THE IMPLICATIONS OF GENERATIVE ARTIFICIAL INTELLIGENCE FOR MATHEMATICS EDUCATION

Candace Walkington
Department of Teaching and Learning, Southern Methodist University
cwalkington@smu.edu

Generative Artificial Intelligence has become prevalent in discussions of educational technology. These AI models can engage in human-like conversation and generate answers to complex questions in real-time, with education reports accentuating their potential to make teachers' work more efficient and improve student learning. In this paper, I provide a review of the current literature on generative AI in mathematics education, focusing on four areas: generative AI for mathematics problem-solving, generative AI for mathematics tutoring and feedback, generative AI to adapt mathematical tasks, and generative AI to assist mathematics teachers in planning. I then discuss ethical and logistical issues that arise with the application of generative AI in mathematics education, and close with some observations, recommendations, and future directions for the field.

Generative Artificial Intelligence has taken the world by storm since the release of ChatGPT in November of 2022. This release marked an important milestone in the development of conversational Artificial Intelligence agents, driven by ChatGPT's ability to engage in human-like conversation and answer complex questions. Stakeholders immediately began imagining how these tools might be applied to education. It has been nearly two years since ChatGPT's release, and research is rapidly emerging on its implications for education. In this paper, I seek to summarize current trends and issues related to GenAI in mathematics education, since the release of ChatGPT.

Artificial Intelligence (AI), according to the U.S. Department of Education, is "automation based on associations" (Cardona et al., 2023, p.1). Generative AI (GenAI) is a class of AI that is capable of generating new data and outputs by learning patterns from training data. Large Language Models (LLMs) are one type of GenAI that "build sophisticated statistical predictors by identifying patterns in a massive set of human-curated training data" (NCTM, 2023, p. 1). What was so revolutionary about GenAI models like ChatGPT when they were released was the ability of a human user to *respond back* to the AI model and ask it to change elements of its response – this practice is known as *prompting*. This ability gave rise to *prompt engineering,* which is the process of constructing inputs for LLMs to elicit precise, coherent, and pertinent responses (Liu et al., 2021). This process allows users to iteratively refine the kinds of output the LLMs provide, customizing the LLM's work to their needs and context.

The U.S. Department of Education gives many possible benefits of AI in education – from increasing the adaptivity of learning materials to students' needs, to providing teachers support via automated teaching assistants, to better customizing learning resources to meet local demands (Cardona et al., 2023). NCTM's (2023) AI Position Statement further expands on these affordances – describing how GenAI can allow for the quick development of multiple problem versions to illustrate a mathematics concept, can efficiently design engaging, personally relevant

Kosko, K. W., Caniglia, J., Courtney, S., Zolfaghari, M., & Morris, G. A., (2024). *Proceedings of the forty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Kent State University.

questions tailored to individual students' experiences, and can generate rich mathematical explanations that adapt to students' current level of expertise. Both of these reports also detail the incredible risks that GenAI poses – including issues of bias towards marginalized groups, concerns related to privacy and surveillance, and GenAI's tendency to hallucinate and provide incorrect information.

I structure this paper by first discussing affordances and use cases of GenAI in mathematics education in four broad areas – focusing on mathematics problem-solving, mathematics tutoring and feedback, adapting mathematical tasks to learner needs, and supporting mathematics teachers in planning. I then move to discussing important ethical, theoretical, and practical issues to consider when implementing GenAI in education. I close by providing some observations and recommendations for the future of GenAI in mathematics education.

Generative AI for Mathematics Problem-Solving

GenAI programs like ChatGPT can have a wide variety of mathematics problems inputted into them and can not only generate an answer to these problems, but also give a detailed explanation of how to get to that answer. The latest version of the LLM GPT-4 (OpenAI, 2023) integrates computer vision, such that the AI can examine mathematical diagrams in addition to the problem's text. GPT-4 scores 700 out of 800 on the mathematics portion of the SAT (OpenAI, 2023), which is in the 89th percentile. This is an improvement on a score of 590 (70th percentile) that was achieved by GPT-3.5, its predecessor. GPT-4 also scores a 4 out of 5 on the AP Calculus BC exam, while its predecessor scored a 1. GPT-4 scores in the 80th percentile on the GRE Quantitative exam, with its predecessor in the 25th percentile. And while GPT-4o scored only 13% on the qualifying exam for the Math Olympiad, the new GPT-o1 model designed for complex reasoning scored an impressive 83% (OpenAI, 2024), although it is slower and more costly than its alternatives. These results paint an impressive picture of the capability of contemporary LLMs for mathematics problem-solving.

LLMs also seem to have a relatively easy time with typical K-12 mathematics word problems used in open-source curricula. For example, GPT-4 solved and generated explanations for middle school mathematics word problems from ASSISTments with only a 4% error rate in its mathematical explanations (Wang et al., 2024a). Interestingly, none of these errors in explanations were associated with incorrect answers, allowing researchers to conclude that the AI was relatively safe for use with middle school students. Other researchers have used GPT-4 to solve more difficult graduate-level mathematics problems, to test whether GenAI can assist mathematicians in their professional activities (Frieder et al., 2023). They found that while GPT-4 could solve undergraduate mathematics problems, it performed poorly on graduate-level work.

