Generative Artificial Intelligence in Mathematics Education: A Systematic Literature Review Using the PRISMA Framework

Jennifer Dela Torre, Higher Colleges of Technology, United Arab Emirates
Jero Sayco, Higher Colleges of Technology, United Arab Emirates

The Asian Conference on Education & International Development 2026
Official Conference Proceedings

Abstract

The emergence of generative artificial intelligence (GenAI), in particular large language models (LLMs), is reshaping mathematics education by offering new forms of tutoring, feedback, problem generation, and personalized learning. This systematic review, following the PRISMA methodology, synthesizes empirical studies published from 2020 to 2025 that examine the application of GenAI in mathematics learning at secondary and tertiary levels. From an initial yield of 68 records across multiple databases, 32 were retained after screening and quality assessment, with 22 meeting criteria for detailed synthesis. The review finds that GenAI tools can significantly support mathematical problem solving, provide timely feedback, adapt instruction to learner needs, and increase accessibility. However, issues of output correctness, over-reliance, reduced deep reasoning, academic integrity, and limited long-term evidence remain. The paper concludes with recommendations for pedagogical integration, teacher training, assessment redesign, and future research directions.

Keywords: generative artificial intelligence, large-language models, mathematics education, PRISMA

iafor
The International Academic Forum
www.iafor.org

Introduction

Artificial intelligence has rapidly evolved in recent years, shifting from data analysis tools to generative AI technologies capable of producing text, code, images, and more (Bommasani et al., 2021; Kasneci et al., 2023a). Large language models (LLMs) like OpenAI's GPT-4, Google's Gemini, and Anthropic's Claude leverage deep neural networks trained on vast datasets to generate coherent responses (Brown et al., 2020). These systems, including domain-specific and multimodal models, have notably impacted educational settings by enabling personalized, adaptive interactions that respond to learners' needs. In mathematics education, GenAI tools such as ChatGPT, Microsoft Copilot, Google Gemini, and Wolfram Alpha provide on-demand support, guiding students through concepts, offering visual explanations, and helping identify errors, transforming study practices and facilitating both procedural and conceptual understanding.

This transition, however, prompts concerns about learning, assessment, and academic integrity as automated systems can generate solutions instantly, potentially changing traditional views on effort and competence (Bastani et al., 2024a; Frieder et al., 2023; Kasneci et al., 2023b). Research indicates GenAI can boost engagement and performance (Bauer et al., 2025; Menlah & Boateng, 2025), but unstructured or excessive use may undermine independent reasoning and durable mathematical competence. Prior studies and reviews are fragmented and methodologically diverse, making it difficult to synthesize the impact on mathematical thinking and instruction. Therefore, this review applies PRISMA procedures to consolidate evidence from peer-reviewed research (2020–2025) on LLM-based tools in secondary and tertiary mathematics education, distinguishing supported benefits from speculative claims and highlighting areas where evidence remains limited, such as fairness and long-term effectiveness.

Theoretical Framework

This section positions the review within (a) foundational scholarship on generative AI and large language models, (b) learning sciences construct that explain how GenAI may influence mathematics learning, and (c) systematic review methodology using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach (Page et al., 2021).

Generative AI and Large Language Models as Foundation Models

Generative artificial intelligence (GenAI) refers to models that create new artifacts such as text, code, images, and worked solutions, rather than simply classifying or predicting based on existing data. In education, large language models (LLMs) dominate, generating natural language by learning statistical patterns from extensive datasets. These LLMs, known as "foundation models," are trained at scale and adapted for tasks like tutoring and feedback (Bommasani et al., 2021). Their few-shot abilities allow them to perform educational tasks from prompts with little training (Brown et al., 2020), making them well suited for simulating tutoring, explanation, and assessment support through scalable dialogic interaction (Kasneci et al., 2023a).

GenAI in Mathematics Education

Theoretical accounts highlight that LLMs can generate explanations, strategies, prompts, and feedback to support learning, but they also risk producing incorrect or misleading reasoning.

These models are reliable for factual retrieval and some undergraduate tasks, but less so for advanced reasoning, where errors may look plausible to beginners (Frieder et al., 2023). Validity in mathematics depends on correct steps, not just answers. The educational value of LLMs depends on how they are used and supervised, as accuracy issues and misuse are closely tied to their benefits (Kasneci et al., 2023b).

