# Key concepts

## Text generation models

OpenAI's text generation models (often referred to as generative pre-trained transformers or "GPT" models for short), like GPT-4 and GPT-3.5, have been trained to understand natural and formal language. Models like GPT-4 allows text outputs in response to their inputs. The inputs to these models are also referred to as "prompts". Designing a prompt is essentially how you "program" a model like GPT-4, usually by providing instructions or some examples of how to successfully complete a task. Models like GPT-4 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 text generation guide and in our prompt engineering guide.

## Assistants

Assistants refer to entities, which in the case of the OpenAI API are powered by large language models like GPT-4, that are capable of performing tasks for users. These assistants operate based on the instructions embedded within the context window of the model. They also usually have access to tools which allows the assistants to perform more complex tasks like running code or retrieving information from a file. Read more about assistants in our Assistants API Overview.

## 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

Text generation 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 text generation 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 text generation and embeddings model can be found in the model index.