Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.16, No.5, 2025

Large Language Models in Tertiary Mathematics Education: A Systematic Literature Review

Jennifer Dela Torre[1]* Jero Sayco[2]
1. General Academic Requirement Division, Higher Colleges of Technology – Fujairah Campuses, UAE
2. Faculty of Computer Information System, Higher Colleges of Technology – Abu Dhabi Campuses, UAE
*E-mail of the corresponding author: jdelatorre@hct.ac.ae

Abstract
Large language models (LLMs) have quickly become a focal point, sparking both excitement and questions within higher education, particularly concerning mathematics instruction. Our systematic literature review (SLR) explored peer-reviewed research published from 2020 through 2025 to understand how LLMs, including tools like GPT, are being used in tertiary mathematics education. The findings reveal a range of applications: serving as digital tutors, providing learner support, automating assessments, assisting with content creation, and aiding curriculum planning. These models show significant potential to enhance teaching and learning. Looking at how they function, LLMs can deliver detailed step-by-step explanations, create practice problems and materials, and offer personalized support to students. They are also valuable for instructors, assisting with tasks like feedback and grading. Studies point towards effective LLM use potentially leading to better student engagement, motivation, and problem-solving skills. Furthermore, educators are starting to adopt these tools, finding them helpful for streamlining. However, challenges persist. LLMs may produce errors, foster student over-reliance, or raise academic integrity issues. Ethical concerns, such as bias and responsible use, underscore the need for clear institutional policies and thoughtful integration. This review identifies key trends and gaps, including the lack of longitudinal classroom research and professional development. With proper oversight, LLMs offer significant potential to support personalized, innovative mathematics education without replacing the critical role of human educators.
Keywords: large language models, mathematics education, tertiary mathematics education, systematic literature review, LLMs, ChatGPT, higher education
DOI: 10.7176/JEP/16-5-10
Publication date: May 30th 2025

1. Introduction

The landscape of educational practice is rapidly changing due to artificial intelligence, with large language models (LLMs) standing out as a particularly disruptive development. LLMs are built on deep neural networks trained on massive text corpora, enabling them to understand and generate human-like language (Kasneci et al., 2023). Newer models, like OpenAI's GPT-3 and GPT-4, have showcased remarkable capabilities, producing coherent text and tackling problems in various fields, including complex question answering and displaying reasoning processes (Frieder et al., 2023). These advanced abilities quickly captured significant attention in the education sector, a sector that was, arguably, ill-prepared for the swift public emergence of AI such as ChatGPT towards the end of 2022 (Kasneci et al., 2023). Within mathematics education, a discipline that traditionally emphasizes sequential problem-solving and rigorous reasoning, LLMs present both compelling opportunities and serious challenges.

On one hand, LLM-driven tools offer the possibility of personalized tutoring at scale which is a long-standing goal in education. An LLM-based tutor can interact in natural language, provide hints or full solutions, adapt to a student's queries, and potentially function as a "virtual teaching assistant" available 24/7. Early findings suggest such tools can enhance student engagement and learning. For instance, studies have shown ChatGPT's ability to support mathematical explanations and increase motivation when integrated into the learning process (Zafrullah et al., 2023; Kumar et al., 2023).

Educators also benefit from generative AI by offloading routine tasks such as creating practice problems, generating example solutions, or drafting lesson plans, thereby allowing them to focus on instructional design and pedagogical decisions (Güler et al., 2024). In mathematics, where creating well-structured problems and detailed solutions is time-intensive, LLMs provide valuable support. Researchers have begun to explore how these models might aid in lesson planning and teaching strategy development with promising early outcomes (Hu et al., 2025).

On the other hand, concerns have emerged among educators and administrators about the accuracy and reliability

of LLM-generated mathematical content. Although fluent and persuasive, these models can produce incorrect or misleading solutions known as "hallucinations" which pose risks in mathematics education where precision is vital (Dao & Le, 2023). If students uncritically accept flawed AI-generated answers, their conceptual understanding may suffer. Another major concern involves academic integrity.

The growing accessibility of advanced AI writing systems has transformed the educational landscape, allowing students to create coursework and test answers with minimal personal effort. This development has sparked profound ethical concerns and driven educational institutions to modify their policies accordingly (Kasneci et al., 2023). Since ChatGPT became publicly available in 2022, schools and universities have witnessed a substantial rise in AI-created academic submissions, forcing instructors to reconsider their evaluation methods and approaches to maintaining academic integrity. Scholars and educators continue to debate whether employing these sophisticated AI writing assistants constitutes cheating or if they should be embraced as valuable tools in modern learning environments (Frieder et al., 2023).