They concluded that GPT-4 can best be leveraged to act as a mathematical search engine and query databases of mathematical objects, rather than as a direct solver of advanced problems. Analyses have also been done into the nature of the mistakes GPT makes when solving mathematical tasks. Typical errors made by GPT-3.5 included using incorrect formulas or methods or unclear question definition, along with calculation errors and misinterpretations of the question being asked (Yen & Hsu, 2023). Mathematics problems that have a high number and

diversity of mathematical operations needed to solve them tend to be more difficult for LLMs, as are problems that utilize a conversion of a quantity that requires real-world knowledge (Srivasta & Kochmar, 2024). Math word problems with unfeasible solutions (i.e., the analytical solution is not practical with respect to the real-world context), that contain quantities in them that are not needed to solve the problem, or that involve a comparison between quantities are also more difficult for LLMs (Albornoz-De Luise1 et al., 2024). In addition, holding mathematical features constant, longer and more difficult-to-read word problems are harder for LLMs to solve. One reason why understanding the capability of LLMs to solve mathematics problems is important is because these tools are used by students as a form of assistance. Although for primarily symbolic problems computer algebra systems like Maxima and Wolfram Alpha are more accurate, LLMs offer the advantage of communication using natural language, and can explain different steps for problem solving. Thus, LLMs may be preferred by students over alternatives. Integrating computer algebra systems (Matzakos et al., 2023) or other secondary systems that can check the LLM's calculations (Yen & Hsu, 2023) with LLMs will be an important future direction to improve the reliability of these systems.

There is surprisingly little research on how GenAI can be used effectively by *students* to enhance their learning of mathematics. Barana et al. (2023) gave university students combinatorics problems to solve with the help of GPT-3.5. They found that although GPT-3.5 did not consistently achieve correct answers to the problems, the output given by GPT-3.5 could be leveraged by the students. The students used the output to generate ideas for how to approach the problem, to compare their reasoning with GPT-3.5's reasoning, to solve smaller parts of the problem, and to evaluate and then correct GPT-3.5's solution paths. This is an important example of how LLMs can support higher-level thinking in mathematics. Research has also been done with pre-service mathematics teachers using ChatGPT to help them solve mathematical modelling tasks (Naresh et al., 2024). The researchers highlighted that incorrect answers from the AI could be opportunities for student learning and could launch important mathematical conversations. The teachers also recognized that their students could self-explain the different steps shown by ChatGPT as an opportunity for deeper learning. In sum, more research on how students can effectively partner with LLMs to confront challenging mathematical tasks, like mathematical modeling tasks, is needed.

Generative AI for Mathematics Tutoring and Feedback

Several different online learning platforms have launched GenAI chatbot mathematics tutors, which communicate with students through text chat to assist them with solving mathematical tasks. The most well-known is Khan Academy's Khanmigo (Khan Academy, n.d.), which is free for teachers through a partnership with Microsoft, but has a monthly charge for students, families, and districts. Khan Academy describes how "Unlike other AI tools such as ChatGPT, Khanmigo doesn't just give answers. Instead, with limitless patience, it guides learners to find the answer themselves." Notably, Khanmigo is student-facing with no human directly in the loop – children interact directly with the GenAI conversational agent, relying on Khan Academy's safeguards to prevent inappropriate interactions. Khanmigo acknowledges that it will sometimes

make mathematical errors, with a disclaimer at the top of the tutoring screen reading "Khanmigo make mistakes sometimes."

Results are starting to emerge on the effectiveness of GenAI tutoring. Bastani et al. (2024) conducted a pre-registered RCT that involved nearly 1000 high school students and compared students learning mathematics over a semester with either: (1) GPT-4, (2) GPT-4 with knowledge of correct solutions and student mistakes, as well as instructions to not give students answers, or (3) a control condition where students used books and notes with no access to devices. They found that both versions of GPT-4 improved immediate performance, but that the version that lacked the safeguards actually harmed later exam performance by 17%. Additional analyses suggested that GPT-4 was being used as a crutch by students, and that they were often simply asking it for the answer without a substantial conversation. The enhanced version of GPT-4 with the safeguards did not offer any benefit on the exam over simply studying the text and notes without devices, and its effect trended slightly negative compared to the control group.

As little research exists on GenAI tutors, we can look to research on chatbot tutors that were built predating the rise of contemporary LLMs. A study that compared adults learning mathematics with a chatbot to adults learning with Khan Academy videos did not find significant differences in learning (Grossman et al., 2019), suggesting that the chatbot was generally not effective. However, a math tutoring chatbot for a younger population of elementary students showed some evidence of positive results for engagement and learning above a control condition with no support (Ruan et al., 2020). A follow-up study (Ruan et al., 2024) of elementary students found no differences in overall learning compared to a control condition with no support, but some suggestion of increased learning for students with lower pretest scores in the chatbot condition. At the secondary level, a chatbot implemented in ASSISTments was compared to students simply receiving static hints, and researchers found no differences in learning (Cheng et al., 2024). However, students who had interacted with the chatbot actually had *lower* confidence in their problem-solving, due to potentially becoming reliant on the chatbot's high degree of assistance.

None of these studies compared chatbots to human tutors, and overall, the evidence base for chatbot tutors does not seem particularly promising. However, we should not discount that many marginalized learners may not have access to human tutoring, and that LLMs' abilities to communicate in different languages may have important affordances. Butgereit and van Staden (2023) report on an implementation in South Africa of adult learners receiving mathematics tutoring through a version of GPT-4 configured for tutoring interactions. The tutoring was delivered in several different languages, including less-resourced African languages that typically perform less well in LLMs.

Given that there is little research on GenAI chatbots, I next look to research on whether LLMs can give actionable feedback to students on their mathematical problem-solving. In online learning platforms, generating text and images for explanations and hints to be administered when students need assistance can be time-consuming for curriculum developers. As a result, many curriculum developers are looking to LLMs to help with this process. Research suggests that GPT-4 has a tendency to over-identify instances of students making mathematical errors

(Karkarla et al., 2024). Other studies have examined the quality of GPT-generated explanations through ratings of explanation quality. Wang et al. (2024a) asked ten undergraduate mathematics majors to evaluate either explanations for middle school mathematics problems written by GPT-4 or explanations for the same problems previously written by educators. They found that the perceived quality of the explanations was higher for GPT-4 than for teacher-written explanations, potentially because the GPT explanations were seen as having a clearer, step-by-step approach. Prior research had shown that GPT-3's explanations for mathematics problems were rated lower than teacher-generated explanations (Prihar et al., 2023). Research comparing pre-service teachers' reactions to educator-generated hints versus hints generated by GPT-4 found that educator-generated hints were preferred in some cases, as they incorporated visuals, while the LLM-generated hints were preferred in other cases, as they often were more thorough and detailed (Gattupalli et al., 2023a). Other research on GPT-4 suggests that the hints it gives may be too procedurally-focused and are not always written appropriately to support students' reading needs (Gattupalli et al., 2023b). The rated quality of GPT-4-written explanations for middle school mathematics problems can be improved if the LLM integrates previous annotations of the student's work from experts into its reasoning (Wang et al., 2024b). This blending of human and GenAI capabilities may be a promising approach.

