Region:

East US

Currency:

United States – Dollar ($) USD

Azure OpenAI Service pricing

Request a pricing quote (/en-us/contact/pricing/)

Try Azure for free (/en-us/free/)

AI Services Family

Overview

Pricing table

Purchase options

Resources

FAQ

Azure OpenAI Service

Free account (https://azure.microsoft.com/en-us/free/)

Chat with Sales

Region:

Azure OpenAI Service pricing overview

Currency:

Azure OpenAI Service delivers enterprise-ready generative AI featuring powerful models from OpenAI, enabling organizations to innovate with text, audio, and vision capabilities. Beyond the cutting-edge models, companies choose Azure OpenAI Service for built-in data privacy, regional/area/global flexibility, and seamless integration into the Azure ecosystem including Fabric, Cosmos DB and Azure AI Search. Companies of all sizes can confidently scale AI solutions to enhance customer experience, automate workflows, and unlock creative potential, driving measurable impact and competitive differentiation.

To help customers in the journey, we offer pricing and cost management solutions to meet your needs. including:

- **Standard (On-Demand):** Pay-as-you-go for input and output tokens.
- **Provisioned (PTUs):** Allocate throughput (https://docs.microsoft.com/en-us/azure/ai-services/openai/concepts/provisioned-throughput) with predictable costs, with monthly and annual reservations available to reduce overall spend.
- **Batch API:** Language models are also now available in the Batch API for global deployments and three regions (https://aka.ms/aoai-batch-how-to), that returns completions within 24 hours for a 50% discount on Global Standard Pricing.

You can choose from the following deployment types for Standard and Provisioned, which enable greater flexibility and control of pricing and performance. This flexibility helps when there is increasingly more restrictive data processing boundaries and need for increased throughput and lower price.

- **Global Deployment** – Global SKU
- **Data Zone Deployment** – Geographic based (EU or US)
- **Regional Deployment** – Local Region (up to 27 regions)

Region:

Explore pricing options

Currency:

Apply filters to customize pricing options to your needs.

Prices are estimates only and are not intended as actual price quotes. Actual pricing may vary depending on the type of agreement entered with Microsoft, date of purchase, and the currency exchange rate. Prices are calculated based on US dollars and converted using London closing spot rates that are captured in the two business days prior to the last business day of the previous month end. If the two business days prior to the end of the month fall on a bank holiday in major markets, the rate setting day is generally the day immediately preceding the two business days. This rate applies to all transactions during the upcoming month. Sign in to the Azure pricing calculator (/en-us/pricing/calculator/) to see pricing based on your current program/offer with Microsoft. Contact an Azure sales specialist (/en-us/contact/pricing/) for more information on pricing or to request a price quote. See frequently asked questions (/en-us/pricing/) about Azure pricing.

Region:

o3

o3 is a powerful reasoning model from the o-series of reasoning models, pushing the frontier across coding, math, science, and visual perception. It excels in complex queries requiring multi-faceted analysis and performs strongly in visual tasks like analyzing images, charts, and graphics. The model features a 200K token context window and has a knowledge cutoff of June 2024.

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
o3 2025-04-16	Input: **$10** Cached Input: **$2.50** Output: **$40**	N/A

o4-mini

o4-mini is a compact, efficient, and cost-effective reasoning model from OpenAI's o-series. It excels in math, coding, and visual tasks. The model features a 200K token context window and has a knowledge cutoff of June 2024.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
o4-mini 2025-04-16	Input: **$1.10** Cached Input: **$0.28** Output: **$4.40**	N/A

GPT-4.1 series

Region:

GPT-4.1 series is a highly advanced general-purpose model with extensive world knowledge and an enhanced ability to understand user intent, making it particularly adept at creative tasks and agentic planning. The series features a 1 million token context window and has a knowledge cutoff of June 2024.

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4.1-2025-04-14	Input: **$2** Cached Input: **$0.50** Output: **$8**	N/A
GPT-4.1-mini-2025-04-14	Input: **$0.40** Cached Input: **$0.10** Output: **$1.60**	N/A
GPT-4.1-nano-2025-04-14	Input: **$0.10** Cached Input: **$0.03** Output: **$0.40**	N/A

Sora in Azure OpenAI

Sora is a multimodal generative AI model now available in Azure AI Foundry, designed to help creative teams bring ideas to life through seamless API-first integration. Built on Azure’s enterprise-grade infrastructure, it offers secure, scalable deployment for transforming concepts into high-quality visual content.