2. Related Literature and Studies

In areas like tertiary mathematics education, which demand abstract reasoning and structured problem-solving, the arrival of LLMs has sparked intense scholarly interest. Over the last five years, particularly with models such as OpenAI's GPT-3 and GPT-4 becoming available, we've observed a substantial increase in research. Systematic literature reviews (SLRs) and empirical studies are becoming more frequent. These investigations are actively exploring precisely how these powerful tools are reshaping educational processes within this domain. Systematic reviews examining LLMs in education began surfacing prominently in 2023 and 2024 as researchers sought to consolidate the effects of tools like ChatGPT on pedagogy, curriculum, and student engagement. A comprehensive SLR by Dong et al. (2024) notably mapped the evolving terrain of LLM use across educational contexts and indicated their promise for enhancing teaching and learning processes. Nevertheless, the authors pointed out a significant dearth of research into domain-specific applications, such as tertiary mathematics, urging for more focused investigation here. In a similar vein, Albadarin et al. (2023), reviewing empirical studies on ChatGPT, underlined the shortage of robust research in higher education mathematics, despite the discipline's unique dependence on step-by-step reasoning and symbolic manipulation. Their findings powerfully emphasized the necessity of exploring how AI can actively aid, rather than potentially impairing mathematical understanding at advanced levels.

Cho et al. (2024) provided another relevant SLR, focusing on knowledge tracing and the role of LLMs in modeling student learning. Even though their analysis didn't focus solely on mathematics, insights offered into how LLMs can personalize instruction and adapt feedback are highly relevant. This aligns well with the crucial goal of individual mastery of concepts in tertiary mathematics education. Across these reviews, a consistent point emerges while LLMs' general applications in education are being explored, their specific implications for STEM fields like mathematics are still significantly underexplored in the current literature. Within mathematics education, emerging empirical studies have begun investigating the use of LLMs for instructional support. Hu et al. (2025) demonstrated the pedagogical potential of ChatGPT through a simulated teaching experiment, where the model generated teacher-student dialogues to refine high school mathematics lesson plans. Though situated at the secondary level, the methodological insights carry over to tertiary contexts where complex content and pedagogical strategies require thoughtful design. The AI-generated lessons were rated comparably to those designed by experienced educators, suggesting LLMs can be effective tools in the early stages of instructional planning.

In higher education specifically, Meissner et al. (2024) developed "ItemForge," a GPT-4-powered system for automatically generating assessment items in university mathematics courses. Their study highlighted the capacity of LLMs to produce curriculum-aligned problems using structured prompts rooted in Bloom's taxonomy. However, they also cautioned that AI-generated solutions sometimes lacked mathematical precision, underscoring the importance of human oversight. These findings reinforce a recurring theme in the literature: LLMs can enhance efficiency but must be deployed under careful scrutiny to ensure accuracy and pedagogical soundness.

Automated assessment is another area receiving growing attention. Henkel et al. (2025) evaluated GPT-4's ability to grade open-response mathematics problems on scale. Their experiment with over 53,000 responses revealed that chain-of-thought prompting significantly improved the model's grading accuracy on complex,

previously ungradable responses. This not only enhanced the platform's feedback quality but also improved the accuracy of student mastery predictions in adaptive learning systems. In large tertiary mathematics courses, providing truly individualized feedback is notoriously difficult. Findings suggest large language models (LLMs) could revolutionize assessment in this context, offering a potential solution. However, concerns about grading consistency and bias have surfaced too, highlighting a clear need for more work on how LLMs are calibrated and how their decisions are interpreted.

Several SLRs and empirical commentaries have addressed the ethical and practical challenges of LLM deployment in education. Yan et al. (2023) conducted a scoping review identifying key obstacles such as data privacy, model transparency, and the need for teacher training. The authors stressed that without institutional support and clear usage guidelines, the benefits of LLMs could be unevenly distributed or misused. This concern is echoed by Kasneci et al. (2023), who argued that AI integration must be accompanied by critical pedagogy and ethical literacy to ensure informed, equitable usage. They also highlighted risks such as over-reliance and the potential erosion of students' critical thinking skills if LLMs are used uncritically. The apprehension regarding these issues finds a parallel in the work of Kasneci et al. (2023). They strongly argue that bringing AI into education must go together with cultivating critical pedagogy and ethical literacy, which they see as essential for ensuring its informed and equitable application. Among the notable hazards they pointed to were students becoming overly dependent on these tools and the potential for their critical thinking abilities to diminish if LLMs are used without careful consideration.

Güler et al. (2024) investigated how mathematics instructors use ChatGPT in pedagogical planning. Teachers reported using the tool to generate examples, anticipate student misconceptions, and brainstorm alternative explanations. While responses were generally positive, teachers also expressed a need for more concrete training and best-practice models. This aligns with findings from Truong (2023), who emphasized that LLMs function best when teachers are actively involved in guiding student interaction with the tool. Professional development is thus a critical component in realizing the full potential of AI in mathematics education.

