Just ‘42’ tells us nothing.

Thomas Kahle
Otto-von-Guericke Universität Magdeburg
thomas.kahle@ovgu.de

Mateusz Michałek
Universität Konstanz
mateusz.michalek@uni-konstanz.de

July 22, 2026

Everybody knows that the answer is 42, and we are looking for the questions. This joke from Douglas Adams’ *The Hitchhiker’s Guide to the Galaxy* foreshadows mathematics in the age of AI agents: What is the value of an answer when producing it is no longer the distinctive skill of a mathematician?

About a year ago, one of us and Christian Stump wrote about mathematics in the context window. At the time, much of the mathematical community met AI with ridicule or indifference. The modest aim of that post was to urge mathematicians not to look away while something important was happening in their field. In only one year the landscape has changed significantly. AI-assisted research now leads to groundbreaking work such as the recent counterexamples to the Erdős unit distance conjecture and the Jacobian conjecture. Humans are being outcounterexampled as Kevin Buzzard put it. Tasks that recently looked safely beyond the machines no longer do. What will mathematics be now and how do we structure our community?

Output versus competence

We have already all seen this for our students, and now we see it among each other: Just because somebody produces good written results, it need not mean that they are good.

AI started out as text generation but by means of coding agents, text generation turns into action generation because the generated text now is commands typed into a computer. Thereby many common activities such as running computer experiments, web searches, reading PDFs, and thus entire research pipelines can be automated. And what is it really, what we do when we do mathematics research? Is it not mostly just trying the next most reasonable thing? We definitely produce a sequence of actions, often involving computers. LLMs do that too, and they can be directed by good mathematicians or maybe just anybody who knows how to steer agents.

Now the question arises: Should we give the Fields Medals or the best jobs to people who produce the best results, no matter how they produced these? Good research output is not the consequence of being a good mathematician (whatever that means!) anymore, at least not if we apply old standards.

Until recently, a research paper quietly carried two claims. 1. The explicit mathematical claim: here is a theorem and here is why it is true. 2. An implicit biographical claim: the authors acquired the understanding and judgment needed to produce it. Journals examined the first claim and hiring and promotion committees used it as evidence for the second. This was always an imperfect but quite useful proxy. AI breaks the connection between those two. A nonexpert can already produce text that looks expert, and an expert can use agents to multiply their output. Knowing how to direct models, writing and reading code, organizing data, and recognizing promising answers, all these are now very useful competences, sometimes more useful than being able to crack the mathematical nut, that is finding the proofs.

Only a year ago, most AI-assisted mathematics was either dull or just wrong, but now mathematicians around us all report the same: trying the latest models on their hard problems yields significant new insight. Researchers who have long left academia are also coming back and submitting the solutions to the open problems of their PhD theses from 10 years ago. To say it again: These submissions are not slop! They are often correct, overcome major technical challenges, discover new connections and are mathematically very exciting. Some even come with Lean formalizations now. As for attribution, some authors openly describe their AI use, but there are also many more papers without any AI declaration where one wonders how exactly now the authors came to all these new insights.

Truth, usefulness, authorship, and academic credit are all important questions that arise in this context. If most of the decisive work was done by a computer system, what exactly does the existence of a result or paper tell us about the people whose names are on it?

Professional road cycling after the arrival of EPO offers an uncomfortable analogy. This drug significantly increases the number of red blood cells and was basically undetectable for years. Everybody was using it and then the ranking of the Tour de France no longer reflected only who was the strongest cyclist. It became almost more important whose body reacted best to the drug, and who had the best doctors to secretly administer it and deal with the side effects.

AI is not EPO, and we should not ban it (which would also be impossible), but there is structural similarity. Without changes, academic rankings soon all select for access to systems and skill at operating them while pretending to measure something else.

The photograph and the mountain

Some may have called it a mistake all along, but we have confused the record of mathematics (= papers) with mathematics itself. A paper is a radically compressed account and usually removes the failed approaches, badly chosen definitions, misleading examples, half-formed intuitions, confusion, frustration, and many other all-too-human aspects. What remains is the Instagram version of reality: only sunshine and happiness all the time.

A final proof is like a photograph from the top of a mountain. The photograph shows that the summit was reached and lets others see part of the view. But the photograph was never the point of climbing. But wait, for Instagram influencers the photo is the entire point! But they also like to take the helicopter up if they can afford it.

Research changes the researcher. When wrestling with a problem, we develop a taste for which examples matter, which definitions are natural, which patterns we find beautiful and insightful. Climbing and problem solving both develop patience with uncertainty, modesty and the ability to revise a flawed plan. Solving problems also produces a dense web of understanding from which the next question can grow. These changes are not side effects of research. They are the main product.

AI can certainly help on this journey. It can suggest an example, check a calculation, expose a gap, retrieve a forgotten lemma, etc. But AI can also transport us directly to an answer and leave us unable to recognize the landscape. Even if we appreciate knowing the answer, we must go back and climb the mountain again, which might be easier after getting a helicopter view of what it entails. Some think the distinction is between using or refusing AI. For us it is between climbing and taking a heli to the top. Are there AI uses that enlarge our agency and let us build expertise to climb even higher mountains? We should be honest about how hard that is. Who, with a free helicopter parked at base camp, keeps climbing on foot? It is the same temptation that the peloton faced with EPO: the shortcut is there, nearly free, and leaves no trace in the photograph at the summit. We also reach for the helicopter these days, often to the limit of what a monthly subscription allows. The pull is real, and pretending otherwise is how one loses the mountain without noticing.

