Tsang et al. *Research and Practice in Technology Enhanced Learning* (2027) 22:27
https://doi.org/10.58459/rptel.2027.22027

Research and Practice in Technology Enhanced Learning

RESEARCH Free and Open Access

Understanding the role of large language models in supporting mathematics education

Henry Pak Yeung Tsang, Dichen Wang, Philip Leung Ho Yu*

*Correspondence:
plhyu@eduhk.hk
Department of Mathematics and Information Technology, The Education University of Hong Kong, 10 Lo Ping Road, New Territories, Hong Kong
Full list of author information is available at the end of the article

Abstract

Large Language Models (LLMs) are redefining mathematics education, including not only student-facing tutors but also educator-centered tools that automate instructional tasks and support professional development. This systematic review synthesizes findings from 131 empirical and technical studies published between January 2022 and February 2025. Guided by PRISMA protocols and BERTopic modeling, we investigated three key questions: (1) How do LLMs perform on mathematical tasks across varying levels of difficulty? (2) In what ways are LLMs being used to support mathematics educators? (3) What are the limitations and concerns of applying LLMs in mathematics education?

BERTopic identified four dominant themes: (a) evaluation of LLM mathematical reasoning, (b) LLM-supported mathematics education, (c) automation of mathematics educational tasks, and (d) LLM capability in mathematical word problems. Across benchmarks (e.g., GSM8K, SVAMP, MATH), LLM performance is strong on many primary and secondary tasks, especially when combined with remediation approaches such as prompt engineering, fine-tuning, and tool/code-based verification, but reliability decreases as problems require abstract reasoning, multi-step planning, or domain-specific formalisms (notably calculus, statistics/probability, geometry/spatial reasoning, and some non-English settings). Synthesizing findings through an Intelligent Tutoring Systems (ITS) lens, we argue that LLMs are most defensible as *generative components* within systems that provide instructional control, verification, and student-model inference, rather than as autonomous authorities for mathematical correctness or pedagogy. Key risks include hallucination, limited pedagogical sensitivity, uneven multilingual robustness, and over-reliance that may weaken learners' independent reasoning. We conclude with implications for evaluation and adoption, emphasizing verifiability, process-sensitive assessment, and educator-centered governance.

Keywords: large language models (LLMs), mathematics education, intelligent tutoring systems (ITS), BERTopic

Introduction

Artificial Intelligence (AI) is reshaping education by transforming how students learn, educators teach, and institutions manage academic processes. From intelligent tutoring systems (ITS) to adaptive learning platforms, chatbots, and automated assessment tools, AI applications collectively referred to as Artificial Intelligence in Education (AIED) have shown the potential to enhance student engagement, optimize teaching practices, and improve administrative efficiency. Among the most transformative innovations in AIED are Large Language Models (LLMs), such as ChatGPT, which have garnered widespread attention for their ability to generate human-like responses, solve problems, and provide real-time feedback. Unlike earlier AI systems, which were limited to domain-specific tasks, LLMs possess the flexibility to support a wide range of educational activities, including generating lesson plans, facilitating collaborative discussions, and automating grading (Lo, 2023; Fu et al., 2025). These models have been rapidly adopted in both PreK-12 and higher education settings, driven by their scalability, accessibility, and versatility. For instance, ChatGPT has been used as a virtual tutor to assist students in understanding complex concepts and as a teacher assistant to streamline instructional design and administrative tasks.

At the same time, evaluating LLMs in mathematics education requires clarity about what "effective teaching" entails. In the Intelligent Tutoring Systems (ITS) literature, a widely cited framing characterizes ITS as an integration of interacting components: an expert (domain) module representing target knowledge and solution strategies; a student model module representing what the learner currently knows, often implemented via an overlay model that treats the learner's knowledge state as a subset of expert knowledge; a tutoring (pedagogical) module that uses information about both the domain and the learner to select instructional actions (e.g., hints, feedback, practice tasks); and a user interface module that mediates communication and enables the system to collect evidence from learner interactions (Nwana, 1990). Recent work suggests that LLMs can be incorporated within this architecture to support adaptive, dialogue-based instruction rather than functioning solely as answer generators. For example, Latif et al. (2026) highlight the growing integration of AI techniques, including LLMs, in tutoring systems to enable more responsive interaction, real-time feedback, and personalization, with many studies reporting improved learning outcomes; similarly, Junpeng (2026) indicates that individualized scaffolding and feedback can promote learning. Accordingly, we use the ITS framework in later sections to evaluate LLM applications in mathematics education and to surface the design conditions and challenges that shape their instructional value.

LLMs may also offer practical benefits for educators and students. For educators, they can automate time-intensive tasks such as drafting assessments or generating feedback, potentially freeing time for higher-value instructional interactions (Wang et al., 2024). For

students, LLMs can provide on-demand explanations and opportunities for guided practice. In some domains, including economics and programming, LLMs have demonstrated strong performance on benchmark assessments and can generate stepwise explanations that support higher-order thinking (Lo, 2023).

Despite their promise, LLMs face significant challenges that raise questions about their broader suitability in education. One critical concern is their susceptibility to hallucination, where models confidently generate incorrect or irrelevant information. Furthermore, LLMs may provide inconsistent adaptive feedback or personalized support, which are key elements of effective teaching. These limitations are compounded by reduced adaptability in non-English contexts and concerns about overreliance on AI tools. For example, Bastani et al. (2025) demonstrated that while students using GPT-4-based tutors performed better on assisted tasks, they struggled significantly on unassisted exams, suggesting that reliance on direct solutions from LLMs discourages deeper engagement with mathematical concepts. Similarly, Lee et al. (2025) found that professional knowledge workers relying on AI tools experienced diminished critical thinking, as cognitive effort shifted from task execution to oversight.

These complexities underscore the need for a nuanced evaluation of LLMs' role in mathematics education, where both the opportunities and limitations of these tools are particularly pronounced. Mathematics, as a discipline, demands logical precision, multi-step reasoning, and problem-solving skills, areas where LLMs face significant challenges. While LLMs may be useful for generating practice problems, explanations, or draft feedback, their accuracy declines in higher-order mathematical domains like calculus, symbolic reasoning, and spatial problem-solving. Consequently, the use of LLMs in mathematics raises critical questions about how AI can support, rather than supplant, human instruction, and how these tools can align with long-term educational goals.

This systematic review aims to address these questions by analyzing 131 articles published between January 2022 and February 2025, focusing on the performance, applications, and limitations of LLMs in mathematics education. Specifically, it seeks to answer the following research questions:

1. How do LLMs perform on mathematical tasks across different levels of difficulty?
2. In what ways are LLMs being applied to support mathematics educators?
3. What limitations and concerns arise when implementing LLMs in mathematics education?

Method

This systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021) to ensure transparency, reproducibility, and rigor in the selection and analysis of studies. The review aimed to evaluate the performance, applications, and impact of Large Language Models (LLMs) in mathematics education by addressing three key research questions.

Data sources and search strategies

To capture a comprehensive range of studies, we searched four major electronic databases: ERIC, Academic Search Ultimate, Education Source (all from the EBSCOhost database), and Scopus. The search was conducted from November 2022 to February 2025, with January 2022 selected as the starting point to coincide with the public launch of ChatGPT-3.5, which marked a significant milestone in the adoption of LLMs in education.

The search terms used were designed to capture studies related to LLMs and their applications in mathematics education. Boolean operators were employed to combine the following terms:

- ("Large Language Model?" OR "ChatGPT" OR "Generative AI")
- AND ("class*" OR "math education" OR "mathematical education")

The question mark (?) was used to account for both singular and plural forms (e.g., "model" and "models"), while the asterisk (*) allowed for variations of the term "math" (e.g., "mathematics").

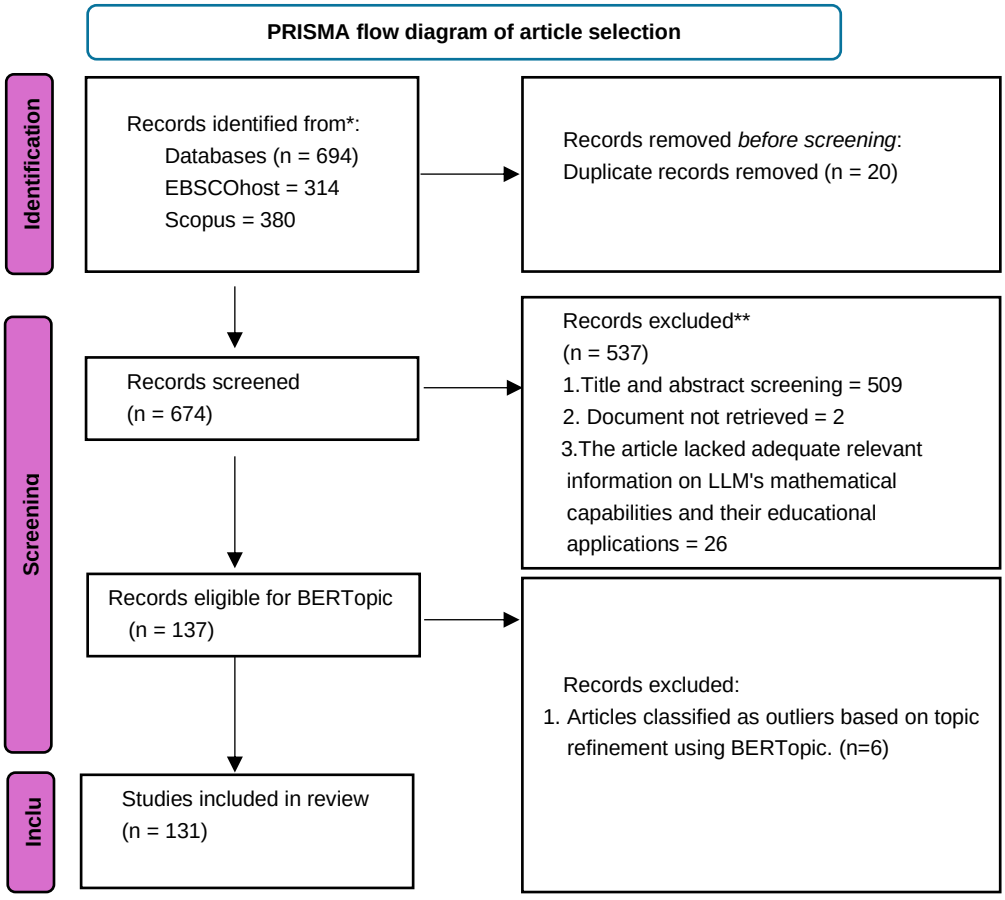
Table 1

Inclusion and exclusion criteria

Factor	Inclusion Criteria	Exclusion Criteria
Date	Studies published from January 2022 to February 2025	Studies published prior 2022
Language	Studies published in the English language	Studies not published in the English language
Literature type	Published journal articles or conference papers	Editorials, letters, protocol papers and systematic reviews.
Evaluation of LLM math performance	• Studies testing LLMs' mathematical abilities on existing or newly proposed datasets. • Studies evaluating LLMs' mathematical performance using established or novel prompting techniques.	• Studies discussing LLMs' strengths or weaknesses in mathematics without empirical data.
LLM application in math education and effectiveness	• Studies detailing how LLMs assist teachers in mathematics education. • Studies assessing the effectiveness of LLMs applications using valid quantitative or qualitative methods.	• Studies discussing LLMs applications and effectiveness based on personal opinions without empirical evidence

Figure 1

PRISMA flow diagram



Thematic distribution and relationships in LLM research for mathematics

To analyze the thematic structure of the included studies, we utilized BERTopic, an advanced topic modeling framework that leverages transformer-based language models for semantic representation. In this approach, dense vector embeddings for each document are generated using the "all-MiniLM-L6-v2" model from the SentenceTransformers library, with each document represented by a concatenation of its title, keywords, and abstract. To address the high dimensionality of these embeddings and facilitate efficient clustering, Uniform Manifold Approximation and Projection (UMAP) was used to project the data into a lower-dimensional space, with adjustable parameters to balance topic granularity and interpretability.

Clustering was performed using the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm, which identifies groups of semantically similar documents and is robust to clusters of varying sizes. After cluster formation, BERTopic extracts representative keywords for each topic using a class-based variant of

TF-IDF (c-TF-IDF), highlighting terms that are particularly distinctive for each topic within the corpus.

We systematically explored a range of parameter settings and selected a configuration that produced interpretable and distinctive topics. Across 100 model configurations, the final solution identified four topics, with an initial outlier rate of 12%. Details of the configuration can be found in Appendix B. Each topic was characterized by its most representative keywords and a set of closely related articles. For example, the article "Complexity-Based Prompting for Multi-step Reasoning" was found to be most thematically similar to:

1. TORA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving (distance: 0.28)
2. Transforming Generative Large Language Models' Limitations into Strengths using Gestalt: A Synergetic Approach to Mathematical Problem-Solving with Computational Engines (distance: 0.35)
3. OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data (distance: 0.37)
4. CONIC10K: A Challenging Math Problem Understanding and Reasoning Dataset (distance: 0.38)
5. MAMMOTH: Building Math Generalist Models Through Hybrid Instruction Tuning (distance: 0.40)
6. MuMath: Multi-perspective Data Augmentation for Mathematical Reasoning in Large Language Models (distance: 0.40)

In this context, the "distance" value reflects the semantic similarity between articles in the reduced 10-dimensional UMAP space, with smaller values indicating greater similarity. Initially, 11 articles were classified as outliers (assigned to topic -1). Since all 137 articles had been pre-screened for relevance, we applied a refinement process to reassign these outliers to appropriate topics, while maintaining the existing assignments for other articles. This reduced the number of outliers from 11 to 6, resulting in a more coherent and accurate thematic structure. The four identified topics were renumbered from 0–3 to 1–4 and labeled based on representative keywords, article content, and relationships visualized through principal component analysis (PCA) of document embeddings.

This topic modeling approach enabled the identification of nuanced clusters and trends in the literature on large language models in mathematics education, supporting both quantitative summaries and qualitative insights for research and practice.

Topic 1: LLM support in math education

This topic centers on the role of large language models in enhancing mathematics education through the development of supportive tools and systems for both students and teachers. Representative keywords—such as students, learning, teachers, chatbots, teaching, tools, results, and potential—underscore the integration of LLMs within educational environments. Articles grouped under this topic frequently discuss practical applications, including lesson planning, intelligent tutoring systems, and chatbot-based interventions designed to improve student engagement and instructional effectiveness. Principal component analysis (PCA) visualization indicates that Topic 1 is conceptually related to Topic 3 as both focus on the educational applications of LLMs.

