What is Azure OpenAI in Azure AI Foundry Models?

07/31/2025

Azure OpenAI provides REST API access to OpenAI's powerful language models including gpt-5 series, o4-mini, o3, gpt-4.1, o3-mini, o1, o1-mini, GPT-4o, GPT-4o mini, GPT-4 Turbo with Vision, GPT-4, GPT-3.5-Turbo, and Embeddings model series. These models can be easily adapted to your specific task including but not limited to content generation, summarization, image understanding, semantic search, and natural language to code translation. Users can access the service through REST APIs, Python/C#/JS/Java/Go SDKs.

Features overview

Expand table

Feature	Azure OpenAI
Models available	gpt-5 series o4-mini & o3 gpt-4.1 **computer-use-preview** **o3-mini & o1** **o1-mini** **GPT-4o & GPT-4o mini** **GPT-4 series (including GPT-4 Turbo with Vision)** **GPT-3.5-Turbo series** Embeddings series Learn more in our Models page.
Fine-tuning	`GPT-4o-mini` (preview) `GPT-4` (preview) `GPT-3.5-Turbo` (0613).
Price	Available here For details on vision-enabled chat models, see the special pricing information.
Virtual network support & private link support	Yes.
Managed Identity	Yes, via Microsoft Entra ID

Feature	Azure OpenAI
UI experience	Azure portal for account & resource management, Azure AI Foundry for model exploration and fine-tuning
Model regional availability	Model availability
Content filtering	Prompts and completions are evaluated against our content policy with automated systems. High severity content is filtered.

Responsible AI

At Microsoft, we're committed to the advancement of AI driven by principles that put people first. Generative models such as the ones available in Azure OpenAI have significant potential benefits, but without careful design and thoughtful mitigations, such models have the potential to generate incorrect or even harmful content. Microsoft has made significant investments to help guard against abuse and unintended harm, which includes incorporating Microsoft's principles for responsible AI use, adopting a Code of Conduct that customers should consider when using Azure OpenAI.

Get started with Azure OpenAI

To get started with Azure OpenAI, you need to create an Azure OpenAI resource in your Azure subscription.

Start with the Create and deploy an Azure OpenAI resource guide.

1. You can create a resource via Azure portal, Azure CLI, or Azure PowerShell.
2. When you have an Azure OpenAI resource, you can deploy a model such as GPT-4o.
3. When you have a deployed model, you can:
   - Try out the Azure AI Foundry portal playgrounds to explore the capabilities of the models.
   - You can also just start making API calls to the service using the REST API or SDKs.

   For example, you can try real-time audio in the playgrounds or via code.

ⓘ Note

A Limited Access registration form is required to access some Azure OpenAI models or features. Learn more on the Azure OpenAI Limited Access page.

Comparing Azure OpenAI and OpenAI

Azure OpenAI gives customers advanced language AI with OpenAI GPT-4, GPT-3, Codex, GPT-image-1 (preview), DALL-E, speech to text, and text to speech models with the security and enterprise promise of Azure. Azure OpenAI co-develops the APIs with OpenAI, ensuring compatibility and a smooth transition from one to the other.

With Azure OpenAI, customers get the security capabilities of Microsoft Azure while running the same models as OpenAI. Azure OpenAI offers private networking, regional availability, and responsible AI content filtering.

Key concepts

Prompts & completions

The completions endpoint is the core component of the API service. This API provides access to the model's text-in, text-out interface. Users simply need to provide an input **prompt** containing the English text command, and the model generates a text **completion**.

Here's an example of a simple prompt and completion:

> **Prompt**: `""" count to 5 in a for loop """`
>
> **Completion**: `for i in range(1, 6): print(i)`

Tokens

Text tokens

Azure OpenAI processes text by breaking it down into tokens. Tokens can be words or just chunks of characters. For example, the word “hamburger” gets broken up into the tokens “ham”, “bur” and “ger”, while a short and common word like “pear” is a single token. Many tokens start with a whitespace, for example “ hello” and “ bye”.

The total number of tokens processed in a given request depends on the length of your input, output, and request parameters. The quantity of tokens being processed will also affect your response latency and throughput for the models.

