and more

Setting Up Azure OpenAI Resource in the Azure Portal

Thanks so much 🥰

Keep your intelligence sharp! 🧠