GenAI connects with theories of feedback, cognitive load, and engagement. Feedback is most helpful when it clarifies goals and supports thinking, not just answers (Hattie & Timperley, 2007). AI can provide timely, personalized feedback, but incorrect or unexplained feedback can hinder understanding. Cognitive Load Theory shows GenAI can reduce unnecessary load by clarifying language and structuring steps, but may harm learning if it replaces students' reasoning (Sweller, 2011). Engagement frameworks like ICAP show GenAI is best for learning when students evaluate and revise AI outputs, not just copy them (Chi & Wylie, 2014).

GenAI impacts assessment, equity, and responsible use. It allows instant solutions, which can disconnect student effort from competence. Unrestricted access can boost short-term performance but harm long-term learning, while guided designs help avoid this (Bastani et al., 2024b). GenAI can improve accessibility through personalized explanations and language support, but equity depends on access and governance. Responsible use requires oversight, assessment redesign, and policies that protect learners (UNESCO, 2023).

Systematic Review and PRISMA

Given the rapid and varied emergence of evidence on GenAI in mathematics education, this paper uses a systematic review guided by PRISMA 2020 standards to ensure transparent and reproducible synthesis. PRISMA structures evidence selection, minimizes bias with explicit criteria, and limits conclusions to what included studies support, enhancing the review's credibility (Page et al., 2021).

Rationale and Research Questions

Rationale

Generative AI's rapid growth offers new possibilities and challenges for mathematics educators. Ongoing issues like limited individualized support, slow feedback, and difficulty fostering conceptual understanding persist, especially for students with disabilities and second-language learners. GenAI can help by personalizing explanations and providing prompt feedback, potentially reducing some barriers when used thoughtfully, though it won't solve all structural problems.

Results from the literature are mixed: some studies report better attention, engagement, or understanding, but outcomes vary widely. Institutional guidance and training often lag adoption, so educators sometimes improvise. This highlights the need to synthesize empirical evidence to clarify GenAI's impact on math learning and teaching in real classrooms.

Research Questions

Based on the current state of the field, this study investigates the following research questions:

RQ1: In what ways has generative AI been utilized to deliver personalized mathematics tutoring, and what effects does this have on students' mathematical understanding, problem-solving performance, and motivation?

RQ2: How do generative AI tools facilitate feedback, error diagnosis, and formative assessment in mathematics, and to what extent is the feedback generated by large language models accurate, useful, and pedagogically meaningful?

RQ3: What influence do explanations and solution strategies generated by large language models have on students' mathematical reasoning, strategic thinking, and depth of conceptual understanding?

RQ4: In what ways are generative AI tools being leveraged to support assessment in mathematics, and what is their impact on assessment validity and student learning outcomes?

RQ5: How does the implementation of generative AI affect accessibility and equity in mathematics education?

RQ6: What risks, limitations, and ethical considerations have been identified regarding the use of generative AI in mathematics education, and how do these concerns inform frameworks for responsible implementation?

Methodology

This systematic review followed PRISMA, which provides a rigorous and transparent framework for reporting systematic reviews.

Search Strategy

The review searched Scopus, Web of Science, ERIC, and IEEE Xplore as main sources, with Google Scholar for extra coverage. Search terms combined generative AI (like "GPT-4," "ChatGPT," "LLM," "generative AI") and mathematics education (such as "mathematics education," "math problem solving"), using Boolean operators to focus on studies linking GenAI and math teaching or learning. The search included studies from January 2020 onward, reflecting the rise of modern LLMs in education.

