You can now train ChatGPT on your own documents via API

Getty Images

reader comments

37 with

On Tuesday, OpenAI announced fine-tuning for GPT-3.5 Turbo—the AI model that powers the free version of ChatGPT—through its API. It allows training the model with custom data, such as company documents or project documentation. OpenAI claims that a fine-tuned model can perform as well as GPT-4 with lower cost in certain scenarios.

In AI, fine-tuning refers to the process of taking a pretrained neural network (like GPT-3.5 Turbo) and further training it on a different dataset (like your custom data), which is typically smaller and possibly related to a specific task. This process builds off of knowledge the model gained during its initial training phase and refines it for a specific application.

So basically, fine-tuning teaches GPT-3.5 Turbo about custom content, such as project documentation or any other written reference. That can come in handy if you want to build an AI assistant based on GPT-3.5 that is intimately familiar with your product or service but lacks knowledge of it in its training data (which, as a reminder, was scraped off the web before September 2021).

“Since the release of GPT-3.5 Turbo, developers and businesses have asked for the ability to customize the model to create

unique and differentiated experiences for their users," writes OpenAI on its promotional blog. "With this launch, developers can now run supervised fine-tuning to make this model perform better for their use cases."

While GPT-4, the more powerful cousin of GPT-3.5, is well-known as a generalist that is adaptable to many subjects, it is slower and more expensive to run. OpenAI is pitching 3.5 fine-tuning as a way to get GPT-4-like performance in a specific knowledge domain at a lower cost and faster execution time. "Early tests have shown a fine-tuned version of GPT-3.5 Turbo can match, or even outperform, base GPT-4-level capabilities on certain narrow tasks," they write.

Also, OpenAI says that fine-tuned models provide "improved steerability," which means following instructions better; "reliable output formatting," which improves the model's ability to consistently output text in a format such as API calls or JSON; and "custom tone," which can bake-in a custom flavor or personality to a chatbot.

Advertisement

OpenAI says that fine-tuning allows users to shorten their prompts and can save money in OpenAI API calls, which are billed per token. "Early testers have reduced prompt size by up to 90% by fine-tuning instructions into the model itself," says OpenAI. Right now, the context length for fine-tuning is set at 4,000 tokens, but OpenAI says that fine-tuning will extend to the 16,000-token model "later this fall."

Using your own data comes at a cost

By now, you might be wondering how using your own data to train GPT-3.5 works—and what it costs. OpenAI lays out a simplified process on its blog that shows setting up a system prompt with the API, uploading files to OpenAI for training, and creating a fine-tuning job using the command-line tool curl to query an API web address. Once the fine-tuning process

is complete, OpenAI says the customized model is available for use immediately with the same rate limits as the base model. More details can be found in OpenAI's official documentation.

All of this comes at a price, of course, and it's split into training costs and usage costs. To train GPT-3.5 costs $0.008 per 1,000 tokens. During the usage phase, API access costs $0.012 per 1,000 tokens for text input and $0.016 per 1,000 tokens for text output.

By comparison, the base 4k GPT-3.5 Turbo model costs $0.0015 per 1,000 tokens input and $0.002 per 1,000 tokens output, so the fine-tuned model is about eight times more expensive to run. And while GPT-4's 8K context model is also cheaper at $0.03 per 1,000 tokens input and $0.06 per 1,000-token output, OpenAI still claims that money can be saved due to the reduced need for prompting in the fine-tuned model. It's a stretch, but in narrow cases, it may apply.

Even at a higher cost, teaching GPT-3.5 about custom documents may be well worth the price for some folks—if you can keep the model from making stuff up about it. Customizing is one thing, but trusting the accuracy and reliability of GPT-3.5 Turbo outputs in a production environment is another matter entirely. GPT-3.5 is well-known for its tendency to confabulate information.

Regarding data privacy, OpenAI notes that, as with all of its APIs, data sent in and out of the fine-tuning API is not used by OpenAI (or anyone else) to train AI models. Interestingly, OpenAI will send all customer fine-tuning training data through GPT-4 for moderation purposes using its recently announced moderation API. That may account for some of the cost of using the fine-tuning service.

And if 3.5 isn't good enough for you, OpenAI says that fine-tuning for GPT-4 is coming this fall. From our experience, that GPT-4 doesn't make things up as much, but fine-tuning

that model (or the rumored 8 models working together under the hood) will likely be far more expensive.

https://arstechnica.com/?p=1962565