Another emergent theme in the literature concerns how LLMs might reshape mathematics curricula. Matzakos et al. (2023) proposed that the availability of AI could allow curricula to de-emphasize routine computation and instead foreground conceptual understanding and critical thinking. Their comparative analysis of LLMs and computer algebra systems suggested that while LLMs are not yet capable of fully replacing traditional tools, they offer unique affordances for scaffolding problem formulation and exploration. Furthermore, scholars like Pavlova (2024) and Frieder et al. (2023) called for the inclusion of AI literacy within mathematics education, noting that students must learn to question and critique AI-generated outputs to develop metacognitive skills.

Despite some positive strides in the field, notable research gaps persist. As one example, Kumar et al. (2023) found that a significant portion of the existing research remains limited to short-term studies conducted solely in experimental or pilot environments. This particular focus leaves considerable questions unanswered concerning long-term learning outcomes, actual student behavior with these tools, and practical integration strategies. Consequently, little is currently known about how LLM use genuinely impacts student learning habits over a full academic term or how AI might be seamlessly and equitably integrated into daily classroom routines. Moreover, most available studies are concentrated in well-resourced, English-speaking contexts. Research from diverse educational environments and multilingual populations is still scarce, as noted in Albadarin et al. (2023). Lastly, there is limited exploration of domain-specific fine-tuning – i.e., customizing LLMs on mathematical content or pedagogical dialogues which may hold the key to improved accuracy and contextual relevance.

In summary, the literature reveals that while LLMs are not yet a panacea, they represent a significant advancement in educational technology, with specific promise in tertiary mathematics instruction. Systematic reviews and empirical studies alike point to their utility in content creation, assessment, and student support, but also underscore the need for human oversight, ethical guidelines, and ongoing research. As these tools become more powerful and accessible, continued inquiry will be vital to ensure they are used to enrich rather than undermine mathematical learning.

3. Rationale and Objectives

Amid the opportunities and challenges, a growing body of research has begun to investigate how LLMs are being deployed and evaluated in mathematics education. This comprehensive analysis of published research

examines work spanning from 2020 through 2025 investigating how advanced AI language systems function within college-level mathematics teaching. Concentrating on university settings, our review aims to clarify how these technologies support both professors and students, their influence on classroom approaches and educational results, and the teaching-related and ethical difficulties they introduce to mathematics education.
This review aims to:

1. Identify how LLMs are being used to support mathematics instruction and student learning in tertiary education contexts.
2. Investigate how AI writing systems affect teaching methods, student performance, and testing procedures in university mathematics courses.
3. Address the obstacles and restrictions, technical shortcomings, teaching challenges, and moral considerations linked to using these AI tools, while pinpointing areas needing further study to guide upcoming research.

Through examining recent scholarly publications, this analysis seeks to uncover key patterns and insights valuable to math teachers, educational technology specialists, and researchers studying AI's place in mathematics teaching. The contribution of this work is to consolidate current knowledge about LLM-based applications in higher math education, highlight effective use cases (e.g. AI-assisted tutoring, automated content creation), and discuss best practices and cautionary lessons. Ultimately, understanding the state of the art will help shape informed strategies for incorporating LLMs into mathematics curricula and pedagogies in a way that maximizes benefits to learning while safeguarding educational values

4. Methodology

Our analysis followed structured research methods consistent with established standards for synthesizing evidence, incorporating the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach for systematic reviews (Page et al., 2021). We focused on scholarly articles published from January 2020 through early 2025 that specifically studied AI language systems in college-level math teaching. For thorough research coverage, we searched numerous academic collections including Google Scholar, ERIC, ProQuest, EBSCOhost, Scopus and IEEE Xplore. We chose these particular databases because they index a wide range of publications spanning education, computer science, and cross-disciplinary research fields (Snyder, 2019).

Keyword searches employed combinations such as “large language model,” “LLM,” “ChatGPT,” “mathematics education,” “higher education,” “teaching,” “learning,” “tutoring,” “assessment,” and “curriculum,” along with terms like “GPT-3,” “GPT-4,” and “generative AI in education” to capture model-specific research. The initial pool of results was screened for duplicates and then filtered using defined inclusion and exclusion criteria. Eligible studies met the following requirements:
(1) publication in a peer-reviewed journal or conference proceeding (or a preprint marked for peer review),
(2) relevance to LLM applications in mathematics instruction at the tertiary level, and
(3) English-language accessibility.
Studies focused solely on other subject areas or offering superficial mentions of AI were excluded, along with opinion pieces or media articles lacking empirical grounding.

Following title and abstract screening, full texts were reviewed for eligibility. A PRISMA flowchart was used to document the selection process, yielding a final sample of 35 studies. These encompassed empirical designs (e.g., experiments, case studies, surveys), as well as theoretical and scoping reviews. Key attributes were extracted from each study, including author(s), publication year, methodology, context or sample, LLM use case, and reported outcomes.