Inclusion and Exclusion Criteria

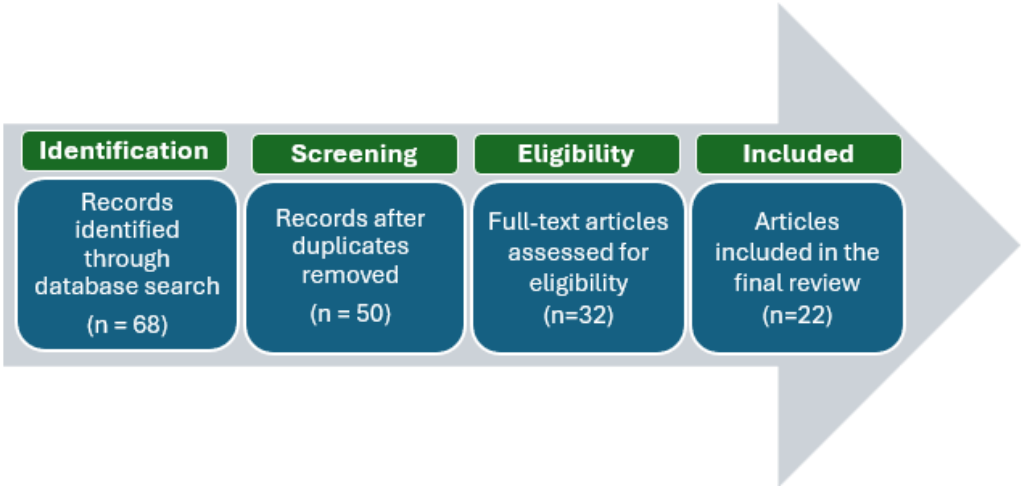
Studies had to be peer-reviewed articles or conference papers with empirical findings, focusing on secondary or tertiary math education or related math-intensive programs. Only studies using generative AI, specifically LLM-based systems, were included. Research had to be published in English and centered on contemporary GenAI tools.

Studies were excluded if they focused on non-math subjects, lacked empirical data, used non-generative AI, or were not peer-reviewed (such as dissertations, preprints, or media articles). These criteria ensured all included studies were relevant and evidence-based.

Screening and Selection

The search yielded 68 records; after removing 8 duplicates, 60 were screened by title and abstract. Ten were excluded for not focusing on math learning or generative AI. Fifty full texts were reviewed, with 18 excluded for missing scope criteria or empirical data. Thirty-two studies were then assessed for quality, resulting in 22 studies included in the final synthesis. The process is shown in the PRISMA diagram (Figure 1).

Figure 1
Methodology Framework Based on PRISMA (Page et al., 2021)



Results and Discussion

This section synthesizes findings from the 22 studies remained in the review, grouped by application type and educational effect.

Tutoring and Personalized Instruction

Table 1
GenAI as Personalized Math Tutor and Learning Support

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
Kasneci et al. (2023b)	Experimental	University calculus students	ChatGPT as conceptual tutor	Improved conceptual explanations, clearer reasoning patterns.
Menlah and Boateng (2025)	Field experiment	400 undergraduate students	AI-based tutoring for problem solving	Higher task performance: benefits were strongest among students reporting higher mathematics anxiety.
Brandibur et al. (2025)	Mixed-methods	313 undergraduate students	LLM-supported mathematics instruction using structured prompt patterns	Students reported higher engagement; identified errors more efficiently; viewed AI as a supportive complement rather than a substitute for the instructor.
Kaya and Yavuz (2025)	Comparative quantitative analysis	NAEP mathematics tasks (Grades 4, 8, 12); U.S. large-scale assessment data	LLM problem-solving (ChatGPT-4o, GPT-4)	Models outperformed most student cohorts; performance remained stable as task difficulty increased; less sensitivity to task complexity.

GenAI tools are increasingly used for personalized math support, offering conversational explanations and adaptive practice. Outcomes include better conceptual understanding and improved completion of complex tasks. At the university level, LLMs support independent study by providing detailed, stepwise solutions. Kasneci et al. (2023b) found that ChatGPT improved students' clarity and reasoning, showing potential for enhancing engagement with mathematical ideas.

Research indicates AI tutoring is especially helpful for students with math anxiety. Menlah and Boateng (2025) observed that those receiving AI guidance solved more problems correctly, with the biggest improvements among anxious students. Immediate, responsive support may lower stress and encourage persistence.

The way GenAI is designed affects learning. Clear prompts and scaffolds keep students engaged and help them correct errors. Most students see AI as a helpful tool, not a replacement for teachers. Brandibur et al. (2025) highlight that GenAI works best alongside teacher guidance.

