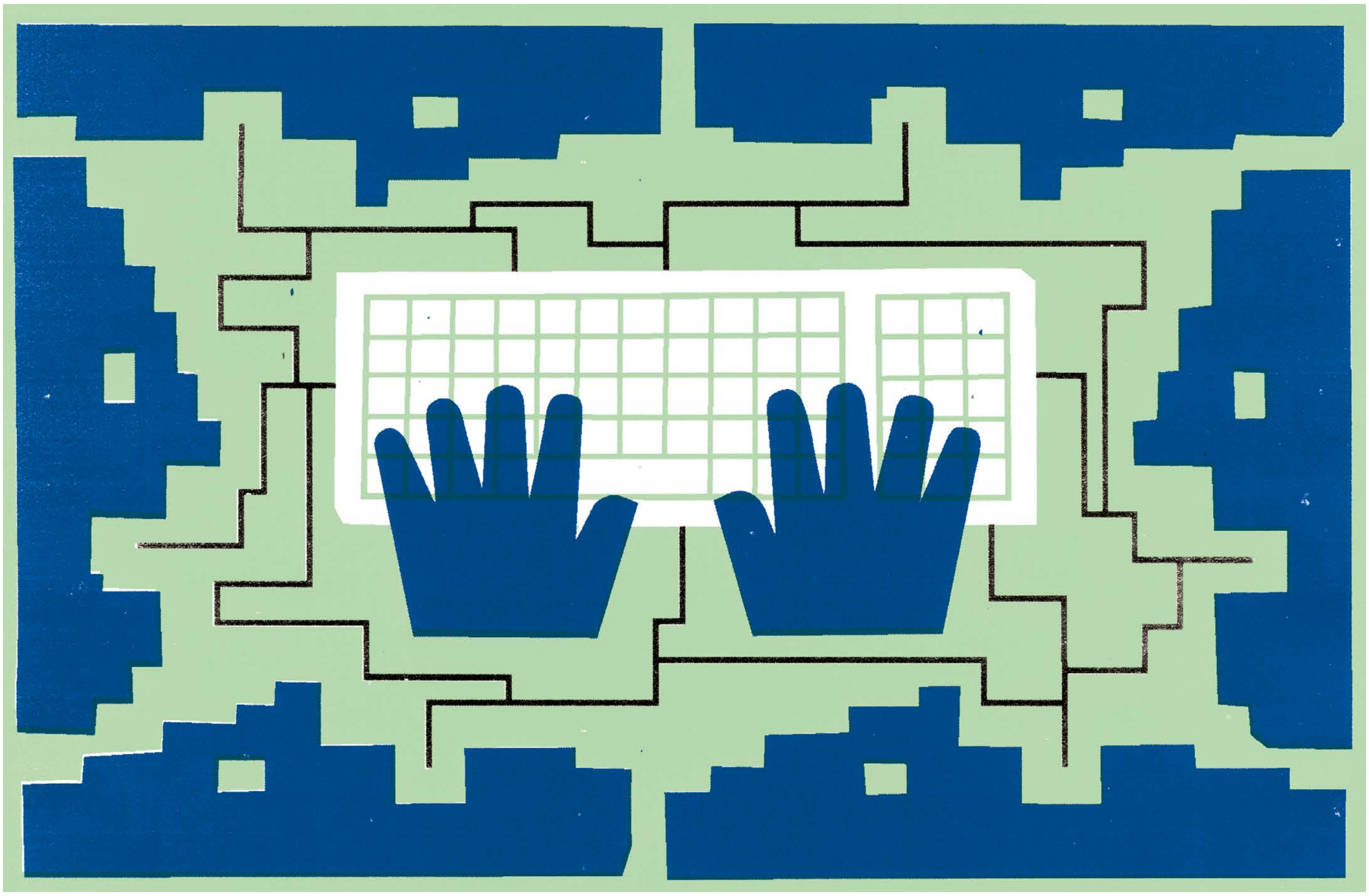


ILLUSTRATION: THE PROJECT TWINS

WHAT'S IT LIKE TO WORK WITH AN AI TEAM OF VIRTUAL SCIENTISTS?

