Introduction

Overview

Looking for ChatGPT? Head to chat.openai.com.

The OpenAI API can be applied to virtually any task that requires understanding or generating natural language and code. The OpenAI API can also be used to generate and edit images or convert speech into text. We offer a range of models with different capabilities and price points, as well as the ability to fine-tune custom models.

Resources

- Experiment in the playground
- Read the API reference
- Visit the help center
- View the current API status
- Check out the OpenAI Developer Forum
- Learn about our usage policies

At OpenAI, protecting user data is fundamental to our mission. We do not train our models on inputs and outputs through our API. Learn more on our API data privacy page.

Key concepts

GPTs

OpenAI's GPT (generative pre-trained transformer) models have been trained to understand natural language and code. GPTs provide text outputs in response to their inputs. The inputs to GPTs are also referred to as "prompts". Designing a prompt is essentially how you "program" a GPT model, usually by providing instructions or some examples of how to successfully complete a task. GPTs can be used across a great variety of tasks including content or code generation, summarization, conversation, creative writing, and more. Read more in our introductory GPT guide and in our GPT best practices guide.

Embeddings

An embedding is a vector representation of a piece of data (e.g. some text) that is meant to preserve aspects of its content and/or its meaning. Chunks of data that are similar in some way will tend to have embeddings that are closer together than unrelated data. OpenAI offers text embedding models that take as input a text string and produce as output an embedding vector. Embeddings are useful for search, clustering, recommendations, anomaly detection, classification, and more. Read more about embeddings in our embeddings guide.

Tokens

GPT and embeddings models process text in chunks called tokens. Tokens represent commonly occurring sequences of characters. For example, the string " tokenization" is decomposed as " token" and "ization", while a short and common word like " the" is represented as a single token. Note that in a sentence, the first token of each word typically starts with a space character. Check out our tokenizer tool to test specific strings and see how they are translated into tokens. As a rough rule of thumb, 1 token is approximately 4 characters or 0.75 words for English text.

One limitation to keep in mind is that for a GPT model the prompt and the generated output combined must be no more than the model's maximum context length. For embeddings models (which do not output tokens), the input must be shorter than the model's maximum context length. The maximum

context lengths for each GPT and embeddings model can be found in the model index.

Guides

Jump into one of our guides to learn more.

Quickstart tutorial
Learn by building a quick sample application

GPT
Learn how to generate text

GPT best practices
Learn best practices for building with GPT models

Embeddings
Learn how to search, classify, and compare text

Speech to text
Learn how to turn speech into text

Image generation
Learn how to generate or edit images

Fine-tuning
Learn how to train a model for your use case