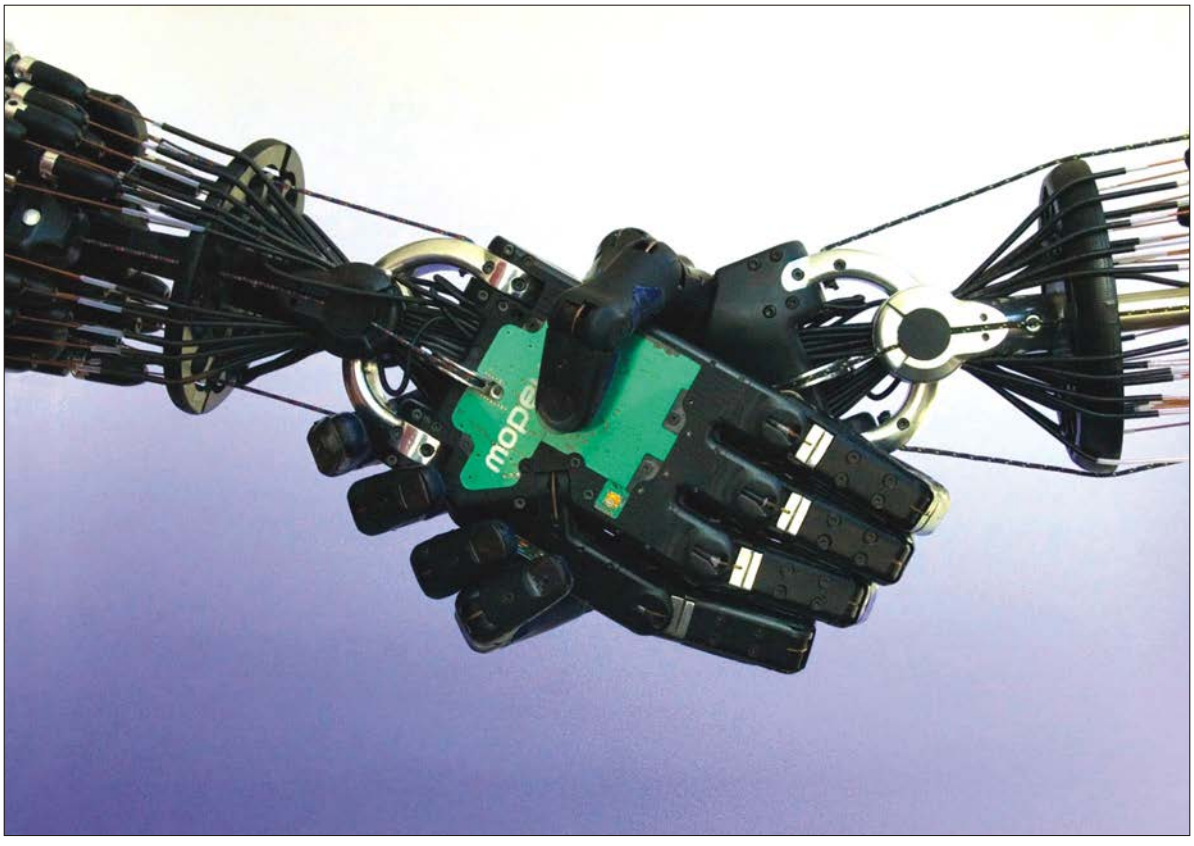

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2026.49.59

Review Article
Large Language Models in Mathematical Intelligent Educational Assessment: A Literature Review

[1]Xu Tong, [1]Razali Yaakob and [2]Sina Abdipoor

[1]Faculty of Computer Science and Information Technology, University Putra Malaysia, Malaysia
[2]Faculty of Engineering and Technology, Sunway University, Malaysia

Abstract

With the in-depth integration of artificial intelligence and education, Large Language Models (LLMs) have become a new technical support for the innovation of mathematical intelligent educational assessment. This literature review systematically sorts out research on the application of LLMs in mathematical intelligent educational assessment, focusing on four core aspects: application status, core technology optimization, existing research achievements and dataset support. A systematic literature retrieval method was adopted. Through critical analysis of relevant literature and comparative study of core journals in the field, this review identifies five key research gaps in current studies, including insufficient subject adaptability of LLMs in mathematics, low feedback quality, imperfect adaptive learning closed loop, single teacher assistance function and insufficient integrated application of mathematical datasets. Finally, based on these gaps, the future research direction of this field is clarified and promotes the in-depth application of LLMs in mathematics education scenarios.

