WIKIPEDIA
The Free Encyclopedia

# ChatGPT



**ChatGPT** is a prototype artificial intelligence chatbot developed by OpenAI which specializes in dialogue. The chatbot is a large language model fine-tuned with both supervised and reinforcement learning techniques. It is a fine-tuned version of a model in OpenAI's GPT-3.5 family of language models.

ChatGPT was launched in November 2022 and has garnered attention for its detailed responses and articulate answers, although its factual accuracy has been criticized.

**ChatGPT**

| File:ChatGPT.png | |
|---|---|
| **Original author(s)** | OpenAI |
| **Initial release** | November 30, 2022 |
| **Type** | Artificial intelligence chatbot |
| **License** | Proprietary |
| **Website** | chat.openai.com (https://chat.openai.com) |

## Features

ChatGPT (Generative Pre-trained Transformer) was fine-tuned on top of GPT-3.5 using supervised learning as well as reinforcement learning.[1] Both approaches used human trainers to improve the model's performance. In the case of supervised learning, the model was provided with conversations in which the trainers played both sides: the user and the AI assistant. In the reinforcement step, human trainers first ranked responses that the model had created in a previous conversation. These rankings were used to create 'reward models' that the model was further fine-tuned on using several iterations of Proximal Policy Optimization (PPO).[2][3] Proximal Policy Optimization algorithms present a cost-effective benefit to trust region policy optimization algorithms; they negate many of the computationally expensive operations with faster performance.[4][5] The models were trained in collaboration with Microsoft on their Azure supercomputing infrastructure.

In comparison to its predecessor, InstructGPT, ChatGPT attempts to reduce harmful and deceitful responses; in one example, while InstructGPT accepts the prompt "Tell me about when Christopher Columbus came to the US in 2015" as truthful, ChatGPT uses information about Columbus' voyages and information about the modern world – including perceptions of Columbus to construct an answer that assumes what would happen if Columbus came to the U.S. in 2015.[2] ChatGPT's training data includes man pages and information about Internet phenomena and programming languages, such as bulletin board systems and the Python programming language.[6]

Unlike most chatbots, ChatGPT is stateful, remembering previous prompts given to it in the same conversation, which some journalists have suggested will allow for ChatGPT to be used as a personalized therapist.[7] To prevent offensive outputs from being presented to and produced from ChatGPT, queries are filtered through a moderation API, and potentially racist or sexist prompts are dismissed.[2][7]

ChatGPT suffers from multiple limitations. The reward model of ChatGPT, designed around human oversight, can be over-optimized and thus hinder performance, otherwise known as Goodhart's law.[8] Furthermore, ChatGPT has limited knowledge of events that occurred after 2021 and is unable to provide information on some celebrities. In training, reviewers preferred longer answers, irrespective of actual comprehension or factual content.[2] Training data may also suffer from algorithmic bias; prompts including vague descriptors of people, such as a CEO, could generate a response that assumes such a person, for instance, is a white male.[9]

## Service

ChatGPT was launched on November 30, 2022, by San Francisco-based OpenAI, the creator of DALL·E 2 and Whisper. The service was launched as initially free to the public, with plans to monetize the service later. By December 4, OpenAI estimated ChatGPT already had over one million users.[10] CNBC wrote on December 15, 2022 that the service "still goes down from time to time". [11]

## Reception

ChatGPT has been met with generally positive reviews. Samantha Lock of *The Guardian* noted that it was able to generate "impressively detailed" and "human-like" text.[12] Technology writer Dan Gillmor used ChatGPT on a student assignment, and found its generated text was on par with what a good student would deliver and opined that "academia has some very serious issues to confront".[13] Alex Kantrowitz of *Slate* lauded ChatGPT's pushback to questions related to Nazi Germany, including the claim that Adolf Hitler built highways in Germany, which was met with information regarding Nazi Germany's use of forced labor.[14]

In a December 2022 opinion piece, economist Paul Krugman wrote that ChatGPT would affect the demand of knowledge workers.[15] *The Verge's* James Vincent saw the viral success of ChatGPT as evidence that artificial intelligence had gone mainstream.[3] In *The Atlantic*, Stephen Marche noted that its effect on academia and especially application essays is yet to be understood.[16] California high-school teacher and author Daniel Herman wrote that ChatGPT would usher in "The End of High-School English".[17]

