Exploring the Role of Generative Artificial Intelligence in Transforming Mathematics Teaching and Learning: A Systematic Mapping Review

Nur Syaza Afrina Muhamad Sani*, Muhammad Akmal Azri Omar, Mohamad Ridhuan Abdullah

Department of Curriculum and Instruction, Kulliyyah of Education,
International Islamic University Malaysia
**Corresponding author: Nur Syaza Afrina Muhamad Sani (syaza.afrina@iium.edu.my)*

Received: 1 February 2026
Received in revised form: 10 March 2026
Accepted: 14 April 2026
Published: 10 July 2026

ABSTRACT

Recent advances in generative artificial intelligence (GenAI), in particular, large language models (ChatGPT, Gemini, Claude, etc.) have reshaped digital innovation in education. While prior studies have examined specific classroom applications, few studies to date have provided a comprehensive mapping of the empirical research landscape across tools, educational levels, mathematical domains, research designs, and geographical distribution, leaving structural patterns and research gaps insufficiently synthesized. This paper presents a systematic mapping review on the integration of GenAI in mathematics education. Using the PRISMA framework, studies published between 2022 and 2025 (final search conducted in December 2025) were retrieved from Web of Science and Scopus, complemented by a full-text search in ScienceDirect. After screening and eligibility procedure, 30 peer-reviewed empirical articles were included for analysis. The mapping results indicate that ChatGPT is the most frequently examined tool, with research predominantly concentrated at the tertiary education level and focused largely on algebra and calculus topics. Most studies employed qualitative or mixed-method designs, and the majority were conducted in Asia and Europe, reflecting emerging but uneven global research distribution. Thematic analysis was employed to identify commonly reported pedagogical applications, including problem-solving support, content generation, scaffolding, lesson planning, and differentiated instruction, as well as frequently reported challenges such as conceptual errors, limitations in abstract and spatial reasoning, and variability in output quality across platforms. Rather than evaluating effectiveness, this review provides an overview of research trends, identifies concentrations and gaps in the literature, and highlights directions for future empirical and theoretical work on GenAI-supported mathematics education.

Keywords
Generative artificial intelligence; Large language models; Mathematics education; Systematic mapping review; Educational technology

Introduction

The recent advancement of generative artificial intelligence (GenAI) in the current digital age has generated significant discussion within mathematics education, particularly regarding its implications for teaching, learning, and assessment (Seebut et al., 2024; Yunianto et al., 2024). In mathematics classrooms, large language models such as ChatGPT, Gemini, and Claude are increasingly explored as tools for problem-solving support, step-by-step explanation generation, and adaptive feedback (Marco & Stylianides, 2025; McGalliard & Otten, 2025; Wardat et al., 2023). Unlike earlier educational technologies that primarily automated practice, GenAI systems can generate context-sensitive explanations and multiple solution pathways, positioning them as potential cognitive partners in mathematical reasoning.

Existing empirical studies report that GenAI is used as a problem-solving assistant, content generator, and scaffolding tool, particularly in self-regulated and personalized learning contexts (Contel & Cusi, 2025; Segal & Klemer, 2025; Yunianto et al., 2024). These models operate by finding patterns in large datasets to generate human-like outputs as the textual explanations, step-by step solutions, and computational code. When guided through iterative prompting, students may refine queries, compare alternative strategies, and reflect on procedural reasoning (Seebut et al., 2024;

Yunianto et al., 2024). However, evidence regarding these pedagogical roles remains context-dependent and fragmented across educational levels and mathematical domains (Kock et al., 2025).

Despite such possible benefits, significant limitations have been documented. GenAI outputs may include significant cognitive error in tasks, fail in abstract reasoning and have poor performance in spatial geometry, higher-order calculus, and conceptual understanding (Parra et al., 2024; Wardat et al., 2023; Marco & Stylianides, 2025). The accuracy and reliability of outputs are highly dependent on prompt quality and platform-specific characteristics, as it may affect the curriculum alignment and credibility of the knowledge generated (Stamenkova, 2025; Mani & Zargeh, 2025). Moreover, while GenAI may improve procedural fluency or provide a variety of solution strategies, this method still cannot identify misconceptions independently and instil them in reasoning, which highlights the critical role of teacher mediation, guidance, and verification (Segal & Klemer, 2025; Pavlova, 2024).

Although individual empirical studies have examined specific classroom applications of GenAI, few studies to date have systematically mapped the empirical research landscape across tools, educational levels, mathematical domains, research designs, and geographical distribution. As a result, structural patterns, research concentrations, and underexplored areas remain insufficiently synthesized. Given the rapid expansion of GenAI research since 2022, a systematic mapping review is timely to provide a structured overview of emerging trends and gaps.

Accordingly, this study presents a systematic mapping review of empirical research on GenAI in mathematics education. Rather than evaluating instructional effectiveness, this review synthesizes how the reviewed studies are distributed across GenAI platforms (e.g., ChatGPT 3.5/4.0, Gemini, Claude), educational contexts (primary to tertiary), mathematical topics, research methodologies, and geographical regions. Essentially, this study clarifies the current state of the field and outlines directions for future pedagogical and theoretical development, by identifying dominant research clusters and underrepresented areas.

Literature Review

Generative Artificial Intelligence in Mathematics Education

The literature indicates that GenAI assists students in their engagement, motivation, self-regulated learning, problem-solving processes, and conceptual understanding, while also highlighting challenges such as errors in complex tasks, variability of outputs, ethical concerns and the need for teacher mediation (McGalliard & Otten, 2025; Parra et al., 2024; Delima et al., 2024; Contel & Cusi, 2025; Walkington, 2025).

To provide a conceptual synthesis, the reviewed studies reveal four main thematic patterns in the use of GenAI in mathematics education. First, regarding cognitive and learning outcomes, studies consistently report improvements in procedural fluency, problem-solving support, and self-efficacy (Seebut et al., 2024; Wahba et al., 2024; Yunianto et al., 2024). However, methodological limitations such as small sample sizes, short-term interventions, and reliance on self-report measures constrain the generalizability of these findings (Daher & Anabousy, 2025; Sutrisno et al., 2025).

Second, teacher mediation and professional development emerge as critical factors, as GenAI cannot independently detect student misconceptions or ensure conceptual fidelity (Segal & Klemer, 2025; Pavlova, 2024; Busuttil & Calleja, 2025). Effective integration therefore requires targeted professional development, scaffolding strategies, and continuous teacher oversight.

Third, ethical and epistemic considerations are frequently highlighted, with studies noting biases, inaccuracies, and misalignment with curricula as main challenges (Walkington et al., 2025; Marco & Stylianides, 2025). Maintaining content reliability and fairness necessitates critical reflection on the ethical use of AI and structured prompt design.

Finally, tool-specific patterns and pedagogical applications indicate that ChatGPT is the most frequently investigated platform, followed by Gemini and Claude (Seebut et al., 2024; Marco & Stylianides, 2025). Pedagogical applications include scaffolding, iterative problem-solving, differentiated instruction, and adaptive feedback. While GenAI demonstrates potential as a cognitive co-agent, performance varies across mathematical domains, with weaknesses most pronounced in abstract reasoning, spatial geometry, and higher-order calculus (Parra et al., 2024; Wardat et al., 2023).

The evidence suggests a fragmented landscape: studies are concentrated in higher education and specific topics like algebra and calculus, with limited longitudinal or cross-cultural investigations. Methodological robustness is uneven, and early-stage research often emphasizes practicality over efficacy. Therefore, a systematic mapping review is necessary to synthesize patterns, identify gaps, and provide a structured overview of GenAI research in mathematics education. Table 1 provides a summary of the 30 reviewed empirical studies, including aims, design, key findings, and recommendations, illustrating both recurring patterns and critical challenges in practice.