Research has also examined the learning implications of AI versus human-generated hints, in addition to preference scores. Pardos and Bhandari (2024) compared GPT-3.5-generated hints for mathematics problems in the OpenStax curriculum to human tutor-generated hints. They found that adult learners had higher learning gains in the GPT condition compared to a control condition with no hints, while the difference between human-generated hints and the control condition did not reach significance. However, they found that 32% of the hints generated by GPT were initially disqualified for inaccuracy, and that this percentage was reduced through the use of an LLM hallucination reduction technique. Overall LLMs seem to be improving in their ability to generate hints and feedback but work still needs to be done on ensuring the hints are accurate, conceptually-focused, and do not lead to over-reliance on the LLM.

Generative AI To Adapt Mathematical Tasks to Learner Needs

GenAI can also be used to adapt learning tasks to meet different learner needs. For example, students often struggle to read the text of mathematics word problems (Walkington et al., 2018), and LLMs have the potential to adapt problems to assist emerging readers. Norberg et al. (2024a) showed that having GPT-4 rewrite middle school mathematics problems to improve their readability resulted in similar effects on student performance as having humans rewrite the problems. They also found that compared to the original problems that had not been rewritten, the problems rewritten for improved readability using GPT-4 could in some cases improve students' mastery rates. Using earlier LLMs, like GPT-3, for the same kind of task resulted in less impressive results, where outputs had more error and noise (Patel et al., 2023).

LLMs can also adapt mathematics problems based on students' interests in popular culture areas like sports or music, or career areas like nursing or engineering – this is often called *context personalization* (Walkington, 2013). GPT-3.5 was used in a research study to rewrite probability

and statistics problems to correspond to undergraduates' career interests in areas like pharmacology or economics (Einarsson et al., 2023). The study found that the problems sometimes required revisions due to mathematical issues, and that while some students liked the relevant contexts, others did not like the extra length and complexity the personalized contexts added. However, the authors recognize that better results for accuracy with respect to mathematical issues may have been found with GPT-4. Indeed, another small study where GPT-4 was used to personalize secondary mathematics word problems to correspond to students' interests in areas like TikTok, results showed that GPT did not change the difficulty, intent, or values in the problem (Yadav et al., 2023). In a unique approach to personalization, Hwang et al. (2024) used GPT-3.5 to pose mathematical problems based on camera-captured images of real-world geometric objects, personalizing mathematics problems to objects in the learners' environment. They found that 5th grade students using the system outperformed a control group.

While research shows some effects of personalization and readability on student outcomes, it is also important to examine the perspectives of teachers. An interview study with teachers who taught 8th grade math in urban settings asked about the possibilities of using GenAI to personalize mathematics problems (Walkington and Bainbridge, under review). The study found that teachers felt it would be an effective way to draw upon students' real-world knowledge, activate interest, and allow for sense-making around mathematical problems. One teacher described how "If it's talking about a place, thing, or situation that they're actually familiar with, that they've actually had hands-on experience with, or have seen with their own eyes then, of course, it's gonna be a little bit easier for them to comprehend the problem," while another teacher said, "I think them being able to have a little bit of background knowledge makes word problems a little less scary, sometimes, too, where they feel like they understand it better." However, the teachers had concerns that LLM-generated problems would create greater reading burdens for students, that students still lacked important fundamental math knowledge, and that having different problem versions would translate into additional preparation time for teachers and/or create difficulty when going over problems as a class. One teacher described how "But if they're struggling in math, even giving them what they're interested in, it still may pose a challenge." Overall teachers showed some enthusiasm about the approach, mixed with concern that it would not solve the fundamental issues they were experiencing with their instruction.

Research has also examined partnering students with LLMs to engage in mathematical problem-posing activities (Silver, 1994), as a way to create personalized versions of story situations written by students (e.g., Walkington et al., 2024a). Norberg et al. (2024b) engaged middle school students in authoring their own personalized problems relating to probability and ratios and based on their interests, using GPT-4 to assist students. They found that students preferred more control over the personalization system and found slight increases in students' sense of belonging in mathematics. Zhang et al. (2024) conducted a study of 4–8-year-olds writing math stories while partnering with a GPT-4 agent to support their storytelling. They found that compared to a human partner, the LLM's assistance was actually better in helping students comprehend mathematical definitions, and comparable for supporting mathematics language learning and the generation of quality math stories. However, children interacting with

the LLM were less likely to provide substantial responses in the conversation and more likely to give uncertain responses, compared to when interacting with a human.

Another study examined middle school girls posing mathematics story problems with the assistance of GPT-3.5, as part of a project-based unit to create a pitch for a mathematics video game (Walkington et al., 2024b). The study found that the girls used prompting to fine-tune the mathematics story problems for their game and their pitch, and typically gave length and style parameters to the LLM. The girls found that using GPT-3.5 created story problems they felt were fun and engaging, was an efficient process, and that the problems had the potential to increase mathematical understanding. However, further analyses found that the quality of the mathematics story problems, in terms of their realism and correspondence to actual real-world objects and events, was relatively low and problems could contain mathematical errors or inconsistencies. Overall, partnering students with GenAI to accomplish complex mathematical tasks that include problem-posing activities may be an important future direction of GenAI research, if students have an appropriate understanding of the issues involved with using GenAI.

