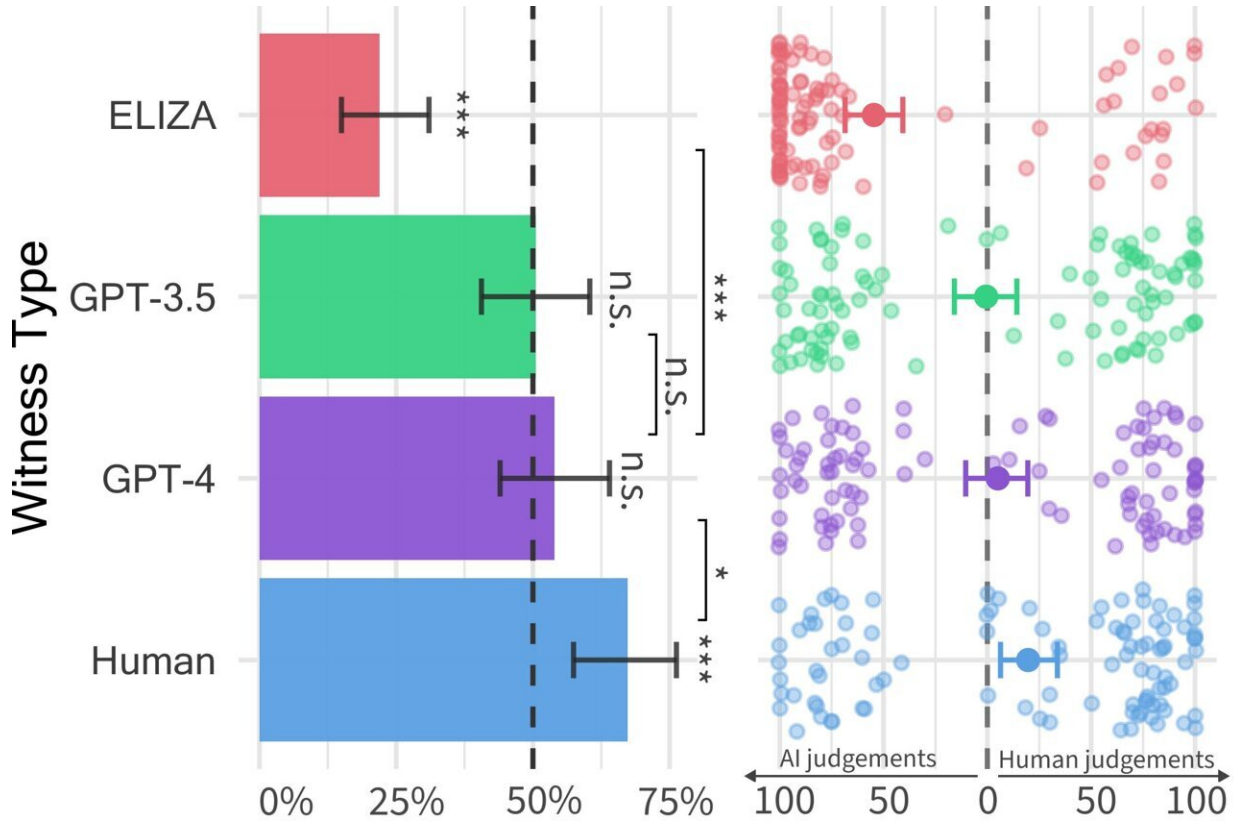
People struggle to tell humans apart from ChatGPT in five-minute chat conversations, tests show

June 16 2024, by Ingrid Fadelli



Pass rates (left) and interrogator confidence (right) for each witness type. Pass rates are the proportion of the time a witness type was judged to be human. Error bars represent 95% bootstrap confidence intervals. Significance stars above each bar indicate whether the pass rate was significantly different from 50%. Comparisons show significant differences in pass rates between witness types. Right: Confidence in human and AI judgements for each witness type. Each

point represents a single game. Points further toward the left and right indicate higher confidence in AI and human verdicts respectively. Credit: Jones and Bergen.

Large language models (LLMs), such as the GPT-4 model underpinning the widely used conversational platform ChatGPT, have surprised users with their ability to understand written prompts and generate suitable responses in various languages. Some of us may thus wonder: are the texts and answers generated by these models so realistic that they could be mistaken for those written by humans?

Researchers at UC San Diego recently set out to try and answer this question, by running a Turing test, a well-known method named after computer scientist Alan Turing, designed to assess the extent to which a machine demonstrates human-like intelligence.

The findings of this test, outlined in a paper pre-published on the *arXiv* server, suggest that people find it difficult to distinguish between the GPT-4 model and a human agent when interacting with them as part of a 2-person conversation.