Topic 2: Evaluation of LLM math reasoning

This topic centers on the assessment of large language models' reasoning capabilities in mathematical contexts. Representative keywords—including *reasoning*, *prompting*, *chain-of-thought (cot)*, *problems*, *tasks*, *dataset*, *solving*, and *datasets*—indicate a focus on logical reasoning and systematic evaluation. Articles classified under this topic investigate the performance of LLMs across various benchmarks and datasets, with particular emphasis on their ability to solve mathematical problems and demonstrate structured reasoning processes. Principal component analysis (PCA) visualization suggests a conceptual relationship between this topic and Topic 4 as both address mathematical reasoning; however, Topic 2 encompasses broader evaluations, while Topic 4 specifically targets math word problem solving.

Topic 3: Automation of mathematics educational tasks

This topic addresses the automation of educational processes in mathematics using large language models. Representative keywords—such as *tutoring*, *learning*, *students*, *questions*, *problems*, *assessment*, *word*, *feedback*, and *classification*—highlight tasks including automated tutoring, assessment, and feedback generation. Articles grouped under this topic frequently explore how LLMs can be employed to streamline various educational functions, such as generating metadata, automating scoring systems, and providing timely feedback to learners. While this topic shares thematic overlap with Topic 1 in its focus on educational applications, Topic 3 places greater emphasis on automation, distinguishing it from the supportive tools and systems highlighted in Topic 1.

Topic 4: LLM capability in solving mathematical word problems

This topic focuses on evaluating the ability of large language models to solve mathematical word problems (MWPs). Representative keywords—including *word*, *problems*, *dataset*, *problem*, *performance*, *numbers*, *MWP*, *difficulty*, and *MWPs*—highlight the specific emphasis on MWPs. Articles within this topic examine LLM performance across various datasets, exploring how effectively these models reason through and solve word-based mathematics questions. While there is some conceptual overlap with Topic 2 due to their shared focus on reasoning, Topic 4 is distinguished by its narrower scope, concentrating specifically on MWPs rather than broader mathematical reasoning tasks.

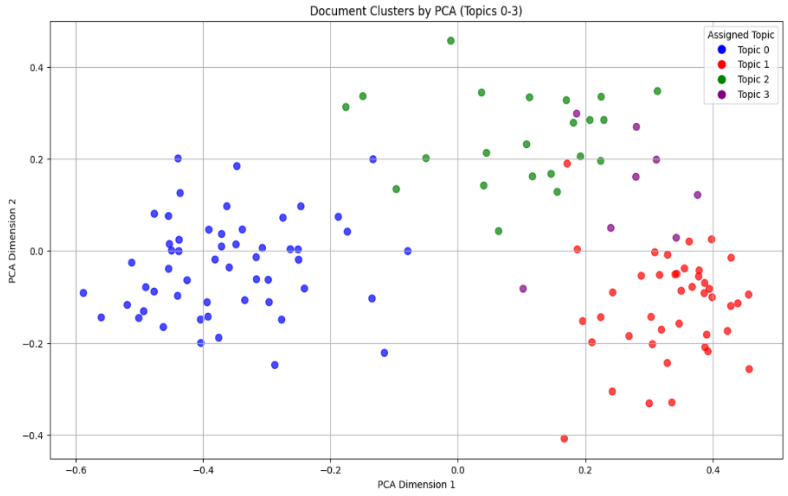
Table 2

BERTopic results. For readability, the original BERTopic cluster labels (0–3) are presented in this article as Topics 1–4.

Topic	Count	Name	Representation
-1	6	(Outliers)	['geometry', 'geometric', 'discovery', 'geogebra', 'natural', 'performance', 'gold', 'small', 'gap', 'automatic']
0	54	LLM Support in Math Education	['students', 'learning', 'teachers', 'chatbots', 'teaching', 'teacher', 'thinking', 'preservice', 'tools', 'results']
1	43	Evaluation of LLM Math Reasoning	['reasoning', 'prompting', 'cot', 'problems', 'complex', 'tasks', 'performance', 'dataset', 'propose', 'datasets']
2	25	Automated Math Tasks	['tutoring', 'learning', 'students', 'problems', 'questions', 'assessment', 'problem', 'quality', 'classification', 'content']
3	9	LLM capability in solving mathematical word problems	['word', 'problems', 'dataset', 'problem', 'performance', 'numbers', 'place', 'mwp', 'difficulty', 'mwps']

Figure 2

PCA projection plot. For readability, the original BERTopic cluster labels (0–3) are presented in this article as Topics 1–4.



The results of the refined BERTopic analysis are presented in Appendix A. To address our research questions, we developed a coding scheme to categorize the articles as follows:

Table 3

Coding scheme

Code	Article in Topic	Description
Automated Tasks	3	Studies focusing on task automation in mathematics education.
Teaching development and pre-service teaching training	1	Research exploring the use of LLMs in teacher development and training programs.
Challenges and impacts of LLM applications in mathematics	1	Analysis of the challenges and broader educational implications of LLMs in mathematics.
Evaluation of LLM math reasoning performance	2,4	Studies assessing the reasoning capabilities of LLMs in math tasks.

Results and Discussion

1. Trends and thematic focus in LLM research for mathematics

To analyze the trends in LLM research for mathematics, we examined studies published between 2022 and February 2025. This section provides an overview of the foundational trends, including the volume of research over time, the balance between evaluation- and application-oriented studies, and the distribution of research across thematic areas. These insights establish the broader context of LLM research in mathematics education,

highlighting shifts in research priorities and emerging patterns. This overview serves as a foundation for the detailed thematic analysis presented in subsequent sections.

Figure 3 illustrates the progression from evaluation-focused studies to application-driven research. By 2023, 22 evaluation studies and 20 application studies reflected a balanced interest in assessing LLM capabilities and exploring their practical use. In 2024, application-driven research surged to 51 studies, significantly outpacing the 28 evaluation studies, signaling a stronger emphasis on real-world implementations. Early 2025 continues this trajectory, with seven application studies already recorded.

Figure 4 highlights the thematic distribution of research. By 2023, growth was evident across all topics, particularly in "Evaluation of LLM Math Reasoning" (17 studies) and "LLM Support in Math Education" (13 studies). Narrower areas, such as "Automated Math Tasks" (7 studies) and "LLM Capability in Solving Math Word Problems" (5 studies), received comparatively less attention. In 2024, research activity peaked, with substantial growth in "LLM Support in Math Education" (34 studies) and "Evaluation of LLM Math Reasoning" (24 studies), while automation gained momentum with 17 studies. Although fewer studies target mathematical word problems (MWPs) than general reasoning benchmarks, this focus is theoretically significant for mathematics education and ITS-aligned evaluation. Here, mathematical thinking refers to forming, connecting, and interpreting mathematical representations to explain relationships and justify conclusions (Li & Schoenfeld, 2019). Word problems have long been a central emphasis in mathematics education, particularly at the elementary level, because translating a contextual situation into an appropriate mathematical expression requires more than reading comprehension: it also develops modelling habits, metacognitive monitoring, and critical judgment about what a problem is asking (Verschaffel, Greer & De Corte, 2000). MWPs are well suited to study such thinking because they require contextual interpretation and modelling, not only procedural execution. In MWPs, learners must interpret a contextualized situation, identify relevant quantities and relations, formulate an appropriate mathematical representation, and interpret results, rather than only execute procedures (Boonen et al., 2016; Fitzpatrick et al., 2020; Verschaffel et al., 2020). A classic illustration is the “ship captain” item, “If a ship had 26 sheep and 10 goats on board, how old is the ship’s captain?” which shows that students may mechanically extract numbers and perform calculations without evaluating whether the question is meaningful or solvable (Selter et al., 2000). From an ITS perspective, MWPs are especially informative because they engage multiple knowledge components represented in the expert model (e.g., problem representation and strategy selection) and can reveal distinct error sources that an overlay student model aims to diagnose (e.g., comprehension or mathematization errors versus procedural execution errors). Consequently, examining LLMs’ performance on MWPs helps clarify whether LLM-based systems can support fine-grained inference about learner

understanding and enable adaptive pedagogical actions (e.g., targeted prompts and scaffolds), instead of functioning primarily as answer generators.

These trends reflect a clear shift from foundational evaluation to practical applications and automation, driven by improvements in LLMs' performance. As their capabilities advance, researchers increasingly focus on integrating LLMs into mathematics education.

Figure 3

Research trends in evaluation vs. application 2022–2025 (Feb)

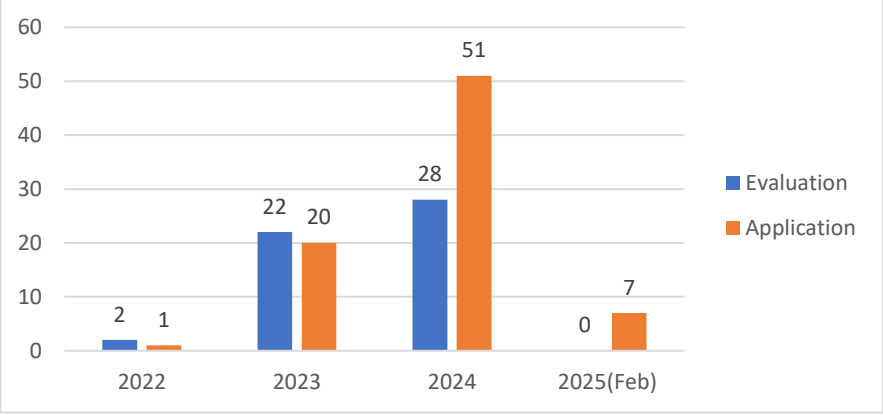
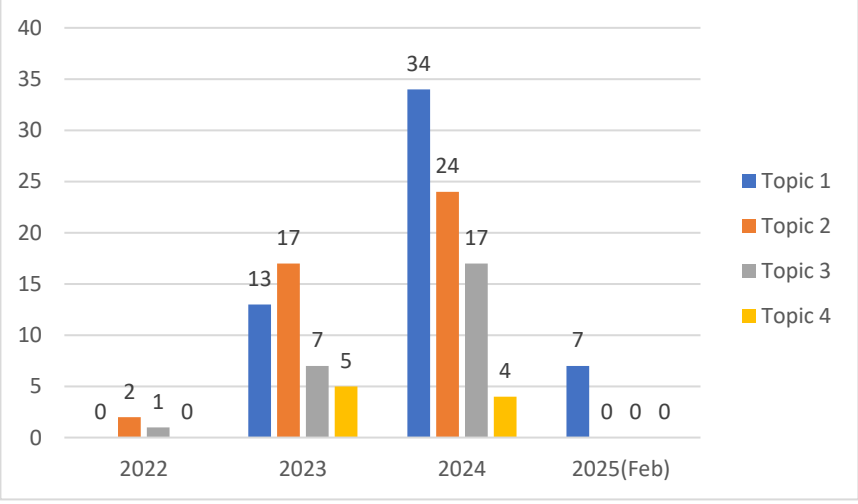


Figure 4

Topic distribution



2. LLM applications for supporting math education through automated tasks

Building on these trends, the following sections delve into the thematic areas identified in the analysis, starting with the use of Large Language Models for automating educational tasks. This theme explores how LLMs are employed to streamline processes such as problem generation, feedback automation, and reasoning evaluation.

To avoid conflating capabilities with instructional authority, we distinguish three roles when interpreting LLM-based applications in mathematics education. In this review, LLMs are treated primarily as generative components that can produce candidate explanations, hints, items, or classifications from natural-language inputs. By contrast, instructional control including maintaining an expert model of valid solution constraints, estimating learner mastery through a student/overlay model, selecting next tasks, and enforcing pedagogical policies belongs to the Intelligent Tutoring System (ITS) architecture when such a system is present (Nwana, 1990). This distinction is particularly consequential for word problems, where expert teaching often targets higher-order habits (e.g., modelling choices, justification, and plausibility checking) through thoughtful task design, strategic questioning, and guided scaffolding. An ITS can support these goals by sequencing tasks and embedding prompts for reflection, justification, and plausibility checking, enabling scaffolded higher-order thinking rather than mere answer generation. Teachers remain responsible for curricular alignment, pedagogical appropriateness, and accountability in high-stakes decisions (e.g., grading, placement, and consequential feedback), and they provide contextual and socio-emotional supports that are not reliably inferable from text alone. Table 4 summarizes these responsibilities and non-responsibilities for LLMs, ITS components, and teachers.

Table 4

Differentiating roles in LLM-supported mathematics education

	Role	What it should do well	What it should not be solely responsible for
LLM	Generation and reasoning assistant	• Generate candidate instructional moves (hints, prompts, explanations) aligned to an assigned strategy. • Produce multiple problem variants with controlled difficulty or features. • Summarize student work into structured evidence. • Suggest targeted follow-up questions.	• Certifying mathematical validity/correctness of final answers or solution steps. • Unsupervised scoring of constructed responses in high stakes contexts. • Deciding progression or issuing consequential feedback without external verification and policy constraints.
ITS	Instructional control and inference	• Enforce instructional policies (hint timing, solution withholding, allowed tools) • Represent the expert model (constraints, solution space, step validity checks) and call verifiers • maintain and update student/overlay estimates from logged evidence (attempts, step errors, time, help use) • select next tasks/hints using mastery and uncertainty; • Specify learning objectives, success criteria, and acceptable solution methods	• Setting curriculum priorities and instructional goals • Making equity or ethics decisions. • Interpreting learner needs that depend on situational context not in the logs (motivation, affect, accessibility) • Overriding teacher professional judgment about appropriateness of interventions.
Teacher	Pedagogical authority and accountability	• Validate mathematical correctness and pedagogical appropriateness of AI-supported material. • Interpret student thinking using domain knowledge plus classroom context. • Set norms for AI use and academic integrity • Make final decisions on grading, placement, and interventions. • Provide equity-aware supports (accommodations, culturally responsive examples) and socio-emotional scaffolding.	• Offloading consequential instructional decisions to automated outputs without review. • Relying on AI-generated analytics as a substitute for formative assessment and professional interpretation. • Permitting unrestricted AI use that undermines intended learning outcomes.

The integration of LLMs into mathematics education has significantly reduced the workload for educators while enhancing learning experiences for students. **Table 5** provides an overview of key automated tasks enabled by LLMs, along with their applications and supporting studies. These tasks cover a wide range of functions, including generating personalized hints and explanations, automating question creation, scoring complex responses, and designing instructional materials.

