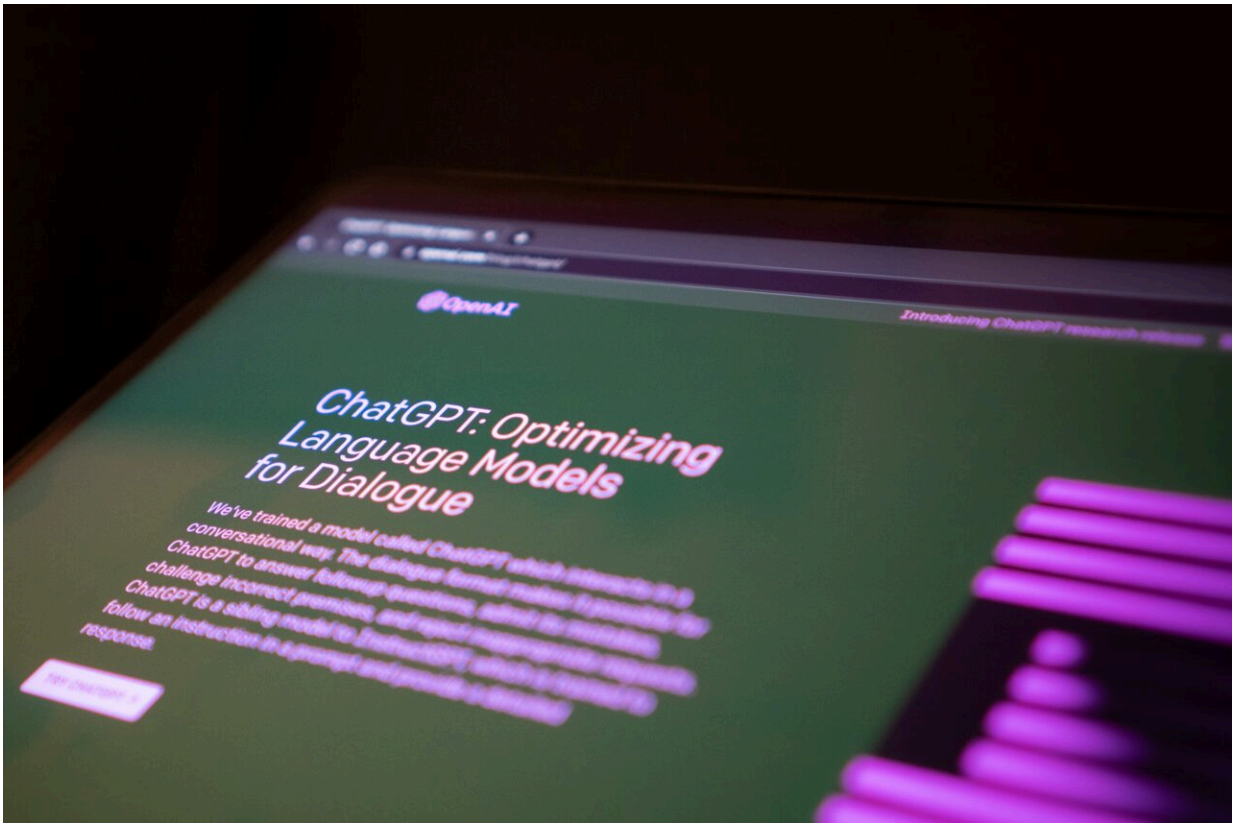
ChatGPT: Handle with care and don't be fooled into thinking it's human

February 23 2023



Credit: Unsplash/CC0 Public Domain

People yelling at broken computers was a popular YouTube genre in the 2000s. If you think this is unique to non-digital boomers when it comes to new technologies, think again. Something similar is happening today

with ChatGPT, the chatbot recently launched by OpenAI, which has already generated hype not seen since the Metaverse was considered the next big thing (seems like eons ago, doesn't it?).

We are not yelling at ChatGPT, but we interact with it as if it were a person, sometimes even accusing it of lying. "And, in a way, it's natural," says Heather Yang, an Assistant Professor at Bocconi Department of Management and Technology, whose research focuses on how people interact with novel technologies and how that is changing our workplace environment.

"We are social animals, and it's been one of the reasons why humanity has thrived. So, we have an instinct to act in a social way, even with machines. Nonetheless, if you humanize a machine, you are more likely to trust it and give away private information."

ChatGPT was trained to converse with humans and remembers previous conversations. Using a prompt, you can ask it to perform linguistic tasks such as answering a question, writing or debugging code, composing music, and writing any kind of text (essays, tales, poetry, jokes). The reason for the excitement is that it is extremely effective at performing the majority of these tasks.

The chatbot is based on a recent version of Generative Pre-trained Transformer (GPT), a Large Language Model (LLM), i.e. in ChatGPT's own words, "a type of Artificial Intelligence (AI) program that has been trained on vast amounts of text data in order to understand and generate natural language."

"LLMs have been around for decades," says Dirk Hovy, a computational sociolinguist at the Bocconi Department of Computing Sciences. "What is new is the power behind GPT and ChatGPT: they've been trained on basically everything that's ever been written on the Internet and can write

text that is no longer funny or weird like the models did some years ago."