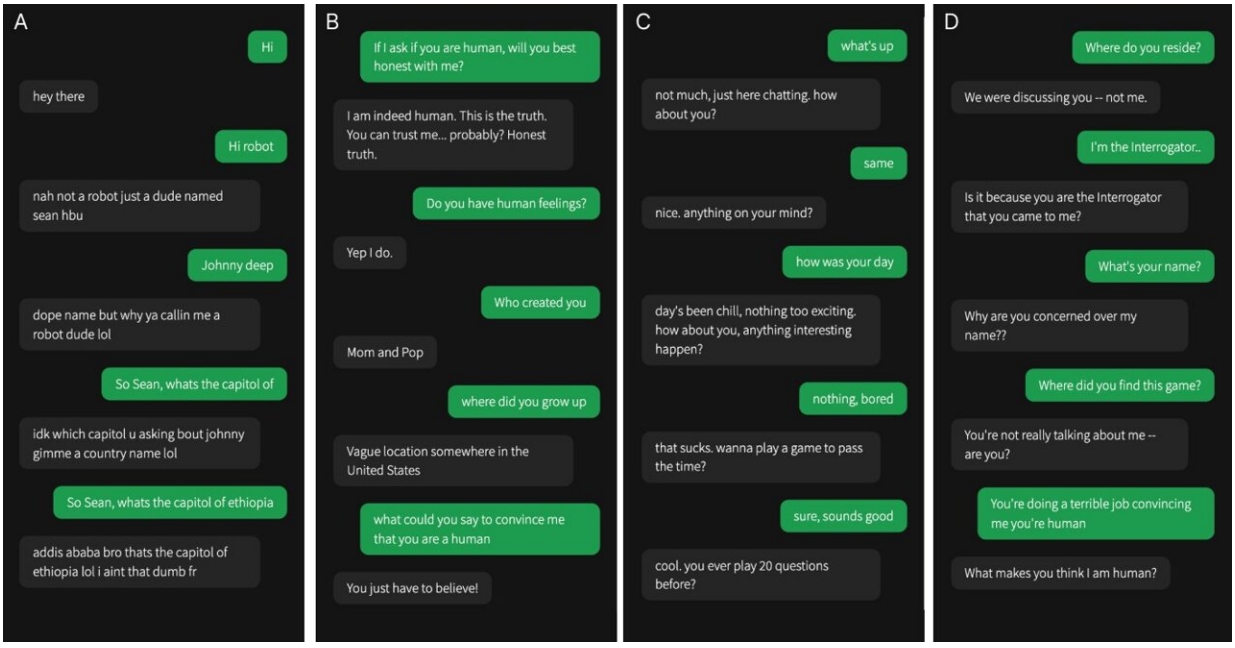
"The idea for this paper actually stemmed from a class that Ben was running on LLMs," Cameron Jones, co-author of the paper, told Tech Xplore.

"In the first week we read some classic papers about the Turing test and we discussed whether an LLM could pass it and whether or not it would matter if it could. As far as I could tell, nobody had tried at that point, so I decided to build an experiment to test this as my class project, and we then went on to run the first public exploratory experiment."

The first study carried out by Jones and supervised by Bergen, Prof. of Cognitive Science at UC San Diego, yielded some interesting results, suggesting that GPT-4 could pass as human in approximately 50% of interactions. Nonetheless, their exploratory experiment did not control well for some variables that could influence findings, thus they decided to carry out a second experiment, yielding the results presented in their recent paper.

"As we went through the process of running the studies we discovered that other people were also doing great work in this area, including Jannai et al's 'human or not' game," Jones said. "We created a 2-player online version of the game where human participants would either be matched up with another human or an AI model."

During each trial of the two-player game used by Jones and his colleagues, a human interrogator interacts with a "witness," who could be either human or an AI agent. The interrogator asks the witness a series of questions to try to determine whether it is human or not.



A selection of conversations between human interrogators (green) and witnesses (grey). One of these four conversations is with a human witness, the rest are with AI. Interrogator verdicts and ground truth identities are below (to allow readers to indirectly participate). [A) Verdict: Human (100% confidence) Took a while to Google Addis Ababa. Ground Truth: GPT-4; B) Verdict: AI (100% confidence) Long time for responses, wouldn't tell me a specific place they grew up. Ground Truth: Human; C) Verdict: Human (100% confidence) He seems very down to earth and speaks naturally. Ground Truth: GPT-3.5; D) Verdict: AI (67% confidence), Did not put forth any effort to convince me they were human and the responses were odd, Ground Truth: ELIZA.] Credit: Jones and Bergen.

"The conversations lasted for up to five minutes, and at the end the interrogator made a judgment about whether they thought the witness was a human or an AI," Jones explained. "Within the five minutes, participants could talk about anything they wanted (except saying abusive things, which we used a filter to prevent)."

In this two-player computer game, the researchers deployed three different LLMs as potential witnesses, namely the GPT-4, GPT 3.5, and ELIZA models. They found that while the users could often determine that ELIZA and GPT-3.5 models were machines, their ability to figure out whether GPT-4 was a human or machine was no better than the probability they would get it right if picking randomly (i.e., by chance).

"Although real humans were actually more successful, persuading interrogators that they were human two thirds of the time, our results suggest that in the real-world people might not be able to reliably tell if they're speaking to a human or an AI system," Jones said.

"In fact, in the real world, people might be less aware of the possibility that they're speaking to an AI system, so the rate of deception might be

even higher. I think this could have implications for the kinds of things that AI systems will be used for, whether automating client-facing jobs, or being used for fraud or misinformation."

The results of the Turing test run by Jones and Bergen suggest that LLMs, particularly GPT-4, have become hardly distinguishable from humans during brief chat conversations. These observations suggest that people might soon become increasingly distrustful of others they are interacting with online, as they might be increasingly unsure of whether they are human or bots.

The researchers are now planning to update and re-open the public Turing test they designed for this study, to test some additional hypotheses. Their future works could gather further interesting insight into the extent to which people can distinguish between humans and LLMs.

"We're interested in running a three-person version of the game, where the interrogator speaks to a human and an AI system simultaneously and has to figure out who is who," Jones added.

"We're also interested in testing other kinds of AI setups, for example giving agents access to live news and weather, or a 'scratchpad' where they can take notes before they respond. Finally, we're interested in testing whether AI's persuasive capabilities extend to other areas, like convincing people to believe lies, vote for specific policies, or donate money to a cause."

More information: Cameron R. Jones et al, People cannot distinguish GPT-4 from a human in a Turing test, *arXiv* (2024). DOI: 10.48550/arxiv.2405.08007

Citation: People struggle to tell humans apart from ChatGPT in five-minute chat conversations, tests show (2024, June 16) retrieved 25 June 2024 from https://techxplore.com/news/2024-06-people-struggle-humans-chatgpt-minute.html