Models

Flagship models

GPT-4o

Our high-intelligence flagship model for complex, multi-step tasks

Text and image input, text output
128k context length
Optimized for intelligence, higher price per token

GPT-4o mini

New

Our affordable and intelligent small model for fast, lightweight tasks

Text and image input, text output
128k context length
Optimized for speed, lower price per token

Model pricing details

Models overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

Model	Description
GPT-4o	Our high-intelligence flagship model for complex, multi-step tasks
GPT-4o mini	Our affordable and intelligent small model for fast, lightweight tasks
GPT-4 Turbo and GPT-4	The previous set of high-intelligence models

GPT-3.5 Turbo	A fast, inexpensive model for simple tasks
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

For GPT-series models, the context window refers to the maximum number of tokens that can be used in a single request, inclusive of both input and output tokens.

We have also published open source models including Point-E, Whisper, Jukebox, and CLIP.

Continuous model upgrades

`gpt-4o`, `gpt-4o-mini`, `gpt-4-turbo`, `gpt-4`, and `gpt-3.5-turbo` point to their respective latest model version. You can verify this by looking at the response object after sending a request. The response will include the specific model version used (e.g. `gpt-3.5-turbo-1106`). The `chatgpt-4o-latest` model version continuously points to the version of GPT-4o used in ChatGPT, and is updated frequently.

With the exception of `chatgpt-4o-latest`, we offer pinned model versions that developers can continue using for at least three months after an updated model has been introduced. With the new cadence of model updates, we are also giving developers the ability to contribute evals to help us improve the model for different use cases. If you are interested, check out the OpenAI Evals repository.

Learn more about model deprecation on our deprecation page.

GPT-4o

GPT-4o ("o" for "omni") is our most advanced model. It is multimodal (accepting text or image inputs and outputting text), and it has the same high intelligence as GPT-4 Turbo but is much more efficient—it generates text 2x faster and is 50% cheaper. Additionally, GPT-4o has the best vision and performance across non-English languages of any of our models. GPT-4o is available in the OpenAI API to paying customers. Learn how to use GPT-4o in our text generation guide.

Model	Description	Context window	Max output tokens	Training data

gpt-4o	**GPT-4o**: Our high-intelligence flagship model for complex, multi-step tasks. GPT-4o is cheaper and faster than GPT-4 Turbo. Currently points to `gpt-4o-2024-05-13` [1].	128,000 tokens	4,096 tokens	Up to Oct 2023
gpt-4o-2024-05-13	`gpt-4o` currently points to this version.	128,000 tokens	4,096 tokens	Up to Oct 2023
gpt-4o-2024-08-06	Latest snapshot that supports Structured Outputs	128,000 tokens	16,384 tokens	Up to Oct 2023
chatgpt-4o-latest	Dynamic model continuously updated to the current version of GPT-4o in ChatGPT. Intended for research and evaluation [2].	128,000 tokens	16,384 tokens	Up to Oct 2023

[1] We will give a 3-week notice before updating `gpt-4o` to point to the new snapshot `gpt-4o-2024-08-06`.

[2] We are releasing this model for developers and researchers to explore OpenAI's latest research. For production use, OpenAI recommends using dated GPT models, which are optimized for API usage.

GPT-4o mini

GPT-4o mini (“o” for “omni”) is our most advanced model in the small models category, and our cheapest model yet. It is multimodal (accepting text or image inputs and outputting text), has higher intelligence than `gpt-3.5-turbo` but is just as fast. It is meant to be used for smaller tasks, including vision tasks.

We recommend choosing `gpt-4o-mini` where you would have previously used `gpt-3.5-turbo` as this model is more capable and cheaper.

Model	Description	Context window	Max output tokens	Training data
gpt-4o-mini	**New** **GPT-4o-mini** Our affordable and intelligent small model for fast, lightweight tasks. GPT-4o mini is cheaper and more capable than GPT-3.5 Turbo. Currently points to	128,000 tokens	16,384 tokens	Up to Oct 2023

	`gpt-4o-mini-2024-07-18`.			
gpt-4o-mini-2024-07-18	`gpt-4o-mini` currently points to this version.	128,000 tokens	16,384 tokens	Up to Oct 2023

GPT-4 Turbo and GPT-4

GPT-4 is a large multimodal model (accepting text or image inputs and outputting text) that can solve difficult problems with greater accuracy than any of our previous models, thanks to its broader general knowledge and advanced reasoning capabilities. GPT-4 is available in the OpenAI API to paying customers. Like `gpt-3.5-turbo`, GPT-4 is optimized for chat but works well for traditional completions tasks using the Chat Completions API. Learn how to use GPT-4 in our text generation guide.

