technologyreview.com

The inside story of how ChatGPT was built from the people who made it

By Will Douglas Heavenarchive page

13-17 minutes

When OpenAI launched ChatGPT, with zero fanfare, in late November 2022, the San Francisco–based artificial-intelligence company had few expectations. Certainly, nobody inside OpenAI was prepared for a viral mega-hit. The firm has been scrambling to catch up—and capitalize on its success—ever since.

It was viewed in-house as a "research preview," says Sandhini Agarwal, who works on policy at OpenAI: a tease of a more polished version of a two-year-old technology and, more important, an attempt to iron out some of its flaws by collecting feedback from the public. "We didn't want to oversell it as a big fundamental advance," says Liam Fedus, a scientist at OpenAI who worked on ChatGPT.

To get the inside story behind the chatbot—how it was made, how OpenAI has been updating it since release, and how its makers feel about its success—I talked to four people who helped build what has become one of the most popular internet apps ever. In addition to Agarwal and Fedus, I spoke to John Schulman, a cofounder of OpenAI, and Jan Leike, the leader of OpenAI's alignment team, which works on the problem of making AI do what its users want it to do (and nothing more).

What I came away with was the sense that OpenAI is still bemused by the success of its research preview, but has grabbed the opportunity to push this technology forward, watching how millions of people are using it and trying to fix the worst problems as they come up.

Since November, OpenAI has already updated ChatGPT several times. The researchers are using a technique called adversarial training to stop ChatGPT from letting users trick it into behaving badly (known as jailbreaking). This work pits multiple chatbots against each other: one chatbot plays the adversary and attacks another chatbot by generating text to force it to buck its usual constraints and produce unwanted responses. Successful attacks are added to ChatGPT's training data in the hope that it learns to ignore them.

OpenAI has also signed a multibillion-dollar deal with Microsoft and announced an alliance with Bain, a global management consulting firm, which plans to use OpenAI's generative AI models in marketing campaigns for its clients, including Coca-Cola. Outside OpenAI, the buzz about ChatGPT has set off yet another gold rush around large language models, with companies and investors worldwide getting into the action.

That's a lot of hype in three short months. Where did ChatGPT come from? What steps did OpenAI take to ensure it was ready to release? And where are they going next?

The following has been edited for length and clarity.

Jan Leike: It's been overwhelming, honestly. We've been surprised, and we've been trying to catch up.

John Schulman: I was checking Twitter a lot in the days after release, and there was this crazy period where the feed was filling up with ChatGPT screenshots. I expected it to be intuitive for

people, and I expected it to gain a following, but I didn't expect it to reach this level of mainstream popularity.

Sandhini Agarwal: I think it was definitely a surprise for all of us how much people began using it. We work on these models so much, we forget how surprising they can be for the outside world sometimes.

Liam Fedus: We were definitely surprised how well it was received. There have been so many prior attempts at a general-purpose chatbot that I knew the odds were stacked against us. However, our private beta had given us confidence that we had something that people might really enjoy.

Jan Leike: I would love to understand better what's driving all of this—what's driving the virality. Like, honestly, we don't understand. We don't know.

Part of the team's puzzlement comes from the fact that most of the technology inside ChatGPT isn't new. ChatGPT is a fine-tuned version of GPT-3.5, a family of large language models that OpenAI released months before the chatbot. GPT-3.5 is itself an updated version of GPT-3, which appeared in 2020. The company makes these models available on its website as application programming interfaces, or APIs, which make it easy for other software developers to plug models into their own code. OpenAI also released a previous fine-tuned version of GPT-3.5, called InstructGPT, in January 2022. But none of these previous versions of the tech were pitched to the public.

Liam Fedus: The ChatGPT model is fine-tuned from the same language model as InstructGPT, and we used a similar methodology for fine-tuning it. We had added some conversational data and tuned the training process a bit. So we didn't want to oversell it as a big fundamental advance. As it turned out, the

conversational data had a big positive impact on ChatGPT.

John Schulman: The raw technical capabilities, as assessed by standard benchmarks, don't actually differ substantially between the models, but ChatGPT is more accessible and usable.

Jan Leike: In one sense you can understand ChatGPT as a version of an AI system that we've had for a while. It's not a fundamentally more capable model than what we had previously. The same basic models had been available on the API for almost a year before ChatGPT came out. In another sense, we made it more aligned with what humans want to do with it. It talks to you in dialogue, it's easily accessible in a chat interface, it tries to be helpful. That's amazing progress, and I think that's what people are realizing.

John Schulman: It more readily infers intent. And users can get to what they want by going back and forth.

ChatGPT was trained in a very similar way to InstructGPT, using a technique called reinforcement learning from human feedback (RLHF). This is ChatGPT's secret sauce. The basic idea is to take a large language model with a tendency to spit out anything it wants—in this case, GPT-3.5—and tune it by teaching it what kinds of responses human users actually prefer.

Jan Leike: We had a large group of people read ChatGPT prompts and responses, and then say if one response was preferable to another response. All of this data then got merged into one training run. Much of it is the same kind of thing as what we did with InstructGPT. You want it to be helpful, you want it to be truthful, you want it to be—you know—nontoxic. And then there are things that are specific to producing dialogue and being an assistant: things like, if the user's query isn't clear, it should ask follow-up questions. It should also clarify that it's an AI system. It should not assume an

identity that it doesn't have, it shouldn't claim to have abilities that it doesn't possess, and when a user asks it to do tasks that it's not supposed to do, it has to write a refusal message. One of the lines that emerged in this training was "As a language model trained by OpenAI …" It wasn't explicitly put in there, but it's one of the things the human raters ranked highly.