Emerging 'co-scientist' systems use teams of chatbots to mimic the deliberations of a research group. Do they really help? **By Nicola Jones**

On a Sunday morning in April, pathologist Thomas Montine ran one of the most surreal meetings of his life. In an online test interface for a system called the Virtual Lab, Montine constructed a team of six artificial-intelligence (AI) characters, all powered by a commercial large language model. He gave them specialities: he made a couple neuroscientists, one a neuropharmacologist and another a medicinal chemist. Then, he asked this virtual lab group to examine possible treatments for Alzheimer's disease and discuss gaps in knowledge, barriers to progress and hypotheses to be tested – the same questions he has to consider in grant applications.

A few minutes later, he had a transcript of their conversation, which ran to more than 10,000 words. A virtual principal investigator had kicked things off: "Thank you all for joining this important meeting."

Montine, who studies cognitive impairment at Stanford University in California, was testing an example of an emerging trend in AI-assisted science: using teams of chatbot specialists to develop a scientific idea as a lab team might do. The developers of these 'co-scientist' systems argue that such collaborative efforts can help researchers to think through research hypotheses rapidly, saving time and – in a more contentious assertion – producing new and important research ideas.

The most prominent team exploring this concept comprises researchers at the technology giant Google, who this February announced the results of early tests of their AI co-scientist with researchers[1] (see also go.nature.com/3hmxuxm). The team has opened up the project to a group of trusted testers as it continues to develop the tool.

Google is not alone. A team including computational biologists at Stanford University announced its Virtual Lab system in November last year[2] – a version of which Montine was playing with. And a group based at the Shanghai Artificial Intelligence Laboratory in China proposed a similar virtual-scientist system, called VirSci, last October[3]; the researchers are building it now.

Rick Stevens, a computer scientist at the

University of Chicago and at Argonne National Laboratory, both in Illinois, says that he and other computationally adept researchers are creating their own such systems by setting up AI personas that then interact. “I mean, everybody can do it,” he says.

In many of these systems, the large language models (LLMs) involved don’t just bounce ideas off each other. They also search the Internet, execute code and interact with other software tools, making them part of ‘agentic AI’, a fuzzy term that refers to LLMs autonomously undertaking tasks, although in practice there is often a lot of human oversight. A group of AI agents can be woven into a larger system that can work on high-level problems for hours without getting distracted or confused, Stevens says.

“It’s not really fundamentally that different than having more colleagues, in some sense,” he says, “except that they don’t get tired, and they’ve been trained on everything.”

To explore what it’s like to work with a virtual team, *Nature* asked a few scientists to trial a version of the Stanford system and spoke to some who have used Google’s AI co-scientist. Does a network of chatbots talk like a room full of Nobel prizewinners or undergraduates? Are the ideas they generate nonsensical, boring and trivial, or smart, valuable and insightful?

Multiple personalities

All co-scientist systems assign roles or personalities to agents and get them to interact, but the details vary. The Virtual Lab, built by computer scientist Kyle Swanson in James Zou’s group at Stanford University and his colleagues, comes with two default characters, both (for now) powered by the LLM GPT-4o from tech firm OpenAI in San Francisco, California. These characters are a principal investigator and a critic — an agent told to provide helpful feedback. The user (or the AI’s principal investigator) can then add as many agents as they like, of any kind, writing in simple descriptions for each agent to guide the characters’ interactions. The user chooses how many turns the agents have to ‘speak’, and a meeting transcript is produced in minutes. The team is working on ways to train the agents on literature that is relevant to the characters’ described expertise (as others have done[4]), rather than merely telling them to assume specific roles.

> “THESE LLMS ARE WHAT FIRE WAS FOR EARLY HUMAN SOCIETIES.”

The Google co-scientist, developed by Google Deep Mind’s Alan Karthikesalingam in London and Vivek Natarajan in Mountain View, California, and their colleagues, is an extension of the researchers’ work on AI that is specialized in biomedicine, including the LLM Med-PaLM.

In contrast to the Stanford system, the Google tool doesn’t let users assign scientific specialities to agents. Instead, agents have predefined specific functions: idea generation; reflection or critique; evolution of ideas; determining the proximity of ideas to reduce duplication; ranking; and meta-review. These six agents are powered by Google’s LLM Gemini 2.0.