Who will ask the questions?

The opening joke was that the answers are cheap and the questions are scarce. Current systems are extraordinarily good at answering short questions. When a question is underspecified, an LLM often attempts some answer instead of becoming curious. This is not surprising given that its training data is the Instagram version of math: papers with success stories.

Could models be trained to ask interesting questions and to challenge their own assumptions? Many mathematicians today say that the machine can never be creative, just as yesterday they said that the machine could never solve a really interesting problem. We also thought LLMs were too stochastic to ever reliably use a computer, and here we are with coding agents that wrap the model in enough deterministic scaffolding to keep it on track for hours without major error. So AI may become better than us even at activities we now regard as distinctively human, including some forms of beauty and creativity. That makes it more urgent, not less, to say why those activities matter to us, and in what sense they belong to us.

There is a further scary possibility. Chess and Go programs first learned from human games, but their superhuman play emerged only once they trained beyond the human record. Imagine a mathematical system solving a few open problems with genuinely new methods. A successor trains on those methods before people have absorbed them, and solves a few more. Repeat the cycle not once but a thousand times. The result might be a body of effective mathematics whose methods and concepts are inaccessible to us, perhaps even hard to express in human terms.

This is speculation, but it needs no machine consciousness and no ominous AGI; the relevant selection pressures need not even sit inside a model. If machine mathematics does pull away, two aspects of mathematics that appeared inseparable become different: extending the collection of true mathematical statements and extending human understanding. There is a chance that we can value the first without giving up the second, but it will be very hard. And will mathematics still be fun if it comes down to only finding the questions to ready-made answers?

What should we pass on?

This brings us to education and to the next generation. Mathematics used to train people for sustained, precise, independent thought. A society should want some of its members to be able to build abstractions, remain with a difficult problem, detect a seductive error, and reason beyond what its most convenient tools already provide.

We therefore need to teach students both to use AI and to survive without it. They should learn to interrogate and verify claims. They should also have protected occasions to struggle before asking for help, to construct examples by hand, and to explain an argument in their own words. The ordering matters: a tool used after an honest attempt can extend thought, while the same tool used before the problem has taken shape in the mind can prevent that thought from beginning. If the show *Jeopardy!* has taught us anything, then finding the question for a known answer yields dumb questions.

We believe that the students' argument, that if the professor uses AI, they should also be allowed to, is flawed! Today's professors had the chance to develop crystallized intelligence before AI was available. It is a hard reality that developing any skills will require extreme focus on exactly that development rather than working goal-oriented towards the completion of projects. Again, if all you want to reach is the top, the helicopter will always be easier.

And then, how do we assess who has made progress? Who deserves a prize or a position? Hiring will have to change. Counting publications was never good enough and it will get a lot worse. Reading only the best paper of a candidate will soon be insufficient too! One could try day-long mathematical assessments during the hiring process, but who has time and energy for that? We do not need to turn academic hiring into an Olympiad, but we certainly need more direct evidence of expertise. This can be gathered in extended mathematical conversations, serious seminar questions, explanations of unfamiliar ideas, diagnosis of plausible but incorrect arguments, and transparent accounts of how a piece of work was done. Publications will remain evidence, but they will no longer interpret themselves. We need a new social network on top of just the publications. We can imagine that reference letters could become more important again as they convey trust. But then, direct assessment starts a new reputation game or in the worst case the old-boy networks in which we know who is good because the right people say so. Social science has studied such patterns. In a study on hiring bias, evaluators shown otherwise comparable CVs reliably favored the man and justified it by whichever credential he happened to hold; the bias disappeared when they had to fix their criteria, and their weights, before seeing any candidate! So we should not have more metrics but commit in advance, what competence we want and how much a piece of evidence counts.

This will not be easy. Building this social network requires an honesty and modesty that seems very challenging. Our moral standards might also need to grow and this happens at a time in which doctoral programs and mathematics departments are already under financial and political pressure. On top of that, the AI companies advertise left and right that they can provide computed expertise that makes human expertise obsolete. It is hard to grasp, but the same AI companies who provide us with powerful tools also advertise for our abolishment.

If universities see mathematicians merely as producers of proofs, then a machine that produces proofs will look like a reason for cuts in the math departments. We need to make a better case: mathematics departments sustain human mathematical culture and thereby a culture of critical thought and general expertise. Math departments train people who can understand, criticize, transmit, and eventually redirect any given subject.

Journey above destination

We are not arguing for the preservation of some romanticized version of the past. Mathematicians were happy to stop factoring large integers by hand, and we should be happy when AI removes further drudgery. We also look forward to answers to questions that we thought we might never live to see resolved. We are very excited about the new insights we ourselves gain from AI systems. The opportunity seems enormous. But mathematicians who tried to publish hand-computed factorizations after the advent of computers would rightly struggle to do so. There is turmoil ahead for the publication system. Some mathematicians, especially those outside academia, might decide simply to ignore the arcane publication system altogether. Even greater dangers would arise if we threw established quality-control systems overboard. To keep in control, mathematicians also need to build expertise in developing and running large-scale AI models. We recommend that people try to run their own open-weights models and maybe even train some. Understanding how understanding works might be a big challenge, but mathematics should be at the center of it.

The right measure of AI collaboration is not only the quality of the answer! We should also ask what happened to the people involved. Do they understand more? Can they explain why the question matters, recognize when the proposed route fails, build on the idea, teach it to someone else? Are they more capable of beginning the next journey?

The journey to a result becomes the part we must learn to value explicitly. The destination may be a theorem. But it is the journey that makes a mathematician.