A

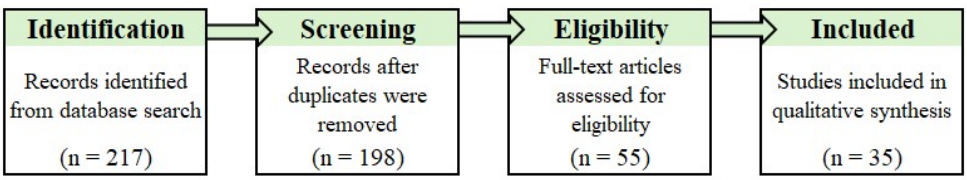

Figure 1 – Flowchart of the methodology based on PRISMA framework (Page et.al., 2021)

thematic analysis was conducted to identify dominant patterns and emerging insights across studies. Coding involved categorizing findings into clusters such as instructional use, assessment practices, learner impact, and ethical concerns. This process was iterative, involving comparative reviews and validation by a secondary reviewer to enhance reliability (Nowell et al., 2017). Disagreements in coding or inclusion decisions were resolved through discussion and consensus.

The results of the analysis are presented thematically in the next section, alongside a synthesized table (Table 1) highlighting representative studies. The discussion that follows interprets the findings, identifies contributions and gaps, and suggests directions for future research. All cited sources are included to ensure transparency and traceability.

5. Results and Discussion

Our literature search confirms that the period 2020-2025 witnessed a rapid emergence of research on LLMs in mathematics education, with most studies appearing in 2023 and 2024. Early in the decade (2020-2021), few if any studies addressed large language models in math education, reflecting the nascent state of the technology. The landscape began changing around 2022 with the introduction of advanced AI writing tools like GPT-3. This period saw researchers starting to publish initial theoretical ideas and conduct preliminary investigations into their potential. However, the late 2022 public launch of ChatGPT triggered a truly explosive surge in interest. This is evident in the research output: over 80% of the studies included in our review were published throughout 2023 alone, covering everything from controlled experiments to comprehensive analyses and various scholarly perspectives. This publication surge highlights the rapidly developing nature of this research area, where math educators and academics are simultaneously exploring exciting possibilities and serious concerns about using these powerful AI systems in mathematics education. Though this research history spans just a few years, we have identified several key patterns across multiple studies, which we present in our findings below.

Table 1 summarizes a subset of representative studies from the full set of 35 included in this review. These examples were selected to reflect a range of methodologies, contexts, and applications of LLMs in mathematics education that are illustrative of the broader findings.

The selection process for Table 1 aimed to highlight studies demonstrating the primary ways LLMs are being explored or used in tertiary mathematics education, aligning with the review's objectives to identify support roles, reported impacts, and challenges. Specifically, the representative studies were chosen to showcase diverse LLM applications, including their use as tutoring aids for step-by-step explanations (e.g., Kumar et al., 2023), as supplementary problem-solving tools in specific courses like linear algebra (e.g., Karjanto, 2023), as general learning tools influencing motivation (e.g., Zafrullah et al., 2023), in innovative pedagogical models like flipped learning (e.g., Pavlova, 2024), for evaluating their performance on mathematical tasks (e.g., Dao & Le, 2023), for conceptual comparison with existing tools (e.g., Matzakos et al., 2023), for the automated generation of assessment items (e.g., Meissner et al., 2024), and for the automated grading of open-response student work (e.g., Henkel et al., 2025). The selection also highlighted varied research designs, incorporating experimental designs, case studies (both classroom-based and qualitative), computational evaluations, theoretical or comparative analysis, and design/evaluation studies of new tools, a breadth reflecting the different approaches researchers are taking to investigate LLMs in this domain. Furthermore, the studies presented key themes and outcomes, with findings related to improved student performance, enhanced motivation and engagement, support for critical thinking, insights into LLM accuracy and limitations on specific math tasks, potential shifts in curriculum focus, and the effectiveness and challenges of automated assessment generation and grading. By presenting these examples, the review aims to provide concrete illustrations of the state of research and ground the subsequent thematic analysis in specific studies, offering a clearer picture of the findings synthesized from the full set of included articles.

Table 1. Representative Studies on LLM Applications in Higher Education Mathematics (2020–2025)