Generative AI To Support Mathematics Teachers in Planning

Research suggests that teachers work an average of 50 hours per week, and that only 49% of this time is in direct interaction with students. The rest of this time involves preparing lessons, giving feedback, doing administrative work, and engaging in professional development (Cardona et al., 2023). There has thus been interest in leveraging GenAI to make teachers more efficient during the 51% of time they are not directly interacting with students. Many tools have arisen that use GenAI to help teachers plan their lessons and accomplish logistical classroom tasks. One of the most well-known tools is MagicSchool (powered by GPT-4, among other models), currently advertised to be used by 2 million educators worldwide (MagicSchool, n.d.). The MagicSchool suite has over 70 AI tools for educators that "help you lesson plan, differentiate, write assessments, write IEPs, communicate clearly, and more." The base version of MagicSchool is currently free for teachers, with more advanced plans having recurring charges. However, there are many other GenAI tools for teachers, with Khanmigo, for instance, having a similar suite of free teacher tools (Khan Academy, n.d.). Gemini for Google Workspace (Google for Education, 2024) has also arisen as a major player in the "GenAI for Teachers" field. Gemini has easy integration with Google tools that are widely used in schools already like Docs, Sheets, Slides, and Gmail, as well as integration with Google's Gemini chatbot.

Research on pre-service and in-service mathematics teachers using MagicSchool (Beauchamp & Walkington, 2024) has examined the use of various tools to make mathematics tasks more relevant to students. This study found that teachers felt the tools had the potential to support students' motivation to learn mathematics and that the tools could increase the efficiency with which the teachers could generate tasks. However, the teachers found the tools limited in their support for English Learners and felt that some of the tasks did not accurately or deeply reflect elements of students' real-world experiences or had mathematical issues. Research has also examined pre-service elementary mathematics teachers using Khanmigo as a support for their mathematics learning related to number theory (Yilmaz et al., 2024). This study found that

teachers using Khanmigo appreciated the individualized learning and were comfortable sharing questions and struggles with Khanmigo. However, in some cases they found Khanmigo responses confusing or of questionable reliability and missed the human interaction of someone who gets to know them as students.

Research has also examined pre-service mathematics teachers using ChatGPT for lesson planning (Berryhill et al., 2024; Broutin, 2024; Kwom & Ko, 2024; Naresh et al., 2024). One study found that teachers asked ChatGPT for both mathematical and pedagogical knowledge, that they used ChatGPT as an assistant in developing and organizing lessons, they had ChatGPT simulate possible student responses, and they asked ChatGPT to validate or comment on their ideas relating to teaching (Broutin, 2024). The teachers used continuous prompting to adjust the output and ideas that ChatGPT generated, and they extensively modified the output from ChatGPT to meet their needs. Similarly, a study on pre-service mathematics teachers using ChatGPT highlighted that it can be used to anticipate student misconceptions and approaches to problems, and that ChatGPT can simulate being an age-appropriate student to assist teachers in planning (Naresh et al., 2024). Teachers can also use ChatGPT to generate culturally relevant word problems, with research suggesting that the LLM can be a helpful thought partner through the use of iterative prompting and revision (Berryhill et al., 2024). GPT can further be used to generate mathematics assessment items. Secondary mathematics teachers using GPT-3.5 to generate statistics assessment items felt variably in their desire to actually use the GPT-generated problems (Kwon & Ko, 2024). The teachers appreciated the creativity, efficiency, and specificity of GPT, in addition to its ability to produce anticipated student solutions. However, concerns were raised about GPT's mathematical errors, security and copyright issues, GPT's lack of transparency, its inability to know teachers' students and classrooms, as well as issues with item difficulty and discrimination. Overall, LLMs have some functions that will be useful to teachers in lesson planning, as long as the output can be modified and enhanced by the teachers themselves to best fit their needs.

Issues with the Use of Generative AI in Mathematics Education

A myriad of important ethical issues and concerns arise when applying GenAI technologies to education. Bender et al.'s (2021) groundbreaking paper describes some of these issues, highlighting the environmental and financial cost of increasingly complex and accurate GenAI that require more and more computing power (see also Li et al., 2023). In addition, the training data for GenAI is from large internet datasets that overrepresent people in positions of power in society, that show bias towards the inclusion of marginalized groups, and that include derogatory associations and stereotypes towards these groups (Bender et al., 2021).

Issues with training data may be of particular concern in mathematics education, as common textbooks (including open access textbooks) that GenAI is drawing from have been found to be culturally-biased. An analysis of the top 9 textbooks for 8th grade mathematics on EdReports. found that the majority of the problems in these texts were situated in White, middle-class American culture (Pruitt-Britton & Walkington, 2022). Many of the activities described in the story problems in these texts required wealth or transportation to participate in – such as a story

problem about renting a jet-ski or vacationing in the Poconos. The analysis also found problems with specialized non-mathematical language that would be challenging for English Learners. If these are the kinds of data that GenAI is trained from, then GenAI is likely to show these same issues and biases when asked to generate problems.

There are also concerns about LLMs having stereotypical negative associations with the subject matter of mathematics itself, given the commonness of mathematics anxiety and negative reactions to mathematics that are prevalent in society and thus in the LLM's training data. Abramski et al. (2023) studied the associations that GPT-4 makes with the academic subject of mathematics, and the degree to which these associations are negative or positive. They found that 10% of sentiments associated with mathematics were negative for GPT-4, compared to a surprising 50% in GPT-3.5. However, the authors still found that in GPT-4, "Math was associated with frustrating, anxiety, fearful, intimidating, confusing, and struggle. These negative associations were not found in the semantic frame of physics, whose negative associates were related to domain knowledge (e.g., chaos, nuclear)" (p. 15).

