Where LLMs Are Falling Short

Art via Mike Sullivan

By Stephanie Palazzolo

Share

Jul 21, 2025, 7:00am PDT

I'm back from the **International Conference on Machine Learning** in Vancouver, one of the biggest annual meetups for artificial intelligence researchers. And this year, the conference underscored all the ways in which large language models are still falling short of everyone's expectations, despite the immense progress that got us to this point.

Researchers even called into question some of the most promising techniques that have gained popularity over the last year, such as **chain-of-thought reasoning**, or asking the models to describe how they arrived at an answer—their "thoughts," so to speak.

For instance, one research paper presented at the event explained how chain-of-thought can actually hurt model performance on certain tasks due to "overthinking." In one example, models were shown strings of letters that followed some rule that the model didn't know. Then the models were shown another string of letters and asked whether it followed the unknown rule or not.

Humans typically perform better on this task when they're told to go with their gut feeling. However, the models performed worse when asked to explain their reasoning. Models are great at finding patterns, but when there are *so many* possibilities for what the unknown rule or pattern might be, they tend to overthink and end up at the wrong answer, the paper argued.

Several papers at ICML pointed out how today's large language models are rooted in text and that much more work is needed in getting the AI to understand and generate images, videos and audio, otherwise known as multimodality.

One paper from **Google DeepMind** explained how models that generate audio can only generate tens of seconds of speech at a time before devolving into gibberish because of memory limitations.

The Google DeepMind researchers came up with a new type of model that's able to generate nearly 20 minutes of speech in one go. Still, it's crazy that even with all the advances in LLMs, speech AI is far behind, especially with the number of startups that want to build AI personal assistants to converse with us.

Another presentation showed how multimodal LLMs fall short when it comes to answering questions where they need to understand images, like diagrams in physics questions.

For instance, in a question that asks the model to predict what a sheet of paper might look like after folding and cutting it in certain ways, the LLM tries to describe (in text form) how it would solve the problem and gets confused. Humans typically arrive at the right answer more easily by visually imagining folding and cutting the paper, according to the research paper.

In another sign of how AI research has gotten more practical and commercial in recent years, ICML also featured lots of papers and presentations about benchmarks, or ways to measure the performance of LLMs on real-world tasks.

These include SWE-Lancer, which measures models' ability to code by having them complete real tasks on Upwork; WindowsAgentArena, which specifically tests AI agents' ability to navigate a Windows operating system and apps like Excel and PowerPoint; and LOB-Bench, a benchmark that measures how well AI models can produce certain kinds of financial data. These benchmarks highlight how AI research has gotten more practical and less theoretical in recent years.

Speaking of agents for Windows…

Making a 'Better Siri'—For Windows PCs

Apple's postponement of AI upgrades to **Siri** earlier this year was a tacit acknowledgement that training LLMs to automate tasks like navigating menus and filling out forms is hard to do.

Jeffrey Lai, who spent more than a decade developing Siri as an Apple engineer, earlier this year cofounded a startup, **IrisGo**, to make what he calls a "better Siri" that can automate tasks within any app that's open on a personal computer screen. (IrisGo contains the word Siri spelled backward though Lai says that was unintentional!)

The catch: The software is only available for **Microsoft Windows** PCs, which Lai believes is the best platform for IrisGo's target audience: white-collar professionals who work in an office

environment and want to save time on tasks like writing emails and working with spreadsheets.

IrisGo's marquee feature is called "watch and learn." It allows a user to record their screen to teach the software how to automate certain actions such as opening a web browser, typing in a website address and clicking through different parts of a website to make a reservation.

To accomplish this, IrisGo uses the federally mandated accessibility features built into Windows to gain broad access to the operating system.

IrisGo says its software will be preinstalled on select models of a large PC maker by the fourth quarter of this year, and it raised funding from **AI Fund**, the investment firm started by **Andrew Ng**, a former executive at **Baidu** and **Google Brain**.