Key words: Large Language Models (LLMs), mathematical intelligent educational assessment, intelligent education, prompt engineering, dataset integration

Citation: Tong, X., R. Yaakob and S. Abdipoor, 2026. Large language models in mathematical intelligent educational assessment: A literature review. J. Artif. Intel., 19: 49-59.

Corresponding Author: Razali Yaakob, Faculty of Computer Science and Information Technology, University Putra Malaysia, Malaysia

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

With the in-depth integration of Artificial Intelligence (AI) and education, intelligent educational assessment has become a core component of the transformation from traditional auxiliary teaching tools to personalized, intelligent and closed-loop intelligent education systems. This literature review focuses on the application of Large Language Models (LLMs) in mathematical intelligent educational assessment, aiming to systematically sort out the research status, summarize existing achievements and deficiencies, clarify research gaps and provide theoretical support.

First, it is necessary to clarify the core theoretical concepts involved in this paper. Large Language Models (LLMs) refer to large-scale pre-trained language models based on deep learning, which have strong capabilities in semantic understanding, logical reasoning and text generation and are represented by OpenAI's GPT series, Google's Gemini, Anthropic's Claude and domestic models such as DeepSeek and Doubao[1-3]. Intelligent educational assessment refers to the use of AI technology to realize automated evaluation, error identification, personalized feedback and learning status tracking of students' learning outcomes, which is different from the traditional manual evaluation model that is inefficient and lacks personalization[4,5]. Mathematical intelligent educational assessment, as a branch of it, focuses on the characteristics of mathematics with strong logical reasoning and strict problem-solving norms and needs to solve the pain points such as the difficulty in identifying subject-specific errors and the inability to analyze thinking paths[6,7].

The research background of this topic is closely linked to the current development bottleneck of intelligent education assessment. On the one hand, traditional educational assessment models face prominent problems: Low efficiency of manual homework correction, lagging feedback, difficulty in realizing personalized adaptation to students' learning status; teachers need to spend a lot of time counting students' error data and cannot quickly focus on class high-frequency errors to carry out targeted intervention; the assessment process mostly stays at "answer judgment" and it is difficult to deeply analyze students' thinking paths and weak links of knowledge points[8-10]. On the other hand, the rapid development of LLMs has provided a new technical path for breaking through these bottlenecks. However, the current application of LLMs in mathematical intelligent educational assessment still has obvious limitations: Insufficient subject adaptability, low feedback quality, imperfect learning closed loop and single teacher assistance function[11-13]. Therefore, it is of great theoretical and practical significance to systematically sort out the relevant research and explore the optimization path of LLMs in mathematical intelligent educational assessment.

MATERIALS AND METHODS

Study design: This study adopted a structured literature review design to systematically examine the application of Large Language Models LLMs in mathematical intelligent educational assessment. A structured review approach was selected because it allows the existing literature to be organized, compared and synthesized according to clearly defined research themes rather than merely summarized descriptively. This method is suitable for identifying research trends, theoretical foundations, methodological patterns, technical approaches, dataset support and existing gaps in an emerging interdisciplinary field[12,13].

The review focused on four major dimensions application scenarios of LLMs in intelligent educational assessment, core technology optimization, existing research achievements and dataset support for mathematical assessment. As the design, technical framework, type of dataset and evaluation objectives of the reviewed studies were not consistent, qualitative synthesis was used instead of meta-analysis. Thus, this study is able to compare the different lines of research on how large language models have been applied to support automated scoring, mathematical error detection, feedback generation, teacher assistance and personalized learning.