When discussing limitations of LLMs, Bender et al. (2021) further describe how, "Text generated by an LLM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that" (p. 616). Although these models can generate human-like responses, they are not reasoning or "thinking" in the way humans do. Indeed, the development of mathematics concepts themselves and the development of students' mathematics learning is situated in their individual and collective interactions with the physical world (Nathan, 2022). However, it has been argued that AI systems are "fundamentally incapable of understanding people's embodied interactions in the ways that humans understand them" (Nathan, 2023, p.1). These systems cannot account for forms of human reasoning that are non-verbal and non-pictorial, like gestures and actions.

In addition, in education particularly, there are concerns about the protection of users' inputs into the LLM, including privacy and issues of ownership of intellectual property (Gómez Marchant & Hardison, 2024). When an LLM collects data about young students to better adapt learning materials to student needs, issues of who sees the data and how it is deleted are paramount (Cardona et al., 2023). The rise of LLMs integrated into educational settings may also involve increasing possibilities for surveillance of both students and teachers, as the LLM collects data from multiple sources in order to best adapt instruction and assist teachers.

Further, research on using ChatGPT in mathematics teacher education has shown that ChatGPT can create developmentally inappropriate learning activities and materials that include mathematical mistakes. ChatGPT may create inappropriate materials, such as a middle school mathematics scenario about someone losing 5 pounds per month in a weight loss program (Sawyer & Aga, 2024). LLM-generated problems can also involve haphazard, rather than purposeful, choice of numbers, and LLM's lack of authentic connections to learners' lived experiences can "demonstrate a dangerous surface-level approach to culturally relevant pedagogy" (Gómez Marchant & Hardison, 2024, p. 3).

Indeed, Walkington et al.'s (2024b) study of middle school girls using GPT-3.5 to create

mathematics story problems, and Beauchamp and Walkington's (2024) study of teachers using MagicSchool to create relevant learning tasks, found that the problems created by these LLMs often involved surface-level connections to students' experiences. For example, one teacher asked MagicSchool to create a lesson relevant to their students' interest in Mexican-American rapper ThatMexicanOT, and it generated the following: "Students will create a Spotify playlist based on songs by ThatMexicanOT. Each song will be represented by a cylinder-shaped object, and students will calculate the volume and surface area of each cylinder. This activity will show students how real-world math concepts are used in creative ways, like organizing playlists based on their favorite music." Obviously, this scenario is nonsensical, as representing songs as cylinders and calculating their volume and surface area makes little sense. Similar issues were found for student-generated math problems in our study of middle school girls – GPT-3.5 generated the problem "Jack, one of the last five remaining humans, is determined to defeat the robot army by factoring the polynomial expression 2x^2 + 5x - 3, representing the robots' central control system. If Jack successfully factors the polynomial into its linear expressions (Ax + B)(Cx + D), where A, B, C, and D are integers, he can exploit the weaknesses in the robots' programming." This again is a shallow connection between the mathematics concepts and the real-world context.

Image-generating GenAI also exhibit significant bias. For example, Figure 1 shows the output that was generated when DALL-E3 was asked to create "An image of a room of mathematics educators attending the Psychology of Mathematics Education - North America conference in Cleveland, Ohio." The lack of diversity in the image is striking. A study of middle school girls using DALL-E2 reported that the girls recognized bias when the GenAI would generate mainly light-skinned images of humans, despite most of the girls being girls of color (Walkington et al., 2024b). One group of girls in this study described how the pictures of the "landlord" character in their game were consistently generated as older, White men. Gómez Marchant and Hardison (2024) further describe how Adobe Firefly's image-generating AI shows negative racial imagery and incorporates an anti-fatness bias. They asked Adobe Firefly to generate images of a mathematics teacher, and all the images were of White adults. In the mathematics textbook analysis mentioned previously (Pruitt-Britton &Walkington, 2022), it was found that the majority of images of humans in mathematics textbooks were of White, able-bodied people. Given that GenAI is largely trained on these kinds of datasets when generating images for mathematical problems, it is not surprising that the generated images lack diversity.

There is also a lack of guidance in schools about how to handle students using GenAI to complete their assignments. This can lead to disciplinary action that may have disproportionate impact marginalized students, specifically special education students (Laird et al., 2023). There is evidence that English Language Learners and neurodivergent students may be disproportionately targeted by AI detection tools (Gegg-Harrison & Quaterman, 2024). Further, there is evidence that Black students are more likely to be false accused of using GenAI tools to cheat (Madden et al., 2024).

Figure 1: DALL-E3 image of attendees of PME-NA 2024

Discussion

NCTM (2023), in their Position Statement on Generative AI, compares GenAI to advances in technology like calculators, search engines, and image-based solving systems like Photomath. They describe how these tools have the potential to reduce an emphasis on computation in mathematics classrooms and increase focus on creative problem-solving. NCTM (2023) also describes how such tools can create positive pressure for teachers and curriculum developers to pose mathematical tasks that are deeper and involve creative thinking and are thus less prone to being solved with LLMs. NCTM further describes how GenAI tools can shift the focus of mathematics instruction from *solving* tasks to both *solving* and *verifying* – an evolutionary change where students must critically examine outputs from LLM and engage in deeper reasoning.

This is a very optimistic and forward-looking account on how GenAI could be used to deliver on its promise to change education. The research that has emerged before and since this Position Statement, however, paints a different picture. There is certainly some important, emerging research happening that leverages GenAI to engage students in rich and meaningful mathematical problem solving – probably far more than is represented in this review, as results may not be published at this early stage. But much of the effort, funding, and emphasis in GenAI in mathematics education is being directed at creating AI chatbots or personalized feedback systems, and then making small incremental enhancements to these systems to more optimally respond to students' errors or assess students' knowledge. This may sound promising, but these

technologies are primarily being developed in contexts where students are solving simple, repetitive, skill-based mathematics tasks individually on a screen. Critiques of GenAI chatbots have been harsh, with Meyer (2024) arguing that Khanmigo "regards math as machine-executable code, a numbered list of steps that a machine can execute one at a time with any error bubbling up the stack and identifying the earliest step that produced it" and "regards students as buggy computers whose errors should be identified and corrected as efficiently as possible." As a result, Meyer describes how "The lie that Khanmigo perpetuates here is that 'math is about a huge number of small ideas.'" This then leads to the important question of, what actual, pressing problem in mathematics education is GenAI suited to solve?