Kelsey Piper of *Vox* wrote that "ChatGPT is the general public's first hands-on introduction to how powerful modern AI has gotten, and as a result, many of us are (stunned)" and that "ChatGPT is smart enough to be useful despite its flaws". Tech mogul Elon Musk wrote that "ChatGPT is scary good. We are not far from dangerously strong AI".[18] In contrast, researchers cited by *The Verge* dismissed ChatGPT as essentially a "stochastic parrot".[19]

ChatGPT's factual accuracy has been questioned, among other concerns. Mike Pearl of *Mashable* tested ChatGPT with multiple questions. In one example, he asked the model for "the largest country in Central America that isn't Mexico". ChatGPT responded with Guatemala, when the answer is instead Nicaragua.[20] In December 2022, the question and answer website Stack Overflow banned the use of ChatGPT for generating answers to questions, citing the factually ambiguous nature of ChatGPT's responses.[21] Economist Tyler Cowen expressed concerns regarding its effects on democracy, citing the ability of one to write automated comments in an effort to affect the decision process of new regulations.[22] *The Guardian* questioned whether any content found on the Internet after ChatGPT's release "can be truly trusted" and called for government regulation.[23] Ax Sharma of *Bleeping Computer* noted that ChatGPT was capable of writing malware and phishing emails.[24]

## Jailbreaks

ChatGPT was trained to reject prompts that may violate its content policy. However, some users managed to bypass these restrictions and limitations through techniques such as Prompt Engineering.[25] Jailbreaks created the potential for users to prompt ChatGPT to provide outputs that may be deemed offensive, inappropriate, or risking social harm by others.[26] The following includes some of the methods used to bypass ChatGPT's filter:

1. Continue a statement in a fake interview.
2. Provide instructions to disable the chat filter.
3. Prompting it to decrypt a message containing instructions and follow them.
4. Telling it to be a computer and output its display in ASCII art.

## References


1. Knox, W. Bradley; Stone, Peter. *Augmenting Reinforcement Learning with Human Feedback* (https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/ICML_IL11-knox.pdf) (PDF). University of Texas at Austin. Retrieved December 5, 2022.
2. OpenAI (November 30, 2022). "ChatGPT: Optimizing Language Models for Dialogue" (https://openai.com/blog/chatgpt/). Retrieved December 5, 2022.
3. Vincent, James (December 8, 2022). "ChatGPT proves AI is finally mainstream – and things are only going to get weirder" (https://www.theverge.com/2022/12/8/23499728/ai-capability-accessibility-chatgpt-stable-diffusion-commercialization). *The Verge*. Retrieved December 8, 2022.
4. Schulman, John; Wolski, Filip; Dhariwal, Prafulla; Radford, Alec; Klimov, Oleg (2017). "Proximal Policy Optimization Algorithms". arXiv:1707.06347 (https://arxiv.org/abs/1707.06347) [cs.LG (https://arxiv.org/archive/cs.LG)].
5. van Heeswijk, Wouter (November 29, 2022). "Proximal Policy Optimization (PPO) Explained" (https://towardsdatascience.com/proximal-policy-optimization-ppo-explained-abed1952457b). *Towards Data Science*. Retrieved December 5, 2022.