Region:

Pricing is not available in the selected region

Currency:

GPT-Image-1

GPT-image-1 enhances DALL·E with better instruction following, accurate text rendering, and support for image input and editing. The model is priced per token, with different pricing for text and image tokens.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-Image-1 Global	Input Text: **$5** Input Image: **$10** Output Image: **$40**	N/A
GPT-Image-1 Regional	Input Text: **$5.50** Input Image: **$11** Output Image: **$44**	N/A

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-Image-1 Data Zone	Input Text: **$5.50** Input Image: **$11** Output Image: **$44**	N/A

Region:

Currency:

GPT-4.5

GPT-4.5-preview is the latest general purpose model with deep world knowledge and better understanding of user intent that makes it good at creative tasks and agentic planning. The model has 128K context and an October 2023 knowledge cutoff.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4.5-Preview-2025-02-27 Global	Input: **$75** Cached Input: **$37.50** Output: **$150**	N/A

o1

o1 is the new reasoning model series for complex tasks. The model has 200K context and an October 2023 knowledge cutoff.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
Region: o1 2024-12-17 Global	Input: **$15** Cached Input: **$7.50** Output: **$60**	N/A
o1 2024-12-17 US/EU – Data Zones	Input: **$16.50** Cached Input: **$8.25** Output: **$66**	N/A
Currency: o1 2024-12-17 Regional	Input: **$16.50** Cached Input: **$8.25** Output: **$66**	N/A
o1 preview 2024-09-12 Global	Input: N/A Cached Input: N/A Output: N/A	N/A
o1 preview 2024-09-12 US/EU – Data Zones	Input: **$16.50** Cached Input: **$8.25** Output: **$66**	N/A
o1 preview 2024-09-12 Regional	Input: N/A Cached Input: N/A Output: N/A	N/A

Plan with the Pricing Calculator (/en-us/pricing/calculator/)

o3 Mini

The o3 mini is the updated version of o1 mini model. o3-mini is a fast, cost-efficient reasoning model tailored to coding, math, and science use cases.

The o3-mini model now boasts an expanded context input window of 200K tokens and a maximum output of 100K tokens, providing ample space for complex and detailed responses. The o1 mini model has 128K context input. Both o3 and o1 models have a knowledge cutoff of October 2023.

Region:

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
o3-mini 2025-01-31 Global	Input: **$1.10** Cached Input: **$0.55** Output: **$4.40**	Input: **$0.55** Output: **$2.20**
o3 mini 2025-01-31-US/EU – Data Zones	Input: **$1.21** Cached Input: **$0.605** Output: **$4.84**	Input: **$0.605** Output: **$2.42**
o3 mini 2025-01-31 Regional	Input: **$1.21** Cached Input: **$0.605** Output: **$4.84**	N/A
o1-mini 2024-09-12 Global	Input: **$1.10** Cached Input: **$0.55** Output: **$4.40**	N/A
o1-mini 2024-09-12 US/EU – Data Zones	Input: **$1.21** Cached Input: **$0.605** Output: **$4.84**	N/A
o1-mini 2024-09-12 Regional	Input: **$1.21** Cached Input: **$0.605** Output: **$4.84**	N/A

Plan with the Pricing Calculator (/en-us/pricing/calculator/)

Audio Models

Region:

Azure OpenAI Service includes the advanced audio models GPT-4o-Transcribe, GPT-4o-Mini-Transcribe, and GPT-4o-Mini-TTS. These models enhance speech-to-text and text-to-speech capabilities, offering high accuracy and customizable speech outputs for various applications. Ideal for customer call centers, live captioning, and interactive voice outputs, they leverage extensive pretraining and advanced distillation techniques for superior performance.

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4o-Transcribe	Text Input: **$2.50** Output: **$10** Audio Input: **$6** Output: N/A	N/A
GPT-4o-Mini-Transcribe	Text Input: **$1.25** Output: **$5** Audio Input: **$3** Output: N/A	N/A
GPT-4o-Mini-TTS	Text Input: **$0.60** Output: N/A Audio Input: N/A Output: **$12**	N/A

Computer-Using Agent (CUA)

Region:

The Computer-Using Agent (CUA) is a specialized AI model that allows AI to interact with graphical user interfaces (GUIs), navigate applications, and automate multi-step tasks—all through natural language instructions. The CUA model can be used as a tool in the Responses API.