Image input tokens

Azure OpenAI's image processing capabilities with GPT-4o, GPT-4o-mini, and GPT-4 Turbo with Vision models uses image tokenization to determine the total number of tokens consumed by image inputs. The number of tokens consumed is calculated based on two main factors: the level of image detail (low or high) and the image’s dimensions. Here's how token costs are calculated:

- **Low resolution mode**
  - Low detail allows the API to return faster responses for scenarios that don't require high image resolution analysis. The tokens consumed for low detail images are:
    - **GPT-4o and GPT-4 Turbo with Vision**: Flat rate of **85 tokens per image**, regardless of size.
    - **GPT-4o mini**: Flat rate of **2833 tokens per image**, regardless of size.
  - **Example: 4096 x 8192 image (low detail)**: The cost is a fixed 85 tokens with GPT-4o, because it's a low detail image, and the size doesn't affect the cost in this mode.
- **High resolution mode**
  - High detail allows the API to analyze images in more detail. Image tokens are calculated based on the image's dimensions. The calculation involves the following steps:

    1. **Image resizing**: The image is resized to fit within a 2048 x 2048 pixel square. If the shortest side is larger than 768 pixels, the image is further resized so that the shortest side is 768 pixels long. The aspect ratio is preserved during resizing.
    2. **Tile calculation**: Once resized, the image is divided into 512 x 512 pixel tiles. Any partial tiles are rounded up to a full tile. The number of tiles determines the total token cost.
    3. **Token calculation**:
       - **GPT-4o and GPT-4 Turbo with Vision**: Each 512 x 512 pixel tile costs **170 tokens**. An extra **85 base tokens** are added to the total.
       - **GPT-4o mini**: Each 512 x 512 pixel tile costs **5667 tokens**. An extra **2833 base tokens** are added to the total.

  - **Example: 2048 x 4096 image (high detail)**:

    1. The image is initially resized to 1024 x 2048 pixels to fit within the 2048 x 2048 pixel square.

2. The image is further resized to 768 x 1536 pixels to ensure the shortest side is a maximum of 768 pixels long.
3. The image is divided into 2 x 3 tiles, each 512 x 512 pixels.
4. **Final calculation**:
    - For GPT-4o and GPT-4 Turbo with Vision, the total token cost is 6 tiles x 170 tokens per tile + 85 base tokens = 1105 tokens.
    - For GPT-4o mini, the total token cost is 6 tiles x 5667 tokens per tile + 2833 base tokens = 36835 tokens.

Image generation tokens

GPT-image-1 generates images by first producing specialized image tokens. Both latency and eventual cost are proportional to the number of tokens required to render an image. The number of tokens generated depends on image dimensions and quality:

Expand table

Quality	Square (1024×1024)	Portrait (1024×1536)	landscape (1536×1024)
Low	272 tokens	408 tokens	400 tokens
Medium	1056 tokens	1584 tokens	1568 tokens
High	4160 tokens	6240 tokens	6208 tokens

Resources

Azure OpenAI is a new product offering on Azure. You can get started with Azure OpenAI the same way as any other Azure product where you create a resource, or instance of the service, in your Azure Subscription. You can read more about Azure's resource management design.

Deployments

Once you create an Azure OpenAI Resource, you must deploy a model before you can start making API calls and generating text. This action can be done using the Deployment APIs. These APIs allow you to specify the model you wish to use.

Prompt engineering

The GPT-3.5, and GPT-4 models from OpenAI are prompt-based. With prompt-based models, the user interacts with the model by entering a text prompt, to which the model responds with a text completion. This completion is the model’s continuation of the input text.

While these models are powerful, their behavior is also sensitive to the prompt. This makes prompt engineering an important skill to develop.

Prompt construction can be difficult. In practice, the prompt acts to configure the model weights to complete the desired task, but it's more of an art than a science, often requiring experience and intuition to craft a successful prompt.

Models

The service provides users access to several different models. Each model provides a different capability and price point.

The image generation models (some in preview; see models) generate and edit images from text prompts that the user provides.

The audio API models can be used to transcribe and translate speech to text. The text to speech models, currently in preview, can be used to synthesize text to speech.

Learn more about each model on our models concept page.

Next steps

Learn more about the underlying models that power Azure OpenAI.