OpenAI's plans according to Sam Altman

Raza Habib

OpenAI’s plans according to Sam Altman

Last week I had the privilege to sit down with Sam Altman and 20 other developers to discuss OpenAI’s APIs and their product plans. Sam was remarkably open. The discussion touched on practical developer issues as well as bigger-picture questions related to OpenAI’s mission and the societal impact of AI. Here are the key takeaways:

1. OpenAI is heavily GPU limited at present

A common theme that came up throughout the discussion was that currently OpenAI is extremely GPU-limited and this is delaying a lot of their short-term plans. The biggest customer complaint was about the reliability and speed of the API. Sam acknowledged their concern and explained that most of the issue was a result of GPU shortages.

The longer 32k context can’t yet be rolled out to more people. OpenAI haven’t overcome the O(n^2) scaling of attention and so whilst it seemed plausible they would have 100k - 1M token context windows soon (this year) anything bigger would require a research breakthrough.

The finetuning API is also currently bottlenecked by GPU availability. They don’t yet use efficient finetuning methods like Adapters or LoRa and so finetuning is very compute-intensive to run and manage. Better support for finetuning will come in the future. They may even host a marketplace of community contributed models.

Dedicated capacity offering is limited by GPU availability. OpenAI also offers dedicated capacity, which provides customers with a private copy of the model. To access this service, customers must be willing to commit to a $100k spend upfront.

2. OpenAI's near-term roadmap

Sam shared what he saw as OpenAI's provisional near-term roadmap for the API.

2023:

- **Cheaper and faster GPT-4 —** This is their top priority. In general, OpenAI's aim is to drive "the cost of intelligence" down as far as possible and so they will work hard to continue to reduce the cost of the APIs over time.
- **Longer context windows —** Context windows as high as 1 million tokens are plausible in the near future.
- **Finetuning API —** The finetuning API will be extended to the latest models but the exact form for this will be shaped by what developers indicate they really want.
- **A stateful API —** When you call the chat API today, you have to repeatedly pass through the same conversation history and pay for the same tokens again and again. In the future there will be a version of the API that remembers the conversation history.

2024:

- **Multimodality —** This was demoed as part of the GPT-4 release but can't be extended to everyone until after more GPUs come online.

3. Plugins "don't have PMF" and are probably not coming to the API anytime soon

A lot of developers are interested in getting access to ChatGPT plugins via the API but Sam said he didn't think they'd be released any time soon. The usage of plugins, other than browsing, suggests that they don't have PMF yet. He suggested that a lot of people thought they wanted their apps to be inside ChatGPT but what they really wanted was ChatGPT in their apps.

4. OpenAI will avoid competing with their customers — other than with ChatGPT

Quite a few developers said they were nervous about building with the OpenAI APIs when OpenAI might end up releasing products that are competitive to them. Sam said that OpenAI would not release more products beyond ChatGPT. He said there was a history of great platform companies having a killer app and that ChatGPT would allow them to make the APIs better by being customers of their own product. The vision for ChatGPT is to be a super smart assistant for work but there will be a lot of other GPT use-cases that OpenAI won't touch.

5. Regulation is needed but so is open source

While Sam is calling for regulation of future models, he didn't think existing models were dangerous and thought it would be a big mistake to regulate or ban them. He reiterated his belief in the importance of open source and said that OpenAI was considering open-sourcing GPT-3. Part of the reason they hadn't open-sourced yet was that he was skeptical of how many individuals and companies would have the capability to host and serve large LLMs.

6. The scaling laws still hold

Recently many articles have claimed that "the age of giant AI Models is already over". This wasn't an accurate representation of what was meant.

OpenAI's internal data suggests the scaling laws for model performance continue to hold and making models larger will continue to yield performance. The rate of scaling can't be maintained because OpenAI had made models millions of times bigger in just a few years and doing that going forward won't be sustainable. That doesn't mean that OpenAI won't continue to try to make the models bigger, it just means they will likely double or triple in size each year rather than increasing by many orders of magnitude.

The fact that scaling continues to work has significant implications for the timelines of AGI development. The scaling hypothesis is the idea that we may have most of the pieces in place needed to build AGI and that most of the remaining work will be taking existing methods and scaling them up to larger models and bigger datasets. If the era of scaling was over then we should probably expect AGI to be much further away. The fact the scaling laws continue to hold is strongly suggestive of shorter timelines.

Discuss on Twitter

[redacted]