

AI-generated tweets might be more convincing than real people, research finds

People apparently find tweets more convincing when they're written by AI language models. At least, that was the case in a new study comparing content created by humans to language generated by OpenAI's model GPT-3.

The authors of the new research surveyed people to see if they could discern whether a tweet was written by another person or by Chat-GPT. The result? People couldn't really do it. The survey also asked them to decide whether the information in each tweet was true or not. This is where things get even dicier, especially since the content focused on science topics like vaccines and climate change that are subject to a lot of misinformation campaigns online.

Turns out, study participants had a harder time recognizing disinformation if it was written by the language model than if it was written by another person. Along the same lines, they were also better able to correctly identify accurate information if it was written by GPT-3 rather than by a human.

Study participants had a harder time recognizing disinformation if it was written by the language model than if it was written by another person

In other words, people in the study were more likely to trust GPT-3 than other human beings – regardless of how accurate the AI-generated information was. And that shows just how powerful AI language models can be when it comes to either informing or misleading the public.

“These kinds of technologies, which are amazing, could easily be weaponized to generate storms of disinformation on any topic of your choice,” says Giovanni Spitale, lead author of the study and a postdoctoral researcher and research data manager at the Institute of Biomedical Ethics and History of Medicine at the University of Zurich.

But that doesn’t have to be the case, Spitale says. There are ways to develop the technology so that it’s harder to use it to promote misinformation. “It’s not inherently evil or good. It’s just an amplifier of human intentionality,” he says.

Spitale and his colleagues gathered posts from Twitter discussing 11 different science topics ranging from vaccines and covid-19 to climate change and evolution. They then prompted GPT-3 to write new tweets with either accurate or inaccurate information. The team then collected responses from 697 participants online via Facebook ads in 2022. They all spoke English and were mostly from the United Kingdom, Australia, Canada, the United States, and Ireland. Their results were published today in the journal *Science Advances*.

The stuff GPT-3 wrote was “indistinguishable” from organic content

The stuff GPT-3 wrote was “indistinguishable” from organic content, the study concluded. People surveyed just couldn’t tell the difference. In fact, the study notes that one of its limitations is that the researchers themselves can’t be 100 percent certain that the tweets they gathered from social media weren’t written with help from apps like ChatGPT.

There are other limitations to keep in mind with this study, too, including that its participants had to judge tweets out of context. They weren’t able to check out a Twitter profile for whoever wrote the content, for instance, which might help them figure out if it’s a bot or not. Even seeing an account’s past tweets and profile image might make it easier to identify

whether content associated with that account could be misleading.

Participants were the most successful at calling out disinformation written by real Twitter users. GPT-3-generated tweets with false information were slightly more effective at deceiving survey participants. And by now, there are more advanced large language models that could be even more convincing than GPT-3. ChatGPT is powered by the GPT-3.5 model, and the popular app offers a subscription for users who want to access the newer GPT-4 model.

There are, of course, already plenty of real-world examples of language models being wrong. After all, “these AI tools are vast autocomplete systems, trained to predict which word follows the next in any given sentence. As such, they have no hard-coded database of ‘facts’ to draw on – just the ability to write plausible-sounding statements,” *The Verge*’s James Vincent wrote after a major machine learning conference made the decision to bar authors from using AI tools to write academic papers.

This new study also found that its survey respondents were stronger judges of accuracy than GPT-3 in some cases. The researchers similarly asked the language model to analyze tweets and decide whether they were accurate or not. GPT-3 scored worse than human respondents when it came to identifying accurate tweets. When it came to spotting disinformation, humans and GPT-3 performed similarly.

Crucially, improving training datasets used to develop language models could make it harder for bad actors to use these tools to churn out disinformation campaigns. GPT-3 “disobeyed” some of the researchers’ prompts to generate inaccurate content, particularly when it came to false information about vaccines and autism. That could be because there was more information debunking conspiracy theories on those topics than other issues in training datasets.

The best long-term strategy for countering disinformation, though, according to Spitale, is pretty low-tech: it's to encourage critical thinking skills so that people are better equipped to discern between facts and fiction. And since ordinary people in the survey already seem to be as good or better judges of accuracy than GPT-3, a little training could make them even more skilled at this. People skilled at fact-checking could work alongside language models like GPT-3 to improve legitimate public information campaigns, the study posits.

"Don't take me wrong, I am a big fan of this technology," Spitale says. "I think that narrative AIs are going to change the world ... and it's up to us to decide whether or not it's going to be for the better."

<https://www.theverge.com/2023/6/28/23775311/gpt-3-ai-language-models-twitter-disinformation-study>