A survey in 2023 of why K-12 teachers are not using GenAI found that the most common reason was "I haven't explored these tools because I have other priorities that are more important" (Klein, 2024). This was echoed by one of the teachers in the Walkington and Bainbridge (under review) interview study. An Algebra 1 teacher with 10 years of experience teaching in a district composed of predominantly marginalized learners, when asked about using GenAI to personalize content to his students' interests, described how:

> It could help, you know - anything is better than nothing, but that's not the issue. The issue is the gaps of what they do and what they don't know based on where they are… We're trying to fill gaps like the city does potholes. If you've ever seen the cities do potholes, man, they just put something over it. But if the car hit that hole, maybe 20-30 minutes later, the pothole gonna be right back there next month. Instead of tearing up the street and starting over. And unfortunately, that's kind of what we're trying to do… we need to be able to kinda almost start over instead of trying to fix their gaps, the gaps are turning into canyons and in doing this, we're kind of wasting a year because we ain't fixing what the real issue is.

These kinds of sentiments relating to teachers having to confront bigger issues than GenAI can solve were also echoed by a first-year mathematics teacher teaching in a district composed primarily of marginalized learners, in the Beauchamp and Walkington (2024) study. During a discussion of using MagicSchool in the classroom, this teacher described how:

> It's so hard, because I feel like coming in, the teaching philosophy was "Oh I want to make sure my kids are well-rounded and critical thinkers." But now, since like the district is like "Why are test scores so bad? Why are test scores so bad?" it's like, my curriculum is going to be test questions basically, to prep them… They still tell us that we need to be stretching our kids thinking, but I'm like, we only have so much time. So I feel like because my district really does want to see test scores higher, I feel like my curriculum really is just test questions.

Discussing the possibilities of GenAI with mathematics teachers can be a reminder that these technologies may not be particularly effective for solving the larger problems they face with mathematics instruction every day. The mixed reviews from teachers we see in the studies of

GenAI we reviewed lends further credence to this point. This leads to the question of whether the hype around GenAI in education, and its transformative potential, is indeed justified.

Conclusion

I close by considering how can GenAI give us opportunities to truly transform the nature of mathematics education, in the way that the advent of calculators and dynamic geometry software was transformative. First, the practice of students engaging critically with the output of LLMs, particularly their function to create an endless amount of worked examples with explanations, could be powerful. This may be especially important for learners who speak diverse languages or learners in low-resourced settings where human tutoring is not possible. This approach may be particularly effective for student learning if the LLM makes mathematical mistakes that learners must grapple with and reason about – but as these LLMs rapidly become more advanced, mistakes happening with regularity seems increasingly unlikely, especially for K-12 mathematics content. Second, students using LLMs as a thought partner for problem-posing or mathematical story-telling activities seems like a promising direction from the existing research – story-telling is one practice that this technology excels at, and mathematics instruction is often missing the integration of rich, compelling stories about quantitative and spatial experiences.

Third, image-generating GenAI still has a long way to go to be ethical and useful. However, a promising way they could be leveraged is to automatically create rich visual representations to accompany mathematical tasks. This could also function to reduce costs associated with the development of high-quality open access materials that are freely available to teachers and districts. Fourth, mathematics teachers using GenAI as a thought partner to help them brainstorm and iterate upon lesson ideas, adapted for their context and needs, certainly has potential, especially if these lessons would be free. However, there are a variety of logistical and structural issues that may prevent this from being possible for individual teachers, and teachers will need to be prepared to modify and adapt the output of LLMs to suit their needs. Depending on the amount of time this modification takes, LLMs may not greatly enhance efficiency, and may instead mainly enhance creativity.

Finally, the power of GenAI to support students with mathematics skill practice, particularly in its ability to adapt to student needs, does have real-world value. These skill-based tasks are ultimately the kinds of mathematical scenarios that students will be held accountable for being able to solve in K-12 settings, and students' mathematical fluency can have high-stakes implications inside of school. However, the field of mathematics education needs to look beyond such applications of GenAI and consider how this technology, coupled with other initiatives, can help us solve the pressing problems teachers actually face with mathematics instruction.

Acknowledgments

This work was supported through funding from Rice University - OpenStax via the Bill & Melinda Gates Foundation. This work was also supported by the National Science Foundation under Grant DRL 2341948. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive network science reveals bias in GPT-3, GPT-3.5 turbo, and GPT-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, *7*(3), 124. http://doi.org/10.3390/bdcc7030124

Albornoz-De Luise, R. S., Arnau, D., Arnau-González, P., & Arevalillo-Herráez, M. (2024, September). Beyond the hype: Identifying and analyzing math word problem-solving challenges for large language models. [Paper presentation]. In S. Balloccu, Z. Kasner, O. Plátek, P. Schmidtová, K. Onderková, M. Lango, O. Dušek, L. Flek, E. Reiter, D. Gkatzia, and S. Mille (Eds.), In Proceeding of The 2nd Workshop on Practical *LLM-assisted Data-to-Text Generation* (pp.1-6), Tokyo, Japan. https://aclanthology.org/2024.practicald2t-1.1/

Asare, B., Arthur, Y. D., & Boateng, F. O. (2023). Exploring the impact of ChatGPT on mathematics performance: The influential role of student interest. *Education Science and Management*, *1*(3), 158-168. http://doi.org/10.56578/esm010304

Barana, A., Marchisio, M., & Roman, F. (2023, October 21-23). *Fostering problem solving and critical thinking in mathematics through generative artificial intelligence* [Paper Presentation]. In D. Sampson, D. Ifenthaler, P.Isaías (Eds), In *International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA)* (pp. 377-385). Madeira Island, Portugal.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., & Mariman, R. (2024). Generative AI can harm learning. *SSRN 4895486*. http://dx.doi.org/10.2139/ssrn.4895486

