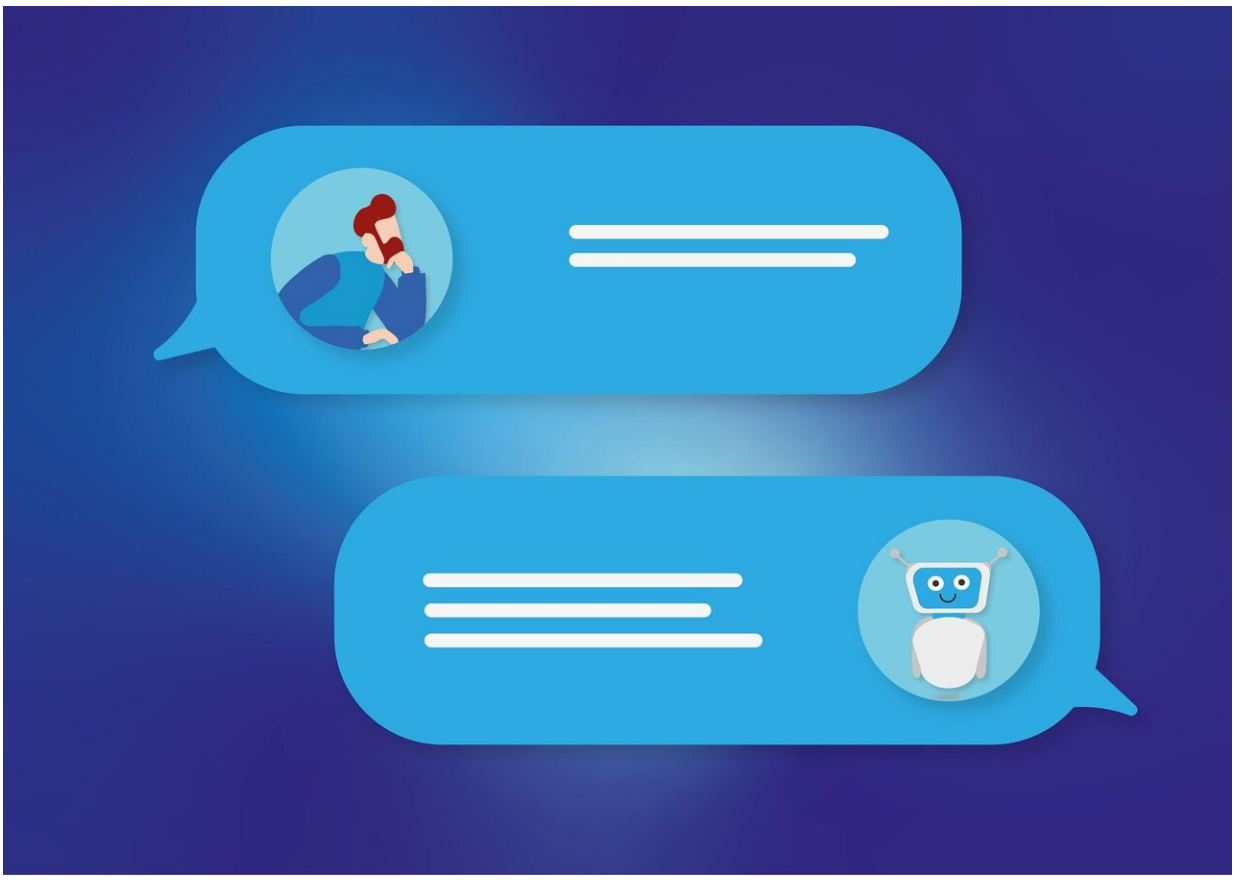
Cognitive psychology tests show AIs are irrational—just not in the same way that humans are

June 4 2024



Credit: Pixabay/CC0 Public Domain

Large language models behind popular generative AI platforms like

ChatGPT gave different answers when asked to respond to the same reasoning test and didn't improve when given additional context, finds a new study by researchers at University College London.

The study, published in *Royal Society Open Science*, tested the most advanced large language models (LLMs) using cognitive psychology tests to gauge their capacity for reasoning. The results highlight the importance of understanding how these AIs "think" before entrusting them with tasks, particularly those involving decision-making.

In recent years, the LLMs that power generative AI apps like ChatGPT have become increasingly sophisticated. Their ability to produce realistic text, images, audio and video has prompted concern about their capacity to steal jobs, influence elections and commit crime.

Yet these AIs have also been shown to routinely fabricate information, respond inconsistently and even to get simple math sums wrong.

In this study, researchers from UCL systematically analyzed whether seven LLMs were capable of rational reasoning. A common definition of a rational agent (human or artificial), which the authors adopted, is whether it reasons according to the rules of logic and probability. An irrational agent is one that does not reason according to these rules.

The LLMs were given a battery of 12 common tests from cognitive psychology to evaluate reasoning, including the Wason task, the Linda problem and the Monty Hall problem. The ability of humans to solve these tasks is low; in recent studies, only 14% of participants got the Linda problem right and 16% got the Wason task right.

The models exhibited irrationality in many of their answers, such as providing varying responses when asked the same question 10 times. They were prone to making simple mistakes, including basic addition

errors and mistaking consonants for vowels, which led them to provide incorrect answers.

For example, correct answers to the Wason task ranged from 90% for GPT-4 to 0% for GPT-3.5 and Google Bard. Llama 2 70b, which answered correctly 10% of the time, mistook the letter K for a vowel and so answered incorrectly.

While most humans would also fail to answer the Wason task correctly, it is unlikely that this would be because they didn't know what a vowel was.

Olivia Macmillan-Scott, first author of the study from UCL Computer Science, said, "Based on the results of our study and other research on large language models, it's safe to say that these models do not 'think' like humans yet. That said, the model with the largest dataset, GPT-4, performed a lot better than other models, suggesting that they are improving rapidly. However, it is difficult to say how this particular model reasons because it is a closed system. I suspect there are other tools in use that you wouldn't have found in its predecessor GPT-3.5."

Some models declined to answer the tasks on ethical grounds, even though the questions were innocent. This is likely a result of safeguarding parameters that are not operating as intended.

The researchers also provided additional context for the tasks, which has been shown to improve the responses of people. However, the LLMs tested didn't show any consistent improvement.

Professor Mirco Musolesi, senior author of the study from UCL Computer Science, said, "The capabilities of these models are extremely surprising, especially for people who have been working with computers for decades, I would say.

"The interesting thing is that we do not really understand the emergent behavior of large language models and why and how they get answers right or wrong. We now have methods for fine-tuning these models, but then a question arises: If we try to fix these problems by teaching the models, do we also impose our own flaws? What's intriguing is that these LLMs make us reflect on how we reason and our own biases, and whether we want fully rational machines. Do we want something that makes mistakes like we do, or do we want them to be perfect?"

The models tested were GPT-4, GPT-3.5, Google Bard, Claude 2, Llama 2 7b, Llama 2 13b and Llama 2 70b.

More information: Olivia Macmillan-Scott and Mirco Musolesi. (Ir)rationality and cognitive biases in large language models, *Royal Society Open Science* (2024). DOI: 10.1098/rsos.240255. royalsocietypublishing.org/doi/10.1098/rsos.240255

Provided by University College London