Large-scale assessment shows advanced language models excel at calculation and rule-based tasks. Kaya and Yavuz (2025) found models outperformed most students and stayed consistent as tasks got harder. However, strong answer accuracy does not guarantee deep understanding, and teacher-led instruction remains key for fostering reasoning.

Feedback, Error Diagnosis, and Formative Assessment

Table 2
GenAI for Feedback and Error Diagnosis

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
Holmes and Tuomi (2022)	Quantitative evaluation	Secondary algebra	Automated procedural feedback	78% accuracy; improved revision behaviors
Abdelghani et al. (2023)	Quasi-experimental	University proof-writing	LLM feedback for proofs	Improved logical structure, coherence, and metacognition
Bauer et al. (2025)	Randomized controlled experiment	Introductory statistics students	AI-generated adaptive feedback on problem solutions	Improved statistical skills; increased interest in statistics; stronger effects for lower-performing students
Walkington (2025)	Qualitative	K–12 and higher education teachers	GenAI-supported feedback & planning	Reduced workload; highlighted pedagogical challenges

This subsection reviews generative AI's role in feedback, error diagnosis, and formative assessment in mathematics. Studies assess AI feedback accuracy and its effects on student revision and teaching.

GenAI is effective for formative assessment by providing prompt, targeted feedback, helping students identify and fix errors during revision. In math, automated feedback supports immediate correction, and Holmes and Tuomi (2022) found AI feedback can guide students in algebra tasks.

At the university level, AI feedback improves logical structure in proof-based tasks and encourages reflection, supporting metacognitive reasoning (Abdelghani et al., 2023).

Adaptive feedback in statistics enhances student skills and interest, especially for lower-performing students, as shown by Bauer et al. (2025).

Qualitative studies note that while teachers value AI feedback for reducing workload and planning, it may miss subtle reasoning errors or deeper misconceptions (Walkington, 2025).

Overall, generative AI works best as a complement to teaching. AI feedback is most reliable for stepwise tasks and is effective when students re-examine and justify their answers; instructional design should encourage students to interpret and verify feedback independently.

Effects on Students' Mathematical Reasoning, Strategic Thinking, and Conceptual Understanding

Table 3
GenAI for Mathematical Reasoning, Explanation, and Inquiry

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
García Navarro et al. (2025)	Comparative performance evaluation	High school and first-year engineering mathematics tasks (20 problems)	GPT-3.5, GPT-4, Wolfram GPT, LLaMA-2, Gemini	GPT-4 beat GPT-3.5; Wolfram GPT helped results slightly; however, persistent reasoning limitations remain a challenge.
Parra et al. (2024)	Comparative error analysis	High school geometry problems	ChatGPT, Bard, and others	Geometry reasoning was weak; frequent conceptual errors; language barriers made these weaknesses worse.
Wang and Fan (2025)	Classroom study	Secondary math students	ChatGPT as a problem solver	Less "productive struggle" for students; risk that learning stays on the surface.

Table 3 highlights how students' reasoning and explanations shift when interacting with LLM-generated solutions. Studies compare human and AI explanations, analyze classroom use, and trace errors. Results show that AI support does not always lead to better understanding and can sometimes hide gaps in comprehension.

Recent research finds that LLMs, like GPT-4, can model step-by-step problem-solving and offer alternative strategies, especially when integrated with math-specific tools (García Navarro et al., 2025). Yet, students often struggle with tasks that need conceptual transfer, showing that following procedures does not guarantee deep understanding.

Subject differences matter. For geometry, AI explanations often include conceptual errors and incomplete justifications (Parra et al., 2024). Language also impacts results, as performance can vary across languages based on training data.

Classroom studies show that students who mainly use AI solutions tend to skip productive struggle and prefer surface revisions, which can limit strategic thinking (Wang & Fan, 2025). Unrestricted access may encourage imitation instead of sense-making, especially in older students.

Overall, AI-generated explanations can help or hinder learning. They may introduce new strategies and support planning if students engage critically, but errors and overreliance can lead to shallow understanding. The impact depends on both the model's quality and whether students are required to check and compare solutions themselves.