Table 5

Applications of large language models in mathematics education and their roles across various tasks

Task	LLM applications	Article
AI-Generated Hints and Explanations	• LLMs are used to generate explanations, walkthroughs, and guidance for mathematics problems, supplementing teacher-created content and thus reducing the labor-intensive efforts of teachers. • AI-generated explanations and tips can act as valuable references for teachers, improving efficiency and enabling more personalized support for students. • Explanations generated by LLMs improve comprehension by clarifying complex reasoning and connecting concepts to prior knowledge. • AI-generated hints provide tailored guidance by breaking problems into smaller, manageable steps, promoting self-directed learning.	[26],[38],[61], [124]
Automated Question and Feedback System	• LLMs are used to automatically generate math questions, reducing the effort required for manual question creation. • LLMs are applied for context-aware and context-unaware math question generation, covering elementary to tertiary education levels. • LLMs are used to generate feedback for formative assessment in mathematical problem-solving, helping both educators and students refine teaching and learning strategies. • LLMs assist in crafting multiple-choice question distractors by generating mathematically valid options, though they struggle to simulate common student misconceptions. • Using LLMs for Socratic questioning supports math problem-solving by generating sequential, guided questions to enhance reasoning and improve solver performance	[18], [32], [53], [59]
AI-Driven Scoring and Diagnosis Tool	• LLMs are used to classify student errors in open-ended math questions, improving the scalability and flexibility of automated feedback systems. • LLMs are explored for detecting incoherent answers in open-ended math responses, helping automate teacher review and reduce their workload. • LLMs are used to automatically score extended math responses, achieving near-human	[6], [17], [135], [137]

	accuracy in assessments like the NAEP Math Scoring Challenge.	
AI-Driven Readability Enhancement	• LLMs, such as GPT-4, are used to revise math word problems for improved readability and accessibility, particularly for younger students. • By automating the revision process, LLMs help educators enhance the clarity of educational content while reducing manual effort.	[70], [111]
Lesson Plan Generation	• LLMs are used to automatically generate customized lesson plans tailored to diverse teaching contexts and student needs. • LLMs assist in instructional design by organizing problem chains, setting teaching objectives, and selecting effective teaching strategies. • LLMs streamline the lesson planning process, significantly reducing the time and effort required from educators.	[19],[121]
Automated Math Annotation Tool	• LLMs can automate the generation of metadata for assessment items, including problem context, math vocabulary, and difficulty level. • LLMs are used for automated annotation and Part-of-Math (POM) tagging of equations, reducing reliance on manual rule-based systems. • LLMs enhance the accessibility and utility of mathematical knowledge in digital libraries by automating annotation and tagging tasks.	[1], [134]
Intelligent Tutoring System	• LLMs are used to help generate and manage the steps and paths (state space) in Intelligent Tutoring Systems, ensuring the tutoring process follows a clear structure. • LLMs add flexibility to tutoring systems by providing dynamic responses while still following predefined teaching strategies to improve learning outcomes.	[20]

3. Applications of LLMs in developing pre-service teaching readiness and teacher professional growth

Large Language Models are being explored for their potential to automate routine educational tasks. Beyond these automated functions, LLMs are increasingly recognized for their ability to support both pre-service and in-service teacher development by fostering critical thinking, refining instructional practices, and enhancing lesson planning. As summarized in Table 6, LLMs contribute to developing pre-service teachers' readiness through scaffolding, multi-strategy problem-solving, and critical thinking, while also fostering professional growth for in-service teachers by improving questioning techniques, lesson planning, and reflective practice.

Table 6

Applications of large language models in supporting teacher development and professional growth

Task	Description	Article
Developing Teacher Readiness	• LLMs support pre-service teachers in developing critical thinking by encouraging them to analyze AI-generated responses and identify logical errors. • ChatGPT supports pre-service teachers in building self-efficacy, higher-order thinking skills, and instructional preparedness by providing scaffolding and feedback during training. • Using ChatGPT for multi-strategy problem-solving helps pre-service teachers apply diverse approaches to solving math problems and adapt teaching strategies to student needs. • LLMs aid pre-service teachers in lesson planning by supporting writing and plan preparation, offering ideas, outlining plots, creating dialogues, and assisting with content creation, while fostering critical thinking through plan evaluation and refinement. • ChatGPT plays a role in helping in-service teachers improve rational questioning and classroom observation skills, enhancing their instructional practices.	[37], [75], [102], [109], [118], [132]
Professional Growth for Teachers	• As a virtual coach, ChatGPT provides actionable feedback on teaching strategies, helping teachers identify areas for improvement and growth. • LLMs help in-service teachers in lesson planning by providing information, developing questions, and enhancing explanations, while enabling content creation and revision. It also enhances critical thinking by analyzing and improving the generated plans.	[57], [75], [76]

4. Bridging potential and challenges

Large Language Models have demonstrated considerable promise in mathematics education, particularly for automating routine tasks, generating worked examples, and supporting teacher planning and professional learning. At the same time, wider adoption makes it essential to evaluate LLMs not only in terms of fluency and speed, but also against the epistemic and pedagogical demands of mathematics, especially in tasks that require

evaluation, synthesis, and application. The limitations summarized in Table 7 therefore reflect two complementary patterns: cross-level vulnerabilities (e.g., arithmetic/algebra execution errors and multi-step reasoning breakdowns that appear from primary to tertiary settings) and domain-concentrated weaknesses that become more salient in upper secondary and college mathematics (notably statistics/probability and calculus). Framed this way, the key issue is not simply that LLMs sometimes "get answers wrong," but that their failures can be systematically tied to the kinds of higher-order reasoning emphasized at later stages of schooling, with direct implications for assessment validity and instructional use. To maintain coherence across the review, we first synthesize these limitation themes from prior studies. Then in a later section, we will use mathematics benchmarks and standardized evaluations that explicitly compare LLMs' performance across grade school, secondary, and college-level item sets.

Table 7

Common concerns and limitations of large language models in mathematical problem-solving

Concern Reported		Description	Articles
Lack of Pedagogical Capabilities		LLMs sometimes fail to adhere to structured teaching strategies or established pedagogical principles, resulting in less effective guidance for learners.	[4], [8], [13], [16], [33], [40], [46], [83], [125]
Arithmetic and Equation-Solving Errors		LLMs occasionally produce incorrect or inconsistent results in basic arithmetic operations or when solving algebraic equations, undermining their reliability in math-focused tasks.	[2], [8], [30], [35], [63], [78], [131]
Logic & Mathematical Reasoning Errors (multi-step / abstract)		Failures in logical inference, strategy selection, proof logic, or validating conclusions against conditions/context; includes contradictions and "answer-correct-but-method-wrong" acceptance.	[2], [7], [8], [16], [36], [90], [116], [130], [131]
Spatial Reasoning Errors		Challenges arise when LLMs attempt to process or solve problems involving spatial relationships, geometry, or visual reasoning.	[2], [8], [35]
Specific Subject weakness	Statistics and Probability	LLMs exhibit variability in performance across different knowledge domains, often struggling with subjects that require nuanced expertise or domain-specific reasoning.	[51], [55]
	Calculus		[7], [35], [36], [55], [130]

Cross-Level vulnerabilities in LLMs' mathematical performance: Arithmetic & algebra execution and multi-Step reasoning failures

Arithmetic and equation-solving errors, alongside logic and mathematical reasoning failures, should be treated as cross-level limitations of large language models rather than as isolated deficiencies associated with any single mathematical topic or educational stage. Across the reviewed evidence base, LLMs' outputs demonstrate that computational correctness is not guaranteed even for ostensibly routine procedures. At earlier levels, this appears in miscomputed arithmetic totals and misinterpreted percentage language,

undermining the reliability of "basic" numeracy and straightforward quantitative comparisons [30]. As mathematical tasks become more formal and multi-step, the same fragility persists in algebraic manipulation and equation solving, including dropped terms during simplification and incorrect handling of valid solution branches, errors that are procedurally consequential even when the problem structure is standard [35], [131]. Importantly, these computational slips do not disappear at the tertiary level; rather, they re-emerge as precision and implementation problems in applied and higher mathematics contexts (e.g., inaccurate high-precision numerical values after a correct analytic setup, or numerically incorrect matrices despite a conventional solution narrative), indicating that "knowing the method" does not reliably translate into correct execution [63], [78]. Taken together, the literature suggests that arithmetic and algebraic reliability remains a persistent vulnerability across educational levels, with the form of the error shifting from basic operations to precision-sensitive and representation-dependent computations.

A parallel and arguably more consequential set of limitations concerns logic and mathematical reasoning, particularly where tasks require method selection, constraint management, and validation of conclusions against context. The reviewed studies document failures that extend beyond calculation into the control of inference itself: using a single familiar rule while neglecting additional conditions, selecting an inappropriate modelling structure, or drawing conclusions that are not licensed by the stated assumptions [2], [8], [90]. In calculus-focused settings, this includes contradictory or mathematically invalid claims about foundational relationships (e.g., continuity and differentiability) and invalid implications that conflate geometric intuitions (e.g., tangent existence) with topological properties (e.g., continuity), revealing breakdowns in conceptual coherence during multi-step explanation [7]. In more formal reasoning tasks, notation-level misunderstandings (e.g., misreading symbols or divisibility relations) can shift the target statement itself and render subsequent argumentation logically irrelevant, while instability across near-identical prompts indicates weak global consistency in proof-oriented reasoning [116]. Notably, reasoning problems also surface in assessment-like interactions where an LLM affirms an answer as "correct" despite an incorrect intermediate method, an error mode that is pedagogically salient because it conflates outcome accuracy with procedural validity and can normalize invalid reasoning in learning environments [130].
Consequently, the literature supports a cautious positioning of LLMs as potentially useful for explanation and exploration, but not as autonomous authorities for mathematical correctness. For educational use across levels, these findings motivate design principles that prioritize verification (e.g., independent checking, constraint auditing, solution-set completeness checks) and process-focused scaffolds that make reasoning steps explicit and assessable, rather than relying on fluent final answers as evidence of mathematical understanding

Domain-concentrated weaknesses in higher-level mathematics: Statistics/probability, and calculus

For statistics and probability, the evidence consistently points to upper secondary and college as key pressure points because many tasks demand interpretive judgement, such as selecting an appropriate method, translating conditions into a correct setup, and checking whether outputs are plausible rather than executing a single procedure. In a comparative evaluation of four chatbots using 12 prompts (5 statistics; 7 calculus) [51] report that weaker models (notably Bard and LLaMA) frequently return incorrect results with low clarity, while even stronger models are not error-free, indicating that reliability varies meaningfully by model and by prompt. This domain sensitivity is reinforced by standardized-test evidence from Korea's CSAT: across 92 mathematics items (2023–2024), [55] reports a large overall performance gap (49.46% accuracy for gpt-4o vs 81.52% for o1-preview) and a pronounced disparity within Probability & Statistics (100% for o1-preview vs 50% for gpt-4o), underscoring that statistical competence is not uniform across LLMs and may remain unstable in high-stakes contexts.

Calculus exhibits a comparable concentration at the senior high school–college transition, but the dominant issues extend beyond "getting the final answer wrong" to include conceptual inconsistency and procedural validity. In a university calculus argumentation task on tangency [7], document how ChatGPT alternates between correct statements (e.g., differentiability implies continuity) and contradictory claims (e.g., asserting "no direct relationship" between continuity and derivative), a pattern that can mislead learners even when working with foundational definitions . Large-scale assessment results similarly suggest calculus can be a persistent weak area for some models: on CSAT items, gpt-4o's calculus accuracy remains 37.5% in both 2023 and 2024, with further declines as item difficulty increases ([55]). Finally, classroom-based evidence highlights a subtler instructional risk: an AI system may validate a numerically correct endpoint result while failing to detect an incorrect derivative procedure that happens to produce the same value at a specific point, reinforcing the need to evaluate methods rather than outcomes alone ([130]).

4.1 Lack of pedagogical capabilities of LLMs in mathematics education

The limitations of large language models extend beyond computational and reasoning errors, as highlighted in the table above. A significant concern, particularly in educational contexts, is their lack of pedagogical capabilities. Effective teaching requires more than accurate problem-solving; it demands structured guidance and adaptive feedback that are conditioned on a model of the learner's current understanding (i.e., a student model) and aligned with established educational principles. However, existing literature ([4], [8], [13], [33], [46], [83], [125]) emphasizes that LLMs often fail to deliver such learner-focused

instruction, limiting their ability to function as intelligent tutors without additional instructional control and teacher oversight. This not only diminishes their effectiveness but also raises questions about their suitability for roles requiring nuanced pedagogical strategies. While this section focuses on pedagogical challenges, the mathematical performance of LLMs, including their reasoning and problem-solving abilities, will be explored in detail in a later section.

One critical limitation of LLMs lies in their inability to scaffold mathematical concepts effectively. Scaffolding involves breaking down complex ideas into manageable steps tailored to a learner's level of understanding. While LLMs can provide direct answers, they often bypass crucial intermediary steps, leaving students with shallow comprehension ([4], [83]). For instance, ChatGPT frequently generates correct solutions without offering the contextual reasoning necessary for students to generalize these solutions to new problems ([83]). This lack of scaffolding hinders students from developing a deep understanding of mathematical concepts and transferring knowledge across domains.

LLMs also fail to foster critical thinking in mathematics education. Unlike human teachers who employ Socratic questioning to prompt reflection and analysis, LLMs provide static responses that do not engage students in deeper reasoning ([33]). This limitation is particularly evident in addressing misconceptions or abstract concepts, where LLMs are unable to anticipate and correct misunderstandings effectively. Consequently, teacher intervention becomes essential to ensure comprehension and fill these pedagogical gaps ([4], [33]).

Another significant challenge is the lack of adaptive feedback. While LLMs can generate generalized explanations and practice problems, they are unable to adjust their responses based on the unique needs of individual learners ([46]). For instance, in creative mathematical writing activities, students often struggled to understand the mathematical ideas embedded in AI-generated content. Teachers reported that these outputs required substantial revision to align with curricular goals and student learning levels, further highlighting LLMs' inability to provide personalized support ([46]).

These limitations underscore the indispensable role of human educators in mathematics education. While AI tools can assist by automating routine tasks and providing supplementary resources, they cannot independently fulfill the pedagogical responsibilities required to foster deep learning, critical thinking, and conceptual mastery. A balanced integration of AI tools alongside teacher-led guidance is essential to ensure that LLMs enhance, rather than hinder, the learning process in mathematics classrooms.

4.2 Educator concerns about future LLM use in mathematics classrooms

Beyond concerns about pedagogy and mathematical performance, educators are increasingly apprehensive about the long-term impact of LLMs on both students and

teachers. Without strong human oversight, these tools risk undermining mathematical reasoning, widening achievement gaps, and raising new ethical challenges. This section explores key concerns, including over-reliance on LLMs, limitations in replicating teacher-led scaffolding, and issues of equity and data ethics, underscoring the need for careful integration of AI into education.