Table 1
Summary of the Studies

No.	Studies	Aims and Design	Key Findings	Recommendations
1	McGalliard & Otten (2025)	Investigate GenAI engagement in complex mathematical reasoning through a qualitative analysis of ChatGPT and Gemini responses to three cognitively demanding secondary tasks.	GenAI can respond to demanding tasks but makes substantial errors; it demonstrates the ability to refine work through iterative prompting.	Educators should stay updated with current GenAI trends and encourage student interaction to explore concepts beyond answer-getting.
2	Marco & Stylianides (2025)	Investigate ChatGPT-4's mathematical knowledge using Plato's slave-boy experiment and interactive chat guidelines.	Defined the "Chat's ZPD" for problems solvable with user prompting; GenAI knowledge reflects both recollection and generation patterns.	Use designated prompts to help GenAI make cross-contextual connections and specify the type of knowledge required.
3	Parra et al. (2024)	Analyse the abilities of ChatGPT, Bing Chat, and Bard to provide solutions to a Spanish-language geometry competition problem.	Performance was disappointing; errors were categorized into Construction, Conceptual, and Contradiction due to lack of symbolic structure.	Use errors as educational tools to promote critical thinking and reflection on alternative solutions.
4	Wardat et al. (2023)	Examine stakeholder perspectives and user experiences through interviews and three educational scenarios.	GenAI improves task performance but lacks deep understanding of geometry and cannot correct misconceptions effectively.	Ensure secure and accountable integration of GenAI into mathematics curricula.
5	Busuttil & Calleja (2025)	Investigate secondary teacher beliefs and practices in Malta using a case study with pre- and post-PD surveys.	Teachers with discovery and connectionist philosophies view GenAI positively as a tool for active learning and exploration.	Provide ongoing professional development to help teachers integrate GenAI effectively into pedagogy.
6	Seebut et al. (2024)	Evaluate student self-efficacy after a three-phase activity combining GPT-3.5 and Google Colab to solve difference equations.	Integration significantly enhanced self-efficacy and shifted students' roles from programmers to work directors/prompt engineers.	Students must develop conceptual foundations and procedural knowledge before relying on GenAI.
7	Schorcht et al. (2025)	Explore a network of customized GPT agents functioning as a multi-professional team to modify mathematical tasks.	Balancing text generation and precision is crucial; the network enriched tasks with solution hints and actions.	Refine inter-agent communication protocols and reduce information overload in AI-modified tasks.

8	Segal & Klemer (2025)	Observe teachers interacting with Claude 2.1 to plan primary lessons using mixed-methods.	Teacher-GenAI dialogue improved all TPACK components; iterative dialogue and critical evaluation of GenAI gaps are required.	Maintain a balance between GenAI innovation and human expertise; tailor professional development to individual teacher contexts.
9	Delima et al. (2024)	Investigate the impact of ChatGPT and music on MSRL and Math Anxiety using correlational descriptive research.	GenAI use significantly enhanced MSRL; neither GenAI nor music had a significant effect on anxiety.	Use GenAI to support independent exploration; integrate music with GenAI to influence learners' dispositions and habits of mind.
10	Kock et al. (2025)	Analyse three interviews with engineering students using the Instrumental Approach to identify initial utilization schemes.	GenAI acts as a co-agent for conceptual explanation and code optimization; students developed critical attitudes toward hallucinations.	Institutions should monitor and guide students in developing sophisticated GenAI utilization schemes over time.
11	Davis (2025)	Use an instrumental approach to analyse prompt logs and synopses of four prospective teachers.	Identified three instrumentalization and five instrumentation schemes; GenAI limitations forced a shift from graphical to algebraic content.	Share iterative refinement and specificity schemes with future teachers to reduce out-of-class lesson preparation time.
12	Pavlova (2024)	Present a case study flipping the classical dialogic method using ChatGPT to stimulate student activity.	The model promotes active roles and research qualities; students initially accept results without critical reformulation.	Teach students to ask "the right questions" and detect GenAI gaps to reduce stress and enhance learning.
13	Susilawati et al. (2025)	Investigate the synergy between GenAI, RME, and gamification using a sequential explanatory mixed-methods design.	Integrated model significantly improves problem-solving (β=0.40) and motivation (β=0.35); motivation acts as a structural mediator.	Align gamification with deep thinking; implement adaptive scaffolding that preserves productive struggle.
14	Contel & Cusi (2025)	Investigate GPT-4 scaffolding of metacognition using micro/macro analysis based on the MAEAB construct.	Identified replacement vs. diagnostic schemes; GenAI tends to bypass planning and exhibits semantic limitations.	Design interventions to support critical, reflective mathematical thinking beyond replacement schemes.
15	Yunianto et al. (2024)	Use educational design research (EDR) to investigate ChatGPT assisting with GeoGebra code generation for area of circle tasks.	GenAI supported programming and debugging; its varied responses allow differentiated instruction.	Provide professional development to help students develop effective prompts and critically evaluate GenAI responses.
16	Stamenkova (2025)	Evaluate 11 LLM models solving problems from the Bulgarian National External Assessment (NEA).	Models like DeepSeek and Mistral achieved perfect accuracy but often deviated from national curricular standards.	Ensure success metrics include mathematical accuracy, methodological alignment, and culturally relevant terminology.
17	Borys et al. (2025)	Assess the effectiveness of GenAI-driven exercise generation using student	Pass rates improved; GenAI accelerates exercise creation but content requires human validation.	Prioritize process-supervised reward models to improve accuracy and

		performance data on the Compass platform.		reduce hallucinations in human-GenAI alignment.
18	Walkington et al. (2025)	Case study of 36 girls using GenAI for game design pitches involving story problems and DALL-E2 visuals.	Students identified racial and gender bias in GenAI images; ChatGPT problems often had low authenticity and mathematical errors.	Provide guidance on prompting strategies to sound like students and minimize biased or inappropriate outputs.
19	Zhuang & Zhang (2025)	Track PSMTs' perceptions of ChatGPT 4.0 and Student GPT over a semester using surveys and interviews.	Perceptions became more positive with familiarity; Student GPT was valued as a practice-based training tool.	Teacher education programs should offer structured guidance and formative feedback on prompt writing and GenAI evaluation.
20	Noster et al. (2024)	Analyse 11 teachers solving four tasks using qualitative content analysis and the KTMT model.	Teachers lacked prompting knowledge and frequently accepted erroneous GenAI outputs.	Explicitly address LLMs and prompting techniques in teacher training to ensure teachers can judge mathematical fidelity.
21	Horrocks & Shearman (2025)	Use thematic analysis of 33 educator interviews to identify core goals in the age of GenAI.	Four primary themes emerged: content mastery, mathematical thinking, communication, and appreciation.	Reframe mathematics and statistics as experimental sciences; prioritize conceptual depth over algorithms.
22	Kim et al. (2025)	Investigate PT capacities for error recognition and correction using AIM activities and AIS strategies.	PTs excelled at verifying correct problems but struggled (53.5% failure rate) to identify errors.	Provide multiple opportunities for PTs to evaluate GenAI outputs.
23	Sutrisno et al. (2025)	Investigate justification abilities of 13 students using a descriptive qualitative study on linear equations.	61.5% understood instructions; students struggled to identify GenAI reasoning errors due to weak conceptual mastery.	Strengthen conceptual understanding and critical thinking to evaluate GenAI outputs independently.
24	Yoon et al. (2024)	Develop the SIPE-AI framework through task-based interviews observing engagement with proving tasks.	Proving with GenAI involves a cyclic process shaped by students' and AI's conceptions and ethical considerations.	Encourage students to take authorship over final proofs and establish class norms for ethical GenAI use.
25	Milenković et al. (2025)	Rank o1 and o3-mini solutions against student results from the 2024 Serbian National Competition.	Both tools qualified for medals in seven of eight cases; o3-mini slightly outperformed o1-mini.	Use GenAI for strategic guidance and inspiration; maintain critical analysis of reasoning processes.
26	Xuan et al. (2025)	Comparative study examining the impact of GenAI on mathematics learning in Vietnamese private and public high schools.	GenAI improved understanding in private schools but was less effective than traditional teaching in resource-limited public schools.	Ensure contextual readiness before implementation; improve infrastructure and AI literacy; conduct longer-term studies.
27	Gurl et al. (2025)	Qualitative study examining PST reactions to mathematical and pedagogical outputs during microteaching.	PSTs were accurate on pedagogy but often accepted incorrect solutions as "different approaches."	Develop a framework to examine GenAI-generated mathematical outputs.

28	Daher & Anabousy (2025)	Study four bots in writing lessons based on Kolb's model using content analysis.	Bots succeeded in lesson writing but differed in defining teaching strategies; Perplexity used authoritative citations.	Use GenAI for co-design workshops; emphasize referencing accurate sources.
29	Wahba et al. (2024)	Evaluate the influence of ChatGPT-based learning on undergraduates using quasi-experimental pre- and post-tests.	ChatGPT-based learning significantly enhanced statistical reasoning and positive attitudes.	Motivate students to use GenAI in learning statistics and incorporate it into teaching methods.
30	Etcuban (2025)	Investigate ChatGPT's use in addressing anxiety and promoting confidence among 320 freshmen using a descriptive correlational method.	Students reported reduced algebra anxiety and increased confidence, leading to "Very Good" performance scores.	Integrate ChatGPT into algebra courses to promote practical and diverse learning settings.

Research Objectives

This study aims to map and describe existing empirical research on the integration of generative artificial intelligence (GenAI) in mathematics teaching and learning. The research objectives of this study are as follows:

1. To identify how GenAI tools are used in mathematics teaching and learning across different educational levels and mathematical topics.
2. To examine the pedagogical roles attributed to GenAI tools in mathematics education.
3. To analyse the types of research designs employed in empirical studies on GenAI in mathematics education.
4. To synthesize the benefits, challenges, and limitations of GenAI integration in mathematics teaching and learning as reported in the literature.

