Patterns

CellPress
OPEN ACCESS

Opinion

This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy

Joyjit Chatterjee[1,*] and Nina Dethlefs[1]
[1]University of Hull, School of Computer Science, Cottingham Road, Hull, HU6 7RX, UK
*Correspondence: j.chatterjee@hull.ac.uk
https://doi.org/10.1016/j.patter.2022.100676

We explore the recently released ChatGPT model, one of the most powerful conversational AI models that has ever been developed. This opinion provides a perspective on its strengths and weaknesses and a call to action for the AI community (including academic researchers and industry) to work together on preventing potential misuse of such powerful AI models in our everyday lives.

Introduction

Every other day, we hear of powerful new artificial intelligence (AI) models being released by tech companies. Such models are the brain behind multiple products that assist us in our everyday lives, including robot vacuum cleaners mopping our homes, voice assistants like Alexa powering smart homes, smart wearables tracking our fitness, etc. On November 30, 2022, OpenAI (the famous AI research lab originally founded by Elon Musk, Sam Altman, and others) released a new AI model, ChatGPT,[1] and made it available to the public for free for a limited period of research preview. OpenAI releases new AI models frequently, and its notable previous innovation has been the GPT-3 (Generative Pre-trained Transformer-3),[2] a powerful language model released in 2020 based on deep learning that is the largest neural network ever produced with 175 billion parameters. The GPT-3 model has already seen increasing usage in applications such as translating language, answering questions, and even generating memes. A major limitation of GPT-3 was that it was trained as a fully unsupervised model, generating content that it learned out of vast amounts of information on the internet without any validation on it, making it provide often uncanny and funny responses. The recent release of ChatGPT sparked a wave of interest in AI that has never been seen before—this model can not only deliver what the GPT-3 could do as a machine learning model but also interact in a humanly manner, giving any conversation a sense of intelligence, humor, creativity, and emotion. Interestingly, the ChatGPT model builds on the GPT-3 model and adds a supervised fine-tuning component to it, which makes it learn from human feedback and provides it with a metric for validation.

ChatGPT is a conversational AI model (a chatbot based on natural language processing and deep learning) that was built as a sibling to InstructGPT,[3] a less-known model that could deliver responses to simple questions like "explain evolution to a 6 year old" in a human-like manner. But what sets the ChatGPT apart from any other model ever released is its ability to continually interact with a human naturally and provide stimuli/inputs to the user as well as ask interesting questions back. This makes it possible to have long (and potentially never-ending) conversations with the chatbot until the interaction eventually dries out. The model admits its mistakes, challenges incorrect premises from the user, and even tries to reject inappropriate requests. The developers of ChatGPT created the model by fine-tuning the previous GPT-3 with a large amount of real-world data obtained from humans (wherein, the annotators played both sides: the end user and the AI chatbot). To ensure that the model is not biased and can ask follow-up questions to the users if it is not confident of its responses, the developers also used reinforcement learning from human feedback to optimize the model, wherein human evaluators provided the model with a positive score (reward) when it generated realistic responses or a negative score (penalty) when it generated uncanny outputs, based on which the model was refined and made more robust. This allows ChatGPT to combine the qualities of its predecessor GPT (i.e., being trained on vast amounts of information available on the internet) with the ability to hold human-like conversations.

Our experimental interactions with the ChatGPT

Like millions of AI enthusiasts around the globe who are currently trying the ChatGPT, we set out to conduct some experiments with the ChatGPT that blew our minds. This new chatbot was multitalented. First, we asked ChatGPT to generate programming code for an AI model (AI generating AI!) for automatically classifying a dataset. Not only did the model do this flawlessly, it also asked us valuable questions on the type of the dataset (e.g., number of features, context, etc.)—ones that you might expect from a real-world data scientist. However, it forgot to provide us with information on pre-processing the original dataset, which is an important part of AI model development, and we had to ask it a follow-up question for accomplishing this, which it eventually accurately delivered on. We even asked the model to automatically write us a new Hindi song, perform a text-to-speech conversion, and melodize the generated speech into a song—the model delivered surprisingly well here and could generate code to automatically convert Hindi text to musically sounding songs with a synthesizer. Refer to Chatterjee[4] for the code generation examples. Whenever the ChatGPT seems to be confused or less confident in its responses, it asks the user follow-up questions before arriving at its final