Over-reliance and the erosion of mathematical thinking

Teachers and empirical studies caution that "instant" explanations can reduce productive struggle and increase students' acceptance of superficially plausible but incorrect reasoning. For example, studies in both secondary and university settings report that students may treat LLMs' confirmations as authoritative and fail to detect hallucinated or flawed steps unless guided to interrogate outputs and refine queries ([58], [130], [66]). This risk appears especially pronounced for lower-performing learners, potentially amplifying rather than closing competence gaps ([66]).

Challenges in replicating teacher-led scaffolding with LLMs

Another recurring theme in the literature is educators' apprehension about the potential replacement of teachers by LLMs. Surveys conducted in Germany ([3]) and the Philippines ([83]) revealed that while most educators appreciated the speed and efficiency of LLMs, they strongly rejected the idea that chatbots could replicate critical pedagogical functions, such as diagnostic questioning, formative feedback, and social-emotional support. Teachers emphasized that LLMs currently lack the capacity to:

1) Tailor prompts to individual learning styles (personalization gap; [83]).
2) Diagnose the underlying misconceptions behind incorrect answers (as seen in geometry tasks; [35]).
3) Provide adaptive follow-up questions calibrated to students' cognitive demands ([117]).

These findings underscore the indispensable role of teachers in providing nuanced, human-centered scaffolding that LLMs are currently incapable of replicating.

Ensuring equitable access to reliable LLM tools

Equity and ethical considerations present significant challenges to the integration of LLMs in mathematics classrooms, particularly in under-resourced regions. Limited infrastructure, including low bandwidth, device shortages, and restrictive data-privacy regulations, continues to hinder the adoption of these tools ([58]). These barriers disproportionately affect underserved communities, potentially exacerbating existing inequities in education. Even in well-resourced environments, concerns about biases in training data, opaque citation practices, and the potential for hallucinated content highlight the challenges of creating equitable access to reliable AI tools. These issues are particularly problematic in

under-resourced regions, where fewer safeguards exist to verify AI outputs in high-stakes educational scenarios, further exacerbating disparities in education.

5. How do LLMs perform on mathematical tasks across different difficulty levels?

In the previous section, we explored how mathematical performance has emerged as a critical concern in the literature, with studies consistently reporting that LLMs, including ChatGPT, often produce fluent yet incorrect reasoning across various mathematical tasks. For instance, errors have been frequently observed in geometry, algebra, and derivative problems ([35], [130]). Recognizing the importance of these challenges, this section delves deeper into the performance of LLMs on mathematical tasks of varying complexity. The analysis builds on findings from studies categorized under label topics 2 and 4, leveraging widely referenced datasets such as GSM8K, SVAMP, and MATH to provide a comprehensive evaluation.

1) **GSM8K**: Developed by Cobbe et al. (2021), this dataset contains 8,500 grade-school math word problems (7,500 for training and 1,000 for testing). It evaluates LLMs' mathematical reasoning capabilities through problems requiring 2 to 8 steps, mainly focusing on basic arithmetic.
2) **SVAMP**: Introduced by Patel et al. (2021), SVAMP includes 1,000 elementary-level arithmetic word problems for grades four and below. These problems, based on variations of the ASDiv-A dataset, test LLMs' reasoning, robustness, and ability to handle structural invariance beyond pattern recognition.
3) **MATH**: Created by Hendrycks et al. (2021), this dataset comprises 12,500 advanced competition-level problems (7,500 for training and 5,000 for testing) from contests like AMC 8, AMC 10, AMC 12, and AIME. Covering seven subjects (e.g., Prealgebra, Geometry, Number Theory), the problems require advanced reasoning and include step-by-step solutions.

The figures (5-7) and tables (8-10) below summarize benchmarks from studies that used these datasets to evaluate LLMs. While researchers often rename models after fine-tuning or modifications, we recorded only the best-performing models in each study, grouped by the foundational LLMs on which they were based.

Figure 5

Scatter plot of the GSM8K benchmark

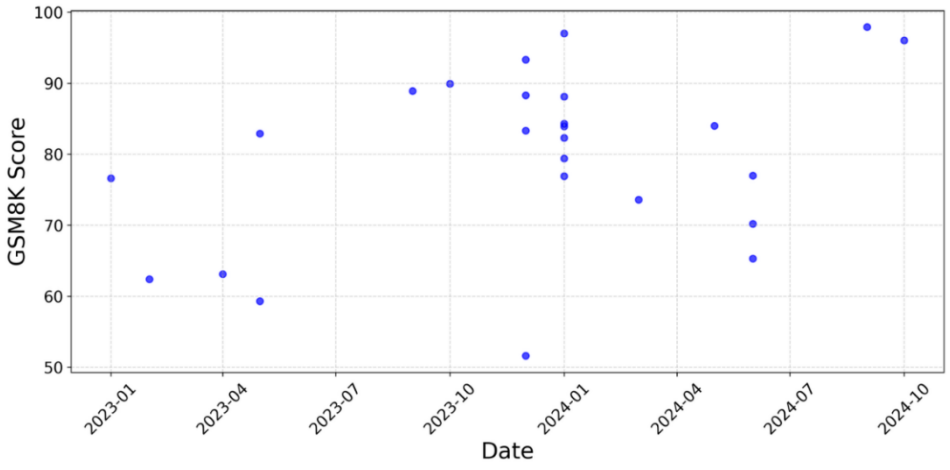


Table 8

GSM8K benchmark

Article ID	Model	GSM8K Score	Date
82	Codex	76.6	2023/01
79	GPT-3	62.4	2023/02
28	Codex	63.1	2023/04
41	Codex	82.9	2023/05
101	GPT-3	59.3	2023/05
73	GPT-3.5-turbo	88.9	2023/09
24	GPT-4	89.9	2023/10
108	LlaMA-33B	51.6	2023/12
77	GPT-4	93.3	2023/12
100	GPT-3.5-turbo	83.3	2023/12
96	LLaMa-2 (70B)	88.3	2023/12
45	GPT-3.5	79.4	2024/01
85	LlaMA-2 (70B)	76.9	2024/01
88	LLaMa-2 (70B)	83.9	2024/01
94	LLaMa-2 (70B)	82.3	2024/01
112	GPT-4	88.1	2024/01
113	GPT-4 Code Interpreter	97.0	2024/01
128	LLaMa-2 (70B)	84.3	2024/01
68	text-davinci-003 (175B)	73.6	2024/03
86	GPT-3.5-turbo	84.0	2024/05
14	Mistral (7B)	70.2	2024/06
93	GPT-3.5-turbo	77.0	2024/06
110	LLaMa-2 (70B)	65.3	2024/06
106	Qwen2.5-Math-7B-Instruct	97.9	2024/09
98	Llama3.1-70B	96.0	2024/10

Figure 6

Scatter plot of SVAMP benchmark

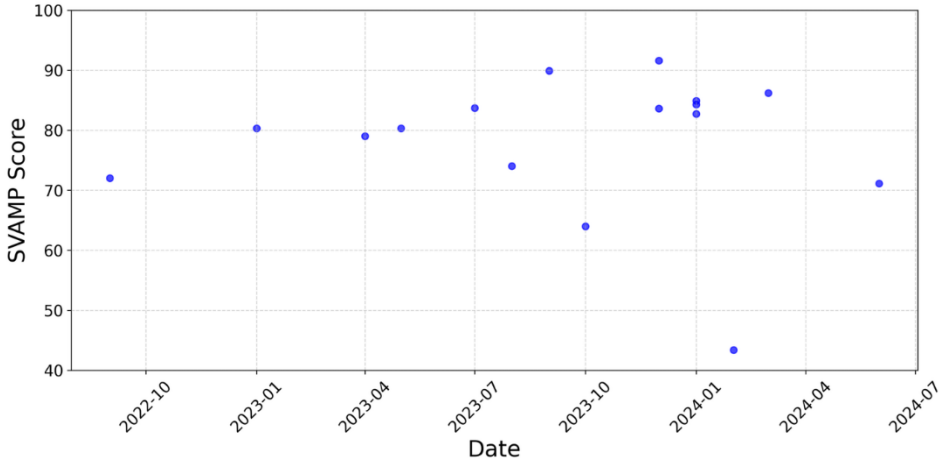


Table 9

SVAMP benchmark

Article ID	Model	SVAMP	Date
119	GPT-3	72.0	2022/09
82	Codex	80.3	2023/01
28	PaLM (540B)	79.0	2023/04
101	GPT-3	80.3	2023/05
44	GPT-3.5-turbo	83.7	2023/07
107	GPT-3	74.0	2023/08
73	GPT-3.5-turbo	89.9	2023/09
10	ChatGPT	64.0	2023/10
77	GPT-4	91.6	2023/12
100	GPT-3.5-turbo	83.6	2023/12
85	CodeLlaMA (34B)	84.3	2024/01
88	LlaMA-2 (70B)	84.9	2024/01
128	LlaMA-2 (70B)	82.7	2024/01
50	RoBERTa Large (355 M)	43.4	2024/02
68	text-davinci-003	86.2	2024/03
110	LlaMA-2 (70B)	71.1	2024/06

Figure 7

Scatter plot of math benchmark

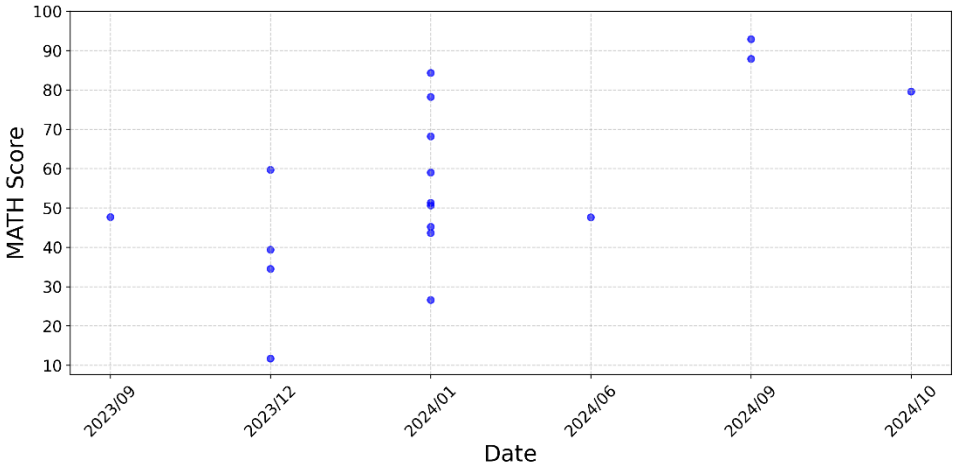


Table 10

Math benchmark

	Model	Math	Date
73	GPT-3.5 turbo	47.7	2023/09
31	ChatGPT	39.4	2023/12
43	ChatGPT	59.7	2023/12
123	GPT-3	11.7	2023/12
96	Llama-2 (70B)	34.5	2023/12
45	GPT-3.5	68.2	2024/01
69	GPT-4	78.2	2024/01
81	CodeLlaMA (34B)	43.6	2024/01
88	CodeLlaMA (34B)	45.2	2024/01
94	LLaMa-2 (70B)	26.6	2024/01
112	GPT-4	51.3	2024/01
113	GPT-4 Code Interpreter	84.3	2024/01
128	CodeLlaMA (34B)	50.6	2024/01
129	GPT-4	59.0	2024/01
93	GPT-3.5 turbo	47.6	2024/06
84	GPT-4 Turbo	87.9	2024/09
106	Qwen2.5-Math-72B-Instruct	92.9	2024/09
98	Llama3.1-70B	79.6	2024/10

Primary and secondary-level mathematics

Large Language Models have demonstrated remarkable progress in primary and secondary-level mathematics, achieving high accuracy on benchmark datasets. For example, by Q2 of 2023, most LLMs scored above 80% on datasets like GSM8K and SVAMP, with top-performing models exceeding 90%. This strong performance underscores the increasing sophistication of LLMs in handling foundational mathematical tasks. However, performance on more challenging datasets, such as MATH, has been more variable. While models like GPT-4 achieved satisfactory result
s early on, many earlier or smaller models struggled to handle competition-level math questions effectively.

Notably, the introduction of the GPT-4 Code Interpreter, an extension of GPT-4 with coding functionalities, marked a significant leap in LLMs' performance. For instance, this model scored an impressive 97% on GSM8K and 84% on the MATH dataset ([113]). By Q4 of 2024, newer models such as GPT-4 Turbo ([84]), Llama-3.1 ([98]), and Qwen2.5 ([106]) further demonstrated substantial advancements across benchmarks, highlighting the rapid evolution of LLMs' capabilities in mathematics. These results set the stage for examining their performance on more advanced, college-level mathematical tasks.

College-level mathematics

While LLMs have shown strong performance in primary and secondary-level mathematics, their results in college-level mathematics have been more varied. Studies indicate that as the complexity of mathematical tasks increases, the accuracy of LLMs tends to decline. For example, the GPT-4 Code Interpreter achieved an overall accuracy of 85.32% on the MMLU dataset, which spans 54 subjects, but its performance dropped to 70–80% in areas such as college mathematics and abstract algebra ([113]). Earlier models, such as GPT-3.5-turbo and LLaMA-2, performed even less effectively, scoring 69.5% and 56.7%, respectively ([93], [85]).

However, more recent models have shown significant promise in addressing these challenges. For instance, Qwen2.5-Math-72B-Instruct ([106]) achieved over 90% accuracy on the MMLU-STEM dataset, demonstrating notable improvements in tackling advanced mathematical tasks. These developments highlight the growing potential of LLMs in higher-level mathematics, even as limitations persist in certain domains. To further understand the scope of these challenges, it is important to consider how language influences LLMs' performance on mathematical tasks.

6. The language used in mathematical questions poses a potential limitation

The performance of LLMs on mathematical tasks is often influenced by the language in which the questions are presented. Most studies evaluating LLMs' performance focus on

English-language datasets, as English dominates the training corpora of these models. However, this linguistic bias introduces significant disparities when testing LLMs on non-English languages, particularly low-resource ones. For example, models such as GPT-3.5, GPT-4, and Bard frequently produced incorrect answers to non-English mathematical questions, often due to errors in reasoning, conceptual understanding, or problem construction.

A study ([42]) assessing LLMs' performance on Chinese high school conic section problems using the CONIC10K dataset illustrates this challenge. While models demonstrated adequate comprehension in semantic parsing tasks, they struggled with mathematical question answering (mathQA). Even GPT-4, the best-performing model, achieved only 15.5% accuracy using zero-shot chain-of-thought prompting under human evaluation. Translating the problems into English modestly improved GPT-4's accuracy to 26.0%, but this remained significantly below the 57.5% accuracy achieved by human experts. Researchers attributed this poor performance primarily to reasoning limitations rather than linguistic barriers, underscoring the challenges LLMs face in non-English domains.