Currency:

Model	Pricing
computer-use-preview Global	Input: **$3**/1M tokens Output: **$12**/1M tokens

Built-in tools

The Responses API and the Assistants API enable seamless interaction with tools like computer use, code interpreter, function calling, and file search, making it easy for developers to build AI agents.

Tool	Input
Computer Use (Responses API only)	Input: **$3**/1M tokens Output: **$12**/1M tokens
File Search Tool Call (Responses API only)	**$2.50**/1K tool calls
File Search (https://docs.microsoft.com/en-us/azure/ai-services/openai/how-to/file-search?tabs=python)*	**$0.10**/GB of vector-storage per day (1 GB free)
Code Interpreter (https://docs.microsoft.com/en-us/azure/ai-services/openai/how-to/code-interpreter?tabs=python)**	**$0.03**/session

*GB refers to binary gigabytes, where 1 gb is 2^30 bytes.

Region:

If your assistant calls Code Interpreter simultaneously in two different threads, this would create two Code Interpreter sessions (2 * $0.03). Each session is active by default for one hour, which means that you would only pay this fee once if your user keeps giving instructions to Code Interpreter in the same thread for up to one hour.

Inference cost (input and output) varies based on the GPT model used with each Assistant. If your assistant calls Code Interpreter simultaneously in two different threads, this would create two Code Interpreter sessions (2 * $0.03). Each session is active by default for one hour, which means that the price is for up to one hour of giving instructions to Code Interpreter in the same thread.

Current

Realtime API

Featured in the Realtime API, the GPT-4o-Realtime-Preview supports multilingual speech-to-speech capabilities. Optimized for real-time, low-latency conversations, it enables natural interactions with minimal delay, ideal for chatbots and conversational AI. GPT-4o is the comprehensive, more powerful version designed for complex tasks, while GPT-4o Mini is a smaller, more affordable option ideal for simpler applications where cost-efficiency and speed are priorities.

Model	Pricing (1M Tokens)
GPT-4o-Realtime-Preview-2024-12-17-Global	Text Input: **$5** Cached Input: **$2.50** Output: **$20** Audio Input: **$40** Cached Input: **$2.50** Output: **$80**

Region:

Currency:

Model	Pricing (1M Tokens)
GPT-4o-Realtime-Preview-2024-12-17-US/EU – Data Zones	Text Input: **$5.50** Cached Input: **$2.75** Output: **$22** Audio Input: **$44** Cached Input: **$2.75** Output: **$88**
GPT-4o-Realtime-Preview-2024-12-17-Regional	Text Input: **$5.50** Cached Input: **$2.75** Output: **$22** Audio Input: **$44** Cached Input: **$2.75** Output: **$88**
GPT-4o-Mini-Realtime-Preview-2024-12-17-Global	Text Input: **$0.60** Cached Input: **$0.30** Output: **$2.40** Audio Input: **$10** Cached Input: **$0.30** Output: **$20**

Region:

Currency:

Model	Pricing (1M Tokens)
GPT-4o-Mini-Realtime-Preview-2024-12-17-US/EU – Data Zones	Text Input: **$0.66** Cached Input: **$0.33** Output: **$2.64** Audio Input: **$11** Cached Input: **$0.33** Output: **$22**
GPT-4o-Mini-Realtime-Preview-2024-12-17-Regional	Text Input: **$0.66** Cached Input: **$0.33** Output: **$2.64** Audio Input: **$11** Cached Input: **$0.33** Output: **$22**
GPT-4o-Realtime-Preview-2024-10-01-Global	Text Input: **$5** Cached Input: **$2.50** Output: **$20** Audio Input: **$100** Cached Input: **$20** Output: **$200**

Model	Pricing (1M Tokens)
GPT-4o-Realtime-Preview-2024-10-01-US/EU – Data Zones	Text Input: **$5.50** Cached Input: **$2.75** Output: **$22** Audio Input: **$110** Cached Input: **$22** Output: **$220**
GPT-4o-Realtime-Preview-2024-10-01-Regional	Text Input: **$5.50** Cached Input: **$2.75** Output: **$22** Audio Input: **$110** Cached Input: **$22** Output: **$220**

Region:

Currency:

Chat Completions API

Featured in the Chat Completions API, the GPT 4o-Audio-Preview model processes and generates audio content. It supports advanced features like speech recognition and audio synthesis, ideal for asynchronous speech interactions and sentiment analysis. GPT-4o is the comprehensive, more powerful version designed for complex tasks, while GPT-4o Mini is a smaller, more affordable option ideal for simpler applications where cost-efficiency and speed are priorities.

