Everything you wanted to know about Azure OpenAI Pricing

Because of the rise of AI, businesses are increasingly turning to cloud platforms to harness the power of AI technologies. Among the frontrunners in this domain is Microsoft Azure. It offers a comprehensive suite of AI services that cater to diverse business needs. As businesses start using Azure AI in their operations, understanding the pricing can be a puzzling challenge. This blog post delves into the intricate world of Azure AI Pricing. From understanding the various pricing models to navigating the plethora of AI service models available, we aim to equip you with the knowledge needed to optimize your investment in Azure.

Table of Contents hide

What is Azure Open AI

Artificial Intelligence (AI)

OpenAI on Microsoft Azure brings powerful AI models into your Azure environment, like GPT-3 and ChatGPT. This integration allows you to run models in the same region as your production code. It enhances performance and maintains data security.

Read our blog for more information

Everything you need to know about Microsoft and OpenAI's Partnership

Dive into our Open AI Case Study with Regeneron

Why choose OpenAI on Azure?

It leverages Azure's robust infrastructure, aligning with familiar tools for developers and benefiting from Microsoft's scalability and global data center network.

Explore Azure Open AI Pricing Options

Like Azure's services, such as Azure Cognitive Services for AI, Azure OpenAI's cost operates on a pay-as-you-go consumption approach. It ensures you pay solely for the resources you utilize. The price per unit is determined by the model's type and size and the quantity of tokens employed in both input and output.

GPT-3.5 models

The current gpt-3.5-turbo has a 4,096-token limit, while the newer gpt-3.5-turbo-16k has a 16,384-token limit. Both are priced at $0.002 per 1,000 tokens for prompts or completions.

GPT-4 models

GPT-4 models come in two versions: the GPT -4 model, which has an 8,192 token limit, and the GPT-4 32k, which has a 32,768 token limit.

The GPT-4 model in prompt mode costs $0.03 for an 8K context and $0.06 for a 32K context per 1,000 tokens. In completion mode, the prices are $0.06 for an 8K context and $0.12 for a 32K context per 1,000 tokens.

Tokens

Tokens can be visualized as fragments of words. Before the API handles prompts, the input is segmented into tokens. These tokens may encompass trailing spaces and even parts of words. Here are some valuable approximations to understand token lengths:

- 1 token is equal to 4 characters in English.
- 1 token is approximately ¾ of a word.
- 100 tokens amount to around 75 words.

Alternatively,

- 1–2 sentences equate to about 30 tokens.
- 1 paragraph comprises roughly 100 tokens.
- 1,500 words correspond to approximately 2048 tokens.

How words are divided into tokens depends on the language. For instance, the phrase ‘Cómo estás’ (‘How are you’ in Spanish) consists of 5 tokens (covering 10 characters). The higher ratio of tokens to characters can increase the cost of using the API for languages other than English.

To explore tokenization further, you can utilize our interactive Tokenizer tool to calculate token numbers and observe how text is segmented. If you prefer programmatically tokenizing text, consider using Tiktoken, a rapid BPE tokenizer designed for OpenAI models. Other libraries, like the transformers package for Python or the GPT-3 encoder package for node.js, are also worth exploring.

Depending on the chosen model, requests can use up to 4,097 tokens shared between prompts and completions. If your prompt uses 4,000 tokens, your completion can have a maximum of 97 tokens.

Fine-tuned model

You can only fine-tune GPT-3 models (Ada, curie, DaVinci, Babbage), collectively known as “base” models.

According to Microsoft Learn, the charges for Azure OpenAI fine-tuned models depend on three factors:

1. Training hours
2. Hosting hours
3. Inference per 1,000 tokens

Consider the hosting hours cost carefully. Once a fine-tuned model is deployed, it accumulates an hourly cost, regardless of its active usage. Monitoring the costs of fine-tuned models is essential.

*** Currently, Azure OpenAI Service does not support fine-tuned models.

Embedding model

Azure OpenAI Service provides an embedding model alongside image and large language models. The standard embedding model, Ada, costs $0.0001 per 1,000 tokens.

DALL-E

Azure OpenAI Service, including image models, operates with pricing determined by the number of images processed, making it relevant for those exploring “Azure OpenAI Service pricing.” The standard image model, DALL-E, costs $2 for every 100 images.

Fueling Microsoft Copilot with Azure OpenAI Service

Copilot, powered by Microsoft Azure OpenAI Service and providing access to Copilot, simplifies designing, operating, optimizing, and troubleshooting apps and infrastructure across the cloud to the edge. It uses language models, the Azure control plane, and insights into your Azure and Arc-enabled assets while prioritizing data security and privacy.

Microsoft Copilot pricing is calculated for specific applications. It varies on the integrated service such as Microsoft 365 apps, Dynamics 365, etc.

Navigate Azure OpenAI Pricing with Expert Guidance

Azure OpenAI is a powerful tool for businesses, integrating seamlessly with Azure for advanced AI capabilities. It automates processes like natural language processing and image recognition, enhancing efficiency. This collaboration between Microsoft and OpenAI has proven impactful globally. Businesses can leverage these advances to automate tasks, improve operations, and unleash limitless potential!

We have investigated the prices for each service, and this page will be regularly updated with the latest information. Feel free to reach out if you have questions or need clarification on using Azure OpenAI services, including their use cases. Additionally, don’t hesitate to contact us for pricing details on specific services mentioned in this blog.

CONTACT OUR AZURE TEAM