Model	Description	Context window	Max output tokens	Training data
gpt-4-turbo	The latest GPT-4 Turbo model with vision capabilities. Vision requests can now use JSON mode and function calling. Currently points to `gpt-4-turbo-2024-04-09`.	128,000 tokens	4,096 tokens	Up to Dec 2023

gpt-4-turbo-2024-04-09	GPT-4 Turbo with Vision model. Vision requests can now use JSON mode and function calling. `gpt-4-turbo` currently points to this version.	128,000 tokens	4,096 tokens	Up to Dec 2023
gpt-4-turbo-preview	GPT-4 Turbo preview model. Currently points to `gpt-4-0125-preview`.	128,000 tokens	4,096 tokens	Up to Dec 2023
gpt-4-0125-preview	GPT-4 Turbo preview model intended to reduce cases of “laziness” where the model doesn’t complete a task. Learn more.	128,000 tokens	4,096 tokens	Up to Dec 2023

gpt-4-1106-preview	GPT-4 Turbo preview model featuring improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. This is a preview model. Learn more.	128,000 tokens	4,096 tokens	Up to Apr 2023
gpt-4	Currently points to `gpt-4-0613`. See continuous model upgrades.	8,192 tokens	8,192 tokens	Up to Sep 2021
gpt-4-0613	Snapshot of `gpt-4` from June 13th 2023 with improved function calling support.	8,192 tokens	8,192 tokens	Up to Sep 2021
gpt-4-0314	**Legacy** Snapshot of `gpt-4` from March 14th 2023.	8,192 tokens	8,192 tokens	Up to Sep 2021

For many basic tasks, the difference between GPT-4 and GPT-3.5 models is not significant. However, in more complex reasoning situations, GPT-4 is much more capable than any of our previous models.

Multilingual capabilities

GPT-4 outperforms both previous large language models and as of 2023, most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark, an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but also demonstrates strong performance in other languages.

GPT-3.5 Turbo

GPT-3.5 Turbo models can understand and generate natural language or code and have been optimized for chat using the Chat Completions API but work well for non-chat tasks as well.

As of July 2024, `gpt-4o-mini` should be used in place of `gpt-3.5-turbo`, as it is cheaper, more capable, multimodal, and just as fast. `gpt-3.5-turbo` is still available for use in the API.

Model	Description	Context window	Max output tokens	Training data

gpt-3.5-turbo-0125	The latest GPT-3.5 Turbo model with higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls. Learn more.	16,385 tokens	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo	Currently points to `gpt-3.5-turbo-0125`.	16,385 tokens	4,096 tokens	Up to Sep 2021

gpt-3.5-turbo-1106	GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Learn more.	16,385 tokens	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-instruct	Similar capabilities as GPT-3 era models. Compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	4,096 tokens	Up to Sep 2021

DALL·E

DALL·E is a AI system that can create realistic images and art from a description in natural language. DALL·E 3 currently supports the ability, given a prompt, to create a new image with a specific size. DALL·E 2 also support the ability to edit an existing image, or create variations of a user provided image.

DALL·E 3 is available through our Images API along with DALL·E 2. You can try DALL·E 3 through ChatGPT Plus.

Model	Description
`dall-e-3`	The latest DALL·E model released in Nov 2023. Learn more.
`dall-e-2`	The previous DALL·E model released in Nov 2022. The 2nd iteration of DALL·E with more realistic, accurate, and 4x greater resolution images than the original model.

TTS

TTS is an AI model that converts text to natural sounding spoken text. We offer two different model variates, `tts-1` is optimized for real time text to speech use cases and `tts-1-hd` is optimized for quality. These models can be used with the Speech endpoint in the Audio API.

Model	Description
`tts-1`	The latest text to speech model, optimized for speed.
`tts-1-hd`	The latest text to speech model, optimized for quality.

Whisper

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. The Whisper v2-large model is currently available through our API with the `whisper-1` model name.

Currently, there is no difference between the open source version of Whisper and the version available through our API. However, through our API, we offer an optimized inference process which makes running Whisper through our API much faster than

doing it through other means. For more technical details on Whisper, you can read the paper.

Embeddings

Embeddings are a numerical representation of text that can be used to measure the relatedness between two pieces of text. Embeddings are useful for search, clustering, recommendations, anomaly detection, and classification tasks. You can read more about our latest embedding models in the announcement blog post.

Model	Description	Output Dimension
`text-embedding-3-large`	Most capable embedding model for both english and non-english tasks	3,072
`text-embedding-3-small`	Increased performance over 2nd generation ada embedding model	1,536
`text-embedding-ada-002`	Most capable 2nd generation embedding model, replacing 16 first generation models	1,536

Moderation

The Moderation models are designed to check whether content complies with OpenAI's usage policies. The models provide classification capabilities that look for content in the following categories: hate, hate/threatening, self-harm, sexual, sexual/minors, violence, and violence/graphic. You can find out more in our moderation guide.

Moderation models take in an arbitrary sized input that is automatically broken up into chunks of 4,096 tokens. In cases where the input is more than 32,768 tokens, truncation

is used which in a rare condition may omit a small number of tokens from the moderation check.