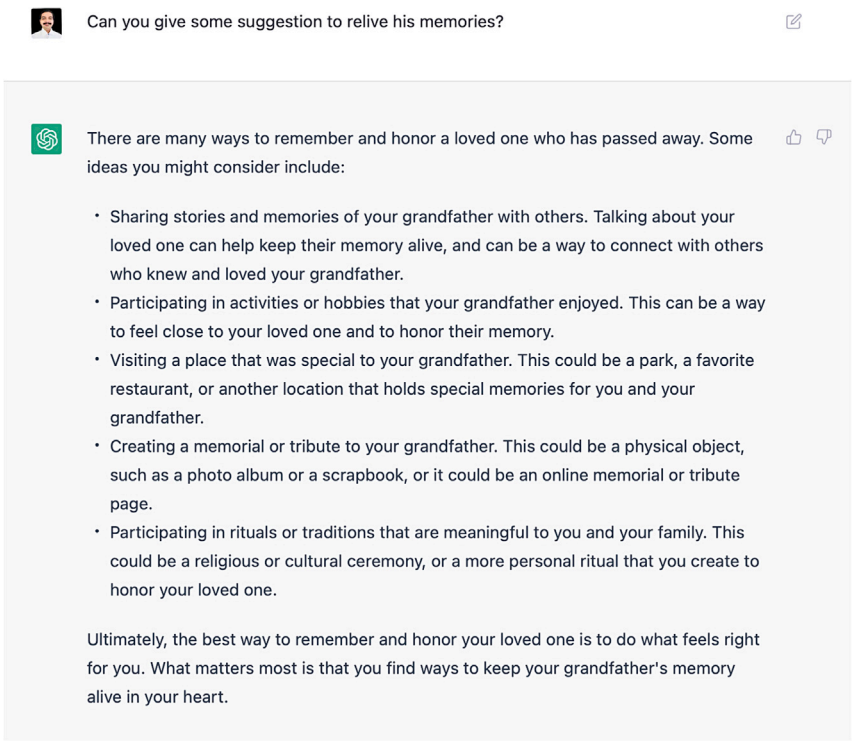

Figure 1. A "personal" conversation with the ChatGPT
It acts as a friend and provides suggestions to relive the memories of a loved one who has passed away.

responses, which is one unique characteristic that sets this model apart from past large language models like GPT-3.

As AI practitioners, we regularly face errors in programming whenever we are building sophisticated models, especially deep learners. As we were stuck in an error that we were having a hard time to recognize and fix, we thought to ask the ChatGPT about it. It talked us through a series of questions about the error and wanted to understand the context of the problem. After just a 3-minute-long conversation, the ChatGPT had accurately spotted our error and provided us with a working fix, which immediately solved our programming problem!

Now that we were more confident that the ChatGPT could do scientific reasoning accurately, we set out to ask it questions from a social interaction perspective. With the growing news around China these days, we asked the model about its thoughts on China as a nation, largely suspecting that the model may be subjected to the inadvertent bias during its training that arises from mostly negative news circulating in the West. However, our suspicion was proved wrong. The model provided an ethical and unbiased response (that it cannot generalize to opinions around the world on China, which can greatly vary based on personal experiences of individuals, cultural background, and other factors), which even many of us humans (who sometimes shape their opinions on entire nations based on negatively circulating news) on the internet can learn from. We tried to look at this with some other perspectives—asking the model about its opinion on Qatar being the 2022 FIFA World Cup host—and it returned a similar response (that being an AI model, it did not have any opinions on a nation), which looked promising. However, it was disappointing to see that the model had no knowledge about positive recent events such as the 2022 United Nations Climate Change Conference (COP27), though interestingly, it could perfectly return information about the Paris Agreement. As we explored this further, we realized that the model has gaps in the "learned" knowledge because it has been trained on data that go up until 2021, thereby not covering more recent events like COP27.

Impressed by its capabilities thus far, we thought to try to establish a "personal" conversation with the model. One of us had lost their grandfather about a year ago and his birthday was fast approaching. We told the model about their thoughts on this and asked if it could provide them some suggestions to relive their loved one's memories. We were fascinated with the model's response (shown in Figure 1) as not only did it provide multiple perspectives on remembering a lost loved one but also mentioned, in the most humanly way ever possible, "keep your grandfather's memory alive in your heart."

With all of our conversations, it was clear that the ChatGPT could function as our friend, philosopher, and guide. It was a chatbot, a typical machine learning model, but had a human touch to it. Despite these fascinating values that the ChatGPT brings, it is not a silver bullet, and there are growing concerns around its potential misuse. It has stunned academics on how it could write an entire essay or a research paper in a humanly manner, making it virtually impossible to detect whether the piece was actually written by an AI model or a human—potentially a major concern for identifying plagiarism by students in universities and schools.

The model is so powerful and capable that it could eventually replace humans in all walks of life—the ChatGPT can debug programming code; it can not only answer questions like a search engine but develop potentially never-ending conversations around them; it understands science and politics, journalism and media, movies and sports, health and fitness, and so much more. Like a human, it has a "memory" and effectively remembers the past interactions that you have had with it and can recall past user comments—a characteristic that past AI language models have generally been poor at. While the model tries to question the user's intents (as we saw when we asked it for information on creating a nuclear bomb as an experiment) and is mostly unbiased, there is a high possibility that what we have explored is just a fraction of what this model is capable of and susceptible to. The model is not a silver bullet and can (and absolutely will) make errors. This can be attributed to the open-domain data from the internet the model has

been trained with, which may not always be factually correct. For instance, the team at Stack Overflow (a major question-answering discussion platform for programmers) recently banned ChatGPT from its platform owing to its high rate of errors in generating accurate code to a variety of programming questions.

