

Mathematics in the Context Window

Thomas Kahle 

thomas.kahle@ovgu.de

Otto-von-Guericke University

Magdeburg

Christian Stump 

christian.stump@rub.de

Ruhr University Bochum

2025-08-15 (v1.0)

ABSTRACT

Starting from successes at math competitions, large language models are slowly making their way into math research. We share some thoughts about this process.

1. Introduction

A revolution is happening in the way math research is conducted – or is it? We believe that the term revolution is too bold for the current AI based advancements, but the scientific evolution is moving forward. And our community should be an active player in the unfolding of this story.

In the past months somewhat outrageous claims have been made about the capabilities of *large language models* (LLMs) and their *large reasoning model* variants (LRMs) for solving research level math problems. We, the two authors, had many discussions about the possible uses of LLMs and LRMs in math research. We structure our thoughts, discuss aspects we can agree on, and dare to envision to how math research changes in the near future.

Before doing so, we start with the following outlook to non-scientific usages of these models. Sam Altman, CEO of OpenAI is speculating about AI accelerating AI in a recent blog post: *There are other self-reinforcing loops at play. The economic value creation has started a flywheel of compounding infrastructure build out to run these increasingly-powerful AI systems. And robots that can build other robots (and in some sense, data centers that can build other data centers) aren't that far off* [1]. It is not hard to find even crazier projections among AI enthusiasts that promise anything from doomsday to heaven on earth [2]. As a consequence, we always need to carefully think about our relationship to AI companies, their political agendas, and their ways to accumulate attention and unfathomable amounts of venture capital, in parts with the promise to become better at math as a step towards the illusive general AI. It is our task to gain agency in these developments.

Many scientists and mathematicians in particular are already thinking about this relationship: in the week this opinion piece is published, a “safeguarded AI” meeting with “50 category theorists and software engineers” is taking place in Bristol [3].

2. Language models “doing” mathematics

In a recent article, the popular science journal Scientific American chose the headline *At Secret Math Meeting, Researchers Struggle to Outsmart AI – The world's leading mathematicians were stunned by how adept artificial intelligence is at doing their jobs* [4]. The basis of this

article was the Frontier Math symposium which took place in May 2025 at UC Berkeley. This was a meeting of representatives of OpenAI with mathematicians with the goal to create a more challenging math problem set for AI benchmark and training. This headline showcases the excitement and hype that evolved around AI “doing” mathematics specifically.

We are naturally lead to questions like what “doing” mathematics is. What is “our job”? Are math competitions “doing” mathematics? On July 21, 2025 Google and OpenAI both reported achieving gold medal level at the International Mathematics Olympiad [5]. To give full credit to the leading high-school students, gold level here really means that the LRMs were involved in a 45-way tie for 27th place. In any case, LLM performance in such competitions is great for flashy headlines, and curiously such news appear to be much less prominently presented for the chemistry olympiad or for poetry slams. There must be something about mathematics that makes it special.

All this media buzz is in stark contrast to the impressions of working mathematicians in their daily interactions with LLMs. Many even despise AI and LLMs, want it to go away, make fun of the hallucinations, etc. Long before all these developments, in his influential essay “On proof and progress in mathematics” William Thurston highlighted that mathematics as a whole is a social endeavor, the collective human effort to further understanding rather than earning “theorem proving credit” [6]. Today, there is also a group of mathematicians who acknowledge and follow the progress that has been made. This group includes prominent mathematicians like Terence Tao. He publishes his comments on the open social media network Mastodon since 2022 [7].

We have all used LLM chat bots, maybe in their newest versions of LRMs such as chatGPT-5 (OpenAI), DeepSeek, Gemini 2.5 (Google), or Claude (Anthropic). Such models automatically query one or more LLMs multiple times to refine a given request and steer the LLM into a certain direction. This has intended similarities with human thought process, although it ultimately relies on statistics of language [8]. LRMs also output their “chains of thought”, as a kind of story for the user to read along. Naturally the resource use of these models is a factor between 10 and 100 higher than those of classical LLMs and the generation time can be several minutes. This way, LRMs can solve many textbook exercises from graduate level math courses. At the same time problems remain. Reasoning models (like humans!) can show significantly reduced math performance when irrelevant information about cat behavior is presented together with a math problem [9].