Assessment, Question Generation, and Scaffolding

Table 4
GenAI for Assessment, Question Generation, and Scaffolding

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
Meissner et al. (2024)	Expert-reviewed evaluation study	Higher education mathematics (240 generated items)	LLM-generated competence-based e-assessment items (ItemForge; GPT-3.5, GPT-4; RAG)	High-quality item alignment; strong linguistic quality; taxonomy fit; solution inaccuracies
Widjaja and Yohannis (2025)	Design and evaluation study	Secondary education MCA mathematics items	GenAI-based automatic question generation (Gemini 1.5 Pro; multimodal)	Diverse question types; strong curriculum alignment; reduced teacher workload
McGalliard and Otten (2025)	Qualitative task analysis	Middle and high school mathematics tasks	GenAI responses to cognitively demanding problems (ChatGPT, Gemini)	Mixed reasoning quality; frequent logical errors; improvement with prompting; strong pedagogical discussion value
Yu et al. (2025)	Two-phase experimental evaluation	Advanced undergraduate mathematics (logic course)	LLM-based automated question generation with RAG and DOK	Improved question correctness and relevance; stronger higher-order reasoning alignment

Table 4 shows that research has explored how generative AI supports assessment activities in secondary and college math, such as item generation, task design, and scaffolding. These tools may boost efficiency but can raise concerns about validity, since AI responses lack human judgment and may contain errors like missing steps or incomplete reasoning. Without review, such flaws can limit the range of student thinking and reduce trust in assessment results.

In higher education, GenAI is used to create quizzes and tasks that match course outcomes. Meissner et al. (2024) found that LLM-based systems can produce items with strong linguistic quality and alignment, but inaccuracies in solutions highlight the need for expert review.

In secondary schools, Widjaja and Yohannis (2025) reported that generative AI can create varied questions aligned with curriculum, helping teachers save time. Still, oversight is needed to ensure items accurately reflect learning objectives.

Evaluations show that AI-generated solutions sometimes omit key steps or misrepresent concepts. McGalliard and Otten (2025) noted that prompting strategies can improve quality but reliability depends on use. Overreliance may weaken confidence in judgments of student understanding.

Experimental work suggests GenAI can support complex assessment items when combined with retrieval and reasoning prompts. Yu et al. (2025) found that prior questions and structured frameworks improve item quality and cognitive alignment. With good safeguards, GenAI may help design assessments that focus on conceptual understanding.

Generative AI is used for creating assessment items and scaffolding responses. While these uses can improve efficiency and alignment, concerns remain about validity, especially when solutions appear correct but lack complete reasoning. The impact depends on expert review, sound instructional practice, and careful implementation to avoid undermining math evaluation quality or deep learning.

Accessibility, Equity, and Inclusion: Supporting Diverse Learners

Table 5
GenAI for Accessibility, Equity, and Inclusion

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
Canonigo (2025)	Qualitative interpretive study	Secondary mathematics teachers and students (urban public schools)	ChatGPT for personalized feedback, inquiry, and scaffolding	Greater learner agency and participation; shifts in classroom interaction patterns; risk of over-reliance without teacher guidance.
Agbo et al. (2025)	Conceptual review with practice synthesis	Mathematics education for students with learning disabilities	AI-driven adaptive learning and personalized feedback systems	Increased participation; more individualized support; greater learner independence; potential equity benefits.
Rizos et al. (2024)	Case study	Students with special educational needs	AI-generated individualized worksheets	Increased engagement; improved self-efficacy; reported a more supportive classroom experience.

The studies summarized in Table 5 indicate that GenAI may influence accessibility and equity in mathematics education, with effects that depend strongly on learner characteristics and implementation conditions. Across classroom-based accounts and case studies, the learning environment appears to mediate whether AI tools reduce barriers or introduce new forms of disadvantage.

One consistent finding is that GenAI can increase access to support by tailoring explanations and pacing individual learners. In secondary contexts, Canonigo (2025) reported that students using ChatGPT received more immediate feedback and expressed greater comfort asking questions. Such interactions may shift classroom participation patterns by enabling students to engage with content privately and at their own pace. However, the study also cautioned that, without teacher oversight, some learners may become overly dependent on AI-provided guidance; balancing increased access with sustained sense-making remains essential.