Model	Pricing (1M Tokens)
GPT-4o-Audio-Preview-2024-12-17-Global Region: Currency:	Text Input: **$2.50** Output: **$10** Audio Input: **$40** Output: **$80**
GPT-4o-Audio-Preview-2024-12-17-US/EU – Data Zones	Text Input: **$2.75** Output: **$11** Audio Input: **$44** Output: **$88**
GPT-4o-Audio-Preview-2024-12-17-Regional	Text Input: **$2.75** Output: **$11** Audio Input: **$44** Output: **$88**
GPT-4o-Mini-Audio-Preview-2024-12-17-Global	Text Input: **$0.15** Output: **$0.60** Audio Input: **$10** Output: **$20**

Model	Pricing (1M Tokens)
GPT-4o-Mini-Audio-Preview-2024-12-17-US/EU – Data Zones	Text Input: **$0.165** Output: **$0.66** Audio Input: **$11** Output: **$22**
GPT-4o-Mini-Audio-Preview-2024-12-17-Regional	Text Input: **$0.165** Output: **$0.66** Audio Input: **$11** Output: **$22**

Region:

Currency:

GPT-4o

GPT-4o is the most advanced multimodal model that’s faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4o-2024-1120 Global	Input: **$2.50** Cached Input: **$1.25** Output: **$10**	Input: **$1.25** Output: **$5**

Region:

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4o-2024-1120 US/EU – Data Zones	Input: **$2.75** Cached Input: **$1.375** Output: **$11**	Input: **$1.375** Output: **$5.50**
GPT-4o-2024-1120 Regional	Input: **$2.75** Cached Input: **$1.375** Output: **$11**	N/A
GPT-4o-2024-08-06 Global	Input: **$2.50** Cached Input: **$1.25** Output: **$10**	Input: **$1.25** Output: **$5**
GPT-4o-2024-08-06 US/EU – Data Zones	Input: **$2.75** Cached Input: **$1.375** Output: **$11**	Input: **$1.375** Output: **$5.50**
GPT-4o-2024-08-06 Regional	Input: **$2.75** Cached Input: **$1.375** Output: **$11**	N/A
GPT-4o-2024-0513 Global	Input: **$5** Output: **$15**	Input: **$2.50** Output: **$7.50**
GPT-4o-2024-0513 US/EU – Data Zones	Input: **$5** Output: **$15**	N/A
GPT-4o-2024-0513 Regional	Input: **$5** Output: **$15**	N/A

Plan with the Pricing Calculator (/en-us/pricing/calculator/)

GPT-4o mini

Region:

GPT-4o mini is the most cost-efficient small model, and has vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Currency:

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
GPT-4o-mini-0718 Global	Input: **$0.15** Cached Input: **$0.075** Output: **$0.60**	Input: **$0.075** Output: **$0.30**
GPT-4o-mini-0718 US/EU – Data Zones	Input: **$0.165** Cached Input: **$0.083** Output: **$0.66**	Input: **$0.083** Output: **$0.33**
GPT-4o-mini-0718 Regional	Input: **$0.165** Cached Input: **$0.083** Output: **$0.66**	N/A

Plan with the Pricing Calculator (/en-us/pricing/calculator/)

Provisioned

You can allocate and manage throughput for deployments, ensuring predictable performance and stable capacity. You are charged an hourly rate per model regardless of usage, but you can also secure additional savings through monthly and annual reservations. Discover how to transition your regional deployments and provisioned reservations to global and data zones on this Learn page (https://aka.ms/pricing-provisioned-transition).

Region:

Currency:

Model	Min PTUs	PTU Hourly pricing	PTU Monthly Reservation Pricing	PTU Yearly Reservation Pricing
GPT-4.1 Global	15	**$1**	**$260**	**$2,652**
GPT-4.1 Data Zones	15	**$1.10**	**$260**	**$2,652**
GPT-4.1 Regional	50	**$2**	**$260**	**$2,652**
GPT-4.1-mini Global	15	**$1**	**$260**	**$2,652**
GPT-4.1-mini US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
GPT-4.1-mini Regional	25	**$2**	**$260**	**$2,652**
GPT-4.1-nano Global	15	**$1**	**$260**	**$2,652**
GPT-4.1-nano US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
GPT-4.1-nano Regional	25	**$2**	**$260**	**$2,652**
o3 Global	15	**$1**	**$260**	**$2,652**
o3 US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
o3 Regional	50	**$2**	**$260**	**$2,652**
o4-mini Global	15	**$1**	**$260**	**$2,652**