The startup plans to generate revenue by selling a Pro subscription to its service, in addition to offering a free version, and also plans to share that revenue with Intel and PC manufacturers that preinstall the software.

While Microsoft has touted **Copilot** AI features in its line of AI-focused PCs, many of these features simply integrate familiar tools like generating text, images and videos.

Microsoft also has shied away from offering features that take advantage of personal user data the way IrisGo would. It still doesn't offer the ability to control and automate actions in apps.

OpenAI, meanwhile, announced a new feature called **ChatGPT** agent last week that can also automate some tasks inside a web browser. It also touted the ability to create and edit spreadsheets and emails, though only using select online services like **Gmail** and **Microsoft Excel**.

Lai said he's encouraged to see growing validation from the market and that his startup plans to double down on features such as context awareness and the control and automation of software.

While IrisGo also answers general knowledge questions and can draft replies like other chatbots, a unique selling point is that much of its computing occurs on device rather than in the cloud, which keeps user data more private, allows for faster response times and reduces IrisGo's server costs.

To accomplish this, IrisGo worked closely with Intel to tailor its software for the chipmaker's latest generation processors, Lai said.

Still, Lai isn't as militant as Apple about processing all AI requests on-device. "The lesson I learned at Apple was that usability is what users care about most," he said. "If they don't find it usable, it

doesn't matter whether it runs on the device or in the cloud."—*Wayne Ma*

A Gold Medal in Math

Late Friday night, **OpenAI** announced that one of its AI models solved problems in this year's **International Math Olympiad** at a level that would have earned a human competitor a gold medal, the highest award in the competition.

That is a much-sought-after achievement for AI developers due to the difficulty of the math problems in the competition. The progress could indicate that language models are good enough at picking up mathematical patterns that they will be able to contribute to mathematical research sooner than previously expected.

Last year, **Google DeepMind** scored a silver medal on the IMO, solving four of the six problems, which first had to be translated into formal mathematical statements that DeepMind's system could understand.

OpenAI's model provided correct proofs for five of the six problems, as graded by former IMO competitors that OpenAI hired. Unlike DeepMind's system, which used a combination of specialized models designed only to solve math problems, OpenAI said it used a general-purpose language model.

Still, not everyone is bowled over by the news. “People are making a big deal over the silver versus gold,” but since there are only six problems, the difference between silver and gold can come down to random noise, said **Elliot Glazer**, lead mathematician at **Epoch AI**, which developed the challenging AI math benchmark Frontier Math.

But “it’s more surprising” that OpenAI was able to achieve its score using a general-purpose LLM, he said. **Gemini 2.5 Pro**, **Google’s** general purpose LLM, held the previous high score on the IMO among language models, but it did not even earn a bronze medal.

While the IMO problems are difficult, “this is a high school competition,” he said. It’s “many miles from the kind of math that really matters to mathematicians.” In contrast, OpenAI’s **o4-mini** solves only 6% of the hardest math problems in Frontier Math. Once AI models can solve all those problems, then they will be ready to undertake their own math research, Glazer said.

Another impressive sign of progress came from the performance of OpenAI’s new ChatGPT agent on **SpreadsheetBench**, a benchmark that tests models’ ability to edit spreadsheets. **ChatGPT agent** was able to complete up to 45.5% of the tasks correctly, significantly higher than Microsoft’s Copilot product in Excel.—*Rocket Drew*

Free Software for Content Moderation

Roost, a nonprofit that launched in February to develop content moderation tools, is releasing software for companies, such as productivity and messaging apps, to detect unwanted content online, including AI-generated material.

One of Roost's tools, **Coop**, reviews content, such as AI-generated videos, and can route potential child sexual abuse material to the National Center for Missing & Exploited Children. Coop is built atop software developed by **Cove**, a startup whose intellectual property Roost acquired.