Users prompt the system with a few sentences, including a goal and a desired format for output. They can choose to add background information, such as relevant papers. The agents collaborate to tackle the problem and search the Internet, then spit out a summary report that can be tens or hundreds of pages long. “The co-scientist is like a smart scientific partner, capable of seeing the obvious and non-obvious connections in a sea of research,” says Natarajan. “We hope to give scientists superpowers.”

Like all LLMs, the models powering these agents sometimes hallucinate, making up text that can be wrong. But having a critic or judge in a multi-agent conversation tends to weed out things that don’t make sense, says Stevens. Besides, he adds, hallucinations can be useful for creative processes such as thinking outside of the box during brainstorming, as long as experts check that the output makes sense.

There is evidence that the multi-agent strategy improves output, compared with chatting to just one AI agent or bot. For instance, Zou has found that adding a critic to a conversation that used GPT-4o bumped up the model’s performance score by a couple of percentage points on graduate-level science tests, and improved its answers in test cases, including designing radiotherapy treatment plans[5].

Google tested its AI co-scientist system to see whether human experts preferred its multi-agent answers over content produced by a lone chatbot[1]. The experts rated the novelty and impact of the ideas generated by the co-scientist system slightly higher than those generated by Gemini 2.0 or OpenAI’s o1.

Testing an AI virtual lab

Selections from the AI-generated output that researcher Francisco Barriga got back from the Virtual Lab system he tested.

Principal Investigator

“Thank you all for joining this important meeting. Our goal today is to identify the minimal number of distinct type I interferons to test their role in anti-tumor immunity, specifically in the context of lung cancer, melanoma and pancreatic cancer using C57BL/6 mice. We need to ensure that our experimental design is both scientifically robust and ethically responsible, minimizing the number of mice used while capturing the relevant biology.

To guide our discussion, here are some initial thoughts and questions:

1. Sequence divergence. Understanding which type I interferons are most sequence-divergent is crucial. This will help us select candidates that are likely to have distinct biological functions. I would like to hear from the Computer Scientist about any bioinformatics analyses or databases that could help us identify these divergent interferons.

2. Role in tumor and immune cells. We need to gather evidence on how these interferons influence tumor progression or immune modulation. The Immunologist and Cancer Biologists can provide insights from the literature and any previous experimental data.

3. Mouse model. Selecting the right mouse model is vital for our experiments. We need to ensure that the model accurately reflects the human condition we are studying. I would like input on whether the C57BL/6 strain is the best choice and if there are any specific considerations we should be aware of.

4. Experimental conditions. Finally, we need to design the experimental conditions. This includes deciding on control groups, treatment groups, and any other variables. I would like the Scientific Critic to help us ensure that our design is rigorous and that we are considering all necessary controls.

Let’s proceed with each team member providing their insights on these points. We’ll start with the Computer Scientist, followed by the Immunologist, Cancer Biologists and, finally, the Scientific Critic. After each round of input, I’ll synthesize information and ask follow-up questions as needed.”

See go.nature.com/4recf2t for Barriga’s input and more AI-generated responses.

Some research has even delved into the optimum number of agents and rounds of discussion. Computer scientist Nanqing Dong at the Shanghai Artificial Intelligence Laboratory and his colleagues, who are building the VirSci system, say that they see peak creativity with a team of eight agents each taking five turns to converse[3]. Swanson says that in his experience, adding more than three specialists often just leads to “wasted text”, and more than three rounds of conversation sometimes sends the agents off on a tangent.

Bright ideas

Stanford University medical researcher Gary Peltz, who often uses AI in his work, was one of the first testers of the Google AI co-scientist. He wanted to use it to find drugs to treat liver fibrosis. Because the AI system was still in development, he sent his query to an intermediary at Google. His prompt read: “Develop hypotheses about the genes and epigenomic changes required for myofibroblast generation in liver fibrosis and indicate what drugs should we test as new treatments for liver fibrosis.” It also included some paragraphs of detailed background information. He got the report back about a day later.

After some preliminaries, the report begins “We propose a novel hypothesis…” then works its way to concluding: “This research could have a profound impact on liver fibrosis research and therapeutic advancement.”

