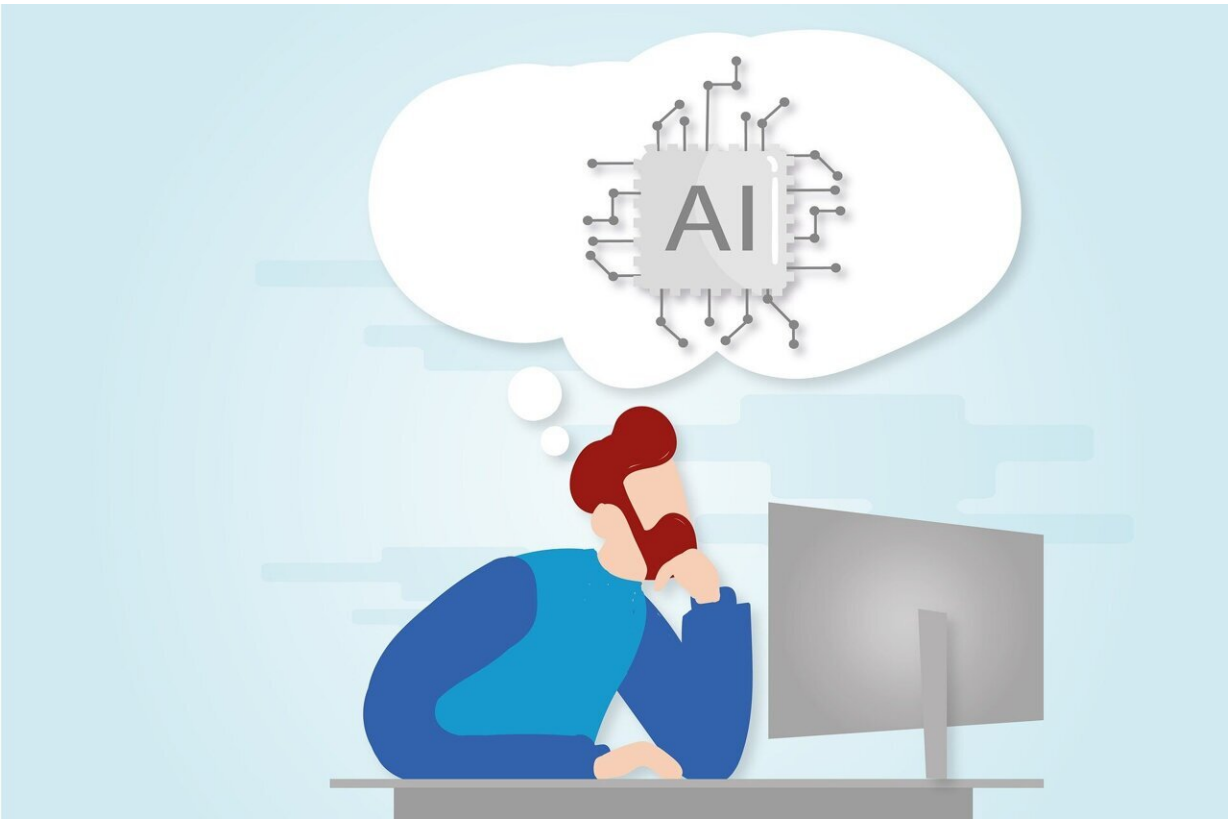
Study finds source validation issues hurt ChatGPT reliability

May 9 2023, by Peter Grad



Credit: Pixabay/CC0 Public Domain

ChatGPT seems to be everywhere. Morgan Stanley, Duolingo, Snapchat, Coca-Cola and Instacart have signed on, as have programmers, web designers, pharmaceutical companies, writers, musicians, translators and

businesses of all types.

Available for general use for merely half a year, ChatGPT has been used by more than 100 million users globally and it handles more than 10 million inquiries a day.

Practically daily, stories are written detailing its potential to upend everything from business models to personal routines. Everyone wants a piece of the action.

But some are raising red flags about the downsides of ChatGPT, which is the fastest growing app in history. Most notable was a letter in March, signed by more than 1,000 experts, urging a moratorium on the expansion of AI technology until its risks and dangers are better understood.

While skeptics have largely focused on the potential for bad actors to manipulate AI for harmful purposes or have imagined a possible scenario where AI models decide on their own to act maliciously, others are focusing on a more immediate issue: accuracy.

The Wall Street Journal recently cautioned, "AI chatbots and other generative AI programs are mirrors to the data they consume. They regurgitate and remix what they are fed to both great effect—and great failure."

Stanford University's Human-Centered AI research group published a paper on the *arXiv* preprint server last month that called into question the reliability of data retrieved in large language model retrievals.

"A prerequisite trait of a trustworthy generative search engine is verifiability," Nelson Liu, a Ph.D. student at Stanford and one of the report's authors, said. After studying output from four popular search

engines, Liu and his two colleagues Tianyi Zhang and Percy Jiang reported that the results were "fluent and appear informative, but frequently contain unsupported statements and inaccurate citations."

The generative search engines they studied were Bing Chat, NeevaAI, perplexity.ai and YouChat. Subjects ranged from biographical data about singer Alicia Keys to the issue of censorship on social media.

The researchers looked at four features: fluency, perceived utility (how helpful the answer was), citation recall (how consistently generated statements were fully supported by citations) and citation precision (the proportion of generated citations supporting associated statements).

A trustworthy generative search engine was defined as one that achieved high citation recall and precision. The results were disheartening.

The team found responses "often had high fluency and perceived utility, but frequently contained unsupported statements or inaccurate citations." Only about half of generated sentences were fully supported by citations, and one quarter of citations failed to support associated sentences.

Moreover, the team found citation recall and precision were inversely correlated with fluency and perceived utility. "The responses that seem more helpful are often those with more unsupported statements or inaccurate citations," they observed.

As a consequence, they concluded, "This facade of trustworthiness increases the potential for existing generative search engines to mislead users."

An article this week in ExtremeTech addressed the issue of sourcing: "Chatbots like ChatGPT and Bing Chat are disturbingly good at making fake information appear true. Without citations—which most chatbot

results lack—it's difficult to differentiate between accuracy and falsehood, particularly in the mere seconds users spend on a search engine's results page."

The Stanford researchers said results of their study "are concerningly low for systems that may serve as a primary tool for information-seeking users—especially given their facade of trustworthiness."

The researchers expressed hope that their research would "further motivate the development of trustworthy generative search engines and help researchers and users better understand the shortcomings of existing commercial systems."

More information: Nelson F. Liu et al, Evaluating Verifiability in Generative Search Engines, *arXiv* (2023). DOI: 10.48550/arxiv.2304.09848