Region:

Currency:

Model	Min PTUs	PTU Hourly pricing	PTU Monthly Reservation Pricing	PTU Yearly Reservation Pricing
o4-mini US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
o4-mini Regional	25	**$2**	**$260**	**$2,652**
GPT-4o Global	15	**$1**	**$260**	**$2,652**
GPT-4o US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
GPT-4o Regional	50	**$2**	**$260**	**$2,652**
Fine-Tuned GPT-4o-Regional	50	**$2**	**$260**	**$2,652**
GPT-4o Mini Global	15	**$1**	**$260**	**$2,652**
GPT-4o Mini US/EU Data Zones	15	**$1.10**	**$260**	**$2,652**
GPT-4o Mini Regional	25	**$2**	**$260**	**$2,652**
Fine-Tuned GPT-4o-Mini Regional	25	**$2**	**$260**	**$2,652**

Plan with the Pricing Calculator (/en-us/pricing/calculator/)

Base models

Region:

Models	Usage per 1,000 tokens
Babbage-002	$0.0004
Davinci-002	$0.002

Currency:

Fine-tuning models

Model		Pricing
o4-mini (Reinforcement fine-tuning)	Region: Regional Currency:	Input: **$1.21**/1M tokens Output: **$4.84**/1M tokens Training: **$110**/1M tokens Hosting: **$1.7**/hour Grader input: o4-mini: **$1.21**/1M tokens 4.1-mini: **$2.20**/1M tokens 4.1: **$2.20**/1M tokens 4.1-nano: **$0.11**/1M tokens o3: **$11**/1M tokens Grader cached input: o4-mini: **$0.303**/1M tokens 4.1-mini: **$0.11**/1M tokens 4.1: **$0.55**/1M tokens 4.1-nano: **$0.55**/1M tokens o3: **$2.75**/1M tokens Grader output: o4-mini: **$4.84**/1M tokens 4.1-mini: **$1.76**/1M tokens 4.1: **$8.80**/1M tokens 4.1-nano: **$0.44**/1M tokens o3: **$44**/1M tokens

Model		Pricing
Region:	Global	Input: **$1.21**/1M tokens Output: **$4.84**/1M tokens Training: **$110**/1M tokens Hosting: **$1.7**/hour
Currency:		Grader input: o4-mini: **$1.21**/1M tokens 4.1-mini: **$2.20**/1M tokens 4.1: **$2.20**/1M tokens 4.1-nano: **$0.11**/1M tokens o3: **$11**/1M tokens Grader cached input: o4-mini: **$0.303**/1M tokens 4.1-mini: **$0.11**/1M tokens 4.1: **$0.55**/1M tokens 4.1-nano: **$0.55**/1M tokens o3: **$2.75**/1M tokens Grader output: o4-mini: **$4.84**/1M tokens 4.1-mini: **$1.76**/1M tokens 4.1: **$8.80**/1M tokens 4.1-nano: **$0.44**/1M tokens o3: **$44**/1M tokens

Model		Pricing
GPT-4.1 Region: Currency:	Regional	Input: **$2.20**/1M tokens Cached Input: **$0.55**/1M tokens Output: **$8.80**/1M tokens Training: **$27.5**/1M tokens Hosting: **$1.7**/hour
	Global	Input: **$2**/1M tokens Cached Input: **$0.50**/1M tokens Output: **$8**/1M tokens Training: **$25**/1M tokens Hosting: **$1.7**/hour
	Developer	Input: **$2**/1M tokens Cached Input: **$0.50**/1M tokens Output: **$8**/1M tokens

Model		Pricing
GPT-4.1-mini Region: Currency:	Regional	Input: **$0.44**/1M tokens Cached Input: **$0.11**/1M tokens Output: **$1.76**/1M tokens Training: **$5.5**/1M tokens Hosting: **$1.7**/hour
	Global	Input: **$0.40**/1M tokens Cached Input: **$0.10**/1M tokens Output: **$1.60**/1M tokens Training: **$5**/1M tokens Hosting: **$1.7**/hour
	Developer	Input: **$0.40**/1M tokens Cached Input: **$0.10**/1M tokens Output: **$1.60**/1M tokens

