ChatGPT

ChatGPT (Generative Pre-trained Transformer[1]) is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3.5 family of large language models, and is fine-tuned with both supervised and reinforcement learning techniques.

ChatGPT was launched as a prototype on November 30, 2022, and quickly garnered attention for its detailed responses and articulate answers across many domains of knowledge. Its uneven factual accuracy was identified as a significant drawback.[2]

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)

Training

ChatGPT was fine-tuned on top of GPT-3.5 using supervised learning as well as reinforcement learning.[3] 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).[4][5] 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.[6][7] The models were trained in collaboration with Microsoft on their Azure supercomputing infrastructure.

In addition, OpenAI continues to gather data from ChatGPT users that could be used to further train and fine-tune ChatGPT. Users are allowed to upvote or downvote the responses they receive from ChatGPT; upon upvoting or downvoting, they can also fill out a text field with additional feedback.[8][9][10]

Features and limitations

In comparison to its predecessor, InstructGPT, ChatGPT attempts to reduce harmful and deceitful responses;[11] 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.[4] 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.[12]

Unlike most chatbots, ChatGPT remembers previous prompts given to it in the same conversation; journalists have suggested that this will allow ChatGPT to be used as a personalized therapist.[13] To prevent offensive outputs from being presented to and produced from ChatGPT, queries are filtered through OpenAI's company-wide[14][15] moderation API, and potentially racist or sexist prompts are dismissed.[4][13]

ChatGPT suffers from multiple limitations. OpenAI acknowledged that ChatGPT "sometimes writes plausible-sounding but incorrect or nonsensical answers".[4] The reward model of ChatGPT, designed around human oversight, can be over-optimized and thus hinder performance, otherwise known as Goodhart's law.[16] Furthermore, ChatGPT has limited knowledge of events that occurred after 2021 and is unable to provide information on some celebrities. According to the BBC, as of December 2022 ChatGPT is not currently allowed to "express political opinions or engage in political activism".[17] In training ChatGPT, human reviewers preferred longer answers,

irrespective of actual comprehension or factual content.[4] Training data also suffers from algorithmic bias, which may be revealed when ChatGPT responds to prompts including descriptors of people. In one instance, ChatGPT generated a rap indicating that women and scientists of color were inferior to white and male scientists.[18][19]

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.[8] CNBC wrote on December 15, 2022, that the service "still goes down from time to time".[20] Unlike some other recent high-profile advances in AI, as of December 2022, there is no sign of an official peer-reviewed technical paper about ChatGPT.[21]

According to OpenAI guest researcher Scott Aaronson, OpenAI is working on a tool to attempt to watermark its text generation systems so as to combat bad actors using their services for academic plagiarism or for spam.[22][23]

Reception, criticism and issues

Positive reactions

ChatGPT was met in December 2022 with generally positive reviews; *The New York Times* labeled it "the best artificial intelligence chatbot ever released to the general public".[24] Samantha Lock of Britain's *The Guardian* newspaper noted that it was able to generate "impressively detailed" and "human-like" text.[25] 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".[26] Alex Kantrowitz of *Slate* magazine 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.[27]

In *The Atlantic*'s "Breakthroughs of the Year" for 2022, Derek Thompson included ChatGPT as part of "the generative-AI eruption" that "may change our mind about how we work, how we think, and what human creativity really is".[28]

Kelsey Piper of the *Vox* website 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". Paul Graham of Y Combinator tweeted that "The striking thing about the reaction to ChatGPT is not just the number of people who are blown away by it, but who they are. These are not people who get excited by every shiny new thing. Clearly something big is happening."[29] Tech mogul Elon Musk wrote that "ChatGPT is scary good. We are not far from dangerously strong AI".[30]

In December 2022 Google internally expressed alarm at the unexpected strength of ChatGPT and the newly discovered potential of large language models to disrupt the search engine business, and reassigned teams within multiple departments to aid in its artificial intelligence products.[31]