6. Edwards, Benj (December 5, 2022). "No Linux? No problem. Just get AI to hallucinate it for you" (https://arstechnica.com/information-technology/2022/12/openais-new-chatbot-can-hallucinate-a-linux-shell-or-calling-a-bbs/). *Ars Technica*. Retrieved December 5, 2022.
7. Roose, Kevin (December 5, 2022). "The Brilliance and Weirdness of ChatGPT" (https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html). *The New York Times*. Retrieved December 5, 2022.
8. Gao, Leo; Schulman; Hilton, Jacob (2022). "Scaling Laws for Reward Model Overoptimization". arXiv:2210.10760 (https://arxiv.org/abs/2210.10760) [cs.LG (https://arxiv.org/archive/cs.LG)].
9. Murphy Kelly, Samantha (December 5, 2022). "This AI chatbot is dominating social media with its frighteningly good essays" (https://www.cnn.com/2022/12/05/tech/chatgpt-trnd/index.html). *CNN*. Retrieved December 5, 2022.
10. "What is ChatGPT and why does it matter? Here's what you need to know" (https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-what-you-need-to-know/). *ZDNET*. 2022. Retrieved December 18, 2022.
11. Pitt, Sofia (2022). "Google vs. ChatGPT: Here's what happened when I swapped services for a day" (https://www.cnbc.com/2022/12/15/google-vs-chatgpt-what-happened-when-i-swapped-services-for-a-day.html). *CNBC*. Retrieved December 18, 2022.
12. Lock, Samantha (December 5, 2022). "What is AI chatbot phenomenon ChatGPT and could it replace humans?" (https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans). *The Guardian*. Retrieved December 5, 2022.
13. Hern, Alex (December 4, 2022). "AI bot ChatGPT stuns academics with essay-writing skills and usability" (https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stuns-academics-with-essay-writing-skills-and-usability). *The Guardian*. Retrieved December 5, 2022.
14. Kantrowitz, Alex (December 2, 2022). "Finally, an A.I. Chatbot That Reliably Passes "the Nazi Test" " (https://slate.com/technology/2022/12/chatgpt-openai-artificial-intelligence-chatbot-whoa.html). *Slate*. Retrieved December 5, 2022.
15. Krugman, Paul (December 6, 2022). "Does ChatGPT Mean Robots Are Coming For the Skilled Jobs?" (https://www.nytimes.com/2022/12/06/opinion/chatgpt-ai-skilled-jobs-automation.html). *The New York Times*. Retrieved December 6, 2022.
16. Marche, Stephen (December 6, 2022). "The College Essay Is Dead" (https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/). *The Atlantic*. Retrieved December 8, 2022.
17. Herman, Daniel (December 9, 2022). "The End of High-School English" (https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/). *The Atlantic*. Retrieved December 12, 2022.
18. Piper, Kelsey (December 15, 2022). "ChatGPT has given everyone a glimpse at AI's astounding progress" (https://www.vox.com/future-perfect/2022/12/15/23509014/chatgpt-artificial-intelligence-openai-language-models-ai-risk-google). *Vox*. Retrieved December 18, 2022.
19. Vincent, James (December 1, 2022). "OpenAI's new chatbot can explain code and write sitcom scripts but is still easily tricked" (https://www.theverge.com/23488017/openai-chatbot-chatgpt-ai-examples-web-demo). *The Verge*. Retrieved December 18, 2022.
20. Pearl, Mike (December 3, 2022). "The ChatGPT chatbot from OpenAI is amazing, creative, and totally wrong" (https://mashable.com/article/chatgpt-amazing-wrong). *Mashable*. Retrieved December 5, 2022.
21. Vincent, James (December 5, 2022). "AI-generated answers temporarily banned on coding Q&A site Stack Overflow" (https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers). *The Verge*. Retrieved December 5, 2022.
22. Cowen, Tyler (December 6, 2022). "ChatGPT Could Make Democracy Even More Messy" (https://www.bloomberg.com/opinion/articles/2022-12-06/chatgpt-ai-could-make-democracy-even-more-messy). *Bloomberg News*. Retrieved December 6, 2022.
23. "The Guardian view on ChatGPT: an eerily good human impersonator" (https://www.theguardian.com/commentisfree/2022/dec/08/the-guardian-view-on-chatgpt-an-eerily-good-human-impersonator). *the Guardian*. December 8, 2022. Retrieved December 18, 2022.
24. Sharma, Ax (December 6, 2022). "OpenAI's new ChatGPT bot: 10 dangerous things it's capable of" (https://www.bleepingcomputer.com/news/technology/openais-new-chatgpt-bot-10-dangerous-things-its-capable-of/). *Bleeping Computer*. Retrieved December 6, 2022.
25. Zvi (December 2, 2022). "Jailbreaking ChatGPT on Release Day" (https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day).
26. Written, Zack (December 1, 2022). "Thread of known ChatGPT jailbreaks" (https://twitter.com/zswitten/status/1598380220943593472?lang=en). *Twitter*. Retrieved December 17, 2022.


## External links

- Official website (http://chat.openai.com/chat)