The Most Sophisticated AIs Are Most Likely to Lie, Worrying Research Finds

As AI chatbots get bigger and more powerful, they are also lying more, instead of declining questions they can't answer.

By Frank Landymore

Sep 28, 2024 06:30 AM · 3 min. read · View original

Beware the smart ones: they seem to have all the answers, but can also weave the most convincing lies.

It seems that this logic also applies to large language models, which are becoming more powerful with each iteration. New research suggests that this smarter crop of AI chatbots are actually becoming *less* trustworthy, because they're more likely to make up facts rather than avoiding or turning down questions they can't answer.

The study, published in the journal *Nature*, examined some of the leading commercial LLMs in the industry: OpenAI's GPT, and Meta's LLaMA, along with an open source model called BLOOM created by the research group BigScience.

While it found that their responses are in many cases becoming more accurate, they were across the board less reliable, giving a higher proportion of wrong answers than older models did.

"They are answering almost everything these days. And that means more correct, but also more incorrect [answers]," study coauthor José Hernández-Orallo, a researcher at the Valencian Research Institute for Artificial Intelligence in Spain, told *Nature*.

Mike Hicks, a philosopher of science and technology at the University of Glasgow, had a harsher assessment.

"That looks to me like what we would call bullshitting," Hicks, who was not involved in the study, told *Nature*. "It's getting better at pretending to be knowledgeable."

The models were quizzed on topics ranging from math to geography, and were also asked to perform tasks like listing information in a specified order. The bigger, more powerful models gave the most accurate responses overall, but faltered at harder questions, for which they had a lower accuracy.

According to the researchers, some of the biggest BS-ers were OpenAI's GPT-4 and o1, which would answer almost any question thrown at them. But all of the studied LLMs appear to be trending this way, and for the LLaMA family of models, none of them could reach a level of 60 percent accuracy for the easiest questions, the study said.

In sum, the bigger the AI models got — in terms of parameters, training data, and other factors — the bigger the percentage of wrong answers they gave.

Still, AI models are getting better at answering more complex questions. The problem, other than their propensity for

BS-ing, is that they still mess up the easy ones. In theory, these errors should be a bigger red flag, but because we're impressed at how the large language models handle sophisticated problems, we may be overlooking their obvious flaws, the researchers suggest.

As such, the work had some sobering implications about how humans perceive the AI responses. When asked to judge if the chatbots' answers were accurate or inaccurate, a select group of participants got it wrong between 10 to 40 percent of the time.

The simplest way to combat the issues, according to the researchers, is to program the LLMs to be less eager to answer everything.

"You can put a threshold, and when the question is challenging, [get the chatbot to] say, 'no, I don't know,'" Hernández-Orallo told *Nature*.

But honesty may not be in the best interests of AI companies looking to woo the public with their fancy tech. If chatbots were reined in to answer only stuff they knew about, it might expose the limits of the technology.

More on AI: *Zuckerberg Says It's Fine to Train AI on Your Data Because It Probably Has No Value Anyway*