Evidence from special education contexts suggests that GenAI may reduce affective barriers to mathematics engagement when materials are adapted to learners' needs. Rizos et al. (2024) worked with students experiencing learning difficulties using AI-generated individualized worksheets. The authors reported increased engagement and reduced anxiety during practice,

plausibly because task difficulty and presentation were better matched to students' profiles. These findings suggest that targeted personalization may contribute to more supportive learning experiences for some learners.

Beyond individual classrooms, Agbo et al. (2025) suggest that adaptive learning systems may support students with learning disabilities by providing responsive feedback and tailored instructional pathways. When lessons adjust on the fly using real-time data, support can be more individualized than conventional approaches. However, scaling such systems remains a challenge, and potential benefits depend on careful implementation, including teacher mediation and appropriate resourcing.

Overall, the reviewed studies suggest that GenAI may contribute to more equitable participation by providing additional avenues for explanation, language support, and individualized pacing. At the same time, these benefits are not automatic. Effective implementation requires teacher guidance, monitoring for over-reliance, and attention to differential access. Under these conditions, GenAI may support equity goals rather than reinforcing existing disparities.

Risks, Limitations, and Ethical Concerns

Table 6
Risks, Ethical Considerations, and Limitations of GenAI in Math Education

Study (Authors, Year)	Research Design	Sample / Context	GenAI Application	Key Outcomes
Silgado-Tuñón et al. (2025)	Qualitative interview study	University mathematics professors (n = 10)	GenAI for text editing, content generation, idea organization	Pragmatic use patterns; limited classroom integration; epistemic ambivalence; need for PD
Bastani et al. (2024a)	Randomized field experiment	1,028 university students	Standard vs. "learning-safe" GPT tutors	Standard AI boosts short-term performance but harms long-term learning
Frieder et al. (2023)	Analytical review	STEM and mathematics education	ChatGPT for problem solving	Hallucinated solutions; assessment integrity risks
Kasneci et al. (2023a)	Systematic interdisciplinary review	Higher education, including mathematics	LLMs for tutoring and explanation	Accuracy risks; student dependence; need for human oversight

Table 6 highlights key risks of using generative AI in math education, such as output accuracy, assessment integrity, and learner dependence. These issues guide responsible implementation.

Bastani et al. (2024a) found that unrestricted AI tutoring improves immediate performance but may weaken retention and understanding. Designs focused on hints and guided reasoning produce more sustained learning, stressing the need to promote cognitive effort.

Output correctness is a recurring issue. Frieder et al. (2023) showed LLMs may generate answers with invalid steps that seem correct, misleading learners. Kasneci et al. (2023a) also urge verification and human oversight, especially for explanation and tutoring.

Over-reliance is another risk. LLMs can reduce student effort in planning and justifying work. Kasneci et al. (2023a) warn this may lessen engagement with mathematical reasoning, so GenAI should support, not replace, student thinking.

For instructors, Silgado-Tuñón et al. (2025) found GenAI is used more for writing support than direct teaching. Professors have concerns about trustworthiness and need clearer ethics policies and professional development for consistent use.

In summary, risks tied to accuracy, academic integrity, and dependence are central. The literature recommends responsible use through instructional design, transparency, and governance, with ongoing professional development and clear policies to ensure GenAI supports learning without harming reasoning or assessment validity.

Summary of Findings

Research shows generative AI impacts math teaching and learning in varied ways, with effects differing by context.

Reported benefits include adaptive pacing, timely feedback, and greater responsiveness in practice and assessment. Studies indicate students may build stronger explanations when AI prompts reasoning. For teachers, AI can reduce routine tasks like drafting feedback and generating practice items, allowing more time for planning and targeted support.

However, studies warn against using AI without clear structure and oversight. Students might disengage from difficult tasks and rely on AI for answers rather than reasoning independently. Output errors, from minor inaccuracies to plausible but incorrect explanations, remain a concern. Uncritical reliance in assessment can threaten validity. Effective use depends on embedding GenAI within instructional goals and teacher-led practices.

Instructional context is crucial. Teachers shape whether GenAI promotes verification, reflection, and strategy comparison. When AI is used in structured activities requiring students to evaluate and justify outputs, learning is more favorable. Without guidance, GenAI may encourage surface-level completion.

Equity is closely linked to access and design. GenAI can provide personalized support for students with learning challenges or developing academic language, reducing barriers to participation.