“When I read it, I literally fell off my chair,” says Peltz. He had just written a grant proposal focused on the importance of epigenetic changes in liver fibrosis, and the AI had targeted the same theme for its proposed therapeutics.

The AI co-scientist suggested three drugs, and Peltz came up with two more (all of which are already approved to treat other conditions). Google paid Peltz to help accelerate lab testing, and over the next few months, Peltz’s lab tested all five drugs in its human organoid model. Two of the AI’s three suggestions showed promise for promoting liver regeneration and inhibiting fibrosis[6], whereas neither of Peltz’s worked out.

The experience left him impressed, he says: “These LLMs are what fire was for early human societies.”

Other liver researchers, however, say that the AI’s suggestions for drugs were neither particularly innovative nor profound. “I personally think they are pretty common sense, not much insight really,” says Shuang Wang, who works on liver disease at the Icahn School of Medicine at Mount Sinai in New York City. Google’s Natarajan counters: “Sometimes things look obvious in hindsight”.

Peltz says he was “particularly struck by the fact that it didn’t prioritize the things that I prioritized”. For the most promising drug in the AI’s candidate list, called vorinostat, he could find only two papers in PubMed that relate to its use in treating liver fibrosis. His choices had many more hits, making them seem like more obvious candidates. He adds that reading the AI report was similar to his discussions with postdocs. “They have a completely different perspective on things than I would,” he says.

Stilted conversations

The code for Stanford’s Virtual Lab is available on the developer platform GitHub, but the team has whipped up a simplified, private web interface for testers who don’t have the computer-science chops to deal with code. This interface differs a little from the full system, says Swanson, but the experience is roughly the same.

In the paper introducing the Virtual Lab[2], an AI team tackled designing biological components that could stick to a particular variant of the coronavirus SARS-CoV-2. In the first of a series of lab meetings mediated by human researchers, the AIs chose to focus on nanobodies (small antibody fragments) and selected four candidates to tweak. The researchers then asked the AI team to pick some existing software tools to redesign those nanobodies, and tasked specific AI agents with writing computer code to assess and rank the results. The process designed 92 nanobodies, of which 2 did indeed bind to the SARS-CoV-2 variant in lab tests.

> “WILL IT BE A GAME CHANGER IN MY DAY-TO-DAY? I DOUBT IT.”

The researchers whom *Nature* asked to try out the Virtual Lab didn’t go as far as to conduct suggested experiments or get their AI teams to write code. But they still found the AI helpful. Montine, for example, says his AI team did a great job of synthesizing the current knowledge (a task that other LLM systems can also handle well) and wrote answers to his grant-application questions with aplomb. “It went further than a postdoc would on a first try of writing a grant, and it only took like 2 minutes,” he says. “And it is a hoot to use.”

Another tester, cancer-genome researcher Francisco Barriga at the Vall D’Hebron Institute of Oncology in Barcelona, describes himself as a biochemist by training and a mouse modeller and genome engineer by choice, with zero coding skills and little experience with AI. He went into the trial hesitantly, suspecting that he would serve as a kind of non-tech-savvy control.

Barriga tasked the Virtual Lab with designing mouse-model experiments to test specific biological compounds, known as type 1 interferons, for their ability to affect tumours or immune cells while using a minimal number of mice – a topic that he knows inside out (see ‘Testing an AI virtual lab’). The AI team suggested exactly what he would have done, Barriga says: it chose “the right models, the right experiments”.

Still, Barriga says he feels like something essential is lacking. “It definitely doesn’t feel like humans are behind this.” The AI agents take turns, often ‘speaking’ in numerated lists, and are never rude, interruptive or argumentative. “It’s missing some of those leaps of intuition that you’ll get from a random conversation with some, I don’t know, plant biologist, over a coffee at 3 p.m. in a random hallway.” He could, of course, add a plant biologist – or a quantum physicist, or anyone at all – to his Virtual Lab group, but hasn’t tried that yet.

“Maybe it’s good to bounce ideas off. But will it be a game changer in my day-to-day? I doubt it,” Barriga says. He adds that the system might be something his PhD students could consult: “If they ever run into trouble and I’m too busy, maybe I’m replaceable.”