Roost's other tool, **Osprey**, helps companies like **Discord**, the messaging service, monitor and remove posts that violate their rules. Discord developed Osprey before donating it to Roost. **Bluesky** said it plans to use the product.

"We cannot expect a thousand companies to build a thousand different safety systems," said **Vinay Rao**, who joined Roost last week after serving as head of the safeguards team at **Anthropic**, which detects and patches vulnerabilities with Anthropic's **Claude** models.

At Anthropic, users would contact Rao's team to report potential bugs in Claude. Most of the time, they identified a behavior that Anthropic already knew about, said Rao, but every two weeks or so,

someone would disclose a bug that warranted a new defense from Anthropic.

Rao aims to make automatic detection tools like the ones Anthropic uses available to other companies through Roost. He says that even **xAI**, which has made a point of developing AI that is willing to say provocative things, still needs these kind of trust and safety tools to prevent child sexual abuse material, crypto scams and unwanted bots.

Roost, which is an acronym for Robust Open Online Safety Tools, is releasing the two products open-source, meaning software developers can download them for free and modify them.—*Rocket Drew*

Deals and Debuts

See The Information's Generative AI Database for an exclusive list of private companies and their investors.

The Gates Foundation and other philanthropists invested $1 billion in **NextLadder Ventures**, which will invest in businesses developing technology to help low-income people manage job loss, health issues and housing insecurity.

Q.Ant, which designs light-based AI chips, raised $72 million in a Series A funding round led by **Cherry Ventures**, **UVC Partners** and **Imec.xpand**.

BrightAI, which uses AI to provide monitoring for essential services such as water and gas, raised $51 million in a Series A funding round led by **Khosla Ventures** and **Inspired Capital**.

OpenAI announced a $50 million fund to support nonprofit and community organizations, implementing a key recommendation of its nonprofit advisory commission.

GeologicAI, which uses AI to help mining companies analyze rock samples, raised $44 million in a Series B funding round led by **Blue Earth Capital**.

QpiAI, which develops quantum computers designed to work with AI, raised $32 million in a Series A funding round led by **Avataar Ventures** and the Indian government's **National Quantum Mission**.

Icounter, which helps companies defend against AI-enabled cyber attacks, raised $30 million in a Series A funding round led by **SYN Ventures**.

Alix, a San Francisco-based AI-powered estate settlement startup, raised $20 million in Series A funding led by **Acrew Capital**.

Anysphere, the company behind **Cursor**, acquired AI-powered customer relationship management startup **Koala**, TechCrunch reported.

OpenAI on Thursday launched features for ChatGPT subscribers to create and edit spreadsheets and presentations, generate reports and automate tasks using web browsers, confirming The Information's earlier report about the product.

ServiceNow's $2.85 billion acquisition of agent startup **Moveworks** is being reviewed over antitrust concerns, Bloomberg reported.

Meta Platforms will not sign the European Union's **code of practice** for artificial intelligence, a voluntary framework meant to help companies comply with the EU's AI Act, Meta Chief Global Affairs Officer Joel Kaplan said.

The White House is preparing an executive order targeting artificial intelligence companies whose models the administration deems too "woke," the Wall Street Journal reported.

Mistral added new features to its **Le Chat** website, including a new deep research mode and image editing capabilities, matching offerings from rivals such **OpenAI**.

Taiwan Semiconductor Manufacturing Company is speeding up the construction schedule of its second and third factories in the U.S. by "several quarters," responding to strong demand from its American customers, CEO C.C. Wei said on an earnings call on Thursday.

Cursor has restricted China-based users from accessing U.S. models offered by its coding assistant software, such as **Claude-4-Sonnet** and **Gemini-2.5-Pro**.

Netflix used AI to generate visual effects for a scene of a building collapsing in the show **The Eternaut**, the first time Netflix has used generative AI for footage in one of its TV shows, BBC reported.

DuckDuckGo, a browser maker, released a new setting that allows users to filter out AI-generated images in search results.