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.

MODELS	DESCRIPTION
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	A set of models that improve on GPT-3 and can understand as well as generate natural language or code
GPT base	A set of models without instruction following that 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
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-3 **Legacy**	A set of models that can understand and generate natural language
Deprecated	A full list of models that have been deprecated

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

Visit our model index for researchers to learn more about which models have been featured in our research papers and the differences between model series like InstructGPT and GPT-3.5.

Continuous model upgrades

> Based on developer feedback, we are extending support for gpt-3.5-turbo-0301 and gpt-4-0314 models in the OpenAI API until at least June 13, 2024. We've updated our June 13 blog post with more details.

With the release of `gpt-3.5-turbo`, some of our models are now being continually updated. `gpt-3.5-turbo`, `gpt-4`, and `gpt-4-32k` point to the latest model version. You can verify this by looking at the response object after sending a ChatCompletion request. The response will include the specific model version used (e.g. `gpt-3.5-turbo-0613`).

We also offer static 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.

The following models are the temporary snapshots, we will announce their deprecation dates once updated versions are available. If you want to use the latest model version, use the standard model names like `gpt-4` or `gpt-3.5-turbo`.

MODEL NAME	DISCONTINUATION DATE	REPLACEMENT MODEL
`gpt-3.5-turbo-0301`	at earliest 2024-06-13	`gpt-3.5-turbo-0613`
`gpt-4-0314`	at earliest 2024-06-13	`gpt-4-0613`
`gpt-4-32k-0314`	at earliest 2024-06-13	`gpt-4-32k-0613`

Learn more about model deprecation on our deprecation page.

GPT-4

> GPT-4 is currently accessible to those who have made at least one successful payment through our developer platform.

GPT-4 is a large multimodal model (accepting text inputs and emitting text outputs today, with image inputs coming in the future) that can solve difficult problems with greater accuracy than any of our previous models, thanks to its broader general knowledge and advanced reasoning capabilities. 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 GPT guide.

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`gpt-4`	More capable than any GPT-3.5 model, able to do more complex tasks, and optimized for chat. Will be updated with our latest model iteration 2 weeks after it is released.	8,192 tokens	Up to Sep 2021
`gpt-4-0613`	Snapshot of `gpt-4` from June 13th 2023 with function calling data. Unlike `gpt-4`, this model will not receive updates, and will be deprecated 3 months after a new version is released.	8,192 tokens	Up to Sep 2021
`gpt-4-32k`	Same capabilities as the standard `gpt-4` mode but with 4x the context length. Will be updated with our latest model iteration.	32,768 tokens	Up to Sep 2021
`gpt-4-32k-0613`	Snapshot of `gpt-4-32` from June 13th 2023. Unlike `gpt-4-32k`, this model will not receive updates, and will be deprecated 3 months after a new version is released.	32,768 tokens	Up to Sep 2021
`gpt-4-0314 (Legacy)`	Snapshot of `gpt-4` from March 14th 2023 with function calling data. Unlike `gpt-4`, this model will not receive updates, and will be deprecated on June 13th 2024 at the earliest.	8,192 tokens	Up to Sep 2021
`gpt-4-32k-0314 (Legacy)`	Snapshot of `gpt-4-32` from March 14th 2023. Unlike `gpt-4-32k`, this model will not receive updates, and	32,768 tokens	Up to Sep 2021

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
	will be deprecated on June 13th 2024 at the earliest.		

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.

GPT-3.5

GPT-3.5 models can understand and generate natural language or code. Our most capable and cost effective model in the GPT-3.5 family is `gpt-3.5-turbo` which has been optimized for chat using the Chat completions API but works well for traditional completions tasks as well.

