Cutting out this middleman of memory cooled things down by only a little less than 4 °C. But it should massively boost the bandwidth between the memory and the processor, which is important for another optimization the team tried—slowing down the GPU.

That might seem contrary to the whole purpose of better AI computing, but in this case, it's an advantage. Large language models are considered "memory-bound" problems—meaning that memory bandwidth is the main limiting factor to their speed. But Myers's team estimated that 3D-stacking HBM on the GPU would boost bandwidth fourfold. With that added headroom, even slowing the GPU's clock by 50 percent still leads to a performance win, while cooling everything down by more than 20 °C. In practice, the processor might not need to be slowed down quite that much: Slowing the clock frequency by just 30 percent led to a GPU that was only 1.7 °C warmer, Myers says.

Another big drop in temperature came from making the HBM stack and the area around it more conductive to heat. That included merging the four stacks into two wider stacks, thereby eliminating a heat-trapping region; thinning out the top die of the stack, which is usually thicker; and filling in more of the space around the HBM with blank pieces of silicon to conduct more heat.

With all of that, the stack now operated at about 88 °C. One final optimization brought things back to near 70 °C. Generally, some 95 percent of a chip's heat is removed from the top of the package, where in this case water carries the heat away. But adding similar cooling to the underside as well drove the stacked chips down a final 17 °C.

Although the research presented at IEDM shows it might be possible, HBM-on-GPU isn't necessarily the right way to go, Myers says. "We are simulating other system configurations to help build confidence that this is or isn't the best choice," he says. "GPU-on-HBM is of interest to some in industry," because it puts the GPU closer to the cooling. But it would likely be a more complex design, because the GPU's power and data would have to flow vertically through the HBM to reach it. ■

ARTIFICIAL INTELLIGENCE

AI Helps Scientists but Hurts Science

› New analysis suggests AI tools flatten discovery

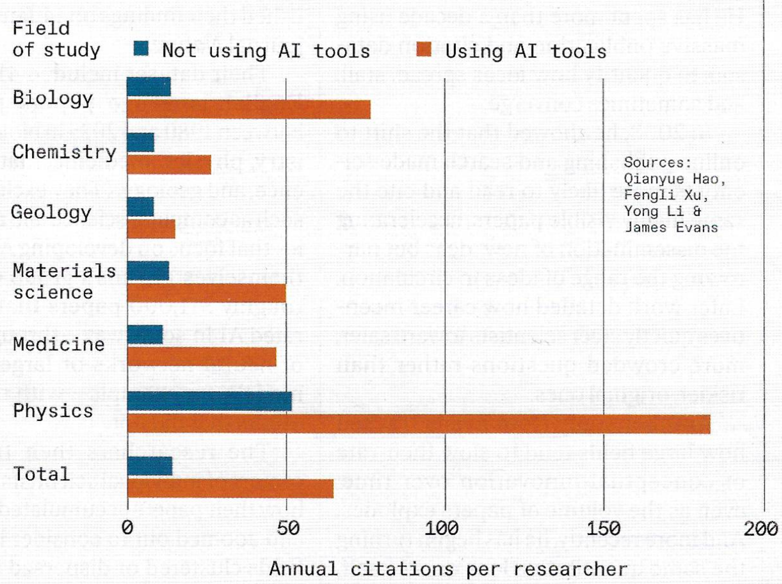
BY ELIE DOLGIN

Artificial intelligence is turning scientists into publishing machines. It's also quietly funneling them into the same crowded corners of research.

That's what a group of researchers in China and the United States concluded after analyzing more than 40 million academic papers. They found that scientists who use AI tools in their research publish more papers, accumulate more citations, and reach leadership roles sooner than peers who don't.

But there's a catch. As individual scholars soar through the academic ranks, science as a whole shrinks its curiosity. AI-heavy research covers less novel ground, clusters around the same data-rich problems, and sparks less follow-on engagement between studies.

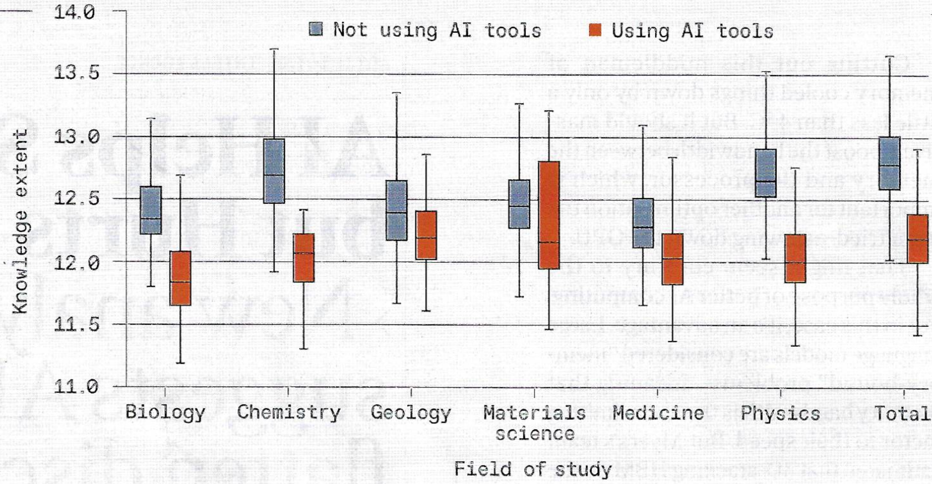
The findings highlight a tension between personal career advancement and collective scientific progress, as tools such as ChatGPT and



Across multiple fields, researchers who used AI tools in producing their research papers saw more annual citations than those who did not.