Despite these challenges, recent advancements in multilingual capabilities have shown promise. As highlighted in OpenAI's technical report (2023), GPT-4 achieved competitive multilingual performance on the MMLU benchmark, a dataset comprising 57 subjects translated into 26 languages. In English, GPT-4 scored 86.4% accuracy, and in high-resource languages such as Spanish, French, and German, accuracy exceeded 85%. Furthermore, GPT-4 performed well in low-resource languages, achieving 80.1% in Latvian, 80.0% in Welsh, and 79.9% in Swahili. These results mark a significant milestone in addressing the linguistic limitations of earlier models, narrowing the performance gap between English and other languages.

The improvements in multilingual performance can be attributed to GPT-4's enhanced training methodologies, which likely incorporate more diverse and representative multilingual datasets. However, challenges remain, particularly with low-resource and morphologically rich languages. Nonetheless, these advancements represent a critical step toward developing globally inclusive AI systems that address linguistic and mathematical disparities.

7. Remediation techniques for improving mathematical performance

Across the reviewed studies, remediation techniques (e.g., prompt engineering, fine-tuning, and external tool integration) consistently boost the accuracy of LLMs on mathematical benchmarks. Beyond higher scores, these gains often reflect improved solution reliability: reduced arithmetic errors and fewer hallucinated intermediate steps or unjustified conclusions due to self-verification and tool-assisted checking. In an ITS framing, improving reliability is consequential because it reduces measurement noise in both (i) the

expert model layer (e.g., generating correct solution paths, verifying steps, and selecting strategies) and (ii) the student-model/overlay layer, where the system attempts to infer which knowledge components are mastered or missing. When the LLMs are less prone to random slips or hallucinations, incorrect outputs are more likely to reflect genuine gaps in comprehension, representation, or strategy selection, supporting more trustworthy diagnosis and, in turn, more targeted scaffolding.

To provide a comprehensive overview, Table 11 summarizes the remediation techniques employed across the reviewed articles, while Table 12 highlights the top-performing models ([84], [98], [106], [113]), offering a concise description of the applied techniques alongside their respective performance improvements.

Table 11

Overview of remediation techniques across studies

ID	Prompt engineering	Fine-tuning	External tool integration
14		✓	
21	✓	✓	
28	✓		
31	✓		
41	✓		
43	✓		✓
45	✓		
50		✓	
52	✓		✓
68	✓		
73	✓		
79	✓		
81		✓	
82		✓	
84	✓		
85		✓	✓
88			✓
92	✓		✓
93	✓		✓
94		✓	
95	✓		✓
96		✓	
98		✓	
100	✓		✓

101	✓			
104	✓			
106		✓		✓
107	✓			
108	✓			
110	✓			
112	✓			
113	✓			✓
115	✓			
119	✓			
123	✓			
128	✓			✓
129	✓			✓

Table 12

Top-performing models, remediation techniques, and performance improvements

Model	Remediation techniques	Key features	Performance
GPT-4 Turbo ([84])	• Role-Based Prompting	Utilizes a multi-agent system with Thinker, Judge, and Executor roles to iteratively refine tasks and eliminate logical errors.	Improves GPT-4 Turbo's accuracy on MATH dataset from 72.78% to 87.92% (+15.14%)
OpenMathInstruct-2 ([98])	• Concise Chain-of-Thought (CoT) Reasoning, • Pre-Training with Synthetic Data	Introduces 40% shorter CoT formats to reduce verbosity, pre-trains on 14M diverse question-solution pairs with synthetic augmentation and uses reward models with Group Relative Policy Optimization (GRPO) for iterative refinement.	Achieves +15.9% accuracy on MATH (51.9% → 67.8%) and +10.5% improvement through synthetic data.
GPT-4 Code Interpreter ([113])	• Zero-Shot Code-Based Self-Verification (CSV) • Python Integration for Code-Based Verification	Dynamically generates and verifies solutions using Python code, iteratively adjusts reasoning when flagged during self-verification, and integrates external tools for enhanced computational accuracy.	Improves MATH dataset accuracy from 53.9% to 84.3% (+30.4%) by leveraging code-based verification and self-debugging.
Qwen2.5-Math ([106])	• Tool-Integrated Reasoning Prompts • Pre-Training with High-Quality Datasets • Supervised Fine-Tuning (SFT) • Python Integration	Embeds Python-based reasoning prompts to offload computations, pre-trains on 1T tokens (Qwen Math Corpus v2) with symbolic reasoning tasks, fine-tunes on 2.5M CoT problems and 395K Tool-Integrated Reasoning samples, and leverages Python interpreters for symbolic reasoning and computation.	Achieves 91.6% on GSM8K, 67.8% on MATH, and near-perfect accuracy on AMC 2023 (solving 39/40 problems in TIR mode).

These gains are promising for instructional settings because they improve the model's accuracy and reliability in solving mathematics problems. That said, higher accuracy alone does not guarantee misconception diagnosis; ITS-aligned use requires that remediation pipelines also elicit interpretable intermediate representations (e.g., problem parsing, variable definitions, and strategy justification) so the overlay model can distinguish conceptual errors from execution failures. To connect these gains to ITS use, we organize remediation techniques by the primary lever they modify: (1) inference-time guidance that can elicit interpretable intermediate representations (prompt engineering), (2) parameter-

level adaptation that improves consistency and domain alignment (fine-tuning), and (3) execution-time verification that reduces arithmetic and symbolic slips (external tool integration).

Conclusion

This systematic review synthesizes recent research on Large Language Models in mathematics education across four themes: (1) assessing mathematical performance, (2) supporting teaching and learning, (3) automating educational tasks, and (4) mathematical word problems. Overall, LLMs perform well on many primary and secondary tasks, especially when supported by fine-tuning, prompting strategies, or external tools. However, reliability drops as problems require abstract reasoning, multi-step planning, or domain-specific formalisms (e.g., calculus, symbolic manipulation, and spatial reasoning).

Across educational practice, LLMs are most mature as assistive tools: they can draft lesson materials, generate practice items, summarize student work, and provide preliminary feedback, which may reduce teacher workload. They also show promise for pre-service teacher education and professional development through coaching and simulation activities. However, these uses require human oversight to ensure mathematical correctness, curricular alignment, and pedagogical appropriateness.

Hallucinations remain a key limitation, uneven performance across languages and contexts, limited pedagogical awareness, and risks of over-reliance that may weaken students' independent reasoning and problem-solving. Emerging agentic systems may increase convenience by automating multi-step tasks, but they also heighten concerns about authority, accountability, and the preservation of core mathematical practices.

Future work should prioritize improving robustness and verifiability, developing pedagogy-aware designs that support learning rather than answer substitution, and addressing equity, privacy, and access, particularly in under-resourced settings and diverse linguistic and cultural contexts. With careful governance and human-centered integration, LLMs can expand scalable support in mathematics education while complementing, not replacing, teacher expertise and students' development of mathematical thinking.

Limitations

This review has several limitations. The rapid pace of advancements in artificial intelligence and large language models makes it challenging to capture all developments in mathematical performance and educational applications, as new models and updates are introduced faster than academic studies can evaluate them. Additionally, most studies reviewed focus on short-term outcomes, leaving gaps in understanding the long-term effects of LLMs on foundational skills and educational practices. The review also has a linguistic bias, prioritizing studies published in English and datasets in English, which limits insights into LLMs' performance in non-English or low-resource languages.

Furthermore, the benchmarks used, such as GSM8K, SVAMP, and MATH, may not fully represent the diversity of mathematical problems encountered in real-world classrooms, and many studies lack detailed pedagogical context, offering limited guidance on how LLMs integrate into teaching strategies and curricula. Lastly, there is limited exploration of how educators and LLMs can collaborate effectively, including teacher training and user trust, which are critical for successful implementation.

Appendix A

Table 13

List of articles assigned after refining BERTopics results

ID	Article	In-text Citation	Assigned Topic
1	A Case Study Using Large Language Models to Generate Metadata for Math Questions	(Bainbridge et al., 2023)	2
2	AI Chatbots as Math Algorithm Problem Solvers: A Critical Evaluation of Its Capabilities and Limitation	(Dahal et al., 2023)	0
3	AI in STEM education: The relationship between teacher perceptions and ChatGPT use	(Beege et al., 2024)	0
4	AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention	(Hashem et al., 2024)	0
5	AI-Based Mathematics Learning Platforms in Undergraduate Engineering Studies: Analysis of User Experience	(Vintere et al., 2024)	0
6	Algebra Error Classification with Large Language Models	(McNichols et al., 2023)	2
7	An argumentation experience regarding concepts of calculus with ChatGPT	(Urhan et al., 2024)	0
8	An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam	(Korkmaz Guler et al., 2024)	0
9	An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP)	(Shakarian et al., 2023)	3
10	Analyzing ChatGPT's Mathematical Deficiencies: Insights and Contributions	(Cheng & Zhang, 2023)	3
11	Analyzing Student Attention and Acceptance of Conversational AI for Math Learning: Insights from a Randomized Controlled Trial	(Li et al., 2024)	0
12	Applications of Artificial Intelligence and Their Relationship to Spatial Thinking and Academic Emotions Towards Mathematics: Perspectives from Educational Supervisors	(Elsayed, 2023)	0
13	Are Lesson Plans Created by ChatGPT More Effective? An Experimental Study	(Karaman & Göksu, 2024)	0
14	Arithmetic Reasoning with LLM: Prolog Generation & Permutation	(Yang et al., 2024)	1
15	Artificial Intelligence in Elementary Math Education: Analyzing Impact on Students Achievements	(Bešlić et al., 2024)	0
16	Assessing Concepts, Procedures, and Cognitive Demand of ChatGPT-generated Mathematical Tasks	(Sapkota & Bondurant, 2024)	0

17	Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models	(Morris et al., 2025)	2
18	Automatic Generation of Socratic Subquestions for Teaching Math Word Problems	(Shridhar et al., 2022)	2
19	Automatic Lesson Plan Generation via Large Language Models with Self-critique Prompting	(Zheng et al., 2024)	2
20	AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails	(Pal Chowdhury et al., 2024)	2
21	Benchmarking and Improving Generator-Validator Consistency of Language Models	(Li et al., 2024)	1
22	Benchmarking Hallucination in Large Language Models based on Unanswerable Math Word Problem	(Sun et al., 2024)	3
23	Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes	(Wang et al., 2024)	2
24	Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate	(Wang et al., 2023)	1
25	Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish	(Parra et al., 2024)	-1
26	Can Large Language Models Generate Middle School Mathematics Explanations Better Than Human Teachers?	(Wang et al., 2024)	2
27	Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange	(Satpute et al., 2024)	3
28	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models	(Wei et al., 2022)	1
29	Chain-of-Thoughts Prompting with Language Models for Accurate Math Problem-Solving	(Fung, Wong, & Tan, 2023)	0
30	Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5. ChatGPT-4, and Google Bard	(Plevris, Papazafeiropoulos, & Jiménez Rios, 2023)	0
31	ChatCoT: Tool-Augmented Chain-of-Thought Reasoning on Chat-based Large Language Models	(Chen et al., 2023)	1
32	ChatGPT as a Math Questioner? Evaluating ChatGPT on Generating Pre-university Math Questions	(Pham Van Long et al., 2024)	2
33	ChatGPT in didactical tetrahedron, does it make an exception? A case study in mathematics teaching and learning	(Dasari et al., 2024)	0
34	ChatGPT in education: benefits and challenges of ChatGPT for mathematics and science teaching practices	(Taani & Alabidi, 2024).	0
35	ChatGPT: A revolutionary tool for teaching and learning mathematics	(Wardat et al., 2023)	0
36	ChatGPT's performance in university admissions tests in mathematics	(Udias et al., 2024)	0
37	ChatGPT-Based Simulation Helps to Develop the Pre-Service Mathematics Teachers' Critical Thinking.	(Drushlyak et al., 2025)	0
38	ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills	(Pardos & Bhandari, 2024)	2

39	Classifying Tutor Discursive Moves at Scale in Mathematics Classrooms with Large Language Models	(Moreau-Pernet et al., 2024)	2
40	Comparing Traditional Instructional Methods to ChatGPT: A Comprehensive Analysis	(Ramprakash et al., 2024)	0
41	Complexity-Based Prompting for Multi-step Reasoning	(Fu et al., 2023)	1
42	CONIC10K: A Challenging Math Problem Understanding and Reasoning Dataset	(Wu et al., 2023)	1
43	Creator: Tool Creation for Disentangling Abstract and Concrete Reasoning of Large Language Models	(Qian et al., 2023)	1
44	Does ChatGPT Comprehend Place Value in Numbers When Solving Math Word Problems?	(An et al., 2023)	3
45	Don't Trust: Verify- Grounding LLM Quantitative Reasoning with Auto formalization	(Zhou et al., 2024)	1
46	Elementary school students' and teachers' perceptions toward creative mathematical writing with Generative AI	(Song et al., 2025)	0
47	Empowering mathematics teacher educators: Exploring Artificial Intelligence-driven mathematical tasks	(Aqazade, Mauntel, & Atabas, 2025)	0
48	Enhancing mathematics education for students with special educational needs through generative AI: A case study in Greece	(Rizos, Foykas, & Georgakopoulos, 2024)	0
49	Enhancing mathematics education through AI Chatbots in a Flipped Learning Environment	(Martínez-Téllez & Camacho-Zuñiga, 2023)	0
50	Enhancing numerical reasoning performance by augmenting distractor numerical values	(Hwang et al., 2024)	1
51	Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics	(Santandreu Calonge, Smail, & Kamalov, 2023)	0
52	Evaluating and Improving Tool-Augmented Computation-Intensive Math Reasoning	(Zhang et al., 2023)	1

