How OpenAI may keep enterprises from building their own AI models

The term *AI arms race* is often taken to mean the competition among a relatively small group of Big Tech companies and well-funded startups to build large generative AI models, such as the GPT-4 model that powers ChatGPT. But there's another player in the race—the open-source community—and it's looking less and less like a dark horse.

The current thinking is that just a small class of well-resourced research companies such as OpenAI, Google, Anthropic, Cohere, and Midjourney have enough capital, research talent, and compute power to build, train, and safeguard large, complex AI models. For that reason, the thinking goes, these companies will, for at least the foreseeable future, be the ones to develop the highest-performing, envelope-pushing models.

But that state of affairs may not hold. Last week a document allegedly written by a Google researcher was discovered on a Discord channel showing that at least some within these large, well-monied research companies perceive open source as a major threat.

"We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?" the researcher writes, adding: "But the uncomfortable truth is, we aren't positioned to win this arms race and neither is OpenAI. While we've been squabbling, a third faction has been quietly eating our lunch. I'm talking, of course, about open source."

Open source surging

The researcher cites the pace at which new models are being developed and posted to open-source sites such as Hugging Face and GitHub. These new models, the researcher says, are often smaller, faster, more customizable, require less development time, and are "pound-for-

pound more capable" than the huge models developed by well-monied players such as Google and OpenAI.

"While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly," the researcher writes.

Judging by the numbers, the open-source AI community is indeed in overdrive.

"We have 15,000 companies that are using us that shared over half a million models, data sets, and demos," says Clem Delangue, CEO of Hugging Face, a major repository of open-source machine learning models, data sets, and tools. Delangue says developers have uploaded more than 100,000 specialized AI models to Hugging Face since the release of ChatGPT last November. "So that confirms this intuition that instead of one model to rule them all, actually every single company is going to have their own specialized, customized models."

All this development is happening at a time when large Fortune 500-type businesses across industries are very keen to test the ability of AI text and image generation models to reinvent key business functions such as content creation, marketing, and customer service. AI models are also being used as the technological foundation of a wave of new businesses such as conversational web search and highly personalized AI assistants.

To build or not to build

Many companies have decided the best way to deploy generative AI is to pay to access models developed and hosted by Google or OpenAI via an API (application programming interface). A company might, for example, decide to tap into the OpenAI language models to make their automated customer service chatbot more human-sounding. This may be better than trying to home-grow their own large language models, which is expensive and time-consuming, and ultimately may not produce as good a model.

On the other hand, companies have good reason to seriously consider building their own models. They may not want to send their data, or their customers' data, outside their firewall through an API to an AI company. After all, that data is valuable.

"They consider it their intellectual property, their competitive advantage in their market," says Ali Ghodsi, CEO of the big data warehousing and processing platform Databricks. Ghodsi says he's seen an uptick in the number of his firm's customers who want to host their own AI models. And when a company decides to do that, they begin by accessing open-source models, training data, and tools.

But whether or not a company can go the open-source route is a nuanced question, and the answer depends a lot on the company's maturity, people, and size of bank account.

Two companies, two answers

The developer collaboration platform Replit balked at the idea of paying every time its platform made a call on an API. Replit offers its users a generative AI coding assistant called Ghostwriter, which is similar to Microsoft/Github's Copilot, except that Ghostwriter runs on Replit's home-build model, while Copilot is powered by OpenAI's GPT models (Microsoft owns a major stake in OpenAI). If Replit relied on someone else's large language model it would have had to pay for an API call every time one of its users asked Ghostwriter to generate some code.

"With 20 million users it's actually cheaper in the long run to train your own custom models and host them yourself, says Reza Shabani, Replit's head of AI. "We want to bring that type of AI capability to everyone but it's not scalable if you're just going to pay OpenAI for the API."