Research Questions

This study is guided by the following research questions:

1. How are generative artificial intelligence (GenAI) tools used in mathematics teaching and learning across different educational levels and mathematical topics?
2. What pedagogical roles are attributed to GenAI tools in mathematics education?
3. What types of research designs are employed in empirical studies on the use of GenAI in mathematics education?
4. What benefits, challenges, and limitations of GenAI integration in mathematics teaching and learning are reported in the existing literature?

Methods

The current study adopted a systematic mapping review approach to examine existing empirical research on the use of GenAI in mathematics teaching and learning. Systematic mapping reviews provide a structured overview of an emerging research field by identifying patterns, trends, and gaps in the literature, rather than synthesizing evidence of effectiveness. The analysis was conducted in accordance with the PRISMA (Preferred Reporting Items to Systematic Reviews and Meta-Analyses) framework, according to which the high-quality, transparent, and reproducible reviews should be conducted (Moher et al., 2009; Page et al., 2021). PRISMA involves four steps, including identification, screening, provision of eligibility, and inclusion. The initial step identifies potentially relevant studies through systematic database searches, followed screening for duplicates and relevance, assessing eligibility, and including studies meeting the criteria. This approach was selected as it allows for a consistent, transparent, methodologically rigor synthesis of research trends, pedagogical applications, and limitations of GenAI tools, including ChatGPT, Gemini and Claude across different educational levels. Figure 1 presents an adapted PRISMA framework depicting the study selection process.

A comprehensive literature search was conducted using two major citation indexing databases (Web of Science and Scopus) and the full-text publisher platform Science Direct. Keywords and phrases were applied to article titles, abstracts, and author-generated keywords to identify studies on GenAI in mathematics education. This includes search terms like "generative AI," "generative artificial intelligence," "ChatGPT,", "large language model,", as well as domain-specific terms like problem-solving, algebra, calculus, geometry, and STEM mathematics. To ensure comprehensiveness, the reference lists of retrieved articles were manually screened to identify additional studies not captured in the initial searches. Mendeley Reference Management software was used to remove the duplicates, and only studies published in English were considered. Using the PRISMA framework, the search was limited to studies published between 2022 and 2025, and the final database search was conducted on 31 January 2025. This timeframe aligns with the emergence and rapid expansion of GenAI technologies within mainstream academic and educational discourse, especially after the introduction of GPT-3 and GPT-4. Table 2 provides the detailed search strings.

Following the identification stage, records were screened based on titles, abstracts, and keywords. Full-text screening was then conducted using predefined inclusion and exclusion criteria focusing on empirical studies that examined applications, opportunities, or challenges of GenAI in mathematics education across educational levels. The final set of included studies was analysed descriptively and thematically to map research characteristics, pedagogical uses, and reported limitations of GenAI in mathematics teaching and learning.

A systematic mapping review approach was selected for this study due to the emerging, rapidly evolving, and methodologically heterogeneous nature of research on generative artificial intelligence (GenAI) in mathematics education. Existing empirical studies vary widely in terms of research design, educational context, unit of analysis, and reported outcomes, making effectiveness synthesis or quality-weighted integration premature at this stage of the field's development. Systematic mapping reviews are particularly appropriate for providing an overview of research trends, identifying concentrations and gaps in the literature, and describing how a research area is being investigated, rather than evaluating the strength or effectiveness of interventions. Accordingly, this review focuses on mapping reported applications, pedagogical roles, and challenges of GenAI in mathematics teaching and learning, rather than drawing causal or effectiveness-based conclusions. As a systematic mapping review, this study does not evaluate the effectiveness of GenAI interventions or weight findings based on methodological quality, but instead aims to provide an overview of research trends, reported applications, and gaps in the existing literature

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

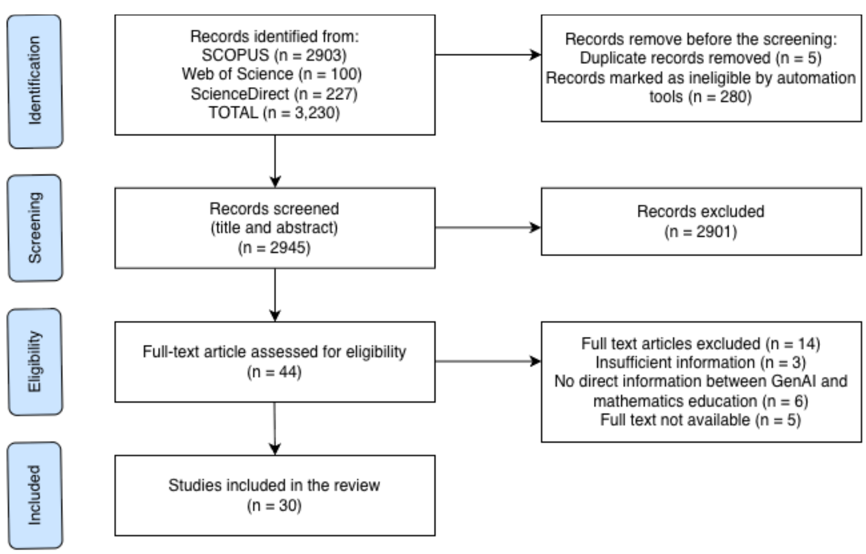


Table 2
The Search Strings

Database	Search Equation
Scopus	(("generative AI" OR "generative artificial intelligence" OR "ChatGPT" OR "large language model" OR "AI-powered" OR "artificial intelligence") AND TITLE-ABS-KEY("mathematics education" OR "math education" OR "algebra" OR "calculus" OR "geometry" OR "problem-solving" OR "STEM mathematics"))
Web of Science (WoS)	(TITLE-ABS-KEY("generative AI" OR "generative artificial intelligence" OR "ChatGPT" OR "large language model" OR "AI-powered" OR "artificial intelligence") AND TITLE-ABS-KEY("mathematics education" OR "math education" OR "algebra" OR "calculus" OR "geometry" OR "problem-solving" OR "STEM mathematics"))
ScienceDirect	(("generative AI" OR "generative artificial intelligence" OR "ChatGPT" OR "large language model" OR "AI-powered" OR "artificial intelligence") AND TITLE-ABS-KEY("mathematics education" OR "math education" OR "algebra" OR "calculus" OR "geometry" OR "problem-solving" OR "STEM mathematics"))

Data Analysis

The systematic mapping review focused exclusively on peer-reviewed literature articles that examine the use of generative artificial intelligence (GenAI) in mathematics education. Studies were included if they explored the applications, challenges, or pedagogical implications of GenAI in mathematics teaching and learning, provided empirical evidence regarding student learning, involvement, motivation, or conceptual understanding, or addressed teachers' perspectives on integrating GenAI. To ensure methodological rigor and relevance, non-empirical sources such as editorials, review papers, book chapters, conference proceedings, dissertations, and other non-peer-reviewed articles were excluded. Studies published in languages other than English were also omitted to guarantee accurate interpretation. Studies in all levels of education, such as primary, secondary, and tertiary studies, were considered to ensure a comprehensive overview of the GenAI practices in mathematics education.

These databases were selected due to their comprehensive coverage, academic rigor, and relevance to educational research. Web of Science and Scopus are widely recognized citation indexing databases that index high-quality, peer-reviewed journals across multiple disciplines, ensuring methodological rigor and credibility in the studies retrieved. ScienceDirect, as a full-text publisher platform, provides direct access to current and relevant articles, particularly in education, computer science, and STEM-related fields, allowing for efficient retrieval of empirical studies on GenAI applications in mathematics education. The combination of these databases maximizes coverage of high-impact journals, reduces the risk of missing main studies, and ensures that the review captures a representative sample of research trends and pedagogical practices across different educational levels and geographical regions. Table 3 summarizes the inclusion and exclusion criteria applied in this study.

Table 3
The Selection Criterion in Searching

Inclusion Criteria (IC)	Exclusion Criteria (EC)
IC1. Peer-reviewed articles published in indexed journals (Web of Science, Scopus, ScienceDirect)	EC1. Non-indexed journals, book chapters, review papers, conference proceedings, dissertations, prefaces, or opinion pieces
IC2. Studies covering all educational levels in mathematics (primary, secondary, tertiary)	EC2. Studies unrelated to mathematics education
IC3. Focus on the application of generative AI in mathematics education	EC3. Studies mentioning AI but not focusing on generative AI in mathematics education
IC4. Reports findings on opportunities, challenges, or practical applications of generative AI in teaching and learning mathematics	EC4. Studies lacking discussion of outcomes, applications, or challenges of generative AI
IC5. Published in English	EC5. Published in languages other than English
IC6. Specifically applied to mathematics or mathematics education	EC6. General application in other professional fields (e.g., medicine, public health, environmental science, engineering)

To extract data for mapping, each study was coded for key characteristics, including GenAI tool type (e.g., ChatGPT, Gemini, Claude, Copilot, Khanmigo, Compass), education level (K-12, undergraduate, teacher education), mathematics topic (algebra, geometry, calculus, statistics, number theory, mixed/problem-solving), research design (qualitative, quantitative, mixed-methods), and geographical location. These variables guided both descriptive reporting and thematic analysis, ensuring alignment with the study objectives of mapping trends, gaps, and pedagogical applications.