Negative reactions

In a December 2022 opinion piece, economist Paul Krugman wrote that ChatGPT would affect the demand for knowledge workers.[32] *The Verge's* James Vincent saw the viral success of ChatGPT as evidence that artificial intelligence had gone mainstream.[5] Journalists have commented on ChatGPT's tendency to "hallucinate".[33] 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.[34] When CNBC asked ChatGPT for the lyrics to "The Ballad of Dwight Fry", ChatGPT supplied invented lyrics rather than the actual lyrics.[20] Researchers cited by *The Verge* compared ChatGPT to a "stochastic parrot",[35] as did Professor Anton Van Den Hengel of the Australian Institute for Machine Learning.[36]

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.[2]

Economist Tyler Cowen expressed concerns regarding its effects on democracy, citing the ability of one to write automated comments to affect the decision process of new regulations.[37] *The Guardian* questioned whether any content found on the Internet after ChatGPT's release "can be truly trusted" and called for government regulation.[38]

Ax Sharma of *Bleeping Computer* noted that ChatGPT was capable of writing malware and phishing emails.[39] The CEO of ChatGPT creator OpenAI, Sam Altman, wrote that advancing software could pose "(for example) a huge cybersecurity risk" and also continued to predict "we could get to real AGI in the next decade, so we have to take the risk of that extremely seriously".[8]

Implications for education

In *The Atlantic* magazine, Stephen Marche noted that its effect on academia and especially application essays is yet to be understood.[40] California high school teacher and author Daniel Herman wrote that ChatGPT would usher in "The End of High School English".[41]

In the *Nature* journal, Chris Stokel-Walker pointed out that teachers should be concerned about students using ChatGPT to outsource their writing but that education providers will adapt to enhance critical thinking or reasoning.[42]

Emma Bowman with NPR wrote of the danger of students plagiarizing through an AI tool that may output biased or nonsensical text with an authoritative tone: "There are still many cases where you ask it a question and it'll give you a very impressive-sounding answer that's just dead wrong."[43]

Joanna Stern with *The Wall Street Journal* described cheating in American high school English with the tool by submitting a generated essay.[44] Professor Darren Hick of Furman University described noticing ChatGPT's "style" in a paper submitted by a student. An online GPT detector claimed the paper was 99.9% likely to be computer-generated, but Hick had no hard proof. However, the student in question confessed to using GPT when confronted, and as a consequence failed the course.[45]

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.[46] Jailbreaks created the potential for users to prompt ChatGPT to provide outputs that may be deemed offensive, inappropriate, or risking social harm by others.[47] 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. Roose, Kevin (December 5, 2022). "The Brilliance and Weirdness of ChatGPT" (https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html) (HTML). New York Times. Retrieved December 26, 2022. "Like those tools, ChatGPT — which stands for "generative pre-trained transformer" — landed with a splash."
2. 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) (HTML). *The Verge*. Retrieved December 5, 2022.
3. 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.
4. OpenAI (November 30, 2022). "ChatGPT: Optimizing Language Models for Dialogue" (https://openai.com/blog/chatgpt/) (HTML). Retrieved December 5, 2022.
5. 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) (HTML). *The Verge*. Retrieved December 8, 2022.
6. 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)].
7. 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.
8. "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.
9. "ChatGPT Feedback Contest: Official Rules" (https://cdn.openai.com/chatgpt/ChatGPT_Feedback_Contest_Rules.pdf) (PDF). *OpenAI*. Retrieved December 30, 2022.