Search strategy: Through the above major academic databases and scholarly search platforms, a literature search was carried out in Web of Science, Scopus, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, ERIC and Google Scholar. These databases were selected because they cover research in artificial intelligence, educational technology, intelligent tutoring systems, learning analytics and assessment of mathematics education.

The three sets of keyword groups for the search terms have been made. The first group of words included "large language models", "LLMs", "generative AI", "ChatGPT", "GPT-4", "Gemini" and "Claude". The second group is about educational assessment and includes "intelligent educational assessment", "automated assessment", "automated scoring", "formative feedback", "personalised feedback", "learning analytics" and "knowledge tracing". The third group focused on mathematics education and included "mathematical reasoning", "mathematics assessment", "math error diagnosis", "mathematical feedback", "math problem-solving" and "mathematical dataset".

Boolean operators were used to combine the search terms. For example, search strings included combinations such as "large language models" AND "mathematical assessment", "LLMs" AND "automated scoring", "generative AI" AND "formative feedback", "large language models" AND "math reasoning" and "LLMs" AND "educational dataset". The search mainly focused on studies published from 2020 to 2026, while several classical studies on formative assessment, self-regulated learning and literature review methodology were also retained because they provide theoretical and methodological support for this review.

Selection criteria: The retrieved literature was screened according to predefined inclusion and exclusion criteria. Studies were included if they met one or more of the following conditions:

- They examined the application of LLMs or generative AI in education
- They focused on intelligent educational assessment, automated scoring, feedback generation, or learning diagnosis
- They involved mathematical reasoning, mathematics education, or mathematical problem-solving assessment
- They discussed technical methods such as prompt engineering, Retrieval-Augmented Generation RAG , multi-agent systems, knowledge tracing, or domain adaptation
- They introduced or evaluated datasets or benchmarks related to mathematical reasoning, educational assessment, or multimodal mathematical tasks

Studies were excluded if they were unrelated to education or assessment, focused only on general AI applications without relevance to LLMs, lacked clear methodological or technical descriptions, or were duplicate records. Studies with weak relevance to mathematical intelligent educational assessment were also excluded. Priority was given to peer-reviewed journal articles, high-quality conference papers, systematic reviews and influential benchmark studies. Recent studies were emphasized to ensure that the review reflects the rapid development of LLMs after the release of ChatGPT and GPT-4[14,15].

Data extraction and analysis: Based on the above screening results, the selected studies were qualitatively synthesized. From all the studies, the following key information has been extracted: Year of publication, research goal, educational context, LLM type, evaluation task, technical method, dataset or benchmark used, evaluation index, main results and reported limitations. The above extractions were arranged in a few sections of analysis.

The selected studies were then compared in five ways. First, application scenarios were explored to determine how LLMs are being applied in automated scoring, mathematical error diagnosis, personalised feedback, teacher assistance and adaptive learning. Second, the above technical approaches were compared: Prompt engineering, RAG, multi-agent collaboration, knowledge tracing and domain-specific adaptation. Third, the accuracy of the assessment function, the quality of feedback, interpretability and personalisation will be discussed. Fourth, the extent, type, constraints and fit of educational and mathematical datasets were analysed by comparison. Fifth, based on the research gaps identified, the limitations of the present studies and directions for future research have been presented.

This analytical process was intended to move beyond a descriptive summary and provide a structured comparison of existing research. In particular, the analysis focused on whether current LLM-based systems can meet the specific requirements of mathematical intelligent educational assessment, including logical reasoning, step-by-step error identification, thinking-path analysis, syllabus alignment and personalized learning support.

Analytical framework

Self-regulated learning theory (SRL): Self-regulated learning SRL Theory emphasizes that learners actively monitor, regulate and adjust their cognitive, metacognitive, motivational and behavioral processes during learning. In this review, SRL theory was used to analyze whether LLM-based mathematical assessment systems can support students' self-monitoring, error reflection and learning strategy adjustment. Studies related to personalized feedback, thinking-path analysis, learning recommendations and adaptive learning were examined from this perspective.