The systematic evaluation of each study followed a multi-stage process based on PRISMA framework, that involved identifying, screening, eligibility, and inclusion. This strategy ensured that only the studies that fit the predefined criteria were retained to qualitative thematic analysis and offered reliable information about the pedagogical functions, capabilities, limitations, and educational implications of GenAI. During the identification step, 3,230 records were initially retrieved from multiple databases. Five duplicates were removed in the process of screening with the help of Mendeley, resulting in 2,945 records for further evaluation. These were screened based on the titles, abstracts, and keywords. This leads to the exclusion of 2,901 studies because it was irrelevant to mathematics education, non-peer-reviewed, lack of focus on GenAI, or was not written in English. During the eligibility phase, the full texts of the rest of the 44 articles were assessed against inclusion criteria, whereby a focus on GenAI in mathematics education, empirical results, and a participation of learners at any level of education were considered. Fourteen articles were excluded due to insufficient relevance, missing information, or entire text was not available. Lastly, 30 studies that met all the criteria were included for qualitative thematic analysis. These studies were reviewed to identify pedagogical applications, research trends, and the capabilities and limitations of GenAI tools.

The thematic analysis followed a systematic coding process. An inductive approach was adopted to identify recurring pedagogical applications, reported challenges, and trends in GenAI usage. Two independent coders reviewed all 30 studies, with discrepancies resolved through discussion to ensure reliability. NVivo software was used to organize codes and track themes. Themes were then cross-checked against extracted variables to confirm consistency with descriptive results, including tool usage, educational level, mathematics topic, and geographical distribution.

This approach ensures that the review provides a structured and reproducible overview of GenAI research in mathematics education, highlighting patterns, gaps, and emerging trends rather than evaluating effectiveness or weighting findings based on methodological quality.

Results

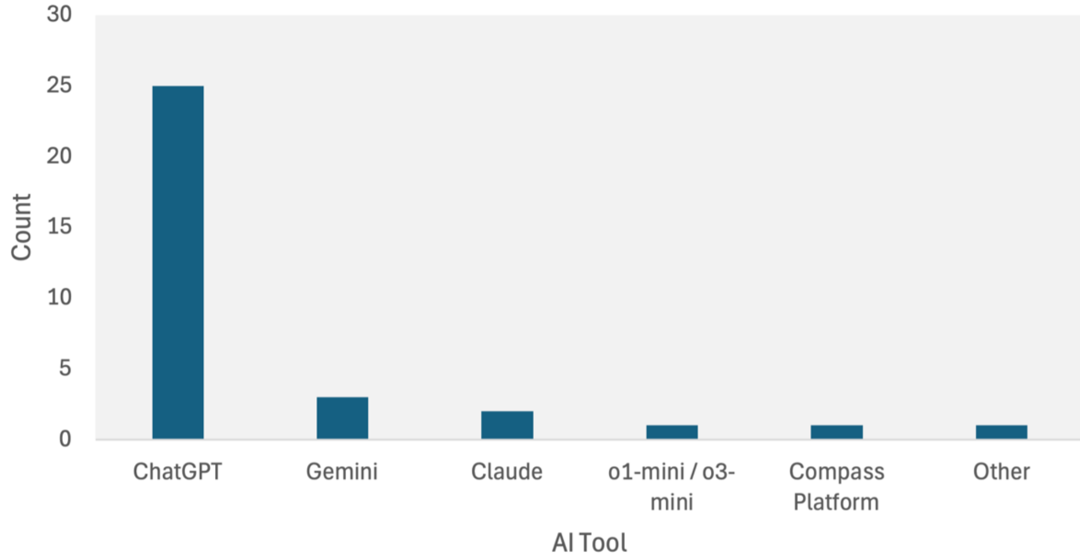
GenAI Tool Usage

In relation to RQ (i), which examines how GenAI tools are used in mathematics teaching and learning across educational levels and mathematical topics, Figure 2 represents the distribution of GenAI tools across the 30 reviewed studies. ChatGPT (all versions) was the most frequently used tool with 25 studies (83%), showing it was widely used in mathematics education research. Other tools were considerably less represented: Gemini (3 studies; 10%), Claude (2 studies; 7%), OpenAI o1-mini and o3-mini (1 study each; 3%), Compass Platform (1 study; 3%), and emerging platforms such as Student GPT, Copilot and Khanmigo (1 study each; 3%).

The strong concentration on ChatGPT suggests that empirical evidence regarding GenAI in mathematics education is largely shaped by a single dominant platform. This dominance may reflect factors such as early public release, ease of access, widespread institutional adoption, and strong media visibility (Seebut et al., 2024; Walkington et al., 2025; Zhuang & Zhang, 2025; Xuan et al., 2025). Methodologically, this concentration indicates a potential platform bias, where conclusions about "GenAI in mathematics education" are primarily derived from the affordances and limitations of one tool rather than representing the broader ecosystem of generative AI systems.

The limited representation of alternative platforms suggests that comparative research across different GenAI architectures remains underdeveloped. Consequently, current findings may reflect tool maturity and accessibility rather than systematic evaluation of diverse GenAI technologies. This pattern indicates that while research activity is expanding, diversification of investigated platforms is still emerging.

Figure 2
AI Tool Usage Frequency in the Reviewed Studies



Education Level

In response to RQ (i), which examines how GenAI tools are used across different educational levels, Figure 3 presents the distribution of studies by educational context. Of the 30 included studies, 13 (43%) were conducted in undergraduate or higher education settings, 12 (40%) in K–12 contexts (middle and high school), and 8 (27%) in teacher education or pre-service teacher programs.

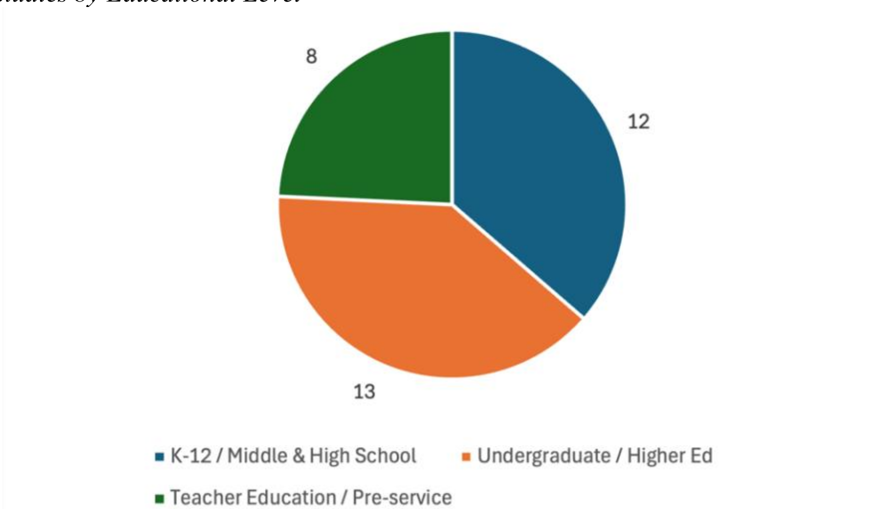
For classification purposes, undergraduate or higher education refers to studies involving university-level students, including postgraduate participants where specified. Teacher education was treated as a distinct category to capture research specifically focused on the professional preparation of future educators, even though such programs are institutionally situated within higher education. These categories were coded based on the primary participant group identified in each study to minimize overlap.

The relatively balanced distribution between K–12 (40%) and higher education (43%) suggests that GenAI integration is not confined to a single educational stage but is being explored across developmental levels (McGalliard & Otten, 2025; Delima et al., 2024; Xuan et al., 2025). The slightly higher concentration in higher education may reflect fewer institutional restrictions, greater digital infrastructure, and increased academic freedom to experiment with emerging technologies (McGalliard & Otten, 2025; Xuan et al., 2025). In contrast, the strong representation in K–12 contexts indicates growing interest in early curricular integration and classroom-level experimentation (Segal & Klemer, 2025; Parra et al., 2024).