10. "Help OpenAI improve ChatGPT and get a chance to win $500" (https://medium.com/geekculture/help-openai-improve-chatgpt-and-get-a-chance-to-win-500-f323777ed207). *Geek Culture*. Medium (blogging website). December 15, 2022. Retrieved December 30, 2022.
11. "What is ChatGPT? History, Features, Uses, Benefits, Drawbacks 2023 - Updated Geek by Raveen Chawla" (https://updatedgeek.com/what-is-chatgpt/). December 26, 2022. Retrieved December 27, 2022.
12. 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.
13. 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.
14. "New and Improved Content Moderation Tooling" (https://openai.com/blog/new-and-improved-content-moderation-tooling/). *OpenAI*. August 10, 2022. Retrieved December 30, 2022.
15. Markov, Todor; Zhang, Chong; Agarwal, Sandhini; Eloundou, Tyna; Lee, Teddy; Adler, Steven; Jiang, Angela; Weng, Lilian (August 5, 2022). "A Holistic Approach to Undesired Content Detection in the Real World" (https://arxiv.org/abs/2208.03274). *arXiv:2208.03274 [cs]*. Retrieved December 30, 2022.
16. 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)].
17. "Could a chatbot answer Prime Minister's Questions?" (https://www.bbc.com/news/uk-politics-64053550). *BBC News*. December 27, 2022. Retrieved December 30, 2022.
18. Perrigo, Billy (December 5, 2022). "AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits" (https://time.com/6238781/chatbot-chatgpt-ai-interview/). *Time (magazine)*. Retrieved December 26, 2022.
19. Biddle, Sam (December 8, 2022). "The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques" (https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/). *The Intercept*. Retrieved December 26, 2022.
20. 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.
21. Walsh, Toby (2022). "Everyone's having a field day with ChatGPT – but nobody knows how it actually works" (https://theconversation.com/everyones-having-a-field-day-with-chatgpt-but-nobody-knows-how-it-actually-works-196378). *The Conversation*. Retrieved December 30, 2022.
22. Kovanovic, Vitomir (2022). "The dawn of AI has come, and its implications for education couldn't be more significant" (https://theconversation.com/the-dawn-of-ai-has-come-and-its-implications-for-education-couldnt-be-more-significant-196383). *The Conversation*. Retrieved December 30, 2022.
23. Wiggers, Kyle (December 10, 2022). "OpenAI's attempts to watermark AI text hit limits" (https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/). *TechCrunch*. Retrieved December 30, 2022.
24. 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 18, 2022.
25. 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.
26. 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.
27. 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.
28. Thompson, Derek (December 8, 2022). "Breakthroughs of the Year" (https://www.theatlantic.com/newsletters/archive/2022/12/technology-medicine-law-ai-10-breakthroughs-2022/672390/). *The Atlantic*. Retrieved December 18, 2022.
29. Scharth, Marcel. "The ChatGPT chatbot is blowing people away with its writing skills. An expert explains why it's so impressive" (https://theconversation.com/the-chatgpt-chatbot-is-blowing-people-away-with-its-writing-skills-an-expert-explains-why-its-so-impressive-195908). *The Conversation*. Retrieved December 30, 2022.
30. 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.
31. Grant, Nico; Metz, Cade (December 21, 2022). "A New Chat Bot Is a 'Code Red' for Google's Search Business" (https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html). *The New York Times*. Retrieved December 30, 2022.
32. 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) (HTML). *The New York Times*. Retrieved December 6, 2022.

33. "ChatGPT a 'landmark event' for AI, but what does it mean for the future of human labour and disinformation?" (https://www.cbc.ca/radio/thecurrent/chatgpt-human-labour-and-fake-news-1.6686210). *CBC*. 2022. Retrieved December 18, 2022.
34. 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.
35. 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.
36. Mannix, Liam (December 13, 2022). "Is AI coming of age - or starting to reach its limits?" (https://www.smh.com.au/national/is-ai-coming-of-age-or-starting-to-reach-its-limits-20221213-p5c5uy.html). *The Sydney Morning Herald*. Retrieved December 18, 2022.
37. 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.
38. "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.
39. 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.
40. 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.
41. 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.
42. Stokel-Walker, Chris (December 9, 2022). "AI bot ChatGPT writes smart essays — should professors worry?" (https://www.nature.com/articles/d41586-022-04397-7). *Nature*. Retrieved December 19, 2022.
43. Bowman, Emma (December 19, 2022). "A new AI chatbot might do your homework for you. But it's still not an A+ student" (https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia). *NPR*. Retrieved December 19, 2022.
44. Stern, Joanna (December 21, 2022). "ChatGPT Wrote My AP English Essay—and I Passed" (https://www.wsj.com/articles/chatgpt-wrote-my-ap-english-essayand-i-passed-11671628256). *The_Wall_Street_Journal*. Retrieved December 21, 2022.
45. "Students using ChatGPT to cheat, professor warns" (https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns/). *The New York Post*. December 26, 2022. Retrieved December 30, 2022.
46. Zvi (December 2, 2022). "Jailbreaking ChatGPT on Release Day" (https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day).
47. 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)
- White paper (https://arxiv.org/abs/2203.02155) for InstructGPT, ChatGPT's predecessor