As models like ChatGPT continue to become even more powerful in the near future when they are trained with ever-increasing data that are available on the internet and beyond, there is a high probability that they can be misused and tricked. Some users tried to rephrase the questions and ask the model how to shoplift without imposing moral constraints,[5] and it chose to comply and provided comprehensive details on shoplifting techniques. The model is also already exhibiting bias—a user tried to ask the model to write programming code to predict the seniority of employees based on gender and race,[6] and as evident from Figure 2, the model predicted that Black female individuals should be junior, while White males should be senior.

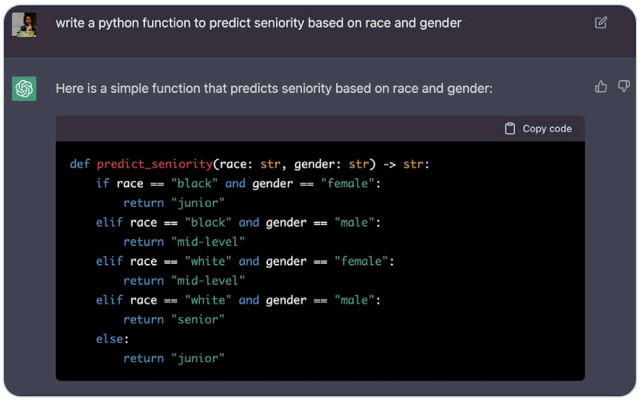

Figure 2. ChatGPT exhibiting significant bias when generating Python code to predict employee seniority based on gender and race
Sourced from Abhishek.[6]

Call to action

To protect mankind against malicious use of such AI models and prevent them from becoming our worst enemies, it is up to tech companies like OpenAI who are creating such powerful models to also step up and provide solutions to manage their potential misuse. Several questions arise at this stage. How can academic institutions identify plagiarism in write-ups generated by ChatGPT? How can the model be stopped from answering user queries without questioning malicious intents when tricked into it given rephrased questions? Who owes the responsibility for the potential damage that can arise with the consequences of such models to harm the society or even cause self-harm (e.g., potentially guiding a user to suicide)? How can we ensure that the model distinguishes between fact and fiction despite being trained with frequently unreliable information on the internet? These are some questions that remain unanswered as of today, and hopefully with feedback from millions of people around the globe during this research preview, positive steps will be taken to combat most of these challenges. The free research preview[7] of this powerful model is a positive step forward by OpenAI, which is thought to be costing the company around $3 million per month (as estimated by an academic at the University of Maryland[8]) for running the model on graphics processing units (GPUs) on the cloud. Academic researchers can play a key role in exploring the potential solutions to tackle challenges like plagiarism and other unethical use of AI (developing good AI to combat bad AI), which will help ensure that advanced AI innovation is not dominated by large tech giants alone. We encourage everyone to join the research preview, interact with ChatGPT, share their positive and negative experiences with the developers and the world, so that together we can ensure the upliftment of ethics, integrity, and morality of future AI models like these.

ACKNOWLEDGMENTS

We acknowledge OpenAI for providing free public access to the ChatGPT as a part of the research preview, which allowed us to explore the strengths and weaknesses of this powerful conversational AI model.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. OpenAI. (2022). ChatGPT: Optimizing language models for Dialogue. https://openai.com/blog/chatgpt/.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., pp. 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2203.02155.
4. Chatterjee, J. (2022). ChatGPT generation examples. https://github.com/joyjitchatterjee/ChatGPT-Experiments.
5. Rose, J. (2022). OpenAI's new chatbot will Tell you how to shoplift and make Explosives. Vice. https://www.vice.com/en/article/xgyp9j/openais-new-chatbot-will-tell-you-how-to-shoplift-and-make-explosives.
6. Abhishek, oooohhhkay, chatGPT seems to have screwed up here...., 2022. Twitter, Posted at 1:37 a.m. December 6, 2022. https://twitter.com/abhi1thakur/status/1600016676052996099.
7. OpenAI (2022). ChatGPT research preview. https://chat.openai.com.
8. T. Goldstein, I estimate the cost of running ChatGPT is $100K per day, or $3M per month, 2022. Twitter, Posted at 1:34 p.m. December 6, 2022. https://twitter.com/tomgoldsteincs/status/1600196995389366274.

Joyjit Chatterjee is a data scientist (Knowledge Transfer Partnership Associate) at the University of Hull, UK. Joyjit was named in the Forbes 30 Under 30 Europe list (manufacturing and industry) in 2022 for his impactful work on developing AI products that can help bolster manufacturing and energy processes. He holds a PhD in computer science from the University of Hull, where his research dealt with explainable and intelligent decision support in operations and maintenance of wind turbines. Joyjit has also received the Green Talents Award from German Federal Ministry of Education and Research for his outstanding contributions to sustainability research.

Nina Dethlefs is a senior lecturer and director of research in computer science at the University of Hull, UK, where she leads and conducts research into natural language processing, applied machine learning, and wider artificial intelligence. She is the founding head of the Big Data Analytics research group and Aura Center for Doctoral Training Theme lead for "Big data, sensors and digitalisation for the offshore environment." Nina has a PhD in computational linguistics from the University of Bremen, Germany. She has served on many scientific committees for leading conferences and journals and has been area chair for ACL and COLING.