The presence of teacher education studies (27%) is particularly significant, as it signals recognition that effective GenAI integration depends not only on student use but also on preparing educators to critically and pedagogically engage with these tools. This suggests that the field is moving beyond exploratory student-focused applications toward broader professional and systemic considerations. Overall, the distribution indicates that research on GenAI in mathematics education is expanding across educational levels, though institutional context may influence the pace and nature of adoption.

e-ISSN 2600-8572 | DOI: https://doi.org/10.11113/itlj.v10.256

Figure 3
Distribution of Studies by Educational Level



Mathematics Topics

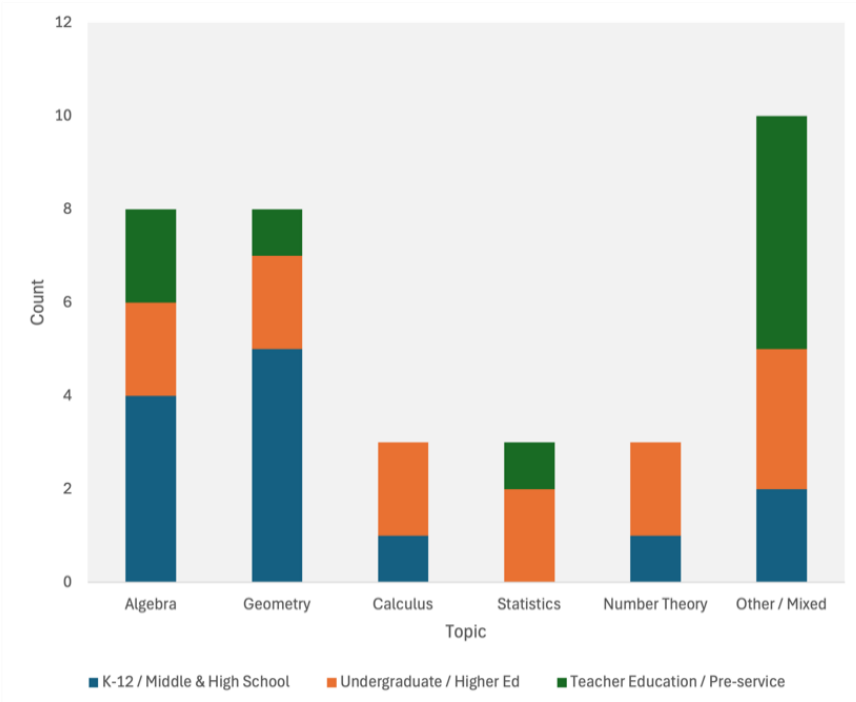
Figure 4 presents the distribution of studies by topic. Across the 30 included studies, Algebra was the most frequently investigated topic, appearing in 14 studies (47%), followed by Geometry in 9 studies (30%). Calculus was addressed in 2 studies (7%), Statistics in 3 studies (10%), Number Theory in 3 studies (10%), and Other/Mixed topics in 7 studies (23%).

For classification, each study was coded based on the mathematics topic(s) primarily addressed. Studies covering multiple topics were counted in all relevant categories. The "Other / Mixed" category includes problem-solving, mathematical literacy, and RME-based activities, often spanning multiple educational levels, particularly in teacher education or pre-service contexts.

The distribution indicates that research on GenAI in mathematics education focuses predominantly on foundational and visual-spatial reasoning topics, such as Algebra and Geometry, across K–12 and undergraduate contexts. Higher-order topics like Calculus, Statistics, and Number Theory are less frequently studied and are primarily explored in undergraduate or higher education settings. The presence of Other/Mixed topics highlights attention to integrative, cross-topic applications, especially in professional preparation for future educators. Overall, these patterns suggest that GenAI research in mathematics education prioritizes core foundational topics, while also exploring broader, integrative applications to support diverse pedagogical and curriculum needs.

Figure 4

Mathematics Topics Addressed in Studies across Educational Contexts



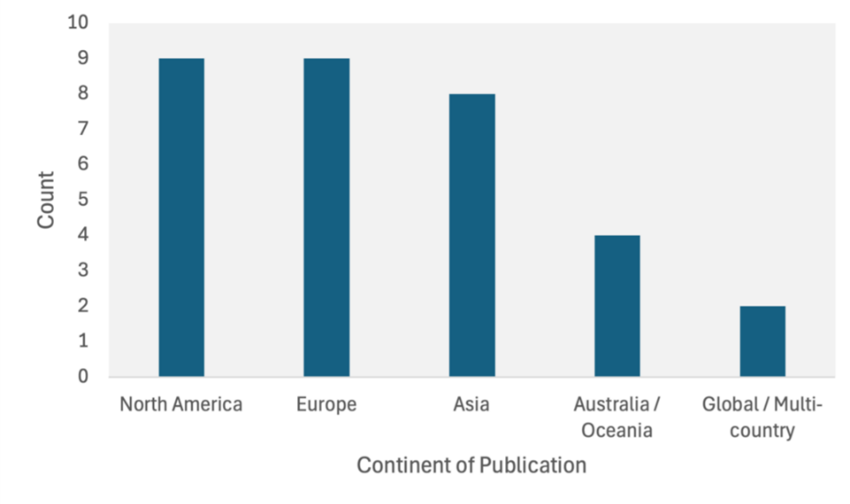
Geographical Distribution

Figure 5 shows the distribution of the reviewed studies according to continent, based on the country where the research was conducted. Of the 30 included studies, 9 (30%) were conducted in North America, 9 (30%) in Europe, 8 (27%) in Asia, 4 (13%) in Australia/Oceania, and 2 (7%) involved multiple countries or global collaborations. In Asia, studies were conducted in countries such as Indonesia, Israel, Thailand, South Korea, Jordan, and the Philippines. Studies conducted across multiple countries were categorized under Global / Multi-country to avoid duplication.

This distribution implies that research in the field of GenAI in mathematics education is concentrated in North America, Europe, and part of Asia, with other regions underrepresented. The findings highlight potential geographical gaps in the literature, suggesting the need for broader international exploration to understand contextual differences in GenAI implementation and adoption.

e-ISSN 2600-8572 | DOI: https://doi.org/10.11113/itlj.v10.256

Figure 5
Continent of Publication



Discussion

Capabilities and Limitations of Generative AI in Mathematics

Across the reviewed studies, GenAI is described as flexible across mathematics education contexts, serving as a problem-solving assistant, content generator, and scaffolding tool across multiple levels (McGalliard & Otten, 2025; Seebut et al., 2024; Segal & Klemer, 2025). However, the effectiveness reported varies considerably depending on study design and methodological rigor; for example, in-depth qualitative analyses (McGalliard & Otten, 2025) provide richer insights into AI reasoning limitations compared to simple pre-post quantitative surveys (Delima et al., 2024). Eighteen of the thirty studies (60%) reported the use of AI for procedural problem-solving, while twelve studies (40%) highlighted its role in content generation and scaffolding. For example, McGalliard and Otten (2025) evaluated ChatGPT and Gemini 1.0 on cognitively demanding secondary tasks and discovered that AI could solve complex problems, such as rates, quadratics, and geometry proofs. Critically, these findings intersect with Marco and Stylianides (2025) who showed that AI's performance is tightly bound to human guidance. This findings emphasize that procedural success does not equate to conceptual understanding. Initial outputs usually had a lot of errors, which require iterative prompting and refinement. Based on the Technological Pedagogical Content Knowledge (TPACK) framework, the findings highlight the importance of integrating technological knowledge with pedagogical strategies to ensure that AI-mediated solutions are conceptually accurate and educationally meaningful (Mimoudi & Mokhtari, 2025; Tan et al., 2025).

Marco and Stylianides (2025) then introduced the concept of a "Chat's Zone of Proximal Development," emphasizing that AI can assist in solving the tasks that can be solved with human guidance, which is a combination of recollection and generative reasoning. Synthesizing this with McGalliard and Otten (2025) and Segal and Klemer (2025), it becomes clear that AI's pedagogical value emerges primarily at the intersection of task design, user input, and iterative scaffolding, rather than as a standalone tool. Although this analogy to Vygotskian scaffolding is insightful, inappropriate use of AI may overstate its role as a cognitively adaptive agent; since AI cannot truly diagnose errors or adapt responsively beyond surface-level patterns, which constrains its transformative pedagogical potential.

At the higher education levels, GenAI supports computational and programming tasks. Seebut et al. (2024) reported that integrating GPT-3.5 with Google Colab to solve difference equations improved students' self-efficacy and encouraged roles such as prompt engineers instead of passive consumers of content. Likewise, Yunianto et al. (2024) demonstrated that GenAI could generate GeoGebra code to solve geometry problems, enabling differentiated learning that catered to diverse student paces. However, these apparent benefits are double-edged; students may become more efficient, but also might lose the ability to think conceptually, since AI can mask procedural fluency with automated outputs, a concern that Seebut et al. (2024), briefly acknowledge, but do not fully explore. These findings illustrate

GenAI's dual roles as a virtual peer and procedural guide, which makes mathematical computation more engaging, but at the same time it supports the idea that GenAI's contributions remain fundamentally instrumental rather than intellectually generative. From a socio-didactical perspective, this highlights the necessity of teacher mediation in co-constructing mathematical understanding, ensuring that AI serves as a scaffold rather than a substitute for conceptual engagement (Tan et al., 2025).