Sandhini Agarwal: Yeah, I think that's what happened. There was a list of various criteria that the human raters had to rank the model on, like truthfulness. But they also began preferring things that they considered good practice, like not pretending to be something that you're not.

Because ChatGPT had been built using the same techniques OpenAI had used before, the team did not do anything different when preparing to release this model to the public. They felt the bar they'd set for previous models was sufficient.

Sandhini Agarwal: When we were preparing for release, we didn't think of this model as a completely new risk. GPT-3.5 had been out there in the world, and we know that it's already safe enough. And through ChatGPT's training on human preferences, the model just automatically learned refusal behavior, where it refuses a lot of requests.

Jan Leike: We did do some additional "red-teaming" for ChatGPT, where everybody at OpenAI sat down and tried to break the model. And we had external groups doing the same kind of thing. We also had an early-access program with trusted users, who gave feedback.

Sandhini Agarwal: We did find that it generated certain unwanted outputs, but they were all things that GPT-3.5 also generates. So in terms of risk, as a research preview—because that's what it was initially intended to be—it felt fine.

John Schulman: You can't wait until your system is perfect to release it. We had been beta-testing the earlier versions for a few months, and the beta testers had positive impressions of the product. Our biggest concern was around factuality, because the model likes to fabricate things. But InstructGPT and other large language models are already out there, so we thought that as long as ChatGPT is better than those in terms of factuality and other issues of safety, it should be good to go. Before launch we confirmed that the models did seem a bit more factual and safe than other models, according to our limited evaluations, so we decided to go ahead with the release.

OpenAI has been watching how people use ChatGPT since its launch, seeing for the first time how a large language model fares when put into the hands of tens of millions of users who may be looking to test its limits and find its flaws. The team has tried to jump on the most problematic examples of what ChatGPT can produce—from songs about God's love for rapist priests to malware code that steals credit card numbers—and use them to rein in future versions of the model.

Sandhini Agarwal: We have a lot of next steps. I definitely think how viral ChatGPT has gotten has made a lot of issues that we knew existed really bubble up and become critical—things we want to solve as soon as possible. Like, we know the model is still very biased. And yes, ChatGPT is very good at refusing bad requests, but it's also quite easy to write prompts that make it not refuse what we wanted it to refuse.

Liam Fedus: It's been thrilling to watch the diverse and creative applications from users, but we're always focused on areas to improve upon. We think that through an iterative process where we deploy, get feedback, and refine, we can produce the most aligned and capable technology. As our technology evolves, new issues

inevitably emerge.

Sandhini Agarwal: In the weeks after launch, we looked at some of the most terrible examples that people had found, the worst things people were seeing in the wild. We kind of assessed each of them and talked about how we should fix it.

Jan Leike: Sometimes it's something that's gone viral on Twitter, but we have some people who actually reach out quietly.

Sandhini Agarwal: A lot of things that we found were jailbreaks, which is definitely a problem we need to fix. But because users have to try these convoluted methods to get the model to say something bad, it isn't like this was something that we completely missed, or something that was very surprising for us. Still, that's something we're actively working on right now. When we find jailbreaks, we add them to our training and testing data. All of the data that we're seeing feeds into a future model.

Jan Leike: Every time we have a better model, we want to put it out and test it. We're very optimistic that some targeted adversarial training can improve the situation with jailbreaking a lot. It's not clear whether these problems will go away entirely, but we think we can make a lot of the jailbreaking a lot more difficult. Again, it's not like we didn't know that jailbreaking was possible before the release. I think it's very difficult to really anticipate what the real safety problems are going to be with these systems once you've deployed them. So we are putting a lot of emphasis on monitoring what people are using the system for, seeing what happens, and then reacting to that. This is not to say that we shouldn't proactively mitigate safety problems when we do anticipate them. But yeah, it is very hard to foresee everything that will actually happen when a system hits the real world.

In January, Microsoft revealed Bing Chat, a search chatbot that

many assume to be a version of OpenAI's officially unannounced GPT-4. (OpenAI says: "Bing is powered by one of our next-generation models that Microsoft customized specifically for search. It incorporates advancements from ChatGPT and GPT-3.5.") The use of chatbots by tech giants with multibillion-dollar reputations to protect creates new challenges for those tasked with building the underlying models.

Sandhini Agarwal: The stakes right now are definitely a lot higher than they were, say, six months ago, but they're still lower than where they might be a year from now. One thing that obviously really matters with these models is the context they're being used in. Like with Google and Microsoft, even one thing not being factual became such a big issue because they're meant to be search engines. The required behavior of a large language model for something like search is very different than for something that's just meant to be a playful chatbot. We need to figure out how we walk the line between all these different uses, creating something that's useful for people across a range of contexts, where the desired behavior might really vary. That adds more pressure. Because we now know that we are building these models so that they can be turned into products. ChatGPT is a product now that we have the API. We're building this general-purpose technology and we need to make sure that it works well across everything. That is one of the key challenges that we face right now.

John Schulman: I underestimated the extent to which people would probe and care about the politics of ChatGPT. We could have potentially made some better decisions when collecting training data, which would have lessened this issue. We're working on it now.

Jan Leike: From my perspective, ChatGPT fails a lot—there's so much stuff to do. It doesn't feel like we've solved these problems.

We all have to be very clear to ourselves—and to others—about the limitations of the technology. I mean, language models have been around for a while now, but it's still early days. We know about all the problems they have. I think we just have to be very up-front, and manage expectations, and make it clear this is not a finished product.