Study (Authors, Year)	Research Design	Sample/Context	LLM Application	Key Outcomes
Kumar et al. (2023)	Experimental, between-subjects	Online participants practicing high school math problems	GPT-3 based explanations provided after or during practice	LLM-generated step-by-step explanations significantly improved subsequent test performance, especially when students first attempted problems unassisted; even when LLM explanations contained errors, they yielded some learning gains over just seeing the final answer.
Karjanto (2023)	Classroom case study (mixed methods)	Undergraduate linear algebra course	ChatGPT used alongside CAS (SageMath) for problem solving	Use of ChatGPT as a supplementary problem-solving tool facilitated students' work on linear algebra tasks and improved their critical thinking and understanding of concepts (e.g. matrix factorization).
Zafrullah et al. (2023)	Descriptive study (survey)	Mathematics education undergraduates (Indonesia)	ChatGPT employed as a learning tool (open AI usage)	Students' learning interest and motivation in mathematics increased by 80.33% after using ChatGPT, indicating that the AI tool served as a positive stimulus for engagement.
Pavlova (2024)	Qualitative case study	High school mathematics & informatics classes (case in Ukraine)	ChatGPT integrated in flipped "dialogic" learning model	Flipped dialogic instruction with AI (students interacting with ChatGPT as part of lesson) improved ease of access to learning materials, reduced learner stress, and stimulated students' research skills and problem-solving abilities.
Dao & Le (2023)	Computational evaluation study	Vietnamese national exam questions (N=250, various subjects)	ChatGPT answering math multiple-choice questions	ChatGPT answered easy math questions correctly at a high rate (83% on lowest-difficulty items) but struggled with hard questions (10% accuracy at highest difficulty); it excelled in certain topics (exponential and logarithmic functions) but underperformed on others (e.g. derivatives, spatial geometry).
Matzakos et al. (2023)	Comparative analysis (theoretical)	Higher education mathematics teaching (literature-based)	LLMs vs. existing educational tools (conceptual comparison)	Large language models offer innovative possibilities for university mathematics instruction, with the potential to influence curriculum design and assessment strategies; however, realizing these benefits requires further research and careful implementation.
Meissner et al. (2024)	Design and evaluation study	Higher ed math courses (concepts from curriculum); 3 expert reviewers	GPT-4 powered item generator ("ItemForge") for math	The LLM-based system could automatically generate high-quality competency-aligned math questions (covering both formative

			assessments	and summative assessment) that matched targeted cognitive levels. Expert review found the content appropriate and challenging, though the AI's provided solutions were occasionally incomplete or inaccurate, indicating the need for human oversight in verification.
Henkel et al. (2025)	Experiment (model performance & learning analytics)	53,000 student answers from online math platform (middle/high school level content; focusing on hardest 1%)	GPT-3.5/4 with chain-of-thought prompting to grade open-response math answers	The best LLM approach (chain-of-thought) accurately graded 97% of previously ungradable "edge-case" answers, improving overall grading accuracy from 96% to 97%. This led to more reliable estimates of student mastery in the platform's learning model (reducing misclassification of mastery from 6.9% to 2.6%), suggesting LLMs can enable wider use of open-ended questions by providing robust automated grading.

RQ1. Ways LLMs support mathematics instructors and learners

1.1 LLMs as teaching assistants and content generators

LLMs are increasingly used to assist instructors in lesson planning, content generation, and instructional design. A notable example is Hu et al. (2025), who used an LLM to simulate teacher-student dialogue and refine math lesson plans. The AI-enhanced plans were rated by experts as comparable in quality to those made by experienced educators, demonstrating LLMs' potential to support pre-class preparation.

Beyond full simulations, ChatGPT has been employed as a brainstorming tool for generating examples, analogies, and alternative strategies to explain mathematical concepts. Güler et al. (2024) reported that math teachers used ChatGPT to enhance lesson delivery and prepare for student inquiries, reducing their planning workload.

LLMs have also been applied to assessment development. Meissner et al. (2024) introduced the ItemForge system, which uses GPT-4 to create competence-based questions aligned with university-level math curricula. While expert reviewers found the questions accurate and well-targeted, minor flaws in AI-generated solutions underscored the need for instructor oversight.

Overall, LLMs offer meaningful support in content creation, effectively freeing educators to concentrate on deeper pedagogical tasks. It is crucial to note, however, that human validation remains absolutely essential to ensure both accuracy and contextual relevance (Meissner et al., 2024; Hu et al., 2025; Güler et al., 2024).

1.2 LLMs for student learning and tutoring

LLMs are increasingly serving as virtual tutors, providing students with personalized and immediate support in learning mathematics. Studies have shown that students who receive step-by-step explanations from LLMs, such as GPT-3, perform better on math tasks than those who only see final answers (Kumar et al., 2023). Learning is particularly enhanced when students attempt problems independently before reviewing the AI's explanation, suggesting LLMs are most effective as scaffolding tools rather than direct solution providers.

In classroom settings, LLMs have also been linked to increased student motivation and engagement. An interesting finding emerged from Zafrullah et al. (2023), their study indicated a substantial uptick in undergraduate math student motivation. A remarkable 80% reported feeling more motivated after leveraging ChatGPT as a learning aid. Meanwhile, Truong (2023) made a parallel observation, noting that ChatGPT appeared to cultivate a more personalized and interactive learning space. This seemed to actively bolster

students' conceptual understanding and problem-solving abilities.

LLMs also support critical thinking in higher education contexts. Karjanto (2023) documented improved reasoning in a linear algebra course where students used ChatGPT to clarify complex topics. Pavlova (2024) showed that using ChatGPT in a flipped classroom promoted self-regulated learning and reduced student anxiety. Despite these benefits, the effectiveness of LLMs depends on the quality of their output. Incorrect responses may mislead students, and over-reliance can hinder independent learning. Frieder et al. (2023) caution that current LLMs perform well on undergraduate-level tasks but are not suitable for advanced mathematical problem-solving.