The API cost wasn't the only reason. Reza says it's extremely valuable to his company to capture and leverage all the data that its users input on its platform, including requests for Ghostwriter. The prompts its users input can be used to train its own language models, as opposed to sharing that data with a third party such as OpenAI, which may use it to train its own models.

Other companies, especially in regulated industries like banking or healthcare, may have real concerns about sending sensitive data to a third-party model.

However, for many companies it still makes sense to pay for access to a large language model via an application programming interface (API). In fact, a legion of small companies has appeared over the past few months that have built new businesses on top of the OpenAI language models. For them, the decision was easy—the development and operation of a large language model was completely out of the question because of the costs involved. Far better for them to pay for the API and build specialized services around it and a cool user interface on top of it.

Perplexity AI is a good example. The small San Francisco-based company pays for the OpenAI API but adds a web index and a clever UX to it so that it can offer consumers a conversational web "answer engine." Aravind Srinivas, Perplexity cofounder and CEO, says he thought carefully about the possibility of using open-source tools to build and host a large language model, but it quickly became obvious that buying the API was a much better option.

Even if a small company like Perplexity could use open-source tools to build a state-the-art LLM or image generator, it would be worthless to that company if it didn't also have the resources to deploy and maintain it. That includes human resources: For smaller companies, the cost of recruiting and retaining highly talented engineers and researchers needed to develop and maintain large language models can be substantial.

Then there's infrastructure—the whole stack of software necessary to do things like load-balance the queries of users so that they're spread evenly among the servers running the model. And, of course, the cost of the servers themselves is very high. Nvidia's newest H100 GPUs, which were designed for AI processing, can cost as much as $30,000. UBS analyst Timothy Arcuri reported that OpenAI used 10,000 Nvidia GPUs to train the model that powers ChatGPT. One estimate says it cost the company $100 million.

How OpenAI protects its big lead

That's a big barrier to entry for companies that might want to build their own models. Even with the hardware, building models is hard. Implementing small changes may require retraining the whole model, and many times the changes don't end up improving the model's performance. The software tools and platforms used to build the model are often dictated by the kind of servers used for training and deployment, or on the cloud where the models will be hosted.

But these inefficiencies open the door for new startups with innovative ways of making AI development faster and cheaper. OpenAI is very aware of this, and of the evolving economics of AI model development that will alter the calculus of companies considering its API.

The company seems ready to make adjustments to its services to push enterprises toward relying on OpenAI models. It has addressed some of the concerns about data privacy. For example, it has stopped using its API customers' interactions with its models by default to train future versions of OpenAI models; customers now have to opt in for that.

OpenAI says it plans to launch something called "ChatGPT Business" in the coming months—a new flavor of the consumer chatbot service that will let corporate users erase their conversation histories and choose to withhold their interactions with the model from use in training.

It's likely that OpenAI will do more to further entice reluctant corporate users. This could include additional security and privacy features, such as the option of encrypting the content of all calls to the OpenAI servers via the API.

For many companies the decision between building models and using an API comes down to how much customization and fine-tuning they can get from their AI service provider. Many companies may want to fine-tune the GPT model with input data from end users and the corresponding output data generated by the model.

One enterprise source says that OpenAI already does a certain amount of fine-tuning for API customers, but that the functionality isn't all there yet, and it isn't easy to use. OpenAI will likely roll out a more polished and functional fine-tuning service, says the source, who spoke on condition of anonymity. This would involve OpenAI hosting a smaller model that's trained on the customer's own prompts.

That service will likely cost significantly more, the source says, but it may be enough to keep potential API customers from taking the plunge and building from scratch with open source. At least for a while.

ABOUT THE AUTHOR

Mark Sullivan is a San Francisco-based senior writer at *Fast Company* who focuses on chronicling the advance of artificial intelligence and its effects on business and culture. He's interviewed luminaries from the emerging space including former Google CEO Eric Schmidt, Microsoft's Mustafa Suleyman, and OpenAI's Brad Lightcap More

---