53	Evaluating ChatGPT-Generated Linear Algebra Formative Assessments	(Rigaud Téllez, Rayón Villela, & Blanco Bautista, 2024)	2
54	Evaluating language models for mathematics through interactions	(Collins et al., 2024)	2
55	Evaluating Mathematical Problem-Solving Abilities of Generative AI Models: Performance Analysis of o1-preview and gpt-4o Using the Korean College Scholastic Ability Test	(Oh, 2025)	0
56	Evaluation of Large Scale Language Models on Solving Math Word Problems with Difficulty Grading	(He, He, Gao, & Sun, 2023)	3
57	Examining Mathematics Teachers Noticing the Rationality: Scenario-Based Training with AI Chatbot	(Galiç et al., 2025)	0
58	Exploring the Integration of Artificial Intelligence-Based ChatGPT into Mathematics Instruction: Perceptions, Challenges, and Implications for Educators	(Egara & Mosimege, 2024)	0
59	Exploring Automated Distractor Generation for Math Multiple-choice Questions via Large Language Models	(Feng et al., 2024)	2
60	Exploring Mathematics Teacher Candidates' Instrumentation Process of Generative Artificial Intelligence for Developing Lesson Plans	(Tapan Broutın, 2024)	0
61	Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning	(Gattupalli et al., 2023)	2
62	Exploring the integration of artificial intelligence in math education: Preservice Teachers' experiences and reflections on problem-posing activities with ChatGPT	(Kim, Park, & Joung, 2025)	0
63	Exploring the performance of ChatGPT for numerical solution of ordinary differential equations.	(Koceski, Koceska, Koceva Lazarova, Miteva, & Zlatanovska, 2025)	0
64	Flipped dialogic learning method with ChatGPT: A case study	(Palvova, 2024)	0
65	Fostering problem solving and critical thinking in mathematics through generative artificial intelligence	(Barana et al., 2023)	0
66	From Helping Hand to Stumbling Block: The ChatGPT Paradox in Competency Experiment	(Kim & Moon, 2024)	0
67	From Large to Tiny: Distilling and Refining Mathematical Expertise for Math Word Problems with Weakly Supervision	(Lin et al., 2024)	-1
68	Get an A in Math: Progressive Rectification Prompting	(Wu et al., 2024)	1
69	GOLD: Geometry Problem Solver with Natural Language Description	(Zhang & Moshfeghi, 2024)	-1
70	Improving mathematics assessment readability: Do large language models help?	(Patel et al., 2023)	2
71	Improving Mathematics Tutoring With A Code Scratchpad	(Upadhyay et al., 2023)	2
72	Inclusive Education in Rural Settings: Leveraging Technology for Improved Math Learning Outcomes Among Students With Disabilities	(Lin & Riccomini, 2025)	0

73	Instructing Large Language Models to Identify and Ignore Irrelevant Conditions	(Wu, Shen, & Jiang, 2024)	1
74	Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement	(Wu et al., 2025)	0
75	Integration of ChatGPT in mathematical story-focused 5E lesson planning: Teachers and pre-service teachers' interactions with ChatGPT	(Şimşek, 2025)	0
76	Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction	(Wang & Demszky, 2023)	0
77	It Ain't Over: A Multi-aspect Diverse Math Word Problem Dataset	(Kim et al., 2023)	3
78	Learning Mathematics with Large Language Models: A Comparative Study with Computer Algebra System	(Matzakos, Doukakis, & Moundridou, 2023)	0
79	Least-to-most Prompting Enables Complex Reasoning in Large Language Models	(Zhou et al., 2023)	1
80	Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation	(Liang et al., 2023)	2
81	Let's Verify Step by Step	(Lightman et al., 2024)	1
82	Leveraging Training Data in Few-Shot Prompting for Numerical Reasoning	(Jie & Lu, 2023)	1
83	Levering AI to enhance students' conceptual understanding and confidence in mathematics	(Canonigo, 2024)	0
84	MACM: Utilizing a Multi-Agent System for Condition Mining in Solving Complex Mathematical Problems	(Lei et al., 2024)	1
85	MAMMOTH: Building Math Generalist Models Through Hybrid Instruction Tuning	(Yue et al., 2024)	1
86	Math Problem Solving : Enhancing Large Language Models with Semantically Rich Symbolic Variables	(Narin, 2024)	1
87	MathAttack: Attacking Large Language Models towards Math Solving Ability	(Zhou et al., 2024)	1
88	MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning	(Wang et al., 2024)	1
89	Mathematical Capabilities of ChatGPT	(Frieder et al., 2023)	1
90	Mathematical Modelling Abilities of Artificial Intelligence Tools: The Case of ChatGPT	(Spreitzer et al., 2024)	0
91	Math-LLMs: AI Cyberinfrastructure with Pre-trained Transformers for Math Education	(Zhang et al., 2025)	-1
92	MathPrompter: Mathematical Reasoning using Large Language Models	(Imani, Du, & Shrivastava, 2023)	1
93	MATHSENSEI: A Tool-Augmented Large Language Model for Mathematical Reasoning	(Das et al., 2024)	1
94	MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models	(Yu et al., 2024)	1
95	Multi-tool Integration Application for Math Reasoning Using Large Language Model	(Duan & Wang, 2024)	1
96	MuMath: Multi-perspective Data Augmentation for Mathematical Reasoning in Large Language Models	(You et al., 2024)	1
97	On Using GeoGebra and ChatGPT for Geometric Discovery	(Botana et al., 2024)	-1

98	OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data	(Toshniwal et al., 2024)	1
99	Over-Reasoning and Redundant Calculation of Large Language Models	(Chiang & Lee, 2024)	1
100	Plan, Verify and Switch: Integrated Reasoning with Diverse X-of-Thoughts	(Liu et al., 2023)	1
101	Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models	(Wang et al., 2023)	1
102	Pre-service teachers and ChatGPT in multistrategy problem-solving: implications for mathematics teaching in primary schools	(Getenet, 2024)	0
103	Prof Pi: Tutoring Mathematics in Arabic Language using GPT-4 and Whatsapp	(Butgereit et al., 2023)	0
104	Prompt Incorporates Math Knowledge to Improve Efficiency and Quality of Large Language Models to Translate Math Word Problems	(Wu & Yu, 2023)	1
105	Prompt the problem - investigating the mathematics educational quality of AI- supported problem solving by comparing prompt techniques	(Schorcht et al., 2024)	2
106	QWEN2.5-MATH TECHNICAL REPORT: TOWARD MATHEMATICAL EXPERT MODEL VIA SELFIMPROVEMENT	(Yang et al., 2024)	1
107	Reasoning in Large Language Models Through Symbolic Math Word Problems	(Gaur & Saunshi, 2023)	1
108	Reasoning with Language Model is Planning with World Model	(Hao et al., 2023)	1
109	Reshaping Mathematics Instruction Via Impact of AI Chatbots on Secondary Education Pre-service Teachers	(Luzano, 2024)	0
110	RESPROMPT: Residual Connection Prompting Advances Multi-Step Reasoning in Large Language Models	(Jiang et al., 2024)	1
111	Rewriting Content with GPT-4 to Support Emerging Readers in Adaptive Mathematics Software	(Norberg et al., 2025)	2
112	SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-By-Step Reasoning	(Miao et al., 2024)	1
113	Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-Based Self-Verification	(Zhou et al., 2024)	1
114	Solving the Self-regulated Learning Problem: Exploring the Performance of ChatGPT in Mathematics	(Li et al., 2023)	0
115	Specialized Mathematical Solving by a Step-By-Step Expression Chain Generation	(Zhang et al., 2024)	1
116	Students' use of generative artificial intelligence for proving mathematical statements	(Yoon et al., 2024)	0
117	Students' perceived roles, opportunities, and challenges of a generative AI-powered teachable agent: a case of middle school math class	(Song et al., 2024)	0
118	Supporting Teachers' Professional Development With Generative AI: The Effects on Higher Order Thinking and Self-Efficacy	(Lu et al., 2024)	0
119	Symbolic Math Reasoning with Language Models	(Gaur & Saunshi, 2022)	1
120	Teaching Language Models to Self-Improve through Interactive Demonstrations	(Yu et al., 2024)	-1

121	Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design	(Hu et al., 2024)	2
122	Teaching the Specialized Language of Mathematics with a Data-Driven Approach: What Data Do We Use?	(Fissore et al., 2025)	0
123	The Art of Socratic Questioning: Recursive Thinking with Large Language Models	(Qi et al., 2023)	1
124	The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers	(Jia et al., 2024)	2
125	The Content Analysis of the Lesson Plans Created by ChatGPT and Google Gemini	(Baytak, 2024)	0
126	The impact of ChatGPT-based learning statistics on undergraduates' statistical reasoning and attitudes toward statistics	(Wahba et al., 2024)	0
127	The Students' Mathematics Self-Regulated Learning and Mathematics Anxiety Based on The Use of ChatGPT, Music, Study Program, and Academic Achievement	(Delima et al., 2024)	0
128	TORA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving	(Gou et al., 2024)	1
129	Transforming Generative Large Language Models' Limitations into Strengths using Gestalt: A Synergetic Approach to Mathematical Problem-Solving with Computational Engines	(Dunn & Hashemi Tonekaboni, 2024)	1
130	Updating Calculus Teaching with AI: A Classroom Experience	(Torres-Peña et al., 2024)	0
131	Use of Language By generative AI Tools in Mathematical Problem Solving: The Case of ChatGPT	(Daher & Gierdien, 2024)	0
132	Using ChatGPT for the Development of Critical Thinking in Youth: Example of Inequality Proof	(Drushlyak et al., 2024)	0
133	Using Large Language Models for Math Information Retrieval	(Mansouri & Maarefdoust, 2024)	3
134	Using Large Language Models to Automate Annotation and Part-of-Math Tagging of Math Equations	(Shan & Youssef, 2024)	2
135	Using Large Language Models to Diagnose Math Problem-solving Skills at Scale	(Jin et al., 2024)	2
136	What Make Math Word Problems Challenging for LLMs?	(Srivatsa & Kochmar, 2024)	3
137	Who's the Best Detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers	(Urrutia & Araya, 2024)	2

Appendix B

The final model was selected after a 100-run random search in which we varied the three main UMAP hyper-parameters inside the ranges. For every run we recorded topic-coherence, number of topics, and outlier rate; the configuration shown in Table 14 gave the best trade-off between coherence and cluster interpretability.

Table 14

Detail of Configuration 4

Parameter	Value	Functions
n_neighbors	6	Controls the size of the local neighborhood considered when building the manifold in UMAP. Smaller values lead to finer-grained clusters, potentially increasing the number of clusters.
n_components	10	Determines the target dimensionality of the UMAP-reduced embedding passed to HDBSCAN and PCA. Higher dimensions retain more semantic information but slow down clustering and complicate visualization.
min_dist	0.0857	Sets the minimum allowed distance between two embedded points in UMAP. Controls cluster compactness: smaller values create tighter, denser clusters, while larger values result in more spread-out clusters.
Result		**Interpretation**
Coherence Score	0.4527	c-TF-IDF coherence of the final model. A score closer to 1 indicates better coherence (0 = poor, 1 = perfect).
Outlier Percentage	12.6437	Share of documents labelled as noise (topic = -1) by HDBSCAN before manual reassignment.

Author's contributions

Henry Pak Yeung Tsang conceived the review, conducted the literature search, synthesized the findings, and wrote the first draft of the manuscript.

Dichen Wang contributed to the Discussion and assisted with manuscript review and editing.

Philip Leung Ho Yu provided supervision, intellectual guidance, and critical revision of the manuscript.

Author's information

HT is a Ph.D student in Education.
DW is a Lecturer in elementary mathematics education and special education.
PY is a Professor in statistics, data science, and artificial intelligence.

Funding

This work was partially supported by Hong Kong Private Funds "Research and Development for Multimodal Artificial Intelligence in Education" (C1205) and "Research and Development of Artificial Intelligence in Educational and Financial Technologies" (CB310).

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

All authors are affiliated with the Education University of Hong Kong, Hong Kong

Received: 18 March 2025 Accepted: 2 April 2026
Published online: 1 January 2027 (Online First: 27 July 2026)

References

An, J., Lee, J., & Gweon, G. (2023). Does ChatGPT comprehend place value in numbers when solving math word problems? In D. R. Thomas, J. Lin, & K. R. Koedinger (Eds.), *Proceedings of the workshop Towards the future of AI-augmented human tutoring in math learning, co-located with the 24th International Conference on Artificial Intelligence in Education (AIED 2023)* (Vol. 3491, pp. 49–58). CEUR-WS. https://ceur-ws.org/Vol-3491/paper5.pdf

Aqazade, M., Mauntel, M., & Atabas, S. (2025). Empowering mathematics teacher educators: Exploring artificial intelligence-driven mathematical tasks. *School Science and Mathematics*, *126*(1), 24–43. https://doi.org/10.1111/ssm.18339

Bainbridge, K., Walkington, C., Ibrahim, A., Zhong, I., Basu Mallick, D., Washington, J., & Baraniuk, R. (2023). A case study using large language models to generate metadata for math questions. In S. Moore, J. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, & C. Brooks (Eds.), *Proceedings of the workshop on Empowering Education with LLMs – the Next-Gen Interface and Content Generation 2023 co-located with the 24th International Conference on Artificial Intelligence in Education (AIED 2023)* (Vol. 3487, pp. 34–42). CEUR-WS. https://ceur-ws.org/Vol-3487/short5.pdf

Barana, A., Marchisio, M., & Roman, F. (2023). Fostering problem solving and critical thinking in mathematics through generative artificial intelligence. In D. G. Sampson, D. Ifenthaler, & P. Isaías (Eds.), *Proceedings of the 20th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2023)* (pp. 377–385). IADIS Press. https://doi.org/10.33965/CELDA2023_202306L046

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Baytak, A. (2024). The content analysis of the lesson plans created by ChatGPT and Google Gemini. *Research in Social Sciences and Technology, 9*(1), 329–350. https://doi.org/10.46303/ressat.2024.19

Beege, M., Hug, C., & Nerb, J. (2024). AI in STEM education: The relationship between teacher perceptions and ChatGPT use. *Computers in Human Behavior Reports, 16*, 100494. https://doi.org/10.1016/j.chbr.2024.100494

Bešlić, A., Bešlić, J., & Kamber Hamzić, D. (2024). Artificial intelligence in elementary math education: Analyzing impact on students' achievements. In T. Volarić, B. Crnokić, & D. Vasić (Eds.), *Digital transformation in education and artificial intelligence application: Second international conference, MoStart 2024, Mostar, Bosnia and Herzegovina, April 24–26, 2024, proceedings* (Communications in Computer and Information Science, Vol. 2124, pp. 27–40). Springer. https://doi.org/10.1007/978-3-031-62058-4_3

Boonen, A. J. H., de Koning, B. B., Jolles, J., & Van der Schoot, M. (2016). Word problem solving in contemporary math education: A plea for reading comprehension skills training. *Frontiers in Psychology, 7*, Article 191. https://doi.org/10.3389/fpsyg.2016.00191

Botana, F., Recio, T., & Vélez, M. P. (2024). On using GeoGebra and ChatGPT for geometric discovery. *Computers, 13*(8), 187. https://doi.org/10.3390/computers13080187

Butgereit, L. L., Martinus, H., & Abugosseisa, M. M. (2023). Prof Pi: Tutoring mathematics in Arabic language using GPT-4 and WhatsApp. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)* (pp. 000161–000164). IEEE. https://doi.org/10.1109/INES59282.2023.10297824

