What's new in v2? Beta

April 2024

We are announcing a variety of new features and improvements to the Assistants API and moving our Beta to a new API version, `OpenAI-Beta: assistants=v2`. Here's what's new:

- We're launching an improved retrieval tool called `file_search`, which can ingest up to 10,000 files per assistant - 500x more than before. It is faster, supports parallel queries through multi-threaded searches, and features enhanced reranking and query rewriting.
- Alongside `file_search`, we're introducing `vector_store` objects in the API. Once a file is added to a vector store, it's automatically parsed, chunked, and embedded, made ready to be searched. Vector stores can be used across assistants and threads, simplifying file management and billing.
- You can now control the maximum number of tokens a run uses in the Assistants API, allowing you to manage token usage costs. You can also set limits on the number of previous / recent messages used in each run.
- We've added support for the `tool_choice` parameter which can be used to force the use of a specific tool (like `file_search`, `code_interpreter`, or a `function`) in a particular run.
- You can now create messages with the role `assistant` to create custom conversation histories in Threads.
- Assistant and Run objects now support popular model configuration parameters like `temperature`, `response_format` (JSON mode), and `top_p`.
- You can now use fine-tuned models in the Assistants API. At the moment, only fine-tuned versions of `gpt-3.5-turbo-0125` are supported.
- Assistants API now supports streaming.
- We've added several streaming and polling helpers to our Node and Python SDKs.

See our migration guide to learn more about how to migrate your tool usage to the latest version of the Assistants API.