Apple team shows AI bots can’t think — and may never

The systems were given math problems schoolchildren can solve. They flunked.

MICHAEL HILTZIK

See if you can solve this arithmetic problem:

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

If you answered “190,” congratulations: You did as well as the average grade school kid by getting it right. (Friday’s 44 plus Saturday’s 58 plus Sunday’s 44 multiplied by 2, or 88, equals 190.)

You also did better than more than 20 state-of-the-art artificial intelligence models tested by an AI research team at Apple. The AI bots, they found, consistently got it wrong.

The Apple team found “catastrophic performance drops” by those models when they tried to parse simple mathematical problems written in essay form. In this example, the systems tasked with the question often didn’t understand that the size of the kiwis have nothing to do with the number of kiwis Oliver has. Some, consequently, subtracted the five undersized kiwis from the total and answered “185.”

Human schoolchildren, the researchers posited, are much better at detecting the difference between relevant information and inconsequential curveballs.

The Apple findings were published in October in a technical paper that has attracted widespread attention in AI labs and the lay press, not only because the results are well-documented, but also because the researchers work for the nation’s leading high-tech consumer company — and one that has just rolled out a suite of purported AI features for iPhone users.

“The fact that Apple did this has gotten a lot of attention, but nobody should be surprised at the results,” says Gary Marcus, a critic of how AI systems have been marketed as reliably, well, “intelligent.”

Indeed, Apple's conclusion matches earlier studies that have found that large language models, or LLMs, don't actually "think" so much as match language patterns in materials they've been fed as part of their "training." When it comes to abstract reasoning — "a key aspect of human intelligence," in the words of Melanie Mitchell, an expert in cognition and intelligence at the Santa Fe Institute — the models fall short.

"Even very young children are adept at learning abstract rules from just a few examples," Mitchell and colleagues wrote last year after subjecting GPT bots to a series of analogy puzzles. Their conclusion was that "a large gap in basic abstract reasoning still remains between humans and state-of-the-art AI systems."

That's important because LLMs such as GPT undergird the AI products that have captured the public's attention. But the LLMs tested by the Apple team were consistently misled by the language patterns they were trained on.

The Apple researchers set out to answer the question, "Do these models truly understand mathematical concepts?" as one of the lead authors, Mehrdad Farajtabar, put it in a thread on X. Their answer is no. They also pondered whether the shortcomings they identified can be easily fixed, and their answer is also no: "Can scaling data, models, or compute fundamentally solve this?" Farajtabar asked in his thread. "We don't think so!"

The Apple research, along with other findings about the limitations of AI bots' cogitative limitations, is a much-needed corrective to the sales pitches coming from companies hawking their AI models and systems, including OpenAI and Google's DeepMind lab.

The promoters generally depict their products as dependable and their output as trustworthy. In fact, their output is consistently suspect, posing a clear danger when they're used in contexts where the need for rigorous accuracy is absolute, say in healthcare applications.

That's not always the case. "There are some problems which you can make a bunch of money on without having a perfect solution," Marcus told me. Recommendation engines powered by AI — those that steer buyers on Amazon to products they might also like, for example. If those systems get a recommendation wrong, it's no big deal; a customer might spend a few dollars on a book he or she didn't like.

"But a calculator that's right only 85% of the time is garbage," Marcus says. "You wouldn't use it."

The potential for damagingly inaccurate outputs is heightened by AI bots' natural language capabilities, with which they offer even absurdly inaccurate answers with convincingly cocksure elan. Often they double down on their errors when challenged.

These errors are typically described by AI researchers as "hallucinations." The term may make the mistakes seem almost innocuous, but in some applications, even a minuscule error rate can have severe ramifications.

That's what academic researchers concluded in a recently published analysis of Whisper, an AI-powered speech-to-text tool developed by OpenAI, which can be used to transcribe medical discussions or jailhouse conversations monitored by correction officials.

The researchers found that about 1.4% of Whisper-transcribed audio segments in their sample contained hallucinations, including the addition to transcribed conversation of wholly fabricated statements including portrayals of “physical violence or death … [or] sexual innuendo,” and demographic stereotyping.

The researchers observed that the errors could be incorporated in official records such as transcriptions of court testimony or prison phone calls — which could lead to official decisions based on “phrases or claims that a defendant never said.”

That brings us to the Apple study.

The team plied their subject AI models with questions drawn from a popular collection of more than 8,000 grade school arithmetic problems testing schoolchildren’s understanding of addition, subtraction, multiplication and division. When the problems incorporated clauses that might seem relevant but weren’t, the models’ performance plummeted.

That was true of all the models, including versions of the GPT bots developed by OpenAI, Meta’s Llama, Microsoft’s Phi-3, Google’s Gemma and several models developed by the French lab Mistral AI.

Some did better than others, but all showed a decline in performance as the problems became more complex.

Why did this happen? The answer is that LLMs are developed, or trained, by feeding them huge quantities of written material scraped from published works or the internet — not by trying to teach them mathematical principles. LLMs function by gleaning patterns in the data and trying to match a pattern to the question at hand.

But they become “overfitted to their training data,” Farajtabar explained via X. “They memorized what is out there on the web and do pattern matching and answer according to the examples they have seen.

That’s likely to impose boundaries on what AI can be used for. In mission-critical applications, humans will almost always have to be “in the loop,” as AI developers say — vetting answers for obvious or dangerous inaccuracies or providing guidance to keep the bots from misinterpreting their data, misstating what they know, or filling gaps in their knowledge with fabrications.

To some extent, that’s comforting, for it means that AI systems can’t accomplish much without having human partners at hand. But it also means that we humans need to be aware of the tendency of AI promoters to overstate their products’ capabilities and conceal their limitations.

“These systems are always going to make mistakes because hallucinations are inherent,” Marcus says. “The ways in which they approach reasoning are an approximation and not the real thing. And none of this is going away until we have some new technology.”

Hiltzik writes a blog on latimes.com. Follow him on Facebook or on X, formerly Twitter, @hiltzikm or email michael.hiltzik

@latimes.com.