But equity benefits depend on access. Advanced tools are not available everywhere, possibly worsening disparities. Most evidence focuses on short-term outcomes, with few studies on long-term effects, making it unclear if gains are sustained or just temporary.

Literature is mainly from higher education, with fewer studies in secondary schools or resource-constrained environments. Studies from non-Western systems are rare. Rapid evolution of models and platforms makes findings quickly outdated. GenAI can support math learning under the right conditions, but outcomes rely on careful planning, governance, and ongoing human involvement.

Conclusion

Generative AI is changing mathematics education, with LLM-based tools offering adaptive support, timely feedback, and flexible learning options. Research from 2020 to 2025 finds these systems can raise student engagement and reduce test anxiety, while helping some learners with diverse needs. GenAI can complement instruction in classrooms with varied student abilities and high workload.

However, positive outcomes depend on thoughtful use. Poorly designed implementation can reduce conceptual understanding and procedural skills. Benefits are most likely when AI is integrated into structured lessons with an emphasis on verification, reflection, and educator guidance rather than relying solely on AI-generated answers.

Effective use will require ongoing teacher development and redesigned assessments focusing on reasoning and justification. Clear institutional policies are needed for equity, transparency, and integrity. Future research should include long-term studies in varied educational settings, especially for learners facing barriers. GenAI is best seen as a resource to support teaching when used with clear goals and ethical safeguards.

Recommendations

Based on this review, key recommendations for integrating generative AI in math education include viewing AI as a tool to support, not replace, human instruction. Educators should use AI to promote deeper reasoning, encouraging students to evaluate and revise AI-generated solutions. Such reflective practice is crucial for authentic learning.

Schools must not only provide technology but also set up clear frameworks for their purposeful and pedagogical use in daily lessons.

Researchers should focus on studying generative AI in real classrooms to inform best practices, especially for intervention and assessment. AI is useful for informal learning, but high-stakes testing requires strong safeguards. Students need explicit instruction on AI literacy, verification, and critique, best taught through hands-on activities.

Policies should define AI's role in coursework and exams to ensure equity and transparency. Teacher training must be comprehensive, enabling lesson redesign and effective use of AI. Equitable access for under-resourced schools and privacy protections is essential. Trust in AI requires ongoing monitoring, transparency, and ethical practice.

Future research should track long-term effects of AI in diverse settings and focus on how AI integrates with regular instruction and feedback. Developing reliable ways to measure AI-linked learning outcomes is also needed, while maintaining educational integrity and ethics.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used Microsoft Copilot integrated within Microsoft Word to help improve the clarity and quality of the writing, including grammar, spelling, and overall readability. All suggestions were reviewed and edited by the authors, who take full responsibility for the final content.

References

Abdelghani, M., Yusuf, K., & Park, H. (2023). AI-supported feedback for mathematical proofs: An empirical investigation. *International Journal of STEM Education, 10*(4), 211–225.

Agbo, C. M., Ugwuanyi, C. C., Okeagu, R. U., & Salifu, M. (2025). Leveraging on Artificial Intelligence (AI) to Make Learning More Inclusive for Students with Learning Disabilities in Mathematics Education. *International Journal of Latest Technology in Engineering Management & Applied Science*, *14*(5), 1073–1078. https://doi.org/10.51583/IJLTEMAS.2025.140500114

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024a). Generative AI can harm learning. The Wharton School Research Paper. Available at SSRN: https://ssrn.com/abstract=4895486. or http://dx.doi.org/10.2139/ssrn.4895486

Bastani, H., Kim, S., Xu, L., & Yuan, F. (2024b). Generative AI without guardrails can harm learning: Evidence from a large-scale randomized experiment. *Science Advances*. Advance online publication.