Canonigo, A. M. (2024). Levering AI to enhance students' conceptual understanding and confidence in mathematics. *Journal of Computer Assisted Learning, 40*(6), 3215-3229. https://doi.org/10.1111/jcal.13065

Chen, Z., Zhou, K., Zhang, B., Gong, Z., Zhao, W. X., & Wen, J.-R. (2023). ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 14777–14790). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.985

Cheng, V., & Zhang, Y. (2023). Analyzing ChatGPT's mathematical deficiencies: Insights and contributions. In J.-L. Wu & M.-H. Su (Eds.), *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 188–193). The Association for Computational Linguistics and Chinese Language Processing. https://aclanthology.org/2023.rocling-1.22/

Chiang, C.-H., & Lee, H. (2024). Over-reasoning and redundant calculation of large language models. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 161–169). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-short.15/

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv*. https://doi.org/10.48550/arXiv.2110.14168

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences, 121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Dahal, N., Lamichhane, B. R., Luitel, B. C., & Pant, B. P. (2023). AI chatbots as math algorithm problem solvers: A critical evaluation of its capabilities and limitations. In W. C. Yang, M. Majewski, D. Meade, & W. K. Ho (Eds.), *Proceedings of the 28th Asian Technology Conference in Mathematics* (pp. 429–438). Mathematics and Technology, LLC. https://atcm.mathandtech.org/EP2023/regular/30102.pdf

Daher, W., & Gierdien, F. (2024). Use of language by generative AI tools in mathematical problem solving: The case of ChatGPT. *African Journal of Research in Mathematics, Science and Technology Education, 28*(2), 222–235. https://doi.org/10.1080/18117295.2024.2384676

Das, D., Banerjee, D., Aditya, S., & Kulkarni, A. (2024). MATHSENSEI: A tool-augmented large language model for mathematical reasoning. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 942–966). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.54

Dasari, D., Hendriyanto, A., Sahara, S., Suryadi, D., Muhaimin, L. H., Chao, T., & Fitriana, L. (2024). ChatGPT in didactical tetrahedron, does it make an exception? A case study in mathematics teaching and learning. *Frontiers in Education, 8*, 1295413. https://doi.org/10.3389/feduc.2023.1295413

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

Drushlyak, M., Lukashova, T., Sabadosh, Y., Melnikov, I., & Semenikhina, O. (2024). Using ChatGPT for the development of critical thinking in youth: Example of inequality proof. In S. Babic (Ed.), *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 334–339). IEEE. https://doi.org/10.1109/MIPRO60963.2024.10569759

Drushlyak, M., Lukashova, T., Shamonia, V., & Semenikhina, O. (2025). ChatGPT-based simulation helps to develop the pre-service mathematics teachers' critical thinking. *International Journal of Instruction, 18*(1), 153–172. https://doi.org/10.29333/iji.2025.1819a

Duan, Z., & Wang, J. (2024). Multi-tool integration application for math reasoning using large language model. *arXiv*. https://doi.org/10.48550/arXiv.2408.12148

Dunn, C. W., & Hashemi Tonekaboni, N. (2024). Transforming generative large language models' limitations into strengths using Gestalt: A synergetic approach to mathematical problem-solving with computational engines. In T. X. Bui (Ed.), *Proceedings of the 57th Hawaii International Conference on System Sciences* (pp. 5185–5193). University of Hawaii at Manoa. https://hdl.handle.net/10125/107007

Egara, F. O., & Mosimege, M. (2024). Exploring the integration of artificial intelligence-based ChatGPT into mathematics instruction: Perceptions, challenges, and implications for educators. *Education Sciences, 14*(7), 742. https://doi.org/10.3390/educsci14070742

Elsayed, S. A. M. (2023). Applications of artificial intelligence and their relationship to spatial thinking and academic emotions towards mathematics: Perspectives from educational supervisors. *Eurasian Journal of Educational Research, 107*, 321–342.

Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Otero Ornelas, N., & Lan, A. (2024). Exploring automated distractor generation for math multiple-choice questions via large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3067–3082). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.193

Fitzpatrick, C. L., Hallett, D., Morrissey, K. R., Yıldız, N. R., Wynes, R., & Ayesu, F. (2020). The relation between academic abilities and performance in realistic word problems. *Learning and Individual Differences*, *83–84,* 101942. https://doi.org/10.1016/j.lindif.2020.101942

Fissore, C., Floris, F., Conte, M. M., & Sacchet, M. (2025). Teaching the specialized language of mathematics with a data-driven approach: What data do we use? In B. Steffen (Ed.), Bridging the Gap Between AI and Reality: *Proceedings of the AISoLA 2023 Conference* (pp. 45–60). Springer. https://doi.org/10.1007/978-3-031-73741-1_4

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems 36*. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html

Fu, Y., Weng, Z., & Wang, J. (2025). Examining AI use in educational contexts: A scoping meta-review and bibliometric analysis. *International Journal of Artificial Intelligence in Education, 35*(3), 1388–1444. https://doi.org/10.1007/s40593-024-00442-w

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2023). Complexity-based prompting for multi-step reasoning. In *International Conference on Learning Representations (ICLR).* OpenReview. https://openreview.net/forum?id=yf1icZHC-l9

Fung, S. C. E., Wong, M. F., & Tan, C. W. (2023). Chain-of-thoughts prompting with language models for accurate math problem-solving. In *2023 IEEE MIT Undergraduate Research Technology Conference (URTC).* IEEE. https://doi.org/10.1109/URTC60662.2023.10534945

Galiç, S., Urhan, S., Dost, Ş., & Lavicza, Z. (2025). Examining mathematics teachers noticing the rationality: Scenario-based training with AI chatbot. *Science & Education, 34*, 2759–2790. https://doi.org/10.1007/s11191-025-00618-3

Gattupalli, S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., & Woolf, B. (2023). Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning. In S. Moore, R. Tong, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, C. Brooks, J. Stamper, & C. Cao (Eds.), *Empowering education with LLMs – the Next-Gen interface and content generation 2023* (Vol. 3487, pp. 151–162). CEUR-WS. https://ceur-ws.org/Vol-3487/paper10.pdf

Gaur, V., & Saunshi, N. (2022). Symbolic math reasoning with language models. In *2022 IEEE MIT Undergraduate Research Technology Conference (URTC)* (pp. 1–5). IEEE. https://doi.org/10.1109/URTC56832.2022.10002218

Gaur, V., & Saunshi, N. (2023). Reasoning in large language models through symbolic math word problems. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 5889–5903). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.364

Getenet, S. (2024). Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education, 19*(1), em0766. https://doi.org/10.29333/iejme/14141

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Huang, M., Duan, N., & Chen, W. (2024). ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview. https://openreview.net/forum?id=Ep0TtjVoap

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 8154–8173). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.507

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., & Abu Khurma, O. (2024). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research and Practice in Technology Enhanced Learning, 19*, 023. https://doi.org/10.58459/rptel.2024.19023

He, X., He, J., Gao, H., & Sun, C. (2023). Evaluation of large scale language models on solving math word problems with difficulty grading. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)* (pp. 287–291). IEEE. https://doi.org/10.1109/IEIR59294.2023.10391224

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*. OpenReview. https://openreview.net/forum?id=d7KBjmI3GmQ

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. In J. Vanschoren & S.-K. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1* (pp. 1–9). Curran Associates. https://doi.org/10.48550/arXiv.2103.03874

Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024). Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design. *IEEE Transactions on Learning Technologies, 17*, 1445–1459. https://doi.org/10.1109/TLT.2024.3384765

Hwang, Y.-C., Lim, J., Lee, Y.-J., & Choi, H.-J. (2024). Enhancing numerical reasoning performance by augmenting distractor numerical values. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 242–245). IEEE. https://doi.org/10.1109/BigComp60711.2024.00045

Imani, S., Du, L., & Shrivastava, H. (2023). MathPrompter: Mathematical reasoning using large language models. In S. Sitaram, B. Beigman Klebanov, & J. D. Williams (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)* (pp. 37–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-industry.4

Jia, J., Wang, T., Zhang, Y., & Wang, G. (2024). The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. *Asia Pacific Journal of Education, 44*(1), 8–28. https://doi.org/10.1080/02188791.2023.2286920

Jiang, S., Shakeri, Z., Chan, A., Sanjabi, M., Firooz, H., Xia, Y., Akyildiz, B., Sun, Y., Li, J., Wang, Q., & Celikyilmaz, A. (2024). RESPROMPT: Residual connection prompting advances multi-step reasoning in large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5784–5809). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.323

Jie, Z., & Lu, W. (2023). Leveraging training data in few-shot prompting for numerical reasoning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10518–10526). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.668

Jin, H., Kim, Y., Park, Y. S., Tilekbay, B., Son, J., & Kim, J. (2024). Using large language models to diagnose math problem-solving skills at scale. In D. Joyner, M.-K. Kim, X. Wang, & M. Xia (Eds.), *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 471–475). Association for Computing Machinery. https://doi.org/10.1145/3657604.3664697

Junpeng, P. (2026). Effect of intelligent tutoring system-delivered scaffolding on mathematics progression: A multivariate study by school type and gender. *Social Sciences & Humanities Open, 13*, 102277. https://doi.org/10.1016/j.ssaho.2025.102277

Karaman, M. R., & Göksu, İ. (2024). Are lesson plans created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education (IJTE), 7*(1), 107–127. https://doi.org/10.46328/ijte.607

Kim, D. G., & Moon, A. (2024). From helping hand to stumbling block: The ChatGPT paradox in competency experiment (CESifo Working Paper No. 11002). Munich Society for the Promotion of Economic Research - CESifo GmbH. https://www.ifo.de/en/cesifo/publications/2024/working-paper/helping-hand-stumbling-block-chatgpt-paradox-competency-experiment

Kim, J., Kim, Y., Baek, I., Bak, J., & Lee, J. (2023). It ain't over: A multi-aspect diverse math word problem dataset. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14984–15011). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.927

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, 126*(1), 9–23. https://doi.org/10.1111/ssm.18336

Koceski, S., Koceska, N., Lazarova, L. K., Miteva, M., & Zlatanovska, B. (2025). Exploring the performance of ChatGPT for numerical solution of ordinary differential equations. *Journal of Technology and Science Education, 15*(1), 18–34. https://doi.org/10.3926/jotse.2709

Korkmaz Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research, 9*(2), em0188. https://doi.org/10.29333/pr/14145

Latif, E., Liu, V., & Zhai, X. (2026). A systematic review of intelligent and robot tutoring systems: Evolution, pedagogical design, and AI-driven classification. *Smart Learning Environments, 13*(1), 1. https://doi.org/10.1186/s40561-025-00427-9

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. https://doi.org/10.1145/3706598.3713778

Lei, B., Zhang, Y., Zuo, S., Payani, A., & Ding, C. (2024). MACM: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. In *Advances in Neural Information Processing Systems 37*. Curran Associates. https://doi.org/10.52202/079017-1692

Li, C., Zhu, W., Xing, W., & Guo, R. (2024). Analyzing student attention and acceptance of conversational AI for math learning: Insights from a randomized controlled trial. In S. Joksimovic & A. Zamecnik (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM. https://doi.org/10.1145/3636555.3636895

Li, P.-H., Lee, H.-Y., Cheng, Y.-P., Starčič, A. I., & Huang, Y.-M. (2023). Solving the self-regulated learning problem: Exploring the performance of ChatGPT in mathematics. In Y.-M. Huang & T. Rocha (Eds*.), Innovative Technologies and Learning: Proceedings of the 6th International Conference (ICITL 2023)* (pp. 80–90). Springer. https://doi.org/10.1007/978-3-031-40113-8_8

Li, X. L., Shrivastava, V., Li, S., Hashimoto, T., & Liang, P. (2024). Benchmarking and improving generator-validator consistency of LMs. In *Proceedings of the International Conference on Learning Representations (ICLR 2024).* OpenReview. https://doi.org/10.48550/arXiv.2310.01846

Li, Y., & Schoenfeld, A. H. (2019). Problematizing teaching and learning mathematics as "given" in STEM education. *International Journal of STEM Education, 6*, Article 44. https://doi.org/10.1186/s40594-019-0197-9

Liang, Z., Yu, W., Rajpurohit, T., Clark, P., Zhang, X., & Kalyan, A. (2023). Let GPT be a math tutor: Teaching math word problem solvers with customized exercise generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14384–14396). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.889

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. In *Proceedings of the International Conference on Learning Representations (ICLR 2024).* OpenReview. https://openreview.net/forum?id=v8L0pN6EOi

Lin, Q., Xu, B., & Huang, Z. (2024). From large to tiny: Distilling and refining mathematical expertise for math word problems with weakly supervision. In D.-S. Huang, Z. Si, & W. Chen (Eds.), *Advanced intelligent computing technology and applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part VI* (Lecture Notes in Computer Science, Vol. 14880). Springer. https://doi.org/10.1007/978-981-97-5678-0_22

Lin, T., & Riccomini, P. (2025). Inclusive education in rural settings: Leveraging technology for improved math learning outcomes among students with disabilities. *Journal of Special Education Technology, 40*(2), 219–226. https://doi.org/10.1177/01626434241263045

Liu, T., Guo, Q., Yang, Y., Hu, X., Zhang, Y., Qiu, X., & Zhang, Z. (2023). Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2807–2822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.169

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Lu, J., Zheng, R., Gong, Z., & Xu, H. (2024). Supporting teachers' professional development with generative AI: The effects on higher order thinking and self-efficacy. *IEEE Transactions on Learning Technologies, 17*, 1279–1289. https://doi.org/10.1109/TLT.2024.3369690

Luzano, J. F. P. (2024). Reshaping mathematics instruction via impact of AI chatbots on secondary education pre-service teachers. *International Journal of Studies in Education and Science (IJSES), 5*(3), 233–245. https://doi.org/10.46328/ijses.97

Mansouri, B., & Maarefdoust, R. (2024). Using large language models for math information retrieval. In G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, & Y. Zhang (Eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2693–2697). Association for Computing Machinery. https://doi.org/10.1145/3626772.3657907

Martínez-Téllez, R., & Camacho-Zuñiga, C. (2023). Enhancing mathematics education through AI chatbots in a flipped learning environment. In *2023 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)* (pp. 1–6). IEEE. https://doi.org/10.1109/WEEF-GEDC59520.2023.10343838

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

McNichols, H., Zhang, M., & Lan, A. (2023). Algebra error classification with large language models. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, proceedings* (Lecture Notes in Computer Science, Vol. 13916, pp. 365–376). Springer. https://doi.org/10.1007/978-3-031-36272-9_30

Miao, N., Teh, Y. W., & Rainforth, T. (2024). SelfCheck: Using LLM to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations (ICLR 2024).* OpenReview. https://openreview.net/forum?id=pTHfApDakA

Moreau-Pernet, B., Tian, Y., Sawaya, S., Foltz, P., Cao, J., Milne, B., & Christie, T. (2024). Classifying tutor discursive moves at scale in mathematics classrooms with large language models. In D. Joyner, M.-K. Kim, X. Wang, & M. Xia (Eds.), *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 361–365). ACM. https://doi.org/10.1145/3657604.3664664

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2025). Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education, 35*(2), 559–586. https://doi.org/10.1007/s40593-024-00418-w

Narin, A. E. (2024). Math problem solving: Enhancing large language models with semantically rich symbolic variables. In M. Valentino, D. Ferreira, M. Thayaparan, & A. Freitas (Eds.), *Proceedings of the 2nd Workshop on Mathematical Natural Language Processing @ LREC-COLING 2024* (pp. 19–24). ELRA Language Resources Association.