In summary, LLMs can enhance math learning by providing responsive, individualized tutoring and fostering engagement. However, successful integration requires active student use, instructor guidance, and a clear understanding of the models' limitations.

1.3 LLMs in assessment and feedback

A key application of LLMs in mathematics education is in automating assessment both in generating test items and evaluating student responses. LLMs like GPT-4 can produce math problems aligned with learning objectives and cognitive taxonomies, as shown in Meissner et al.'s (2024) ItemForge system, which was validated by expert reviewers. Though the generated questions were well-structured, human oversight was necessary due to occasional errors in AI-generated solutions.

Beyond item generation, LLMs are increasingly used for formative feedback and grading. For instance, Henkel et al. (2025) explored GPT-4's ability to assess 53,000 open-response math answers, particularly those previously ungradable by rule-based systems. Using chain-of-thought prompting, the model achieved 97% grading accuracy in difficult cases. Even a modest improvement in overall grading (from 96% to 97%) significantly reduced student mastery misclassification from 6.9% to 2.6%, suggesting LLMs can enhance the precision of adaptive learning systems.

However, challenges remain. LLMs can be inconsistent or biased, especially if not carefully prompted. While current studies report minimal grading bias in math tasks, concerns persist about fairness, student manipulation of AI graders, and the need for rigorous prompt engineering. These limitations point to the importance of teacher involvement and quality control.

Looking forward, LLMs have the potential to transform assessment strategies into higher education. Their ability to grade open-ended responses could encourage a shift away from multiple-choice formats toward richer, more meaningful evaluations. Some scholars even advocate integrating LLMs into assessments themselves. Students should be evaluated not just on answers, but on their ability to effectively use and critique AI tools.

RQ2. Curriculum design and pedagogical implications

The increasing integration of large language models into mathematics education is driving a reconsideration of both curriculum design and pedagogical strategies among educators and researchers alike. Emerging literature highlights several key themes on how instruction may evolve alongside AI integration.

2.1 Revisiting curriculum content and sequence

With LLMs capable of handling routine problem-solving, researchers suggest a curricular shift toward deeper conceptual understanding, modeling, and critical interpretation. Matzakos et al. (2023) argue that mathematics instruction could focus less on repetitive drills and more on skills like formulating problems, validating AI outputs, and comparing solution strategies. This aligns with calls to integrate AI literacy into math curricula which includes teaching students how to effectively prompt LLMs, assess their responses, and understand the limits of algorithmic reasoning (Kasneci et al., 2023). Much like the historical integration of calculators, LLMs could become tools students learn to use critically and responsibly.

2.2 Assessment and academic integrity policies

The rise of LLMs is prompting changes in assessment practices and academic integrity policies. Institutions are developing guidelines that clarify acceptable AI use like allowing tools like ChatGPT for practice or study but restricting them in graded tasks unless disclosed (Kasneci et al., 2023). Transparency and ethical use are

emphasized, with students encouraged to use LLMs as learning aids, not as substitutes for original work. Educators are responding by redesigning assessments to be less AI-replicable, including reflective writing or oral defenses to verify student understanding (Matzakos et al., 2023). While formal policy analysis is limited, discussions across recent literature highlight the need for updated assessment strategies that align with responsible AI integration.

2.3 Inclusive and equitable learning

LLMs have the potential to support inclusion in mathematics education by providing accessible, individualized tutoring, especially for students lacking private support or those in large classes. Kasneci et al. (2023) suggest LLMs could assist students with disabilities through voice interaction or help learners with weaker academic backgrounds by simplifying mathematical explanations. Some studies also note that students who struggle academically benefit most from AI tools, as they can ask questions repeatedly without judgment. However, equity concerns remain like access to advanced LLMs often requires internet connectivity or subscriptions, posing barriers for some learners. To ensure equitable integration, curriculum planning must address access gaps by institutionalizing AI availability in classrooms.

2.4 Teacher roles and professional development

In math classrooms enhanced by AI, teachers take on a more pivotal role. They are not merely delivering content but facilitating students' effective use of LLMs. They help interpret AI outputs, address misconceptions, and support critical engagement with technology (Pavlova, 2024). This shift requires targeted professional development. Teachers need training in prompt design, integrating AI tools into instruction, and assessing AI-assisted student work. While early studies show growing interest among educators, many express the need for structured guidance and best-practice models (Güler et al., 2024). Institutions are thus encouraged to embed AI competencies into teacher training programs and curriculum development frameworks.