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`gpt-3.5-turbo`	Most capable GPT-3.5 model and optimized for chat at 1/10th the cost of `text-davinci-003`. Will be updated with our latest model iteration 2 weeks after it is released.	4,097 tokens	Up to Sep 2021
`gpt-3.5-turbo-16k`	Same capabilities as the standard `gpt-3.5-turbo` model but with 4 times the context.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-instruct`	Similar capabilities as `text-davinci-003` but compatible with legacy Completions endpoint and not Chat Completions.	4,097 tokens	Up to Sep 2021
`gpt-3.5-turbo-0613`	Snapshot of `gpt-3.5-turbo` from June 13th 2023 with function calling data. Unlike `gpt-3.5-turbo`, this model will not receive updates, and	4,097 tokens	Up to Sep 2021

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
	will be deprecated 3 months after a new version is released.		
`gpt-3.5-turbo-16k-0613`	Snapshot of `gpt-3.5-turbo-16k` from June 13th 2023. Unlike `gpt-3.5-turbo-16k`, this model will not receive updates, and will be deprecated 3 months after a new version is released.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-0301 (Legacy)`	Snapshot of `gpt-3.5-turbo` from March 1st 2023. Unlike `gpt-3.5-turbo`, this model will not receive updates, and will be deprecated on June 13th 2024 at the earliest.	4,097 tokens	Up to Sep 2021
`text-davinci-003 (Legacy)`	Can do any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models. Also supports some additional features such as inserting text.	4,097 tokens	Up to Jun 2021
`text-davinci-002 (Legacy)`	Similar capabilities to `text-davinci-003` but trained with supervised fine-tuning instead of reinforcement learning	4,097 tokens	Up to Jun 2021
`code-davinci-002 (Legacy)`	Optimized for code-completion tasks	8,001 tokens	Up to Jun 2021

We recommend using `gpt-3.5-turbo` over the other GPT-3.5 models because of its lower cost and improved performance.

OpenAI models are non-deterministic, meaning that identical inputs can yield different outputs. Setting temperature to 0 will make the outputs mostly

deterministic, but a small amount of variability may remain.

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.

LATEST 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

DALL·E

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

The current DALL·E model available through our API is the 2nd iteration of DALL·E with more realistic, accurate, and 4x greater resolution images than the original model. You can try it through the our Labs interface or via the API.

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. Our second generation embedding model, `text-embedding-ada-002` is a designed to replace the previous 16 first-generation embedding models at a fraction of the cost. Embeddings are useful for search, clustering, recommendations, anomaly detection, and classification tasks. You can read more about our latest embedding model in the announcement blog post.

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 to fix the models specific context window.

MODEL	DESCRIPTION
`text-moderation-latest`	Most capable moderation model. Accuracy will be slightly higher than the stable model.
`text-moderation-stable`	Almost as capable as the latest model, but slightly older.

GPT-3 Legacy

GPT-3 models can understand and generate natural language. These models were superseded by the more powerful GPT-3.5 generation models. However, the original GPT-3 base models (`davinci` , `curie` , `ada` , and `babbage`) are current the only models that are available to fine-tune.

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`text-curie-001`	Very capable, faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
`text-babbage-001`	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`text-ada-001`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019
`davinci`	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality.	2,049 tokens	Up to Oct 2019
`curie`	Very capable, but faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
`babbage`	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019
`ada`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019

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/completions`	No	30 days	Yes

ENDPOINT	DATA USED FOR TRAINING	DEFAULT RETENTION	ELIGIBLE FOR ZERO RETENTION
`/v1/chat/completions`	No	30 days	Yes
`/v1/edits`	No	30 days	Yes
`/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/files`	No	Until deleted by customer	No
`/v1/fine_tuning/jobs`	No	Until deleted by customer	No
`/v1/fine-tunes`	No	Until deleted by customer	No
`/v1/moderations`	No	Zero data retention	-

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/audio/transcriptions`	`whisper-1`
`/v1/audio/translations`	`whisper-1`

ENDPOINT	LATEST MODELS
/v1/chat/completions	gpt-4, gpt-4-0613, gpt-4-32k, gpt-4-32k-0613, gpt-3.5-turbo, gpt-3.5-turbo-0613, gpt-3.5-turbo-16k,