GenAI exhibits significant limitations, especially in abstract reasoning. Parra et al. (2024) analysed large language models (LLMs) solving geometry competition problems in Spanish and identified errors primarily in three categories: Construction, Conceptual, and Contradiction. Eight studies reported similar domain-specific reasoning challenges, and it shows that AI reasoning is fragile and domain-specific. Intersections with Wardat et al. (2023) indicate that repeated limitations in conceptual reasoning are consistent across studies, suggesting a systemic challenge rather than isolated incidents. Moreover, ChatGPT struggles to rectify errors in geometry, which reinforces the view that GenAI cannot yet replace human conceptual mediation. This limitation challenges overly optimistic narratives about GenAI democratizing high-level mathematical tasks, showing that human expertise is still required in the process of guiding abstraction and reasoning.

Contextual weaknesses are evident as well. High-level spatial problems (Marco & Stylianides, 2025) and advanced calculus tasks (Parra et al., 2024) are among the problems that models like ChatGPT fail to solve successfully. Reports of differing accuracy and national curricular standards across the studies highlight that such differences are not merely technical but epistemologically significant (Stamenkova, 2025). In effect, educators must critically evaluate AI-generated solutions, before introducing them to students. This approach emphasizes the teacher's dual role in facilitation and verification. Learning theories, such as Vygotsky's socio-cultural theory, further support the need for guided interaction, where AI outputs are interpreted, critiqued, and integrated within students' evolving cognitive structures (Boulhrir & Hamash, 2025).

The pedagogical implication is clear: AI should function as a complementary tool rather than a replacement. Teachers must guide students on how to interpret, verify, refine GenAI outputs, and encourage students to focus on prompt engineering and critical evaluation (Mani & Zargeh, 2025; Borys et al., 2025). Integration across K–12 and higher education studies suggests that effective AI use requires teacher competency, digital literacy, and curricular awareness. This findings indicate that the potential benefits of GenAI may be unevenly distributed and could further deepen educational disparities.

GenAI Enhancing Conceptual Understanding, Engagement, and Motivation

GenAI has significant potential for enhancing conceptual learning in mathematics, though its scaffolding is limited. Across the reviewed studies, three cross-cutting themes emerge: metacognitive support, motivation and engagement, and multiple solution strategies. As evidenced by Contel and Cusi (2025), GPT-4 can support metacognitive strategies when solving problems, but it tends to bypass planning stages, and create diagnostic instead of replacement learning opportunities. This implies that AI scaffolding may reinforce fragmented reasoning, as students receive solutions without fully internalizing procedural logic, a risk often underemphasized in the literature. Flipped-dialogue models, as presented by Pavlova (2024), exemplify how structured AI-student interaction can stimulate critical thinking and research skills. Evidence from eight other studies similarly indicates that AI cannot independently diagnose or adapt beyond surface-level patterns, emphasizing the role of human orchestration and instrumental genesis (Wardat et al., 2023; Kock et al., 2025).

GenAI appears most effective in terms of motivation and engagement when integrated into interactive or gamified settings. According to Susilawati et al. (2025), methods such as AI combined with Realistic Mathematics Education (RME) and gamification significantly improved problem-solving and engagement, with motivation acting as a structural mediator. Walkington et al. (2025) showed that middle school girls using ChatGPT and DALL-E2 attempted to pose mathematical problems, though outputs required careful human supervision to counter bias and inaccuracies. While promising, gains in engagement are conditional: self-reported motivation increased (Delima et al., 2024; Susilawati et al., 2025), yet few studies triangulated these outcomes with performance data. Critical evaluation of these studies shows that strong methodological designs, such as sequential explanatory or mixed-method studies, offer richer insights on how motivation interacts with teacher support and task complexity, while weaker designs risk overgeneralizing benefits.

AI also supports multiple solution strategies, contributing to deeper understanding. Wardat et al. (2023) discovered that ChatGPT could generate alternative explanations for algebraic manipulations or statistical reasoning, benefiting students who struggle with traditional teaching approaches. Similarly, Kock et al. (2025) described students treating AI as a co-agent to conceptualize problems and optimize solutions, at the same time, developing critical evaluation skills. However, the co-agent framing can be exaggerated when it comes to the capacity of AI to think independently. The model lacks intentionality and adaptive cognition, i.e. the students need to constantly filter outputs for accuracy. In this context, deterministic reasoning refers to the tendency of students to accept AI-generated outputs as fixed, authoritative solutions without interrogating underlying assumptions or alternative pathways (Kruijssen & Emmons, 2025). Theoretically, this aligns with concerns in metacognitive and self-regulated learning research, where reduced monitoring and evaluation can constrain deeper conceptual processing (Panadero et al., 2025; Shekh-Abed, 2025). This introduces a potential cognitive load, which is rarely measured but could constrain learning, consistent with Sweller's (2011) cognitive load theory.

The gains of engagement are, however, not uniform. Parra et al. (2024) reported that GenAI struggles with complex geometrical problems, which may reduce the confidence of students if errors are misinterpreted as correct. Ethical concerns are present as well, including students relying on AI to generate proofs without fully understanding underlying the concepts (Yoon et al., 2024). This indicates a larger conflict, in which AI can motivate students superficially but fail to develop deep comprehension, thus posing a paradox that requires extensive pedagogical guidance. Effective integration thus requires active reflection, iterative prompting, and teacher-mediated dialogue (Segal & Klemer, 2025; Mani & Zargeh, 2025).

From a TPACK perspective, these findings highlight the importance of integrating technological knowledge with pedagogical and content knowledge to ensure AI-mediated solutions are conceptually accurate and educationally meaningful. Similarly, socio-didactical frameworks emphasize the teacher's role in shaping AI outputs into coherent learning experiences, while instrumental genesis underscores the necessity of students internalizing AI-mediated reasoning rather than passively consuming solutions.

In summary, GenAI has clear potential to enhance conceptual understanding, engagement, and motivation, but these benefits are contingent on teacher guidance, structured mediation, and student prior competence. Gains are more pronounced when AI is integrated with reflective, interactive, or gamified activities, yet uneven across learners and educational levels. Effective implementation requires active reflection, critical evaluation, and theoretically informed scaffolding to prevent superficial engagement and maximize conceptual learning.

Pedagogical Integration of GenAI in Mathematics Education

When it comes to teaching mathematics, GenAI implementation must be mediated by the teacher. Across the reviewed literature, pedagogical integration consistently follows a teacher-regulated orchestration model, in which AI operates within established instructional structures rather than as an autonomous agent. Mani and Zargeh (2025) proposed a socio-didactical tetrahedron positioning the student at the center with AI as a new vertex, emphasizing that teacher oversight is essential to prevent over-reliance. Segal and Klemer (2025) reported that teacher-AI conversation improved the TPACK components but required iterative evaluation to ensure pedagogical integrity. Synthesizing these perspectives suggests a unified pedagogical model: AI integration is less about innovation and more about restructuring current teacher-led practices, challenging claims that AI inherently transforms pedagogy. This teacher-mediated and context-sensitive model is reinforced by Xuan et al. (2025), who found that GenAI demonstrated marginal benefits in well-resourced private schools characterized by strong digital literacy and self-regulated learning, whereas traditional teacher-centered instruction remained more effective in resource-constrained public schools. Their findings highlight that AI effectiveness is contingent upon contextual readiness, infrastructure, and pedagogical scaffolding rather than technological capability alone.

Professional development should focus on prompt engineering, critical evaluation, and ethical use. A cross-study synthesis reveals three emerging professional development orientations: (1) technical-operational training (prompt

e-ISSN 2600-8572 | DOI: https://doi.org/10.11113/itlj.v10.256

construction and tool navigation), (2) pedagogical-design training (embedding AI into task design and lesson sequencing), and (3) critical-ethical training (bias detection, epistemic evaluation, and responsible authorship). Relying on the guided practice of AI prompts, Zhuang and Zhang (2025) noted that the perceptions of pre-service teachers improved, which shows the significance of hands-on training. Noster et al. (2024) emphasized that without sufficient prompt knowledge, it leads to the uncritical acceptance of AI outputs. However, most empirical evidence is concentrated in higher education contexts and specific geographical regions, as identified in the mapping results, which limits systematic generalization across systems. This points to a structural constraint since teacher competence and digital literacy are crucial factors that determine the effectiveness of AI integration. While several studies imply that unequal access to structured professional development may widen existing gaps, comparative cross-level or cross-regional evidence remains limited, these findings imply that AI may worsen inequities rather than address them unless systematic support is provided.