The progress that AI has made in answering math questions is stunning. But are these really breakthroughs? Are they showing that general intelligence is near? We argue that they are not. There are two principal reasons why mathematics is an extremely lucrative target for the AI companies and their advertising. First, mathematics is perceived as something complicated, something for geniuses, something for which a special talent is necessary. Second, LLMs are trained to produce language. And mathematics uses a very formalized, almost programming-like language. Sticking to the rules of that language is something that humans have to learn laboriously to become able to express their thoughts. Scientists have long been using the writing style and other language clues to smell out bad arguments from bad style [10]. This correlation of correct arguments and correct language is now broken by the language models. Their style and command of the language of mathematics is often flawless and this is entirely independent of whether the output is nonsensical. The formalized language used in mathematics benefits the application of LLMs to mathematics research as opposed to other fields in which creative

writing is more important. Consequently, the AI companies can leave the strongest impression of intelligence by focussing on mathematics rather than essay writing, say.

3. Then what is mathematics?

In contrast to the Math Olympiad, research mathematics is not a problem solving competition. Many problems are hidden in layers and layers of definitions. Consider the following.

Theorem. Every squircle is a squorcle.

Proof. By Lemma 7, every squircle has squocorable minors. Now, applying the squorcle-glueing process from [XYZ42, Cor. 11.4] to the squocorably-distinguished minors implies the result. \square

This is of course nonsense, but structurally this is what many statements in mathematics look like. Layers and layers of definitions have been piled up to make the final desired result expressible in the most simple way. Mathematicians even optimize the line endings, so that the q.e.d.-box “ \square ” neatly fills the last line of the proof. All the work for this theorem has been done before somewhere. The process of mathematical discovery is as much problem solving as it is inventing and refining the language to express our thinking, our ideas, our visions. Mathematical progress is as much in the definitions, as it is in the results and in the proofs. In his book *Mathematica* David Bessis even argues that research in mathematics is all about *building* a mental image and that expressing this image can be very troublesome [11]. In a recent essay, Asvin G highlights the experimental nature of mathematics and claims that definitions are APIs through which we can interact with the mental images [12]. In contrast, results in other sciences are based on experiments and observations of nature. Until Sam Altman’s robots work in labs, such research cannot be conducted by LLMs. This again highlights the quality of mathematics as a playground for LLMs.

Can the structuring of language be taken over by LLMs? And then, what about having ideas in the first place? What do we actually want to prove? AI systems generate output upon input. And it is output that is in a very precise sense from the same probability distribution as the input, heavily correlated with it. What is the trigger that generates new math research? Such a trigger could of course be some conjecture that we desperately need to know the answer to, like the Riemann hypothesis. For many conjectures in mathematics there is literally zero value in knowing that the conjecture is true without any insight why. Mathematics is furthering understanding paired with unexplainable human curiosity to come up with questions like *Is a random triangle more likely to be acute or obtuse?* [13]

AI companies view math through a problem and solution paradigm and this may not come as a surprise as the models are trained on math papers, our necessarily incomplete written records. These reduce mathematics to a stream of unnegotiable true statements, while not capturing the essence of curiosity and the many non fruitful alleys the authors visited along their way to the final form. How should an LLM ever genuinely say “Hmm, I’m not sure” or “We need to modify the definition for property X to be more natural” if there is no part of the “doing math” process in the training data? It would of course be trivial to make the LLMs say such things. But meaning it, and then supporting it with revised definitions based on mental images that makes the theory fall in place is fundamentally different from what LLMs do.

4. And what will mathematics be?

How many math papers does a mathematician read? For most mathematicians that number will be astonishingly low, both compared to the number of papers appearing on the arXiv every day, but even compared to how many they write. The LLMs have read *all* math papers. Mathematics has always been a frontrunner in open science. We have been early to openly share our preprints on the arXiv and, for what it's worth, the Elsevier boycott *The cost of knowledge* has been initiated by mathematician Timothy Gowers. We have no hesitation to share our work-in-progress knowledge in conferences and workshops, because of the abundance of problems. In math, if somebody else solves your problem you are usually happy (of course only if it is not your thesis problem shortly before submission). This openness is another reason the AI companies love mathematics. We produce all the training data and we hand it over for free.

Whom are we writing our papers for? If we want our papers to be read by humans, we have to write for humans. We will have to compete for the order of magnitude 10^1 papers that each of our colleagues reads every year. The hard reality is that most math papers are never read and hardly ever cited. So maybe the future of mathematics holds at least that all research somehow contributes to the progress of mathematics because the machine can read all our technical papers and thus everybody contributes the “machine knowledge”. But will better mathematics arise from such automated language statistics on our papers? And who owns that knowledge? Is it also public like our papers?

Undergraduate and PhD students have already grown up in an environment where LLMs have changed education in good and in bad ways. Students' AI adoption is clearly ahead of that of their teachers. Studying mathematics is about getting to the roots of problems, pondering ideas and trying different approaches, developing mental images and communicating them. And all this in a collaborative and interactive way, in a *social way*. The future generation of mathematicians should not be afraid of LLMs. Nothing will ever replace your genuine understanding and ideas when solving exercises, approaching open problems and identifying interesting research directions. But you might use LLMs to support this process.