The final results from each request to the moderation endpoint shows the maximum value on a per category basis. For example, if one chunk of 4K tokens had a category score of 0.9901 and the other had a score of 0.1901, the results would show 0.9901 in the API response since it is higher.

Model	Description	Max tokens
`text-moderation-latest`	Currently points to `text-moderation-007`.	32,768
`text-moderation-stable`	Currently points to `text-moderation-007`.	32,768
`text-moderation-007`	Most capable moderation model across all categories.	32,768

GPT base

GPT base models can understand and generate natural language or code but are not trained with instruction following. These models are made to be replacements for our original GPT-3 base models and use the legacy Completions API. Most customers should use GPT-3.5 or GPT-4.

Model	Description	Max tokens	Training data

`babbage-002`	Replacement for the GPT-3 `ada` and `babbage` base models.	16,384 tokens	Up to Sep 2021
`davinci-002`	Replacement for the GPT-3 `curie` and `davinci` base models.	16,384 tokens	Up to Sep 2021

How we use your data

Your data is your data.

As of March 1, 2023, data sent to the OpenAI API will not be used to train or improve OpenAI models (unless you explicitly opt in). One advantage to opting in is that the models may get better at your use case over time.

To help identify abuse, API data may be retained for up to 30 days, after which it will be deleted (unless otherwise required by law). For trusted customers with sensitive applications, zero data retention may be available. With zero data retention, request and response bodies are not persisted to any logging mechanism and exist only in memory in order to serve the request.

Note that this data policy does not apply to OpenAI's non-API consumer services like ChatGPT or DALL·E Labs.

Default usage policies by endpoint

Endpoint	Data used for training	Default retention	Eligible for zero retention
`/v1/chat/completions*`	No	30 days	Yes, except (a) image inputs or (b) schemas provided

			for Structured Outputs*
/v1/assistants	No	30 days **	No
/v1/threads	No	30 days **	No
/v1/threads/messages	No	30 days **	No
/v1/threads/runs	No	30 days **	No
/v1/vector_stores	No	30 days **	No
/v1/threads/runs/steps	No	30 days **	No
/v1/images/generations	No	30 days	No
/v1/images/edits	No	30 days	No
/v1/images/variations	No	30 days	No

`/v1/embeddings`	No	30 days	Yes
`/v1/audio/transcriptions`	No	Zero data retention	-
`/v1/audio/translations`	No	Zero data retention	-
`/v1/audio/speech`	No	30 days	Yes
`/v1/files`	No	Until deleted by customer	No
`/v1/fine_tuning/jobs`	No	Until deleted by customer	No
`/v1/batches`	No	Until deleted by customer	No
`/v1/moderations`	No	Zero data retention	-
`/v1/completions`	No	30 days	Yes

* Image inputs via the `gpt-4o`, `gpt-4o-mini`, `chatgpt-4o-latest`, or `gpt-4-turbo` models (or previously `gpt-4-vision-preview`) are not eligible for zero retention. When

Structured Outputs is enabled, schemas provided (either as the `response_format` or in the function definition) are not eligible for zero retention, though the completions themselves are.

** Objects related to the Assistants API are deleted from our servers 30 days after you delete them via the API or the dashboard. Objects that are not deleted via the API or dashboard are retained indefinitely.

For details, see our API data usage policies. To learn more about zero retention, get in touch with our sales team.

Model endpoint compatibility

Endpoint	Latest models
/v1/assistants	All GPT-4o (except `chatgpt-4o-latest`), GPT-4o-mini, GPT-4, and GPT-3.5 Turbo models. The `retrieval` tool requires `gpt-4-turbo-preview` (and subsequent dated model releases) or `gpt-3.5-turbo-1106` (and subsequent versions).
/v1/audio/transcriptions	`whisper-1`
/v1/audio/translations	`whisper-1`
/v1/audio/speech	`tts-1`, `tts-1-hd`
/v1/chat/completions	All GPT-4o, GPT-4o-mini, GPT-4, and GPT-3.5 Turbo models and their dated releases. `chatgpt-4o-latest` dynamic model. Fine-tuned

	versions of `gpt-4o`, `gpt-4o-mini`, `gpt-4`, and `gpt-3.5-turbo`.
/v1/completions (Legacy)	`gpt-3.5-turbo-instruct`, `babbage-002`, `davinci-002`
/v1/embeddings	`text-embedding-3-small`, `text-embedding-3-large`, `text-embedding-ada-002`
/v1/fine_tuning/jobs	`gpt-4o-mini`, `gpt-4`, `gpt-3.5-turbo`, `babbage-002`, `davinci-002`
/v1/moderations	`text-moderation-stable`, `text-moderation-latest`
/v1/images/generations	`dall-e-2`, `dall-e-3`

This list excludes all of our deprecated models.