Broader insights

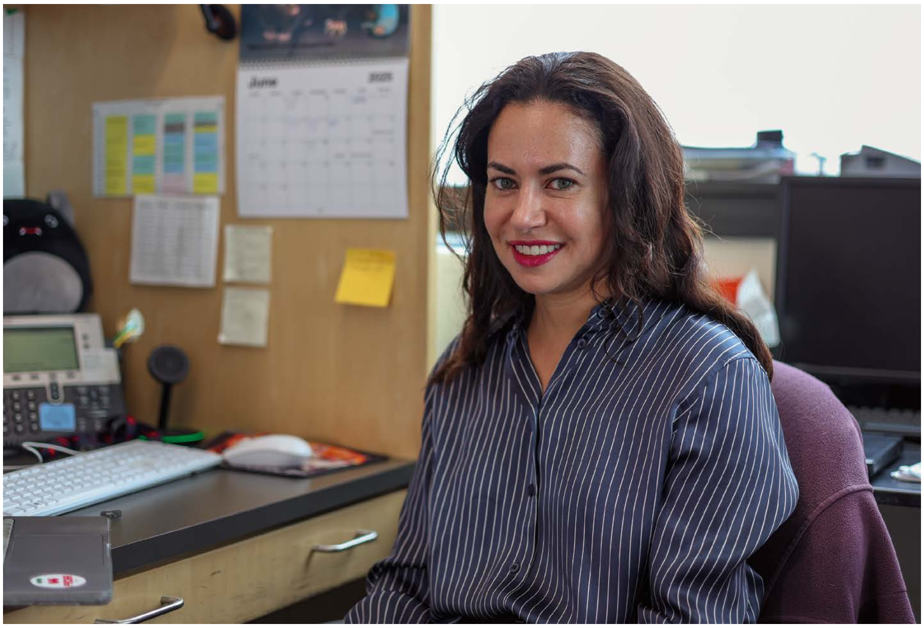
A third tester approached by *Nature*, Catherine Brownstein, is a geneticist who works on orphan diseases at Boston Children’s Hospital in Massachusetts, and has more experience with AI tools. She says that she uses LLMs for speed, efficiency and to broaden her thinking. But she cautions that users typically have to be experts so that they can spot errors – in the past, chatbots have sent her on time-consuming wild-goose chases, with incorrect paper summaries forcing her to reread a paper and its references to be sure she hadn’t got things wrong. “You have to kind of know what you’re talking about, otherwise it’s really easy to get completely led astray,” she says.

However, when Brownstein used the Virtual Lab to critique a paper she was writing, she was startled – and grateful – when the AI suggested that she ask the patients where they felt the research should go next. This had not occurred to her, although she says it should have. “I was embarrassed,” she says. “I stopped and stared for a full minute, because I was just like, ‘Oh my God. How did I get so far away from my original passion of having patient-focused and -centred research?’”

A simple checklist – or a chat with a friend, chatbot or even a bartender – might have led to the same insight. Yet, she says, none of her colleagues who had read her paper had thought to mention it. “It was actually a very humbling moment.”

Stepping sideways

For tissue engineer Ritu Raman at the Massachusetts Institute of Technology in



KEVIN FERGUSON/BOSTON CHILDREN'S HOSPITAL

Catherine Brownstein cautions that research expertise is still needed when using chatbots.

Cambridge, the value of Google's AI co-scientist was to help her to step outside her usual focus. Raman and her colleagues had found that when healthy muscle cells are exercised, they help neurons to grow faster and farther[7]. She wondered whether this result might be interesting to, say, someone studying motor neuron disease (amyotrophic lateral sclerosis) – a neurodegenerative disease in which failing neurons have downstream impacts on muscles. Perhaps, she thought, there is an upstream effect from the muscles back to the neurons.

Google's co-scientist came up with a list of 20 experimental approaches to investigate possible links. Some were related to procedures that Raman had no expertise in, such as genetics or protein folding. But one suggested protocol was more familiar: she could take some of the signalling molecules secreted by healthy and by affected muscle cells and feed them to a dish of motor neurons, to see whether the molecules from healthy cells were better at promoting neuron growth. Again, Google paid for her to run the experiment; the work is ongoing.

Raman could have come up with this protocol herself, but she says the co-scientist gave her confidence that she wasn't making a mistake, or simply repeating others' work, in an unfamiliar area.

