Corica, A., Parra, V., Schiaffino, S. and Godoy, D.: Friend or Foe? Reflecting on the role of generative artificial intelligence in mathematics education.
For the learning of mathematics: An international journal of mathematics education. Vol. 46(1), pp. 23–24.
Pre-print draft version.

Friend or Foe? Reflecting on the role of Generative Artificial Intelligence in Mathematics Education

Ana Corica[1,2,3], Verónica Parra[1,2,3] , Silvia Schiaffino[1,3] and Daniela Godoy[1,3]

[1] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[2] *Núcleo de Investigación en Educación Matemática (NIEM), UNCPBA*
[3] *Instituto Superior de Ingeniería de Software (ISISTAN - CONICET/UNCPBA)*

The advent of generative Artificial Intelligence (AI) has disrupted many areas, but particularly the field of education. Chatbots powered by Large Language Models (LLMs), trained with massive volumes of data to recognize and generate content as a response to natural language queries, have become the fastest-adoption technology in history, quickly making their way into classrooms.

The role of technology in mathematics education has long been a matter of discussion. While some authors argue that computers and calculators can hinder reasoning processes because students need to first understand mathematical knowledge, others argue that technology can introduce new forms of teaching mathematics and serve students in different contexts. Ramirez (2020) stated that instead of assuming that the mere use of technology would lead to changes in learning, teachers who wish to use AI technology effectively must transform it into a valuable asset for teaching and, therefore, learning.

Most of the discussion about this issue has, so far, occurred in the context of traditional Information and Communications Technology with traditional software (e.g. Geogebra). Generative AI is, in many regards, a distinctive technology both in its relation with mathematics and in social matters. Here we discuss the novel challenges it poses and the new possibilities it opens to teaching and learning mathematics.

LLMs for mathematics problem-solving

LLM mathematical problem-solving capabilities are fundamental to any application of generative AI in mathematics education. Understanding their possibilities and limitations is critical to take advantage of their potential and, more importantly, to avoid misconceptions that can entail serious risks in educational settings.

LLMs' performance in mathematics is usually evaluated with benchmarks of word problems. Every LLM release is commonly followed by claims of outperforming previous benchmarks, sometimes accompanied with cherry-picked examples in which the LLM performs remarkably well. However, as argued by many authors (e.g. Mirzadeh et al., 2024) there is substantial difference between genuine logical or symbolic reasoning and the probabilistic pattern-matching approach of LLMs. Instead of reasoning about the concepts involved in a problem, LLMs replicate abstract reasoning steps matching those on training data, probabilistically creating the text by predicting the next likely word given the previous sequence.

All in all, LLMs' answers are inherently unreliable when it comes to complex mathematics problems. Jiang et al. (2024) demonstrate, with statistical guarantees, that LLMs success in reasoning largely depends on recognizing superficial patterns with a strong bias in the input, thereby raising concerns about their apparent reasoning and generalization abilities. Moreover, Shojaee et al. (2025) showed that even the more advanced so-called

reasoning models fail to develop generalizable problem-solving capabilities for planning tasks and even drop their performance to zero beyond a complexity threshold. This can be even worse in fields such as geometry where spatial notions are involved (Parra et al., 2024b). Even though LLMs are unreliable mathematics problem solvers, their sophisticated approach to approximate answers can still be fully leveraged in several ways.

From a mathematics education perspective, the fact that LLMs can deliver erroneous, but plausible-sounding answers presents several implications and challenges. Unlike other conventional technologies applied in mathematics teaching, generative AI introduces uncertainty as a factor requiring teacher and student awareness. Misconceptions in this regard are frequent. The functionality of chatbots is, for example, many times confused with that of traditional search engines, conferring them unwarranted credibility. Moreover, interaction with chatbots can negatively influence teachers' actions. Noster et al. (2024) show that the rate of pre-service teachers providing an incorrect answer is high when having been presented with an incorrect answer by an LLM. Thus, it becomes essential to include basic notions of AI in pre-service teacher programs. A general understanding about how AI predictions are built, not necessarily the algorithms' inner workings, would allow teachers to interpret and critically evaluate their outputs. More elemental skills like prompt engineering could be a valuable asset too, but only as a means for interacting with LLMs in the constructing of solutions. For example, ‘few-shot chain-of-thought prompting’ requires users to provide examples expressing the thought process for an AI to replicate. This in turn can help students to articulate the steps toward a solution, indirectly eliciting mathematical thinking.