Beauchamp, T., & Walkington, C. (2024). Mathematics teachers using generative AI to personalize instruction of students' interests. *AMTE Connections*. https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In M. Elish, W. Isaac, and R. Zernel (Eds.) *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Berryhill, A., Chandler, L., Bondurant, L., & Sapkota, B. (2024). Using ChatGPT as a thought partner in writing relevant proportional reasoning word problems. *AMTE Connections*. https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher

Broutın, M. S. T. (2024). Exploring mathematics teacher candidates' instrumentation process of generative artificial intelligence for developing lesson plans. *Yükseköğretim Dergisi*, *14*(1), 165-176. https://doi.org/10.53478/yuksekogretim.1347061

Butgereit, L., & van Staden, A. (2023, December). Supporting home-language education in Africa with multi-lingual mathematics tutoring using GPT-4. In S. Pudaruth (Ed), In *International Conference on Artificial Intelligence and its Applications* (pp. 44-49). https://doi.org/10.59200/ICARTI.2023.006

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). Artificial intelligence and the future of teaching and learning: Insights and recommendations. Office of Educational Technology, U.S. Department of Education. https://policycommons.net/artifacts/3854312/ai-report/4660267/

Cheng, L., Croteau, E., Baral, S., Heffernan, C., & Heffernan, N. (2024). Facilitating student learning with a chatbot in an online math learning platform. *Journal of Educational Computing Research*, *62*(4), 907-937. https://doi.org/10.1177/07356331241226592

Einarsson, H., Lund, S. H., & Jónsdóttir, A. H. (2023). Application of ChatGPT for automated problem reframing across academic domains. *Computers and Education: Artificial Intelligence*, 100194. https://doi.org/10.1016/j.caeai.2023.100194

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *Advances in neural information processing systems*, *36*. https://doi.org/10.48550/arXiv.2301.13867

Gattupalli, S. S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B. P., & Woolf, B. (2023a). Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning. In *LLM@ AIED* (pp. 151-162).

Gattupalli, S., Maloy, R. W., & Edwards, S. (2023b). Comparing teacher-written and AI-generated math problem solving strategies for elementary school students: Implications for classroom learning. *College of Education Working Papers and Reports Series, 5*. https://doi.org/10.7275/8sgx-xj08

Gegg-Harrison, W., & Quarterman, C. (2024). AI detection's high false positive rates and the psychological and material impacts on students. In *Academic Integrity in the Age of Artificial Intelligence* (pp. 199-219). IGI Global. https://doi.org/10.4018/979-8-3693-0240-8.ch011

Gómez Marchant, C. & Hardison, H. (2024). In the shadows of burgeoning colossi: The whiteness of AI in mathematics teacher education. *AMTE Connections*. https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher

Google for Education (2024). Gemini for Google Workspace. https://workspace.google.com/solutions/ai/

Grossman, J., Lin, Z., Sheng, H., Wei, J. T. Z., Williams, J. J., & Goel, S. (2019). MathBot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*.

Kakarla, S., Thomas, D., Lin, J., Gupta, S., & Koedinger, K. R. (2024). Using large language models to assess tutors' performance in reacting to students making math errors. *arXiv preprint arXiv:2401.03238*. https://doi.org/10.48550/arXiv.2401.03238

Khan Academy (n.d.). Khanmigo by Khan Academy. https://www.khanmigo.ai/

Klein, A. (2024) Top 13 reasons teachers avoid ChatGPT and other AI tools. *Education Week*. https://www.edweek.org/technology/top-13-reasons-teachers-avoid-chatgpt-and-other-ai-tools/2024/02

Kwon, M., & Ko, I. Secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items. https://mathsa.uantwerpen.be/fame/FAME_2024_paper_38.pdf

Laird, E., Dwyer, M., & Grant-Chapman, H. (2023). Off task: edtech threats to student privacy and equity in the age of AI. *Center for democracy & technology*. https://cdt.org/insights/report-off-task-edtech-threats-to-student-privacy-and-equity-in-the-age-of-ai/

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI less "thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:2304.03271*. https://doi.org/10.48550/arXiv.2304.03271

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (arXiv:2107.13586). https://arxiv.org/abs/2107.13586 https://doi.org/10.1145/3560815

Madden, M., Calvin, A., Hasse, A., & Lenhart, A. (2024). *The dawn of the AI era: Teens, parents, and the adoption of generative AI at home and school.* San Francisco, CA: Common Sense.

MagicSchool.ai (n.d.). Magic School: AI built for schools. http://magicschool.ai.

Matzakos, N., Doukakis, S., & Moundridou, M. (2023). Learning mathematics with large language models: A comparative study with computer algebra systems and other tools. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(20), 51-71. https://www.learntechlib.org/p/223774/. https://doi.org/10.3991/ijet.v18i20.42979

Meyer, D. (2024). Khanmigo WANTS to love kids but doesn't know how. *Mathworlds*. https://danmeyer.substack.com/p/khanmigo-wants-to-love-kids-but-doesnt

Nathan, M. J. (2021). *Foundations of embodied learning: A paradigm for education*. Routledge. https://doi.org/10.4324/9780429329098

Nathan, M. J. (2023). Disembodied AI and the limits to machine understanding of students' embodied interactions. *Frontiers in Artificial Intelligence*, *6*, 1148227. https://doi.org/10.3389/frai.2023.1148227

National Council of Teachers of Mathematics (NCTM) (2023). Artificial Intelligence in mathematics teaching: A position of the National Council of Teachers of Mathematics. https://www.nctm.org/standards-and-positions/Position-Statements/Artificial-Intelligence-and-Mathematics-Teaching/

Naresh, N., Yilmaz, Y., & Galanti, T. (2024). Leveraging the potential of AI as a partner in teaching. *AMTE Connections*. https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher

Norberg, K. A., Almoubayyed, H., De Ley, L., Murphy, A., Weldon, K., & Ritter, S. (2024a). Rewriting content with GPT-4 to support emerging readers in adaptive mathematics software. *International Journal of Artificial Intelligence in Education*, 1-40.