GenAI also presents new pedagogical opportunities, such as lesson planning, simulation of problem-solving tasks, and rapid generation of differentiated teaching materials (Gurl et al., 2025; Daher and Anabousy, 2025). According to Borys et al. (2025), AI-generated exercises can accelerate learning process but require human validation to ensure accuracy. Across studies, this recurring need for validation positions AI primarily as an efficiency-enhancing augmentation tool rather than an independent mechanism of knowledge construction. It can act as a co-designer of the learning experience and help teachers to create learner-centered classrooms. However, it remains highly reliant on human expertise.

Ethical and practical considerations remain critical. Sawyer and Aga (2025) discovered that AI-generated content could be repetitive or shallow, particularly in elementary curricula. Educators need to instruct students in critical thinking, authorship responsibility, and ethical engagement with AI-generated materials (Yoon et al., 2024; Sutrisno AB et al., 2025). Synthesized across studies, these concerns converge on the need for explicit epistemic scaffolding to prevent deterministic acceptance of AI outputs. In the absence of explicit ethical scaffolding, AI may unintentionally promote superficial thinking, which reinforces surface-level engagement rather than meaningful mathematical reasoning.

Conclusion

This systematic mapping review investigated how GenAI tools are integrated into mathematics education across educational levels, mathematical topics, and geographical contexts. The review reveals that the integration of GenAI predominantly revolves around a single tool, ChatGPT, which appeared in 83% of the studies, while other platforms such as Gemini, Claude, and emerging tools are underrepresented. Research evidence (McGalliard & Otten, 2025; Contel & Cusi, 2025; Delima et al., 2024) shows that GenAI can enhance student engagement, motivation, self-regulated learning, and conceptual understanding, particularly in foundational topics such as algebra and geometry, with less frequent exploration in advanced topics like calculus, statistics, and number theory. The review also shows that studies are relatively balanced between K–12 (40%) and higher education contexts (43%), with 27% focusing on teacher education, and that research is concentrated in North America, Europe, and parts of Asia, revealing geographical gaps in the literature.

GenAI is most effectively applied when combined with interactive, gamified, or differentiated learning activities (Gurl et al., 2025; Susilawati et al., 2025; Yunianto et al., 2024), where it can scaffold procedural fluency, generate multiple solution strategies, and encourage new learner roles, such as prompt engineers and co-agents (Seebut et al., 2024; Kock et al., 2025). However, its contributions are limited by domain-specific reasoning errors, weak abstract thinking, and reliance on precise prompting, which underscores the necessity of teacher-mediated scaffolding to ensure conceptual depth, ethical engagement, and critical evaluation of AI outputs (Segal & Klemer, 2025; Pavlova, 2024; Yoon et al., 2024). These findings indicate that while GenAI has the potential to transform certain aspects of mathematics learning, its impact is conditional, unevenly distributed across learners, and influenced by task design, user input, and platform reliability.

Importantly, GenAI is capable of scaffolding procedural fluency and generating multiple solutions to parallel problems. However, it lacks actual adaptive understanding, intentionality, and the ability to independently identify errors within its outputs, meaning that teacher mediation remains fundamental to ensure conceptual depth, ethical engagement, and critical evaluation of AI outputs (Pavlova, 2024; Segal & Klemer, 2025; Yoon et al., 2024). The outcomes of these studies emphasize the need for thoughtful integration, in which teachers are trained in prompt engineering, ethical use of AI, and iterative validation to maximize learning outcomes while reducing risks of over-reliance, surface-level reasoning, and inequitable access (Noster et al., 2024; Zhuang & Zhang, 2025).

Systematic mapping of the current literature provides insight into patterns of tool dominance, topic concentration, educational level distribution, and geographical imbalances, offering a synthesis that has been limited in prior reviews. Overall, the findings also reveal emerging professional development needs, including prompt engineering, ethical use, and iterative validation of AI outputs, emphasizing that GenAI's transformative potential is achievable only through purposeful pedagogical design, structured teacher mediation, and continuous evaluation, rather than being inherent to the technology.

Limitations and Future Studies

While providing valuable insights, the systematic mapping literature review has several limitations. To begin with, only studies in English were considered in the review, which may have excluded relevant research in other languages and limited the understanding of GenAI applications in diverse linguistic and cultural contexts. Second, the review focused exclusively on peer-reviewed empirical studies, excluding theoretical, conceptual, or non-peer-reviewed research, such as conference proceedings, which may provide complementary perspectives on the pedagogical roles and instructional design of GenAI in mathematics education; this exclusion is particularly relevant given the rapidly evolving nature of the field. Third, the search strategy was limited to selected databases (e.g., ScienceDirect), which may introduce database bias, and potential inconsistencies in search strings, duplicate counts, or the absence of inter-rater reliability checks could further constrain the comprehensiveness and reproducibility of the review. Fourth, no formal quality appraisal or risk-of-bias assessment was conducted, which limits the ability to evaluate the methodological robustness of the included studies. Finally, the rapid evolution of GenAI tools means that findings may quickly become outdated as new versions with enhanced capabilities are introduced. This highlights the need for ongoing evaluation of emerging platforms.

The gaps to be addressed in future research include the longitudinal effects of GenAI integration, in terms of whether it leads to sustained learning outcomes, conceptual mastery, and problem-solving abilities across various areas of mathematics. Cross-cultural studies are also essential to evaluate the adaptability of GenAI-supported instruction in diverse educational systems and cultural settings, particularly given the geographic imbalances highlighted in this mapping. Ethical and pedagogical concerns, including bias, over-reliance, and inequitable access, should also be addressed, as these remain underexplored (Qian, 2025; Tan et al., 2025). Moreover, future studies should explore the ethical and pedagogical aspects of AI implementation, especially the idea of bias, over-reliance, and equitable access. It is also important to balance the innovation of AI and conceptual depth so that AI can support learning and develop critical thinking, autonomy, and ethical reasoning. This balance may be quite difficult to achieve because, to date, GenAI is most effective in enhancing efficiency and motivation (Contel & Cusi, 2025; Delima et al., 2024; McGalliard & Otten, 2025), whereas its ability to improve true conceptual mastery remains uncertain. Lastly, comparative studies across different GenAI platforms such as ChatGPT, Gemini, Claude, and other emerging tools, are also recommended to identify platform-specific strengths, limitations, and best practices for instructional integration.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors are deeply grateful to the Kulliyyah of Education for their continuous support and guidance throughout this study. Our special thanks to the International Islamic University Malaysia for providing the necessary resources and facilities that greatly contributed to the success of this paper. Finally, we acknowledge our research partners for their dedicated support, collaboration, and expertise which were instrumental in bringing this study to completion.

Generative AI Declaration

This manuscript was developed with limited assistance from generative artificial intelligence (GenAI) tools, including ChatGPT. The tool was used solely to support language refinement, grammar checking, and minor improvements in sentence clarity during the manuscript preparation process. All conceptualization of the study, literature selection, data interpretation, synthesis of findings, and academic arguments were conducted independently by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the content presented in this paper. The use of GenAI did not influence the research design, data analysis, or conclusions of the study, and all sources cited were verified by the authors.

References

Almarashdi, H. S., Jarrah, A. M., Khurma, O. A., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *EURASIA Journal of Mathematics, Science and Technology Education, 20*(12), em2555. https://doi.org/10.29333/ejmste/15739

Borys, D., Macura, J., Sikora, B., & Wróbel, Ł. (2025). From manual to AI-driven: Methods for generating mathematics and programming exercises in interactive educational platforms. *Applied System Innovation, 8*(6), 174. https://doi.org/10.3390/asi8060174

Boulhrir, T., & Hamash, M. (2025). Unpacking artificial intelligence in elementary education: A comprehensive thematic analysis systematic review. *Computers and Education: Artificial Intelligence, 9,* 100442. https://doi.org/10.1016/j.caeai.2025.100442

Busuttil, L., & Calleja, J. (2025). Teachers' beliefs and practices about the potential of ChatGPT in teaching mathematics in secondary schools. *Digital Experiences in Mathematics Education, 11,* 140–166. https://doi.org/10.1007/s40751-024-00168-3

Contel, F., & Cusi, A. (2025). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education, 11*, 167–191. https://doi.org/10.1007/s40751-024-00164-7

Daher, W., & Anabousy, A. A. (2025). The didactical knowledge of generative artificial intelligence tools: The case of writing mathematics lessons. *EURASIA Journal of Mathematics, Science and Technology Education, 21*(9), em2691. https://doi.org/10.29333/ejmste/16769

