What is Azure OpenAI Service?

Article • 12/11/2023

Azure OpenAI Service provides REST API access to OpenAI's powerful language models including the GPT-4, GPT-4 Turbo with Vision, GPT-3.5-Turbo, and Embeddings model series. In addition, the new GPT-4 and GPT-3.5-Turbo model series have now reached general availability. 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 SDK, or our web-based interface in the Azure OpenAI Studio.

Features overview

Expand table

Feature	Azure OpenAI
Models available	**GPT-4 series (including GPT-4 Turbo with Vision)** **GPT-3.5-Turbo series** Embeddings series Learn more in our Models page.
Fine-tuning (preview)	`GPT-3.5-Turbo` (0613) `babbage-002` `davinci-002`.
Price	Available here For details on GPT-4 Turbo with Vision, see the special pricing information.
Virtual network support & private link support	Yes, unless using Azure OpenAI on your data.
Managed Identity	Yes, via Microsoft Entra ID
UI experience	**Azure portal** for account & resource management, **Azure OpenAI Service Studio** 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 will be 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 requiring applicants to show well-defined use cases, incorporating Microsoft's principles for responsible AI use, building content filters to support customers, and providing responsible AI implementation guidance to onboarded customers.

How do I get access to Azure OpenAI?

How do I get access to Azure OpenAI?

Access is currently limited as we navigate high demand, upcoming product improvements, and Microsoft's commitment to responsible AI. For now, we're working with customers with an existing partnership with Microsoft, lower risk use cases, and those committed to incorporating mitigations.

More specific information is included in the application form. We appreciate your patience as we work to responsibly enable broader access to Azure OpenAI.

Apply here for access:

Apply now

Comparing Azure OpenAI and OpenAI

Azure OpenAI Service gives customers advanced language AI with OpenAI GPT-4, GPT-3, Codex, DALL-E, Whisper, 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 will generate 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 tokens (GPT-4 Turbo with Vision)

The token cost of an input image depends on two main factors: the size of the image and the detail setting (low or high) used for each image. Here's a breakdown of how it works:

- **Detail: Low resolution mode**
  - Low detail allows the API to return faster responses and consume fewer input tokens for use cases that don’t require high detail.
  - These images cost 85 tokens each, regardless of the image size.
  - **Example: 4096 x 8192 image (low detail)**: The cost is a fixed 85 tokens, because it's a low detail image, and the size doesn't affect the cost in this mode.

- **Detail: High resolution mode**
  - High detail lets the API see the image in more detail by cropping it into smaller squares. Each square

uses more tokens to generate text.

- The token cost is calculated by a series of scaling steps:

  1. The image is first scaled to fit within a 2048 x 2048 square while maintaining its aspect ratio.
  2. The image is then scaled down so that the shortest side is 768 pixels long.
  3. The image is divided into 512-pixel square tiles, and the number of these tiles (rounding up for partial tiles) determines the final cost. Each tile costs 170 tokens.
  4. An additional 85 tokens are added to the total cost.

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

  1. Initially resized to 1024 x 2048 to fit in the 2048 square.
  2. Further resized to 768 x 1536.
  3. Requires six 512px tiles to cover.
  4. Total cost is `170 × 6 + 85 = 1105` 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, 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 extremely powerful, their behavior is also very 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 DALL-E models, currently in preview, generate images from text prompts that the user provides.

The Whisper models, currently in preview, 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.