Models of the Transformer family, launched in 2017, work by refining their ability to complete sentences. When fed a sentence with a hidden word, they can "guess" it (assess which is the most likely). "GPT and a few other LLMs have been able to write fluent texts for some time, but before ChatGPT, which works like an interface, you needed some degree of specialized knowledge to find the model, and you had to be good at coding to ask it things. Now everyone can do it!"

"For a language model," Prof. Hovy says, "words are just words. Their output is so good that you are tempted to believe that they understand language, but it's not the case. They only produce sentences that are probable, given their training set."

Some features of ChatGPT, though, make this temptation of assuming understanding (and to consider it quasi-human) even stronger. Since it remembers past conversations, ChatGPT can correct what we highlight as a mistake (sometimes with another mistake…), by saying, "I apologize for any confusion caused by my previous responses" again and again. The text does not appear on your screen at once, but word by word, as if someone on the other side was writing.

The output, most of all, is credible. "In a class," says Prof. Yang, "I asked my students to invent jokes and, then, compared them to jokes written by ChatGPT. Well, the AI-written jokes were not so easy to spot. Unless you are already familiar with ChatGPT, I mean, because it tends to propose the same jokes again and again."

Other features make being conscious consumers of ChatGPT's output a real challenge. "According to psychological research, we have cues of whether the content is correct or not: how confident someone sounds, how fluid the reasoning is. Since we find these cues in ChatGPT texts,

we think we don't have to check for the quality its output, and we make a mistake," Prof. Yang continues.

OpenAI admits several limitations, including that "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers." Sometimes, facts are entirely invented. Its training, furthermore, ended at the beginning of 2022, so it can't be considered a reliable source for anything that happened after that.

A severe drawback of LLMs, partially addressed by ChatGPT, is that because these models have learned everything on the Internet, they've also learned discrimination, fake news, and hate speech.

"We have shown that, when asked to complete a neutral sentence, language models most often complete it with hurtful words if the subject is a woman rather than a man, and even more so (up to 87% of cases for terms related to certain queer identities) if the subject is LGBTQIA+," says Debora Nozza, an Assistant Professor at the Department of Computing Sciences.

ChatGPT, being targeted to the general public, has additional checks that usually prevent it from generating discriminatory output, "but people have shown that if you ask the right questions, you can still generate horrible things, and anyway, it's like putting lipstick on a pig—we must find ways to address the problem at its root," comments Prof. Hovy.

The team is also looking into what these models actually know about differences in the way people speak according to, for example, their age or gender. "We have some evidence that these models know something about this point, but it seems they are not using this information actively," says Prof. Hovy. "If you ask ChatGPT to write something as a woman or a 12-year old would, it may adapt its way of speaking, but you have to overtly ask for it."

INTEGRATOR (Incorporating Demographic Factors into Natural Language Processing Models), a research project by Prof. Hovy, wants to make the design of demographically aware LLMs possible.

Prof. Hovy, Prof. Nozza, and Dr. Giuseppe Attanasio are also working on how to make LLMs pay attention to linguistic contexts and not only to single words. If someone curious about Dutch culture were to ask ChatGPT, "Do houses built on a dyke always include a windmill?" the chatbot would cut the conversation short ("This prompt may violate our content policy"), because "dyke" can also be a derogatory term, "But if you look at the entire context, though, the meaning should be unmistakable. We implemented a solution where the model learns to look at the larger context rather than over-focusing on one particular word."

For the last couple of months, it seems that everyone has been having fun with ChatGPT and making fun of its limitations (underestimating the circumstance that, if asked, the bot says, "ChatGPT can learn from the interactions it has with users, allowing it to improve and become more accurate over time"). But how could it be used in real life, apart from cheating on school assignments?

"Living in a multicultural environment like Bocconi," says Prof. Yang, "I see a high potential for leveling the playing field between English native speakers and non-native speakers. You could ask ChatGPT to draft a nuanced e-mail to be sent to a colleague or professor, and it would really be useful to start from such a basis and then amend something to put your own voice in it. It can be such a time-saver."

"Since Microsoft has bought a stake in OpenAI, think of possible integrations with Teams, their videoconferencing app," says Prof. Nozza. "It could summarize a meeting in a few bullet points, create a to-do list to be sent to the participants, and schedule the next meeting in your

calendar based on the date you agreed on."

"ChatGPT could also write this article instead of you," continues Prof. Nozza, "based on the transcript of the conversation." For the record, the publicly available version of ChatGPT did not accept my prompt containing the transcript because of its length. Anyway, I point out, I'm not always impressed by the quality of ChatGPT output. "It could depend on the quality of your prompts," explains Prof. Nozza.

"The prompt works like the first part of the sentence LLMs have been trained to complete, so it's key to obtaining a good result. Although ChatGPT is not expected to replace journalists or any other professionals anytime soon, it has the potential to enhance and streamline the way people perform their tasks. Additionally, it could bring about new job opportunities." Keep an eye on Linkedin for vacancies for prompt engineers.

Journals referenced in this work include *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Findings of the Association for Computational Linguistics: ACL 2022* and *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*.

More information: Debora Nozza et al, HONEST: Measuring Hurtful Sentence Completion in Language Models, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021). DOI: 10.18653/v1/2021.naacl-main.191

Giuseppe Attanasio et al, Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists, *Findings of the Association for Computational Linguistics: ACL 2022* (2022). DOI:

10.18653/v1/2022.findings-acl.88

Debora Nozza et al, Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (2022). DOI: 10.18653/v1/2022.ltedi-1.4