Model		Pricing
GPT-4.1-nano Region: Currency:	Regional	Input: **$0.11**/1M tokens Cached Input: **$0.028**/1M tokens Output: **$0.44**/1M tokens Training: **$1.7**/1M tokens Hosting: **$1.7**/hour
	Global	Input: **$0.10**/1M tokens Cached Input: **$0.025**/1M tokens Output: **$0.40**/1M tokens Training: **$1.5**/1M tokens Hosting: **$1.7**/hour
	Developer	Input: **$0.10**/1M tokens Cached Input: **$0.025**/1M tokens Output: **$0.40**/1M tokens
GPT-4o-2024-08-06	Regional	Input: N/A/1M tokens Cached Input: N/A/1M tokens Output: N/A/1M tokens Training: N/A/1M tokens Hosting: N/A/hour
	Global	Input: N/A/1M tokens Cached Input: N/A/1M tokens Output: N/A/1M tokens Training: use regional Hosting: N/A/hour

Model		Pricing
GPT-4o-mini Region:	Regional	Input: N/A/1M tokens Cached Input: N/A/1M tokens Output: N/A/1M tokens Training: N/A/1M tokens Hosting: N/A/hour
 Currency:	Global	Input: N/A/1M tokens Cached Input: N/A/1M tokens Output: N/A/1M tokens Training: use regional Hosting: N/A/hour
GPT-3.5-Turbo (16K)	Regional	Input: N/A/1M tokens Output: N/A/1M tokens Training: N/A/1M tokens Hosting: N/A/hour

Image models

Models	Quality	Resolution	Price (per 100 images)
Dall-E-3	Standard	1024 * 1024	**$4**
	Standard	1024 * 1792, 1792 * 1024	**$8**
Dall-E-3	HD	1024 * 1024	**$8**
	HD	1024 * 1792, 1792 * 1024	**$12**

Models	Quality	Resolution	Price (per 100 images)
Dall-E-2	Standard	1024 * 1024	**$2**

Region:

Currency:

Embedding models

Models	Per 1,000 tokens
Ada	**$0.0001**
text-embedding-3-large	**$0.00013**
text-embedding-3-small	**$0.00002**

Speech Models

Pricing is not available in the selected region

Legacy Language Models

Region:

Currency:

Models	Context	Input (Per 1M Tokens)	Output (Per 1M Tokens)
GPT-3.5-Turbo-0301	4K	$1.50	$2
GPT-3.5-Turbo-0613	4K	$1.50	$2
GPT-3.5-Turbo-0613	16K	$3	$4
GPT-3.5-Turbo-1106	16K	$1	$2
GPT-3.5-Turbo-0125	16K	$0.50	$1.50
GPT-3.5-Turbo-Instruct	4K	$1.50	$2
GPT-4-Turbo	128K	$10	$30
GPT-4-Turbo-Vision	128K	$10	$30
GPT-4	8K	$30	$60
GPT-4	32K	$60	$120

Azure pricing and purchasing options

Region:

Currency:

Connect with us directly

Get a walkthrough of Azure pricing. Understand pricing for your cloud solution, learn about cost optimization and request a custom proposal.

Talk to a sales specialist (/en-us/contact/pricing/)

See ways to purchase

Purchase Azure services through the Azure website, a Microsoft representative, or an Azure partner.

Explore your options **(/en-us/pricing/purchase-options/)**

Additional resources

Region:

Azure OpenAI Service (/en-us/products/cognitive-services/openai-service/)

Learn more about Azure OpenAI Service features and capabilities.

Currency:

Pricing calculator (/en-us/pricing/calculator/?service=openai-service)

Estimate your expected monthly costs for using any combination of Azure products.

SLA (/en-us/support/legal/sla/cognitive-services/)

Review the Service Level Agreement for Azure OpenAI Service.

Region:

Currency:

Documentation (https://docs.microsoft.com/en-us/azure/cognitive-services/openai/overview)

Review technical tutorials, videos, and more Azure OpenAI Service resources.

Frequently asked questions

Frequently asked questions about Azure pricing (/en-us/pricing/)

How is Azure OpenAI Service priced?

Where is Azure OpenAI Service available?

What is the SLA for Azure OpenAI Service?

Region:

How can I learn more about provisioned throughput units (PTUs)?

Currency:

Talk to a sales specialist for a walk-through of Azure pricing. Understand pricing for your cloud solution.

Request a pricing quote (/en-us/contact/pricing/#contact-sales)

Get free cloud services and a $200 credit to explore Azure for 30 days.

Try Azure for free (/en-us/free/)