Norberg, K. A., Almoubayyed, H., De Ley, L., Murphy, A., Weldon, K., & Ritter, S. (2025). Rewriting content with GPT-4 to support emerging readers in adaptive mathematics software. *International Journal of Artificial Intelligence in Education, 35*(2), 587–626. https://doi.org/10.1007/s40593-024-00420-2

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review, 4*(4), 251–277. https://doi.org/10.1007/BF00168958

Oh, S. (2025). Evaluating mathematical problem-solving abilities of generative AI models: Performance analysis of o1-preview and gpt-4o using the Korean College Scholastic Ability Test. *IEEE Access, 13*, 1227–1235. https://doi.org/10.1109/ACCESS.2024.3523703

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pal Chowdhury, S., Zouhar, V., & Sachan, M. (2024). AutoTutor meets large language models: A language model tutor with rich pedagogy and guardrails. In D. Joyner, M. K. Kim, X. Wang, & M. Xia (Eds.), *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 5–15). ACM. https://doi.org/10.1145/3657604.3662041

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

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

Patel, A., Bhattamishra, S., & Goyal, N. (2021). Are NLP models really able to solve simple math word problems? In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2080–2094). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.168

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

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

Pham, P. V. L., Vu, D. A., Hoang, N. M., Do, X. L., & Luu, A. T. (2024). ChatGPT as a math questioner? Evaluating ChatGPT on generating pre-university math questions. In J. Hong & J. W. Park (Eds.), *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC 2024)* (pp. 65–73). ACM. https://doi.org/10.1145/3605098.3636030

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI, 4*(4), 949–969. https://doi.org/10.3390/ai4040048

Qi, J., Xu, Z., Shen, Y., Liu, M., Jin, D., Wang, Q., & Huang, L. (2023). The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4177–4199). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.255

Qian, C., Han, C., Fung, Y. R., Qin, Y., Liu, Z., & Ji, H. (2023). CREATOR: Tool creation for disentangling abstract and concrete reasoning of large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6922–6939). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.462

Ramprakash, B., Surya Devi, B., Nithyakala, G., Bhumika, K., & Avanthika, S. (2024). Comparing traditional instructional methods to ChatGPT: A comprehensive analysis. *Journal of Engineering Education Transformations, 37*, 612–620. https://doi.org/10.16920/jeet/2024/v37is2/24095

Rigaud Téllez, N., Rayón Villela, P., & Blanco Bautista, R. (2024). Evaluating ChatGPT-generated linear algebra formative assessments. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(5), 75–82. https://doi.org/10.9781/ijimai.2024.02.004

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. https://doi.org/10.30935/cedtech/15487

Santandreu Calonge, D., Smail, L., & Kamalov, F. (2023). Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. *Journal of Applied Learning & Teaching, 6*(2), 1–10. https://doi.org/10.37074/jalt.2023.6.2.22

Sapkota, B., & Bondurant, L. (2024). Assessing concepts, procedures, and cognitive demand of ChatGPT-generated mathematical tasks. *International Journal of Technology in Education (IJTE), 7*(2), 218–238. https://doi.org/10.46328/ijte.677

Satpute, A., Gießing, N., Greiner-Petter, A., Schubotz, M., Teschke, O., Aizawa, A., & Gipp, B. (2024). Can LLM master math? Investigating large language models on Math Stack Exchange. In G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, & Y. Zhang (Eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2316–2320). ACM. https://doi.org/10.1145/3626772.3657945

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education, 9*, 1386075. https://doi.org/10.3389/feduc.2024.1386075

Selter, C. (2000). [Review of the book Making sense of word problems, by L. Verschaffel, B. Greer, & E. de Corte]. *Educational Studies in Mathematics, 42*(2), 211–213. https://doi.org/10.1023/A:1004190927303

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An independent evaluation of ChatGPT on mathematical word problems (MWP). In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, & F. van Harmelen (Eds.), *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023).* CEUR-WS. https://ceur-ws.org/Vol-3433/paper8.pdf

Shan, R., & Youssef, A. (2024). Using large language models to automate annotation and part-of-math tagging of math equations. In A. Kohlhase & L. Kovács (Eds.), *Intelligent Computer Mathematics: 17th International Conference, CICM 2024, Montréal, QC, Canada, August 5–9, 2024, Proceedings* (pp. 3–20). Springer. https://doi.org/10.1007/978-3-031-66997-2_1

Shridhar, K., Macina, J., El-Assady, M., Sinha, T., Kapur, M., & Sachan, M. (2022). Automatic generation of Socratic subquestions for teaching math word problems. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4136–4149). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.277

Şimşek, N. (2025). Integration of ChatGPT in mathematical story-focused 5E lesson planning: Teachers and pre-service teachers' interactions with ChatGPT. *Education and Information Technologies, 30*(8), 11391–11462. https://doi.org/10.1007/s10639-024-13258-x

Song, Y., Kim, J., Liu, Z., Li, C., & Xing, W. (2025). Students' perceived roles, opportunities, and challenges of a generative AI-powered teachable agent: A case of middle school math class. *Journal of Research on Technology in Education,* 1–19. https://doi.org/10.1080/15391523.2024.2447727

Song, Y., Kim, J., Xing, W., Liu, Z., Li, C., & Oh, H. (2025). Elementary school students' and teachers' perceptions toward creative mathematical writing with generative AI. *Journal of Research on Technology in Education*, 1–23. https://doi.org/10.1080/15391523.2025.2455057

Spreitzer, C., Straser, O., Zehetmeier, S., & Maaß, K. (2024). Mathematical modelling abilities of artificial intelligence tools: The case of ChatGPT. *Education Sciences, 14*(7), 698. https://doi.org/10.3390/educsci14070698

Srivatsa, K. A., & Kochmar, E. (2024). What makes math word problems challenging for LLM? In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1138–1148). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.72

Sun, Y., Yin, Z., Guo, Q., Wu, J., Qiu, X., & Zhao, H. (2024). Benchmarking hallucination in large language models based on unanswerable math word problems. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2178–2188). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.196/

Taani, O., & Alabidi, S. (2024). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. *International Journal of Mathematical Education in Science and Technology, 56*(6), 1–30. https://doi.org/10.1080/0020739X.2024.2357341

Tapan Broutın, M. S. (2024). Exploring mathematics teacher candidates' instrumentation process of generative artificial intelligence for developing lesson plans. Yükseköğretim Dergisi / TÜBA *Higher Education Research/Review (TÜBA-HER), 14*(1), 165–176. https://doi.org/10.53478/yuksekogretim.1347061

Torres-Peña, R. C., Peña-González, D., Chacuto-López, E., Ariza, E. A., & Vergara, D. (2024). Updating calculus teaching with AI: A classroom experience. *Education Sciences, 14*(9), 1019. https://doi.org/10.3390/educsci14091019

Toshniwal, S., Du, W., Moshkov, I., Kisacanin, B., Ayrapetyan, A., & Gitman, I. (2024). OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*. OpenReview. https://openreview.net/forum?id=l5FDMofecw

Udias, A., Alonso-Ayuso, A., Alfaro, C., Algar, M. J., Cuesta, M., Fernández-Isabel, A., Gómez, J., Lancho, C., Cano, E. L., Martín de Diego, I., & Ortega, F. (2024). ChatGPT's performance in university admissions tests in mathematics. *International Electronic Journal of Mathematics Education, 19*(4), em0795. https://doi.org/10.29333/iejme/15517

Upadhyay, S., Ginsberg, E., & Callison-Burch, C. (2023). Improving mathematics tutoring with a code scratchpad. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 20–28). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.2

Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive Learning Environments, 32*(10), 7186–7211. https://doi.org/10.1080/10494820.2024.2308093

Urrutia, F., & Araya, R. (2024). Who's the best detective? Large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *Journal of Educational Computing Research, 61*, 1723–1754. https://doi.org/10.1177/07356331231191174

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making Sense of Word Problems*. Swets & Zeitlinger.

Verschaffel, L., Schukajlow, S., Star, J., & van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM Mathematics Education, 52*, 1–16. https://doi.org/10.1007/s11858-020-01130-4

Vintere, A., Safiulina, E., & Panova, O. (2024). AI-based mathematics learning platforms in undergraduate engineering studies: Analyses of user experiences. In *Engineering for Rural Development* (Vol. 23, pp. 1042–1047). https://doi.org/10.22616/ERDev.2024.23.TF216

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

Wang, A., Prihar, E., Haim, A., & Heffernan, N. (2024). Can large language models generate middle school mathematics explanations better than human teachers? In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part II* (Communications in Computer and Information Science, Vol. 2151, pp. 242–250). Springer. https://doi.org/10.1007/978-3-031-64312-5_29

Wang, B., Yue, X., & Sun, H. (2023). Can ChatGPT defend its belief in truth? Evaluating LLM reasoning via debate. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 11865–11881). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.795

Wang, K., Ren, H., Zhou, A., Lu, Z., Luo, S., Zhang, R., & Li, H. (2024). MathCoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. OpenReview. https://openreview.net/forum?id=z8TW0ttBPp

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2609–2634). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.147

Wang, R. E., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024). Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2174–2199). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.120

Wang, R., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023*) (pp. 626–667). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.53

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*, 124167. https://doi.org/10.1016/j.eswa.2024.124167

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems 35* (pp. 24824–24837). Curran Associates. https://doi.org/10.52202/068431-1800

Wu, H., & Yu, X. (2023). Prompt incorporates math knowledge to improve efficiency and quality of large language models to translate math word problems. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)* (pp. 1–5). IEEE. https://doi.org/10.1109/IEIR59294.2023.10391245

Wu, H., Hui, W., Chen, Y., Wu, W., Tu, K., & Zhou, Y. (2023). Conic10k: A challenging math problem understanding and reasoning dataset. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6444–6458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.427

Wu, T., Lee, Y., Chen, H., Lin, J., & Huang, M. (2025). Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement. *Journal of Computer Assisted Learning, 41*(1), e13085. https://doi.org/10.1111/jcal.13085

Wu, Z., Jiang, M., & Shen, C. (2024). Get an A in math: Progressive rectification prompting. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 19288–19296). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v38i17.29898

Wu, Z., Shen, C., & Jiang, M. (2024). Instructing large language models to identify and ignore irrelevant conditions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6799–6819). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.379

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., & Zhang, Z. (2024). Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv*. https://doi.org/10.48550/arXiv.2409.12122

Yang, X., Chen, B., & Tam, Y.-C. (2024). Arithmetic reasoning with LLM: Prolog generation & permutation. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)* (pp. 699–710). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-short.61

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*(7), 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

You, W., Yin, S., Zhao, X., Ji, Z., Zhong, G., & Bai, J. (2024). MuMath: Multi-perspective data augmentation for mathematical reasoning in large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2932–2958). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.185

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J., Li, Z., Weller, A., & Liu, W. (2024). MetaMath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview. https://openreview.net/forum?id=N8N0hgNDRt

Yu, X., Peng, B., Galley, M., Gao, J., & Yu, Z. (2024). Teaching language models to self-improve through interactive demonstrations. In K. Duh, H. Gomez, & S. Bethard (Eds*.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5127–5149). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.287

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., & Chen, W. (2024). MAmmoTH: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations (ICLR 2024).* OpenReview. https://openreview.net/forum?id=yLClGs770I

Zhang, B., Zhou, K., Wei, X., Zhao, W. X., Sha, J., Wang, S., & Wen, J.-R. (2023). Evaluating and improving tool-augmented computation-intensive math reasoning. In E. Denton, J.-W. Ha, & J. Vanschoren (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2023*. Curran Associates. https://openreview.net/forum?id=OB10WTlwmX

Zhang, F., Li, C., Henkel, O., Xing, W., Baral, S., Heffernan, N., & Li, H. (2025). Math-LLMs: AI cyberinfrastructure with pre-trained transformers for math education. *International Journal of Artificial Intelligence in Education, 35*(6), 533–558. https://doi.org/10.1007/s40593-024-00416-y

Zhang, J., & Moshfeghi, Y. (2024). GOLD: Geometry problem solver with natural language description. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 263–278). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.19

Zhang, W., Shen, Y., Hou, G., Wang, K., & Lu, W. (2024). Specialized mathematical solving by a step-by-step expression chain generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 3128–3140. https://doi.org/10.1109/TASLP.2024.3410028

Zheng, Y., Li, X., Huang, Y., Liang, Q., Guo, T., Hou, M., Gao, B., Tian, M., Liu, Z., & Luo, W. (2024). Automatic lesson plan generation via large language models with self-critique prompting. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I* (Communications in Computer and Information Science, Vol. 2150, pp. 163–178). Springer. https://doi.org/10.1007/978-3-031-64315-6_13

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., & Duan, N. (2024). AGIEval: A human-centric benchmark for evaluating foundation models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2299–2314). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.149

Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2024). Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview. https://openreview.net/forum?id=c8McWs4Av0

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2023). Least-to-most prompting enables complex reasoning in large language models. *In Proceedings of the International Conference on Learning Representations (ICLR 2023)*. OpenReview. https://openreview.net/forum?id=WZH7099tgfM

Zhou, J. P., Staats, C., Li, W., Szegedy, C., Weinberger, K. Q., & Wu, Y. (2024). Don't trust: Verify – Grounding LLM quantitative reasoning with autoformalization. In *The Twelfth International Conference on Learning Representations (ICLR 2024).* OpenReview. https://openreview.net/forum?id=V5tdi14ple

Zhou, Z., Wang, Q., Jin, M., Yao, J., Ye, J., Liu, W., Wang, W., Huang, X., & Huang, K. (2024). MathAttack: Attacking large language models towards math solving ability. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 19750–19758). AAAI. https://doi.org/10.1609/aaai.v38i17.29949

Publisher's Note