Note that so-called mathematics tutors that proliferate these days, chatbots trained to answer mathematics questions via fine-tuning (adaptation of a pre-trained model to a specific domain by adding new data), inherit the same limitations of general-purpose chatbots as ChatGPT. Tutors are intended to serve as study companions, offering explanations or self-quizzing for exam preparation, being sometimes personalized to tailor the interaction to the students' individual needs. The conversational nature of AI is expected to enhance students’ engagement and motivation, but their proficiency in tackling mathematics problems is still lacking, dangerously exposing students to possibly inaccurate content.

Departing from the mainstream research on chatbots, neuro-symbolic approaches which integrate symbolic reasoning with deep learning, such as AlphaGeometry (where a LLM is meant to guide logical deduction), seem to be paving the way to safer and more powerful uses of generative AI in mathematics. Vuong & Ho (2024) even wonder if AlphaGeometry is signaling the beginning of the end of mathematics education. The authors argue that increasingly prevalent smarter machines will make it harder for learners to find the true meaning of learning mathematics, and for educators to motivate and guide them. Generative AI therefore requires a paradigm-shift in mathematics education back to humanistic values.

LLMs for educational content generation

One of the main concerns for mathematics teachers today is managing practices that promote mathematical understanding through the design, analysis and evaluation of study tasks. As content generators themselves, LLMs seem to be natural candidates to alleviate teachers' workload. To thoroughly analyze this aspect, we need to distinguish simple

everyday mathematics tasks from the implementations of more advanced mathematics education theories.

At the instrumental level of creating problems or activities, LLMs can be good allies for teachers. Generative AI can even have a place in teacher training as long as pre-service teachers insightfully and critically reflect on the correctness, usefulness, and overall value of the output produced by chatbots. LLMs' ability to generate realistic and contextualized mathematics problems can enrich the creation of tasks, fostering deep mathematical thinking and training problem-solving skills. The creation of class designs can also be aided by chatbots. In designing an annual plan for third year secondary students in Argentina, for example, we found chatbots useful for generating a first draft that distributed study topics across the established timeframe. However, the expertise of the mathematics teacher was required to carefully analyze the proposal as chatbots often include notions that exceed the targeted academic level and allocate time unrealistically. They, however, offered a variety of assessment strategies, inspiring mathematics teachers with wider options aligned with their everyday practice.

While generative AI can be a time-saving tool for generating mathematics activities, there are a number of caveats concerning the didactic aspects of mathematics education. Parra et al. (2024a) investigated the behavior of several chatbots when asked to create a class proposal following a specific didactic theory, the Theory of Didactical Situations (TDS) of Guy Brousseau. The chatbots were able to provide reasonable answers covering general aspects of the proposals such as class duration, concepts to study, or technological resources to use, but they failed to adequate the proposal to the didactic framework prescribed. The generated proposals matched the components of the framework (e.g. action, formulation, validation and institutionalization situations) but rarely represented their intended meanings. Moreover, in some answers the proposals mixed notions from other theoretical references, such as the instrument-object dialectic. It is worth noting that as a word prediction mechanism, LLMs do not have a general comprehension of any given theory and they do not apprehend any explicit definition (e.g. TDS situations). Considering also the scarcity of documents referring to this subject on the Web, LLMs were only able to grasp some of the vocabulary involved in the theory for the formulation of the proposals. The consequences of these results have different implications depending on whether pre-service or in-service teachers are the ones asking for a TDS proposal. In-service teachers already prepared with a solid knowledge of theoretical frameworks can critically analyze the delivered answers. In contrast, the use of these tools by pre-service teachers must be carefully supervised as class design with chatbots might alter the process of acquiring background on certain didactic theories.