2.5 Ethical and critical thinking education

Integrating ethical awareness and critical thinking into mathematics education is essential as LLMs like ChatGPT become more common. Pavlova (2024) and others emphasize that students should be taught to evaluate AI-generated solutions critically rather than accept them at face value. Assignments that involve identifying and correcting errors in AI responses can deepen mathematical understanding and foster awareness of AI limitations. Studies show that students who actively engaged with AI through questioning, verifying, and reflecting demonstrated stronger learning outcomes. This supports the broader goal of maintaining critical thinking and active learning as core components of AI-integrated instruction.

RQ3. Challenges and research gaps

Despite promising developments, several challenges limit the effective integration of LLMs in mathematics education. A central issue is mathematical accuracy. LLMs such as ChatGPT may produce errors in arithmetic, algebra, or interpretation particularly on complex or visual tasks like geometry involving diagrams (Dao & Le, 2023). Their performance remains below graduate-level expectations, even when they appear fluent and confident (Frieder et al., 2023).

Another concern is student over-reliance. When learners depend on LLMs without engaging in active problem-solving, learning may become superficial. Some students also exhibit confirmation bias, accepting AI outputs uncritically due to their authoritative tone (Kumar et al., 2023). Training students in responsible AI use and promoting reflective engagement is critical to mitigating these effects.

Ethical issues include academic honesty, data privacy, and the potential for biased or homogenized solutions. Few studies have explored how LLMs might unintentionally favor certain problem styles or reduce diversity in mathematical thinking which is an area in need of more attention (Kasneci et al., 2023).

There are also significant research gaps. Most studies to date are short-term; there is limited evidence on the long-term impact of LLMs when embedded into full academic semesters. Questions about habit formation, sustained engagement, and performance remain largely unanswered. Additionally, theoretical models of learning such as scaffolding and self-regulated learning need refinement in the context of AI-supported environments. Early reviews suggest AI tools can either support or hinder self-regulation depending on how they are designed (Truong, 2023).

As LLMs rapidly evolve, continuous evaluation is essential. GPT-4, for instance, shows improved reasoning compared to GPT-3.5, and newer models like Gemini or math-specific LLMs may further change the landscape. The literature must remain dynamic and adaptable to these developments (Henkel et al., 2025).

6. Conclusion

This review has examined the state of peer-reviewed research from 2020 through early 2025 on the use of LLMs in tertiary-level mathematics education. In this short span, the emergence of powerful LLMs like GPT-3 and GPT-4 and particularly the public release of ChatGPT which has initiated a new era in which AI can actively participate in educational processes. The collected evidence paints a picture of transformative potential: LLMs can serve as on-demand tutors, assist instructors in creating and grading complex mathematical tasks, and personalize the learning experience in ways previously unimaginable in large-class settings. Empirical studies have documented concrete benefits such as improved student performance on problem-solving tasks when LLM explanations are utilized, increased student engagement and motivation in AI-supported learning environments, and efficiency gains in assessment generation and feedback provision. Equally important, scholarly discussions have begun to map out how curriculum and pedagogy might evolve in many ways like by incorporating AI literacy, redefining assessment strategies, and focusing on higher-order skills that complement AI capabilities.

At the same time, our review underscores that realizing the promise of LLMs in mathematics education requires overcoming significant hurdles. Chief among these are issues of accuracy, trust, and ethics. An LLM that occasionally produces incorrect mathematics can erode confidence or propagate misunderstandings if left unchecked; thus, robust verification mechanisms and teacher oversight are non-negotiable in high-stakes usage. The ease with which students can obtain answers necessitates a recommitment to academic integrity and perhaps a reimagining of what skills assessments should target in the age of AI. There are also broader concerns about equitable access to these technologies, potential biases, and the preparedness of educators and institutions to integrate AI tools effectively. These challenges form a research agenda for the immediate future: continued interdisciplinary efforts are needed to refine LLM technology for educational purposes (making it more reliable and transparent), to develop pedagogical frameworks that incorporate AI in a principled way, and to longitudinally study the impacts on learning outcomes across diverse student populations.

In conclusion, LLMs represent a powerful new ally in the teaching and learning of mathematics at the tertiary level. One that, if harnessed wisely, could enhance educational quality and accessibility. Rather than viewing LLMs as a threat to traditional education, the emerging consensus is to approach them as tools that require thoughtful integration. A key theme from literature is the importance of maintaining the human element. For example, the most successful implementations are those where human educators guide and complement the AI's contributions, and where students are taught to engage critically with the AI rather than accept its output at face value. Mathematics has always been a discipline that values reasoning and insight; these qualities will remain at the forefront, even as the mechanics of problem-solving and information retrieval are increasingly supported by AI. Ultimately, the goal is a synergy where LLMs take over routine or remedial tasks and provide augmented feedback, freeing both teachers and students to focus on deeper understanding, creativity, and the joy of mathematical discovery.

As this review has shown, the groundwork is being laid by early studies, but much remains to be learned. We encourage educators to experiment with LLMs in their practice (with proper safeguards) and share results, and we urge researchers to continue rigorously investigating the pedagogical effects of these tools. By doing so, the field of mathematics education can ensure that it remains proactive in shaping the narrative of AI in education – leveraging LLMs to support the age-old mission of helping students learn and appreciate mathematics, while upholding the standards and integrity that the discipline demands.