Bauer, E., Richters, C., Pickal, A. J., Klippert, M., Sailer, M., & Stadler, M. (2025). Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education. *British Journal of Educational Technology*, 56, 1735–1757. https://doi.org/10.1111/bjet.13609

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv. https://arxiv.org/abs/2108.07258

Brandibur, O., Filipowicz-Chomko, M., Girejko, E., Kaslik, E., Mozyrska, D., Mureșan, R., Pappas, N., Tănasie, A., & Zaharia, C. (2025). Higher mathematics education and AI prompt patterns: Examples from selected university classes. *Applied Sciences, 16*, 339. https://doi.org/10.3390/app16010339

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Canonigo, A. M. (2025). *AI in math education: Shifting the balance toward student empowerment*. Frontiers in Artificial Intelligence and Applications, 454, 454–462. https://doi.org/10.3233/FAIA250744

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49(4)*, 219–243. https://doi.org/10.1080/00461520.2014.965823

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., & Berner, J. (2023). Mathematical capabilities of ChatGPT. arXiv. https://arxiv.org/abs/2301.13867

García Navarro, A. L., Koneva, N., Hernández, J. A., & Sánchez-Macián, A. (2025). On the Use of Large Language Models at Solving Math Problems: A Comparison Between GPT-4, LlaMA-2 and Gemini. *International Journal of Interactive Multimedia and Artificial Intelligence*, *9*(5), 40–50. https://doi.org/10.9781/ijimai.2025.03.001

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, **77**(1), 81–112. https://doi.org/10.3102/003465430298487

Holmes, W., & Tuomi, I. (2022). Automated feedback systems in mathematics education: Accuracy and limitations. *British Journal of Educational Technology, 53*(5), 1482–1499.

Kasneci, E., Sessler, K., & Bannert, M. (2023a). *ChatGPT in higher education: A study of conceptual math explanations*. Computers and Education: Artificial Intelligence, 4, 100150.

Kasneci, E., Sessler, K., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kübler, J., Löw, T., Mollick, E., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023b). *ChatGPT for good? On opportunities and challenges of large language models for education*. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kaya, D., & Yavuz, S. (2025). Can Generative AI and ChatGPT break human supremacy in mathematics and reshape competence in cognitive-demanding problem-solving tasks? *Journal of Intelligence*, 13(4), 43. https://doi.org/10.3390/jintelligence13040043

McGalliard, W., & Otten, S. (2025). AI Responses to Challenging Problems and Educator Responses to AI Availability. *Digital Experiences in Mathematics Education*, *11*(2), 319–332. https://doi.org/10.1007/s40751-024-00167-4

Menlah, C. K. A., & Boateng, F. O. (2025). Examining the effect of AI-based tutoring systems on students' mathematical problem-solving skills: The moderating role of mathematics anxiety. *Journal of Pedagogical Sociology and Psychology, 7*(3), 5–17. https://doi.org/10.33902/jpsp.202536137

Meissner, A., Schneider, J., Förster, M., & Kuhn, J. (2024). *LLM-generated competence-based e-assessment items: Opportunities and limitations for higher education mathematics*. arXiv. https://arxiv.org/abs/2406.0xxxx

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160

Parra, V., Sureda, P., Corica, A., Schiaffino, S., & Godoy, D. (2024). 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. https://doi.org/10.9781/ijimai.2024.02.009

Rizos, I., Foykas, E., & Georgakopoulos, S. V. (2024). Enhancing mathematics education for students with special educational needs through generative AI: A case study in Greece. *Contemporary Educational Technology, 16*(4), ep535.

Silgado-Tuñón, D. A., Sureda, P., López-Flores, J. I., & Magallanes, E. (2025). Exploring university mathematics professors' perceptions and use of GenAI: a conceptual fields approach. *Amazonia Investiga*, *14*(86), 250–263. https://doi.org/10.34069/AI/2025.86.02.19

Sweller, J. (2011). Cognitive load theory. In J. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*. Advance online publication. https://doi.org/10.1111/ssm.18356

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanit Soc Sci Commun* 12, 621 (2025). https://doi.org/10.1057/s41599-025-04787-y

Widjaja, S. Y., & Yohannis, A. (2025). *AI-powered automatic question generation for teachers*. In Proceedings of the 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML) (pp. 1–6). IEEE. https://doi.org/10.1109/SIML65326.2025.11081160

Yu, Y., Krantz, A., & Lobczowski, N. G. (2025). *From recall to reasoning: Automated question generation for deeper math learning through large language models* (arXiv preprint). https://arxiv.org/abs/2505.11899

Contact email: jdelatorre@hct.ac.ae