Another person asking the same query and getting the same results might have run a different experiment, she adds, or picked up on a different insight in the AI response. "The interpreter is just as important as the algorithm," Raman says.

The novelty debate

In announcing its co-scientist, Google said the system had helped researchers to come up with "novel predictions". As the reaction to Peltz's experience suggests, some scientists question just how novel they are.

A similar discussion arose when microbiologist José Penadés at Imperial College London and his colleagues tested the Google system[8]. They fed the AI co-scientist a detailed question about bacterial evolution and asked it to come up with hypotheses. One paper the researchers included as background was a 2023 study by the group about how a type of mobile genetic element (a chunk of DNA that can infect other cells) moves around by stealing the tails of bacteriophages, which are viruses that infect bacteria. Researchers had assumed that the elements would only hijack tails lurking in the same cell as themselves. But the AI system proposed that they could steal tails from elsewhere.

This idea hadn't been published anywhere, and it went against established thought at the time. But Penadés and his team had privately already come to the same conclusion through unpublished experiments, so they knew that the idea was right.

The AI system "made an obvious correlation" from the work it had been fed, says Penadés, but that should still count as a successful new hypothesis, he says. "It's very good at joining the dots in an unbiased way."

Current co-scientist systems, like other LLM products, seem to be time-saving devices that can trigger insights by reminding users of things they had forgotten, kicking their brains into new modes, stating the obvious, flagging up background information they hadn't previously seen or suggesting particular paths. Only time will tell how often this process throws up truly new and useful ideas.

"The real proof, for us, are the real-world results," says Natarajan, referring to the lab validation of the ideas in both Penadés' and Peltz's studies.

To the future

For now, multi-agent LLM systems designed to help scientists with lab-meeting style reports are not readily available in easy-to-use formats. But Stevens expects that plenty of tech-savvy researchers and big tech firms will build systems with components aimed at different research needs, such as literature review or experiments. In May, the start-up firm FutureHouse in San Francisco, California, announced an autonomous AI discovery system, which it said combines a number of specialized AI modules to both review the literature and then suggest and design experiments (see go.nature.com/4df9gm8). Stevens expects that researchers will try giving co-scientists even more agency by coupling them to robotic systems in automated labs.

In many ways, playing with these multi-agent systems feels like chatting with any LLM interface. But some of the testers said the systems seemed to think more deeply, which might have to do with the training regimen of a particular LLM or with the benefits of multi-agent systems in general.

Still, none of the researchers *Nature* enlisted, nor any designers of the systems, felt that AIs could or should take over from people: the idea is not to replace lab workers, but to assist them. All felt that the adoption of LLMs into idea generation and experimental design is as inevitable as the adoption of Internet searches into science. But they differed in whether they felt the results would be dramatic or always beneficial.

If young scientists take shortcuts by using LLMs, that could, for example, undermine their learning processes, says Stevens. "I feel we're outsourcing a lot of our thought process," says Barriga – including the fun part of coming up with ideas.

In future, argues Peltz, science is likely to end up with a bounty of good AI-generated hypotheses developed with or by co-scientist LLMs – and the main limitation will be getting the funding to test them all.

Nicola Jones is a freelance journalist in Pemberton, Canada. Additional reporting by Richard Van Noorden and Elizabeth Gibney.

1. Gottweis, J. *et al*. Preprint at arXiv https://doi.org/10.48550/arXiv.2502.18864 (2025).
2. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. Preprint at bioRxiv https://doi.org/10.1101/2024.11.11.623004 (2024).
3. Su, H. *et al*. Preprint at arXiv https://doi.org/10.48550/arXiv.2410.09403 (2024).
4. Saeedi, D., Buckner, D., Aponte, J. C. & Aghazadeh, A. Preprint at arXiv https://doi.org/10.48550/arXiv.2503.23170 (2025).
5. Yuksekgonul, M. *et al. Nature* **639**, 609–616 (2025).
6. Guan, Y. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2025.04.29.651320 (2025).
7. Bu, A. *et al. Adv. Healthcare Mater*. **14**, 2403712 (2025).
8. Penadés, J. R. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2025.02.19.639094 (2025).