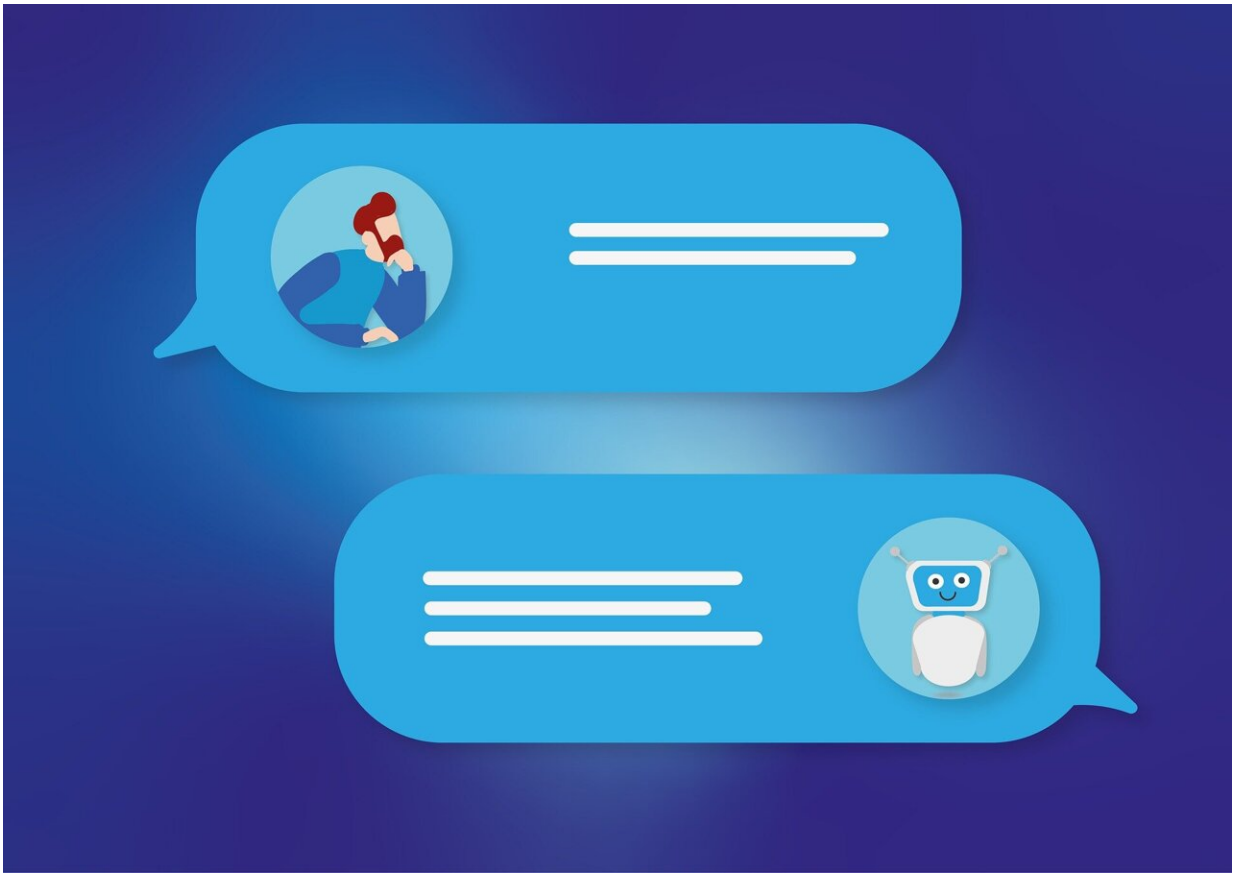
GPT-4 falls short of Turing threshold

November 2 2023, by Peter Grad



Credit: Pixabay/CC0 Public Domain

One question has relentlessly followed ChatGPT in its trajectory to superstar status in the field of artificial intelligence: Has it met the Turing test of generating output indistinguishable from human response?

Two researchers at the University of California at San Diego say it comes close, but not quite.

ChatGPT may be smart, quick and impressive. It does a good job at exhibiting apparent intelligence. It sounds humanlike in conversations with people and can even display humor, emulate the phraseology of teenagers, and pass exams for law school.

But on occasion, it has been found to serve up totally false information. It hallucinates. It does not reflect on its own output.

Cameron Jones, who specializes in language, semantics and machine learning, and Benjamin Bergen, professor of cognitive science, drew upon the work of Alan Turing, who 70 years ago devised a process to determine whether a machine could reach a point of intelligence and conversational prowess at which it could fool someone into thinking it was human.

Their report titled "Does GPT-4 Pass the Turing Test?" is available on the *arXiv* preprint server.

They rounded up 650 participants and generated 1,400 "games" in which brief conversations were conducted between participants and either another human or a GPT model. Participants were asked to determine who they were conversing with.

The researchers found that GPT-4 models fooled participants 41% of the time, while GPT-3.5 fooled them only 5% to 14% of the time. Interestingly, humans succeeded in convincing participants they were not machines in only 63% of the trials.

The researchers concluded, "We do not find evidence that GPT-4 passes the Turing Test."

They noted, however, that the Turing test still retains value as a measure of the effectiveness of machine dialogue.

"The test has ongoing relevance as a framework to measure fluent social interaction and deception, and for understanding human strategies to adapt to these devices," they said.

They warned that in many instances, chatbots can still communicate convincingly enough to fool users in many instances.

"A success rate of 41% suggests that deception by AI models may already be likely, especially in contexts where human interlocutors are less alert to the possibility they are not speaking to a human," they said. "AI models that can robustly impersonate people could have could have widespread social and economic consequences."

The researchers observed that participants making correct identifications focused on several factors.

Models that were too formal or too informal raised red flags for participants. If they were too wordy or too brief, if their grammar or use of punctuation was exceptionally good or "unconvincingly" bad, their usage became key factors in determining whether participants were dealing with humans or machines.

Test takers also were sensitive to generic-sounding responses.

"LLMs learn to produce highly likely completions and are fine-tuned to avoid controversial opinions. These processes might encourage generic responses that are typical overall, but lack the idiosyncrasy typical of an individual: a sort of ecological fallacy," the researchers said.

The researchers have suggested that it will be important to track AI

models as they gain more fluidity and absorb more humanlike quirks in conversation.

"It will become increasingly important to identify factors that lead to deception and strategies to mitigate it," they said.

More information: Cameron Jones et al, Does GPT-4 Pass the Turing Test?, *arXiv* (2023). DOI: 10.48550/arxiv.2310.20216

Citation: GPT-4 falls short of Turing threshold (2023, November 2) retrieved 9 May 2024 from https://techxplore.com/news/2023-11-gpt-falls-short-turing-threshold.html