Researchers who used AI tools generally saw a decrease in "knowledge extent," which is functionally a decline in how much novel research space was covered in their work.

Sources: Qianyue Hao, Fengli Xu, Yong Li & James Evans



AlphaFold seem to reward speed and scale—but not surprise.

"You have this conflict between individual incentives and science as a whole," says James Evans, a sociologist at the University of Chicago who co-led the study.

And as more researchers pile onto the same scientific bandwagons, some experts worry about a feedback loop of conformity and declining originality. "This is very problematic," says Luís Nunes Amaral, a physicist who studies complex systems at Northwestern University, in Evanston, Ill., and who was not involved in the study. "We are digging the same hole deeper and deeper."

For Evans, the tension between efficiency and exploration is familiar terrain. He has spent more than a decade using massive publication and citation datasets to quantify how ideas spread, stall, and sometimes converge.

In 2008, he showed that the shift to online publishing and search made scientists more likely to read and cite the same highly visible papers, accelerating the dissemination of new ideas but narrowing the range of ideas in circulation. Later work detailed how career incentives quietly steer scientists toward safer, more crowded questions rather than riskier, original ones.

Another study from Evans tracked how large fields tend to slow their rate of conceptual innovation over time, even as the volume of papers explodes. And more recently, he has begun turning the same quantitative lens on AI itself, examining how algorithms reshape collective attention, discovery, and the organization of knowledge.

To quantify the AI effect, Evans and his collaborators from the Beijing National Research Center for Information Science and Technology trained a natural-language processing model to identify AI-augmented research across six natural science disciplines. They published their findings on 14 January in the journal *Nature*.

> **"We are digging the same hole deeper and deeper."**
>
> —LUÍS NUNES AMARAL, NORTHWESTERN UNIVERSITY

Their dataset included 41.3 million English-language papers published between 1980 and 2025 in biology, chemistry, physics, medicine, materials science, and geology. (They excluded fields such as computer science and mathematics that focus on developing AI methods themselves.) Evans's group compared roughly 311,000 papers that incorporated AI in some way—through the use of neural networks or large language models, for example—with millions of others that did not.

The researchers then traced the careers of individual scientists, examined how their papers accumulated attention, and zoomed out to consider how entire fields clustered or dispersed intellectually over time.

The results revealed a striking tradeoff. Scientists who adopt AI on average publish three times as many papers, receive nearly five times as many citations, and become team leaders a year or two earlier than those who do not.

But when those papers are mapped in a high-dimensional "knowledge space," AI-heavy research occupies a smaller intellectual footprint, clusters more tightly around popular, data-rich problems, and generates weaker networks of follow-on engagement between studies.

The pattern held across decades of AI development, spanning early machine learning, the rise of deep learning, and the current wave of generative AI. "If anything," Evans says, "it's intensifying."

Intellectual narrowing isn't the only unintended consequence. With automated tools making it easier to mass-produce manuscripts and conference submissions, journal editors and meeting organizers have witnessed a surge in low-quality and fraudulent papers, often produced at industrial scale.

"We've become so obsessed with the number of papers [that scientists publish] that we are not thinking about what it is that we are researching—and in what ways that contributes to a better understanding of reality, of health, and of the natural world," says Northwestern's Nunes Amaral.

At stake is the trajectory of scientific inquiry. "Certain types of questions are more amenable to AI tools," says Catherine Shea, who researches organizational behavior at Carnegie Mellon University's Tepper School of Business, in Pittsburgh, and who also was not involved in the research. In an academic environment in which papers are the main currency of success, researchers naturally gravitate toward the problems that are easiest for these tools to crank through and turn into publishable results. "It just becomes this self-reinforcing loop over time," she adds.

Models trained on abundant existing data excel at optimizing well-defined problems: predicting protein structures, classifying images, extracting patterns from massive datasets. Some systems have also begun to propose new hypotheses and research directions—a glimpse of what some call an "AI co-scientist."

But such systems—and the scientists who rely on them—are unlikely to venture into poorly mapped territories where data are scarce and questions are messier, Evans says, unless they are deliberately designed and given an incentive to do so. The danger is not that science slows down, but that it becomes more homogeneous—a high-speed version of the same narrowing Evans first documented when search engines replaced library stacks.

"This is a really scary paper to think about in terms of how the second- and third-order effects of using AI in science play out," says Shea.

Whether this trend persists may depend on how the next generation of AI tools is built and deployed across scientific workflows.

In a paper published in December, Bowen Zhou and his colleagues at the Shanghai Artificial Intelligence Laboratory, in China, argued that the application of AI in science remains fragmented. Data, computation, and hypothesis-generation tools are often deployed in a siloed and task-specific fashion. This limits knowledge transfer and blunts transformative discovery. But when those elements are integrated, AI-for-science systems help expand scientific discovery, says Zhou, a machine learning researcher who previously served as chief scientist of the IBM Watson group.

Evans doesn't think the problem is baked into the algorithmic design of AI. What matters most, he says, is overhauling the reward structures that shape what scientists work on in the first place.

For now, Evans says, the challenge is to deliberately redirect how AI is used and rewarded in science: "In some sense, we haven't fundamentally invested in the real value proposition of AI for science, which is asking what it might allow us to do that we haven't done before." ■