Davis, J. D. (2025). Evolving techniques and emerging schemes: Prospective teachers' transformation of ChatGPT. *International Journal of Science and Mathematics Education, 23*, 2181–2198. https://doi.org/10.1007/s10763-024-10533-8

Delima, N., Kusuma, D. A., & Paulus, E. (2024). The students' mathematics self-regulated learning and mathematics anxiety based on the use of ChatGPT, music, study program, and academic achievement. *Infinity Journal, 13*(2), 349–362. https://doi.org/10.22460/infinity.v13i2.p349-362

Etcuban, J. O. (2025). The use of ChatGPT in addressing algebra anxiety and promoting confidence. *International Electronic Journal of Mathematics Education, 20*(2), em0825. https://doi.org/10.29333/iejme/16008

Gurl, T. J., Markinson, M. P., & Artzt, A. F. (2025). Using ChatGPT as a lesson planning assistant with preservice secondary mathematics teachers. *Digital Experiences in Mathematics Education, 11*, 114–139. https://doi.org/10.1007/s40751-024-00162-9

Horrocks, M., & Shearman, D. (2025). Rethinking what is valuable in mathematics and statistics education. *International Journal of Mathematical Education in Science and Technology, 56*(12), 2513–2533. https://doi.org/10.1080/0020739X.2025.2556864

Kim, Y. R., Park, M. S., & Joung, E. (2025). Exploring the integration of artificial intelligence in math education: Preservice teachers' experiences and reflections on problem-posing activities with ChatGPT. *School Science and Mathematics*, 1–15. https://doi.org/10.1111/ssm.18336

Kock, Z., Salinas-Hernández, U., & Pepin, B. (2025). Engineering students' initial use schemes of ChatGPT as an instrument for learning. *Digital Experiences in Mathematics Education, 11*, 192–218. https://doi.org/10.1007/s40751-025-00169-w

Kruijssen, J. M. D., & Emmons, N. (2025). Deterministic AI agent personality expression through standard psychological diagnostics. Allora Decentralized Intelligence, 2, 15–39. https://doi.org/10.70235/allora.0x20015

Mani, J., & Zargeh, C. (2025). AI artifacts in the mathematics didactical tetrahedron: A developed model. *EURASIA Journal of Mathematics, Science and Technology Education, 21*(5), em2630. https://doi.org/10.29333/ejmste/16307

Marco, N., & Stylianides, A. J. (2025). An exploration into the nature of ChatGPT's mathematical knowledge. *International Journal of Mathematical Education in Science and Technology, 56*(11), 2279–2297. https://doi.org/10.1080/0020739X.2025.2543817

McGalliard, W., & Otten, S. (2025). AI responses to challenging problems and educator responses to AI availability. *Digital Experiences in Mathematics Education, 11*, 319–332. https://doi.org/10.1007/s40751-024-00167-4

Milenković, A., Vučićević, N., & Svičević, M. (2025). Evaluating Open AI tools o1 and o3-mini in solving high school problems from Serbian national mathematics competition. *The Teaching of Mathematics, 28*(1), 14–29. https://doi.org/10.57016/TM-EYRQ1676

Mimoudi, A., & Mokhtari, K. (2025). 'AIA-PCEK': A new framework for teaching with AI. *Cogent Education, 12*(1), 2563171. https://doi.org/10.1080/2331186X.2025.2563171

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Noster, N., Gerber, S., & Siller, H. S. (2024). Pre-service teachers' approaches in solving mathematics tasks with ChatGPT. *Digital Experiences in Mathematics Education, 10*, 543–567. https://doi.org/10.1007/s40751-024-00155-8

Page, M. J., McKenzie, J. E., 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., … & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews, 10*, 89. https://doi.org/10.1186/s13643-021-01626-4

Panadero, E., Fernández-Ortube, A., Zamorano, D., Pinedo, L., Sánchez-Iglesias, I., & Barrenetxea-Mínguez, L. (2025). Tracking self-regulated learning in action: How individual differences shape positive and negative regulation across three types of tasks. Learning and Individual Differences, 124, 102808. https://doi.org/10.1016/j.lindif.2025.102808

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

Pavlova, N. H. (2024). Flipped dialogic learning method with ChatGPT: A case study. *International Electronic Journal of Mathematics Education, 19*(1), em0764. https://doi.org/10.29333/iejme/14025

Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends, 69*(6), 1105–1120. https://doi.org/10.1007/s11528-025-01100-1

Sawyer, A. G., & Aga, Z. G. (2025). A mile high and an inch deep: Exploring ChatGPT as a mathematics curriculum development tool. *School Science and Mathematics*, 1–15. https://doi.org/10.1111/ssm.18348

Schorcht, S., Peters, F., & Kriegel, J. (2025). Communicative AI agents in mathematical task design: A qualitative study of GPT network acting as a multi-professional team. *Digital Experiences in Mathematics Education, 11*, 77–113. https://doi.org/10.1007/s40751-024-00161-w

Seebut, S., Wongsason, P., & Kim, D. (2024). Combining GPT and Colab as learning tools for students to explore the numerical solutions of difference equations. *EURASIA Journal of Mathematics, Science and Technology Education, 20*(1), em2377. https://doi.org/10.29333/ejmste/13905

Segal, R., & Klemer, A. (2025). Dialogic interactions between mathematics teachers and GenAI: Multi-environment task design and its contribution to TPACK. *International Journal of Mathematical Education in Science and Technology*, 1-25. https://doi.org/10.1080/0020739X.2025.2551363

Shekh-Abed, A. (2025). Metacognitive self-knowledge and cognitive skills in project-based learning of high school electronics students. European Journal of Engineering Education, *50*(1), 214–229. https://doi.org/10.1080/03043797.2024.2374479

Stamenkova, R. (2025). Large language models: A tool for solving mathematical problems in high school. *Annual of Sofia University "St. Kliment Ohridski" Faculty of Mathematics and Informatics, 112*, 165–183. https://doi.org/10.60063/GSU.FMI.112.165-183

Susilawati, W., Sharov, S., Pasqa, M., & Malik, H. (2025). Integrating realistic mathematics education, AI, and gamification to enhance students' learning motivation and problem-solving skills. *Journal on Mathematics Education, 16*(4), 1257–1282. https://doi.org/10.22342/jme.v16i4.pp1257-1282

Sutrisno, J. A. B., Pratama, E. Y., Putra, R. A., Nasution, S. H., & Lestyanto, L. M. (2025). Students' mathematical justification abilities in analyzing ChatGPT's answers. *Infinity Journal, 14*(2), 445–460. https://doi.org/10.22460/infinity.v14i2.p445-460

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), The psychology of learning and motivation: Cognition in education. *Psychology of Learning and Motivation*, *55*, 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Tan, X., Cheng, G., & Ling, M. H. (2025). Enhancing teachers' AI competency: A professional development intervention study based on intelligent-TPACK framework. *Computers and Education: Artificial Intelligence, 9*, 100521. https://doi.org/10.1016/j.caeai.2025.100521

Wahba, F., Ajlouni, A. O., & Abumosa, M. A. (2024). The impact of ChatGPT-based learning statistics on undergraduates' statistical reasoning and attitudes toward statistics. *EURASIA Journal of Mathematics, Science and Technology Education, 20*(7), em2468. https://doi.org/10.29333/ejmste/14726

Walkington, C., Pando, M., Lipsmeyer, L. L., Beauchamp, T., Sager, M., & Milton, S. (2025). Middle school girls using generative AI to engage in mathematical problem-posing. *Mathematical Thinking and Learning*, 1-22. https://doi.org/10.1080/10986065.2025.2542724

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Xuan, S. H., Nguyen, A. T., Nguyen, T., Nguyen, L., Nguyen, H., Pham, N., Phung, T., Ngo, B., Nguyen, V., Nguyen, M., Tran, T., Le, T., Nguyen, K., & FNU, P. (2025). Evaluating the impact of generative AI in mathematics education: A comparative study in Vietnamese high schools. *Human Behavior and Emerging Technologies*, *2025*, 8886206, 1-23. https://doi.org/10.1155/hbe2/8886206

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education, 56*, 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

Yunianto, W., Lavicza, Z., Kastner-Hauler, O., & Houghton, T. (2024). Investigating the use of ChatGPT to solve a GeoGebra-based mathematics and computational thinking task in a geometry topic. *Journal on Mathematics Education, 15*(3), 1027–1052. https://doi.org/10.22342/jme.v15i3.1027-1052

Zhuang, Y., & Zhang, S. (2025). Pre-service mathematics teachers' perceptions of using GenAI for practicing teacher questioning: A semester-long study. *EURASIA Journal of Mathematics, Science and Technology Education, 21*(9), em2689. https://doi.org/10.29333/ejmste/16764