A recent study on the effect that LLM usage has on the brain, has shown that participants who first approached a creative writing task with only their brains and then later used an LLM to question and refine their thinking scored very highly [14]. Their score was much higher than that of the converse group who started with an LLM and later was relying only on their brains. Another outcome was that the LLM users as a group produced much more homogeneous output. This is something to keep in mind for education. There are aspects of mathematics where homogeneity is desired and trained for. We largely want to use the same language, the same phrases to carve out the precision needed to express the logical structure of a complicated proof in text. But when we develop a whole new theory, a new realm of ideas, a new approach to the Riemann hypothesis, then in-distribution LLM generated text will never cut it. We must train our students to think above the LLM level, the same way we are now routinely educating pupils in mental calculation despite the existence of calculators.

Even if some mathematicians wish, language models and AI will not go away. Maybe they lose popularity, maybe progress grinds to a halt, but what is already possible now does have value, even if only for humans to reflect on their own thought processes. LLMs change the way we do mathematics. But such changes have happened several times in the past. Take the prime number theorem. In the 18th century, humans spent lots of time to determine factors of ever

larger integers. For example, in 1777, Austrian Anton Felkel published a table giving complete prime factorizations of all integers not divisible by 2, 3, and 5, from 1 to 408,000. Such data made it possible to see the prime number theorem and eventually prove it. Today a simple computer program solves the same task in under a second and of course chatGPT will happily write such a program for you.

LLMs can identify patterns. Not long ago, the second author solved a lattice path counting problem on the 2d integer grid. After feeding only the definition the LLM found a recursive structure, wrote down the functional equation, solved it and extracted the coefficients to provide a counting formula for the original problem. This is nothing we could not have done ourselves (in a few hours) using standard techniques. This shows that automatization of mathematics can go beyond exact computation. But when does adding all math papers into the training data turn into a “knowledge” reflecting human knowledge and understanding within the mathematics community?

Mathematics changes like everything in the world. We advocate for positivity and openness to change. We look forward to automating standard arguments, writing and implementing algorithms, finding counterexamples, and so on. We look forward to not spending days on writing a simple python script that in theory is super easy to write and run—we want to free our mind and time for creativity. Like many technologies before, (the steam engine, the calculator, the personal computer), LLMs only supercharge what humans have been doing already. LLMs supercharge generating text. We should think about, what such a machine could be good for in our research, what are the dangers, and how do we actively shape the future.

This is about us. We believe that there is an opportunity that mathematics becomes even more about creativity and beauty. We might have argument helpers that further standardize techniques that are often used and let us focus entirely on the novelty. We face the reality of LLMs and better today than tomorrow. Let’s get used to their existence, let’s try them out, let’s laugh about their stupidities and their non-human behavior. Let’s be critical of the economics behind all this, let’s build our own. Let’s focus on the potential for furthering understanding. **And never stop being creative and human.**

Bibliography

1. Sam Altman. [The Gentle Singularity](#). *Sam Altman’s Blog*. 2025.
2. AI Futures Project. [AI 2027](#). 2025.
3. John Carlos Baez. [@johncarlosbaez@mathstodon.xyz](#). Mastodon post. 2025.
4. Lyndie Chiou. [At Secret Math Meeting, Researchers Struggle to Outsmart AI](#). *Scientific American*. 2025.
5. Thang Luong and Edward Lockhart. [Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad](#). *DeepMind Blog*. 2025.
6. William P Thurston. On proof and progress in mathematics. *For the learning of mathematics*, 15(1) pp. 29–37. 1995.
7. Terence Tao. [@tao@mathstodon.xyz](#). Mastodon profile. 2025.
8. Andrés Castro Araújo. [LLMs for Researchers](#). 2025.

9. Meghana Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudan, James Zou, and Nazneen Rajani. [Cats confuse reasoning LLM: Query agnostic adversarial triggers for reasoning models](#). *arXiv preprint arXiv:2503.01781*. 2025.
10. Terence Tao. [Embracing Change and Resetting Expectations](#). Accessed: 2025-07-24. 2023.
11. David Bessis. *Mathematica: A Secret World of Intuition and Curiosity*. Yale University Press. 2024.
12. Asvin G. [The Unreasonable Effectiveness of Mathematical Experiments: What Makes Mathematics Work](#). *preprint, arXiv:2506.19787*. 2025.
13. Stephen Portnoy. A Lewis Carroll pillow problem: Probability of an obtuse triangle. *Statistical Science* pp. 279–284. 1994.
14. Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. [Your brain on chatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task](#). *preprint, arXiv:2506.08872*. 2025.