Norberg, K., Molick, E. S., Almoubayyed, H., De Lay, L., Fisher, J., Murphy, A., Fancsali, S. and Ritter, S. (2024b). A.I. Math Personalization Tool (AMPT): Empowering students through peer-authored math content. *AAAI Workshop on AI4Ed*. https://openreview.net/forum?id=vf6W8Ak90P

OpenAI. (2023, March 14). GPT-4. *Open AI.* https://openai.com/index/gpt-4-research/

OpenAI (2024, September 12). Introducing OpenAI o1-preview. *OpenAI*. https://openai.com/index/introducing-openai-o1-preview/

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos One*, *19*(5), e0304013. https://doi.org/10.1371/journal.pone.0304013

Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. (2023). Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning*, *39*(3), 804-822. https://doi.org/10.1111/jcal.12776

Prihar, E., Lee, M., Hopman, M., Kalai, A. T., Vempala, S., Wang, A., Wickline, G., Murray, A., & Heffernan, N. (2023, June). Comparing different approaches to generating mathematics explanations using large language models. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. Santos, V. Dimitrova (Eds), In *International Conference on Artificial Intelligence in Education* (pp. 290-295). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_45

Pruitt-Britton, T., & Walkington, C. (2022). Are math textbooks really indoctrinating kids? *Education Week*. https://www.edweek.org/teaching-learning/opinion-are-math-textbooks-really-indoctrinating-kids/2022/07

Ruan, S., He, J., Ying, R., Burkle, J., Hakim, D., Wang, A., Yin, Y., Zhou, L., Xu, Q., AbuHashem, A., Dietz, G., Murnane, E., Brunskill, E., & & Landay, J. A. (2020, June). Supporting children's math learning with feedback-augmented narrative technology. In E. Rubegni & A. Vasalou (Eds.), *Proceedings of the Interaction Design and Children Conference* (pp. 567-580). https://doi.org/10.1145/3392063.3394400

Ruan, S., Nie, A., Steenbergen, W., He, J., Zhang, J. Q., Guo, M., Liu, Y., Nguyen, K., Wang, C. Ying, R., Landay, J. & Brunskill, E. (2024). Reinforcement learning tutor better supported lower performers in a math task. *Machine Learning*, *113*, 3023-3048. https://doi.org/10.1007/s10994-023-06423-9

Sawyer, A., & Aga, Z. (2024). Counterexamples to demonstrate artificial intelligence chatbot's lack of knowledge in the mathematics education classroom. *AMTE Connections*. https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher

Silver, E. A. (1994). On mathematical problem posing. *For the learning of mathematics*, *14*(1), 19-28.

Srivatsa, K. V., & Kochmar, E. (2024). What makes math word problems challenging for LLMs?. *arXiv preprint arXiv:2403.11369*. https://doi.org/10.48550/arXiv.2403.11369

Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932-945. https://doi.org/10.1037/a0031882

Walkington, C., Bernacki, M., Vongkulluksn, V., Greene, M., Darwin, T., Leyva, E., Istas, B., Hunnicutt, J., Washington, J., & Wang, M. (2024a). The effect of an intervention personalizing mathematics to students' career and popular culture interests on math interest and learning. *Journal of Educational Psychology, 116(4), 506–531.* https://doi.org/10.1037/edu0000840

Walkington, C., Milton, S., Pando, M., Lipsmeyer, L., Sager, M., & Beauchamp, T. (2024b). Adolescents using generative AI to engage in mathematical problem-posing. *Proceedings of the 15th International Congress on Mathematical Education (ICME-15)*. Sydney, Australia. https://link.springer.com/book/10.1007/978-1-4757-4238-1

Walkington, C., Clinton, V., & Shivraj, P. (2018). How readability factors are differentially associated with performance for students of different backgrounds when solving math word problems. *American Educational Research Journal*, *55*(2), 362-414. https://doi.org/10.3102/0002831217737028

Wang, A., Prihar, E., Haim, A., & Heffernan, N. (2024a, July). Can large language models generate middle school mathematics explanations better than human teachers? In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. Santos, V. Dimitrova (Eds), *International Conference on Artificial Intelligence in Education* (pp. 242-250). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64312-5_29

Wang, R., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024b, June). Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In K. Duh, H. Gomez, S. Bethard

(Eds). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2174-2199). https://doi.org/10.18653/v1/2024.naacl-long.120

Yadav, G., Tseng, Y. J., & Ni, X. (2023). Contextualizing problems to student interests at scale in intelligent tutoring system using large language models. *arXiv preprint arXiv:2306.00190*. https://doi.org/10.48550/arXiv.2306.00190

Yen, A. Z., & Hsu, W. L. (2023). Three questions concerning the use of large language models to facilitate mathematics learning. *arXiv preprint arXiv:2310.13615*. https://doi.org/10.48550/arXiv.2310.13615

Yilmaz, Z., Naresh, N., Galanti, T., & Kanbir, K. (2024, November). Pre-Service teachers' perceptions of exploring number theory concepts using Khanmigo: Benefits and challenges. To appear in *Proceedings of the 46th International Conference for the Psychology of Mathematics Education - North American Chapter – PME-NA 2024*. Cleveland, OH, United States.

Zhang, C., Liu, X., Ziska, K., Jeon, S., Yu, C. L., & Xu, Y. (2024, May). Mathemyths: leveraging large language models to teach mathematical language through Child-AI co-creative storytelling. In F. Mueller, P. Kyburz, J. Williamson, C. Sas, M. Wilson, P.Toups Dugas, I.Shklovski (Eds), In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-23). https://doi.org/10.1145/3613904.3642647