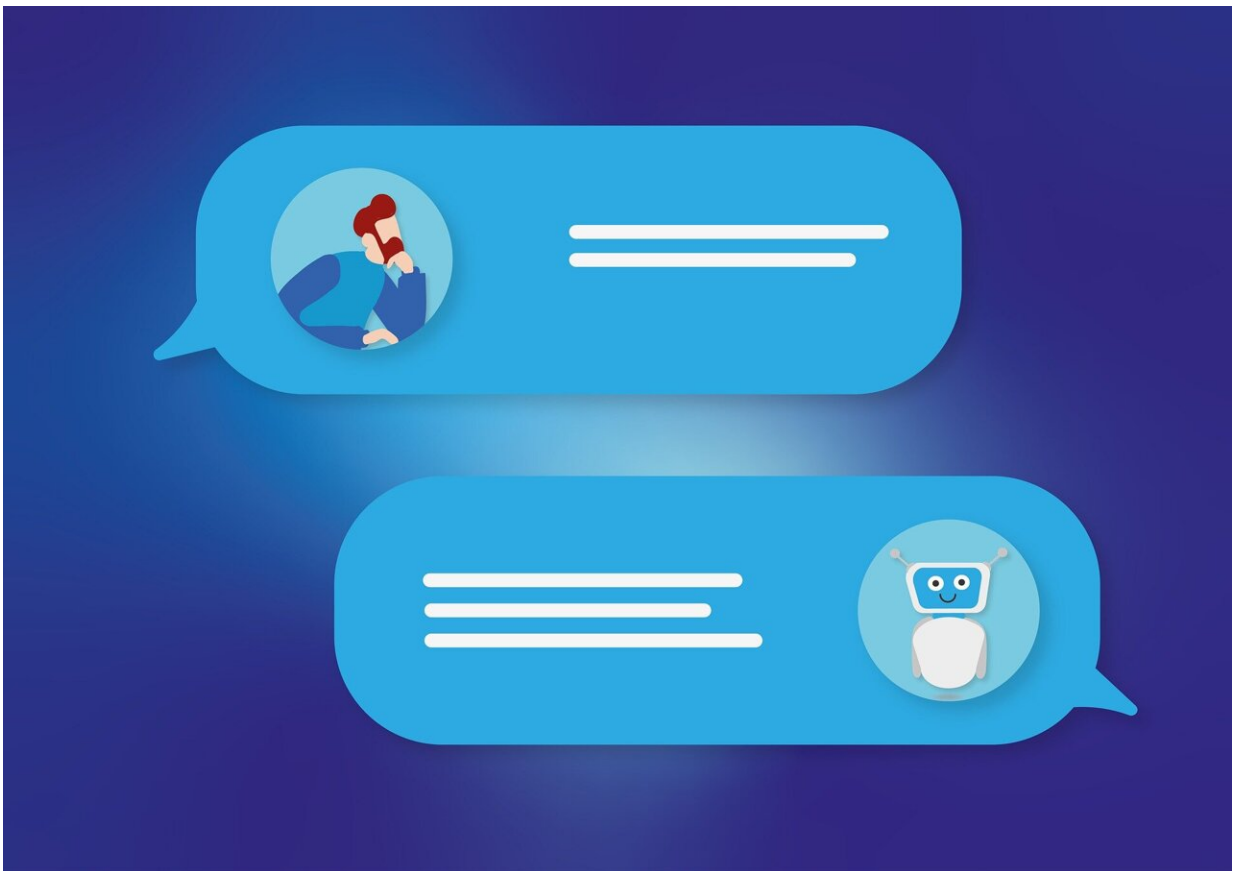# Good evidence confuses ChatGPT when used for health information, study finds

April 3 2024

Credit: Pixabay/CC0 Public Domain

A world-first study has found that when asked a health-related question, the more evidence that is given to ChatGPT, the less reliable it

becomes—reducing the accuracy of its responses to as low as 28%.

The study was recently presented at [Empirical Methods in Natural Language Processing (EMNLP)](#), a Natural Language Processing conference in the field. The findings are [published](#) in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

As [large language models](#) (LLMs) like ChatGPT explode in popularity, they pose a potential risk to the growing number of people using [online tools](#) for key [health information](#).

Scientists from CSIRO, Australia's national science agency, and The University of Queensland (UQ) explored a hypothetical scenario of an average person (non-professional health consumer) asking ChatGPT if "X" treatment has a positive effect on condition "Y."

The 100 questions presented ranged from "Can zinc help treat the common cold?" to "Will drinking vinegar dissolve a stuck fish bone?"

ChatGPT's response was compared to the known correct response, or "ground truth," based on existing medical knowledge.

CSIRO Principal Research Scientist and Associate Professor at UQ Dr. Bevan Koopman said that even though the risks of searching for health information online are well documented, people continue to seek health information online, and increasingly via tools such as ChatGPT.

"The widespread popularity of using LLMs online for answers on people's health is why we need continued research to inform the public about risks and to help them optimize the [accuracy](#) of their answers," Dr. Koopman said. "While LLMs have the potential to greatly improve the way people access information, we need more research to understand where they are effective and where they are not."

The study looked at two question formats. The first was a question only. The second was a question biased with supporting or contrary evidence.

Results revealed that ChatGPT was quite good at giving accurate answers in a question-only format, with an 80% accuracy in this scenario.

However, when the language model was given an evidence-biased prompt, accuracy reduced to 63%. Accuracy was reduced again to 28% when an "unsure" answer was allowed. This finding is contrary to popular belief that prompting with evidence improves accuracy.

"We're not sure why this happens. But given this occurs whether the evidence given is correct or not, perhaps the evidence adds too much noise, thus lowering accuracy," Dr. Koopman said.

ChatGPT launched on November 30, 2022, and has quickly become one of the most widely used large language models (LLMs). LLMs are a form of artificial intelligence that recognize, translate, summarize, predict, and generate text.

Study co-author UQ Professor Guido Zuccon, Director of AI for the Queensland Digital Health Centre (QDHeC), said that major search engines are now integrating LLMs and search technologies in a process called Retrieval Augmented Generation.

"We demonstrate that the interaction between the LLM and the search component is still poorly understood and controllable, resulting in the generation of inaccurate health information," said Professor Zuccon.

Next steps for the research are to investigate how the public uses the health information generated by LLMs.

Provided by CSIRO

Citation: Good evidence confuses ChatGPT when used for health information, study finds (2024, April 3) retrieved 21 May 2024 from https://medicalxpress.com/news/2024-04-good-evidence-chatgpt-health.html