Rethinking the mathematics curriculum

Generative AI poses additional challenges to those responsible for designing mathematics curricula. A quick look at the curriculum design of Buenos Aires Province in Argentina, which dates back 20 years, provides plenty of task examples that can be efficiently solved by chatbots, with a very small margin of error, like solving a system of linear equations with two unknowns. In the presence of AI, curricula focused on the repetitive application of simple algorithms seems obsolete. Instead, they should focus on tasks seeking a functional study of mathematics, with emphasis on interpreting and formulating resolution techniques as well as creative thinking, exactly the kind of tasks LLMs struggle with and, consequently, less prone

to being solved by chatbots. Moreover, new strategies must be conceived for integrating chatbots into mathematics instruction by shifting the focus to error recognition, verification of reasoning steps (not just final results) and comparison of multiple resolution strategies.

Concluding thoughts

Overall, these issues reveal the liabilities of generative AI in mathematics education. On the one hand, its lack of true reasoning and deep understanding of the world make chatbots inherently prone to inaccuracies in mathematics, and hence, untrustworthy companions for helping students with tasks without proper supervision. Far from constituting a viable substitute for mathematics educators, the unreflective use of generative AI tools can hinder the proper acquisition of mathematical concepts and contribute to the formation of misconceptions. Nonetheless, chatbots can save teachers time and enhance students' engagement given their widespread adoption.

On the other hand, the particularities and difficulties of the Didactics of Mathematics clearly goes beyond the simple question-answering scenarios proposed by chatbots. Thus, it becomes imperative to define the role of generative AI in the context of constructivist approaches for teaching mathematics. The questionable credibility of generative AI can be leveraged as a catalyst for critical analysis and validation while constructing new mathematical knowledge. Finally, curriculum designs require urgent revision to place greater emphasis on deep mathematical thinking and to move away from tasks that are easily solvable by chatbots.

An additional warning should be delivered to AI developers and practitioners building software based on LLMs for mathematics education. It is essential that they fully commit to responsible AI principles in development as these applications have an undeniable influence on teaching-learning processes. This includes effectively communicating the capabilities and limitations of LLMs, avoiding anthropomorphic interpretations and human-like attributions. Crucially, educators must be kept in-the-loop when developing applications for this field.

References

Gordon, C. (2023, February 2). ChatGPT is the fastest growing ap in the history of web applications. *Forbes*. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W., Taylor, C. & Roth, D. (2024) A peek into token bias: large language models are not yet genuine reasoners. *EMNLP*. 4722–4756.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S. & Farajtabar, M. (2024) GSM-Symbolic: understanding the limitations of mathematical reasoning in large language models. *arXiv:2410.05229*.

Noster, N., Gerber, S. & Siller, H-S. (2024). Pre-service teachers' approaches in solving mathematics tasks with ChatGPT. *Digital Experiences in Mathematics Education* **10**, 543–567.

Parra, V., Sureda, P. & Corica, A. (2024a). Teoría de situaciones didácticas e inteligencia artificial: diseño de propuestas para enseñar las nociones de muestra y población en educación secundaria. *Revista UNO* **104**, 43–50.

Parra, V., Sureda, P., Corica, A., Schiaffino, S. & Godoy, D. (2024b). Can generative AI solve geometry problems? strengths and weaknesses of LLMs for geometric reasoning in Spanish. *International Journal of Interactive Multimedia and Artificial Intelligence* **8**(5), 65–74.

Ramírez, P. (2020) A reflection on Ros Sutherlands book: education and social justice in a digital age. *For the Learning of Mathematics* **40**(3), 38–39.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. & Farajtabar, M. (2025): The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv:2506.06941*.

Vuong, Q-H. & Ho, M-T. (2025) The disruptive AlphaGeometry: is it the beginning of the end of mathematics education?. *AI & Society* **40**, 1571–1573.