WIKIPEDIA
The Free Encyclopedia

ChatGPT

This is an old revision of this page, as edited by Artem.G (talk | contribs) at 20:28, 9 December 2022 *(Spam, Undid revision 1126521538 by Maria hameed0 (talk))*. **The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.**

ChatGPT is a prototype artificial intelligence chatbot developed by OpenAI that specializes in dialogue. The chatbot is a large language model fine-tuned with both supervised and reinforcement learning techniques. The base model that was fine-tuned was OpenAI's GPT-3 language model.

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

ChatGPT	
Original author(s)	OpenAI
Initial release	November 30, 2022
Type	Artificial intelligence chatbot
License	Proprietary
Website	openai.com /blog/chatgpt/ (https://openai. com/blog/chatg pt/)

Features

ChatGPT was fine-tuned on top of GPT-3 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 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 on 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 its knowledge of Columbus' voyages and its understanding of 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 knowledge of 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, potentially allowing for ChatGPT to be used as a personalized therapist.[7] In an effort 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] 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]

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.[10] 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".[11] 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.[12] In an opinion piece, economist Paul Krugman wrote that ChatGPT would affect the demand of knowledge workers.[13] Writing for *The Verge*, 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.[14]

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, despite Mexico not being a part of Central America. ChatGPT responded with Guatemala, when the answer is instead Nicaragua. However, when asked what the largest country in Central America is, ChatGPT correctly responded with Nicaragua.[15] 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.[16] 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.[17] Ax Sharma of *Bleeping Computer* noted that ChatGPT was capable of writing malware and phishing emails.[18]

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. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.

17. 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.
18. 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.

External links

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

Retrieved from "https://en.wikipedia.org/w/index.php?title=ChatGPT&oldid=1126522739"