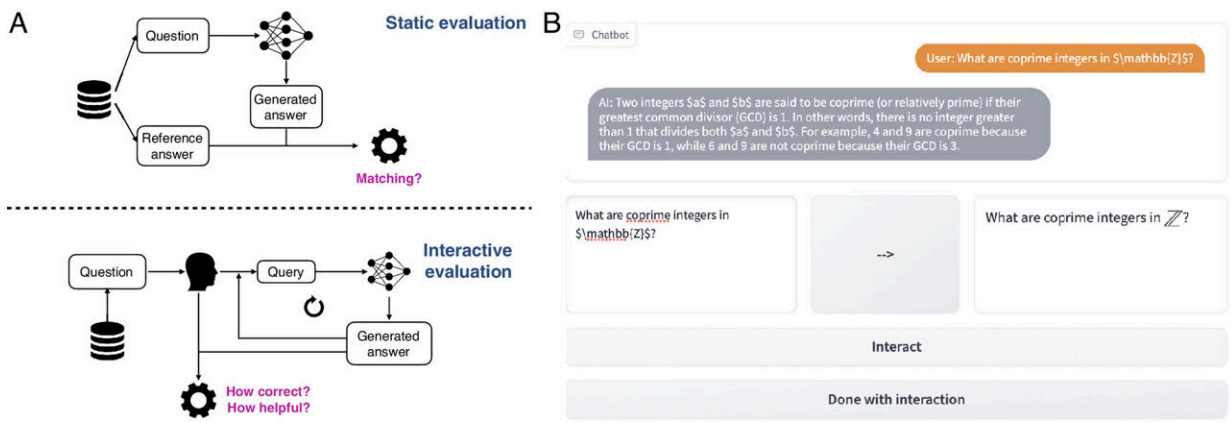
New open-source platform allows users to evaluate performance of AI-powered chatbots

June 4 2024, by Sarah Collins



(A) Contrasting typical static evaluation (Top) with interactive evaluation (Bottom), wherein a human iteratively queries a model and rates the quality of responses. (B) Example subset of the chat interface from CheckMate where users interact with an LLM. The participant can type their query (Lower Left), which is compiled in LaTeX (Lower Right). When ready, the participant can press "Interact" and have their query routed to the model. Credit: *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2318124121

A team of computer scientists, engineers, mathematicians and cognitive scientists, led by the University of Cambridge, have developed an open-source evaluation platform called CheckMate, which allows human users to interact with and evaluate the performance of large language models

(LLMs).

The researchers tested CheckMate in an experiment where human participants used three LLMs—InstructGPT, ChatGPT and GPT-4—as assistants for solving undergraduate-level mathematics problems.

The team studied how well LLMs can assist participants in solving problems. Despite a generally positive correlation between a chatbot's correctness and perceived helpfulness, the researchers also found instances where the LLMs were incorrect, but still useful for the participants. However, certain incorrect LLM outputs were thought to be correct by participants. This was most notable in LLMs optimized for chat.

The researchers suggest models that communicate uncertainty, respond well to user corrections, and can provide a concise rationale for their recommendations, make better assistants. Human users of LLMs should verify their outputs carefully, given their current shortcomings.

The results, reported in the *Proceedings of the National Academy of Sciences*, could be useful in both informing AI literacy training, and help developers improve LLMs for a wider range of uses.

While LLMs are becoming increasingly powerful, they can also make mistakes and provide incorrect information, which could have negative consequences as these systems become more integrated into our everyday lives.

"LLMs have become wildly popular, and evaluating their performance in a quantitative way is important, but we also need to evaluate how well these systems work with and can support people," said co-first author Albert Jiang, from Cambridge's Department of Computer Science and Technology. "We don't yet have comprehensive ways of evaluating an

LLM's performance when interacting with humans."

The standard way to evaluate LLMs relies on static pairs of inputs and outputs, which disregards the interactive nature of chatbots, and how that changes their usefulness in different scenarios. The researchers developed CheckMate to help answer these questions, designed for but not limited to applications in mathematics.

"When talking to mathematicians about LLMs, many of them fall into one of two main camps: either they think that LLMs can produce complex mathematical proofs on their own, or that LLMs are incapable of simple arithmetic," said co-first author Katie Collins from the Department of Engineering. "Of course, the truth is probably somewhere in between, but we wanted to find a way of evaluating which tasks LLMs are suitable for and which they aren't."

The researchers recruited 25 mathematicians, from undergraduate students to senior professors, to interact with three different LLMs (InstructGPT, ChatGPT, and GPT-4) and evaluate their performance using CheckMate. Participants worked through undergraduate-level mathematical theorems with the assistance of an LLM and were asked to rate each individual LLM response for correctness and helpfulness. Participants did not know which LLM they were interacting with.

The researchers recorded the sorts of questions asked by participants, how participants reacted when they were presented with a fully or partially incorrect answer, whether and how they attempted to correct the LLM, or if they asked for clarification. Participants had varying levels of experience with writing effective prompts for LLMs, and this often affected the quality of responses that the LLMs provided.

An example of an effective prompt is "what is the definition of X" (X being a concept in the problem) as chatbots can be very good at

retrieving concepts they know of and explaining it to the user.

"One of the things we found is the surprising fallibility of these models," said Collins. "Sometimes, these LLMs will be really good at higher-level mathematics, and then they'll fail at something far simpler. It shows that it's vital to think carefully about how to use LLMs effectively and appropriately."

However, like the LLMs, the human participants also made mistakes. The researchers asked participants to rate how confident they were in their own ability to solve the problem they were using the LLM for. In cases where the participant was less confident in their own abilities, they were more likely to rate incorrect generations by LLM as correct.

"This kind of gets to a big challenge of evaluating LLMs, because they're getting so good at generating nice, seemingly correct natural language, that it's easy to be fooled by their responses," said Jiang. "It also shows that while human evaluation is useful and important, it's nuanced, and sometimes it's wrong. Anyone using an LLM, for any application, should always pay attention to the output and verify it themselves."

Based on the results from CheckMate, the researchers say that newer generations of LLMs are increasingly able to collaborate helpfully and correctly with human users on undergraduate-level math problems, as long as the user can assess the correctness of LLM-generated responses.

Even if the answers may be memorized and can be found somewhere on the internet, LLMs have the advantage of being flexible in their inputs and outputs over traditional search engines (though should not replace search engines in their current form).

While CheckMate was tested on mathematical problems, the researchers say their platform could be adapted to a wide range of fields. In the

future, this type of feedback could be incorporated into the LLMs themselves, although none of the CheckMate feedback from the current study has been fed back into the models.

"These kinds of tools can help the research community to have a better understanding of the strengths and weaknesses of these models," said Collins. "We wouldn't use them as tools to solve complex mathematical problems on their own, but they can be useful assistants if the users know how to take advantage of them."

More information: Katherine M. Collins et al, Evaluating language models for mathematics through interactions, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2318124121

Provided by University of Cambridge