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-4 Turbo and GPT-4	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
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

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

Continuous model upgrades

`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-0613`).

We also 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 people 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-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	TRAINING DATA
`gpt-4-turbo`	New **GPT-4 Turbo with Vision** 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	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	Up to Dec 2023
`gpt-4-turbo-preview`	GPT-4 Turbo preview model. Currently points to `gpt-4-0125-preview`.	128,000 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. Returns a maximum of 4,096 output tokens. Learn more.	128,000 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. Returns a maximum of 4,096 output tokens. This is a preview model. Learn more.	128,000 tokens	Up to Apr 2023

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-4-vision-preview`	GPT-4 model with the ability to understand images, in addition to all other GPT-4 Turbo capabilities. This is a preview model, we recommend developers to now use `gpt-4-turbo` which includes vision capabilities. Currently points to `gpt-4-1106-vision-preview`.	128,000 tokens	Up to Apr 2023
`gpt-4-1106-vision-preview`	GPT-4 model with the ability to understand images, in addition to all other GPT-4 Turbo capabilities. This is a preview model, we recommend developers to now use `gpt-4-turbo` which includes vision capabilities. Returns a maximum of 4,096 output tokens. Learn more.	128,000 tokens	Up to Apr 2023
`gpt-4`	Currently points to `gpt-4-0613`. See continuous model upgrades.	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	Up to Sep 2021
`gpt-4-32k`	Currently points to `gpt-4-32k-0613`. See continuous model upgrades. This model was never rolled out widely in favor of GPT-4 Turbo.	32,768 tokens	Up to Sep 2021
`gpt-4-32k-0613`	Snapshot of `gpt-4-32k` from June 13th 2023 with improved function calling support. This model was never rolled out widely in favor of GPT-4 Turbo.	32,768 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.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-3.5-turbo-0125`	New **Updated GPT 3.5 Turbo** 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. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo`	Currently points to `gpt-3.5-turbo-0125`.	16,385 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. Returns a maximum of 4,096 output tokens. Learn more.	16,385 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	Up to Sep 2021
`gpt-3.5-turbo-16k`	Legacy Currently points to `gpt-3.5-turbo-16k-0613`.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-0613`	Legacy Snapshot of `gpt-3.5-turbo` from June 13th 2023. Will be deprecated on June 13, 2024.	4,096 tokens	Up to Sep 2021

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-3.5-turbo-16k-0613`	Legacy Snapshot of `gpt-3.5-16k-turbo` from June 13th 2023. Will be deprecated on June 13, 2024.	16,385 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`	New **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`	New **Text-to-speech 1** The latest text to speech model, optimized for speed.
`tts-1-hd`	New **Text-to-speech 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	New **Embedding V3 large** Most capable embedding model for both english and non-english tasks	3,072
text-embedding-3-small	New **Embedding V3 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 image inputs*
`/v1/assistants`	No	Until deleted by customer	No
`/v1/threads`	No	60 days *	No
`/v1/threads/messages`	No	60 days *	No
`/v1/threads/runs`	No	60 days *	No
`/v1/threads/runs/steps`	No	60 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-4-turbo` model (or previously `gpt-4-vision-preview`) are not eligible for zero retention.

* For the Assistants API, we are still evaluating the default retention period during the Beta. We expect that the default retention period will be stable after the end of the Beta.

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-4 and GPT-3.5 Turbo models except `gpt-3.5-turbo-0301` supported. 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`	`gpt-4` and dated model releases, `gpt-4-turbo-preview` and dated model releases, `gpt-4-vision-preview`, `gpt-4-32k` and dated model releases, `gpt-3.5-turbo` and dated model releases, `gpt-3.5-turbo-16k` and dated model releases, fine-tuned versions of `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-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.