7. Limitations of this review

This review is limited by the rapid pace of technological change; recent developments (especially in 2025) may not be fully reflected. Limiting this review strictly to tertiary mathematics education meant we had to leave out insights that could have been quite valuable from K-12 settings. On top of that, most of the research we looked at came from English-speaking or well-resourced areas. This unfortunately resulted in scant coverage of regions and environments that are underrepresented or multilingual. There's also the issue of publication bias to consider; studies with positive findings might appear more often in literature than is truly representative.

8. Recommendations

To effectively leverage LLMs in mathematics education, practitioners are encouraged to use these tools as instructional supports (not replacements) by incorporating them into lesson planning, example generation, and inquiry-based learning. Teaching students AI literacy is essential, including how to prompt, critique, and verify LLM outputs. LLMs should primarily be used in formative settings where feedback can be guided, with caution exercised in summative assessments. Promoting active learning and ensuring equitable access to AI tools are also critical to avoid deepening existing educational divides. For researchers, priorities include conducting long-term classroom studies, refining theoretical frameworks that reflect AI-supported learning, and investigating issues of equity, bias, and implementation strategies. Exploring domain-specific LLMs tailored to mathematics may further improve instructional effectiveness and model precision.

Acknowledgment

We extend our sincere appreciation to the reviewers and editors of the IISTE *Journal of Education and Practice* for their constructive feedback and insightful suggestions, which have significantly contributed to the refinement of this article in its final form.

References

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2023). *A systematic literature review of empirical research on ChatGPT in education* [Preprint]. ResearchGate. https://www.researchgate.net/publication/380892561_A_systematic_literature_review_of_empirical_research_on_ChatGPT_in_education

Cho, Y., AlMamlook, R. E., & Gharaibeh, T. (2024). A systematic review of knowledge tracing and large language models in education: Opportunities, issues, and future research. *arXiv preprint arXiv:2412.09248*. https://arxiv.org/abs/2412.09248

Dao, T., & Le, H. (2023). Evaluating ChatGPT's performance on national mathematics exams. *Vietnam Journal of Educational Technology*, 12(3), 101–114.

Dong, Y., Li, X., & Zhang, H. (2024). *Large language models in education: A systematic review* [Preprint]. ResearchGate. https://www.researchgate.net/publication/382389137_Large_Language_Models_in_Education_A_Systematic_Review

Frieder, S., Pinckaers, M., Griffiths, R.-R., Salvadores, M., Dobrev, T., Huang, W., & Gaspar, J. (2023). *Mathematical capabilities of ChatGPT* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.13867

Güler, C., Almanshadi, H. M., & Altun, H. (2024). Examining the use of ChatGPT by mathematics teachers in pedagogical planning. *Journal of Educational Technology and Practice*, 14(2), 42–56.

Henkel, J., Papoutsaki, A., Rosen, Y., & Salehi, N. (2025). Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy. *Physical Review Physics Education Research*, 21(1), 010126. https://doi.org/10.1103/PhysRevPhysEducRes.21.010126

Hu, Y., Wang, C., & Li, J. (2025). Designing AI-enhanced mathematics lesson plans through classroom simulation with ChatGPT. *Journal of Educational Computing Research*, 63(1), 75–89.

Karjanto, N. (2023). The role of ChatGPT in enhancing understanding of linear algebra. *Asia-Pacific Journal of Mathematics Education*, 9(2), 78–90.

Kasneci, E., Sessler, K., Schmid, U., & Eberle, J. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kumar, A., Ram, K., & Jain, S. (2023). Evaluating the effectiveness of GPT-3 explanations in learning algebra. *Computers & Education: Artificial Intelligence*, 4, 100127.

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*, 18(5), 4–17. https://doi.org/10.3991/ijet.v18i05.42979

Meissner, J., Becker, T., & Krauss, F. (2024). Automatic generation of mathematics assessments using large language models. *British Journal of Educational Technology*, 55(2), 240–263.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1–13. https://doi.org/10.1177/1609406917733847

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

Pavlova, T. (2024). Dialogic pedagogy in the era of artificial intelligence: A case study of flipped learning with ChatGPT. *Ukrainian Journal of Educational Innovation*, 15(1), 29–42.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Truong, H. T. (2023). Using ChatGPT to foster engagement in mathematics learning: A case study. *International Journal of Digital Education and Artificial Intelligence*, 3(1), 33–47.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., & Gašević, D. (2023). *Practical and ethical challenges of large language models in education: A systematic scoping review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.13379

Zafrullah, M., Pradipta, A., & Akbar, R. (2023). Improving learning motivation with ChatGPT: Evidence from undergraduate math education. *Indonesian Journal of Educational Studies*, 17(4), 135–143.