From the SRL perspective, mathematical intelligent assessment should not only judge whether an answer is correct, but also help students understand why an error occurs and how to improve their problem-solving process. Therefore, studies that used LLMs to generate explanatory feedback, identify weak knowledge points, or guide students to revise their reasoning were categorized under this theoretical dimension.

Intelligent assessment theory: Intelligent Assessment Theory emphasizes the transformation of assessment from result-oriented judgment to process-oriented, real-time and

personalized evaluation. It is closely related to formative assessment, which highlights the role of feedback in improving learning rather than merely measuring final performance[16]. In this review, Intelligent Assessment Theory was used to analyze how LLMs contribute to automated scoring, error identification, feedback generation, learning status tracking and teacher decision support.

The reviewed studies were examined according to three dimensions Assessment accuracy, feedback immediacy and personalization. Accuracy refers to whether LLMs can correctly identify mathematical answers, reasoning steps and error types. Immediacy refers to whether feedback can be generated in a timely manner to support learning. Personalization refers to whether feedback and learning suggestions can be adapted to students' individual knowledge levels, learning difficulties and error patterns.

LLM adaptation theory: A LLM Adaptation Theory focuses on the degree to which general-purpose LLMs can be adapted to specific domains, tasks and user needs. Although LLMs have strong language understanding and generation capabilities, direct application in mathematical educational assessment may still face problems such as hallucination, weak domain alignment, unstable reasoning and insufficient sensitivity to subject-specific errors[17,18]. Therefore, adaptation is necessary when LLMs are applied to mathematics education.

In this review, LLM Adaptation Theory was used to evaluate the adaptability of LLMs to mathematical intelligent educational assessment. The analysis focused on four aspects mathematical reasoning adaptation, subject-specific error identification, feedback alignment with curriculum requirements and dataset-based evaluation. Studies involving prompt engineering, RAG, domain-specific knowledge bases, fine-tuning, multi-agent collaboration and mathematical benchmarks were analyzed under this dimension.

Based on this analytical framework, the reviewed literature was synthesized around three guiding questions:

- How do LLMs support students' self-regulated learning in mathematical assessment
- How do LLMs improve the accuracy, immediacy and personalization of intelligent educational assessment
- How can LLMs be adapted to the specific requirements of mathematical reasoning, error diagnosis and feedback generation

RESULTS AND DISCUSSION

Overview of research themes: The reviewed literature shows that the application of Large Language Models (LLMs) in educational assessment has developed rapidly in recent years. Existing studies mainly focus on several research themes: LLM-based educational assessment, automated scoring, feedback generation, mathematical reasoning, personalized learning support and assessment redesign in the context of generative artificial intelligence[19-23]. Therefore, the above themes suggest that LLMs are no longer restricted to general text generation, but are increasingly being explored as tools for educational evaluation, learning diagnosis and feedback support and intelligent tutoring.

Based on the research results in intelligent educational assessment, most studies now focus on learning diagnosis through processes rather than merely checking the results. Earlier intelligent assessment systems mainly focused on whether the final answer was correct, but recent LLM-based studies have paid more attention to reasoning steps, error patterns, feedback quality and learner-specific support[20]. At the same time, when studying mathematics, we need to be correct in terms of answers; logic and the correct procedure for reaching these answers are also required.

According to the above studies, LLM-based assessment is also closely related to human-computer collaboration in education. LLMs can help students get explanations and feedback; at the same time, they can reduce repetitive grading work for teachers and help with interpreting student responses[20-22]. However, the present studies are also uneven. Although there are some studies on automated scoring and feedback generation, fewer have been carried out to study long-term adaptive learning, teacher decision support and mathematics-specific error diagnosis.

Application scenarios of LLMS in mathematical intelligent educational assessment: The first is a representative case of use for LLMs in educational assessment. Most research has used LLMs for automatic scoring, constructing constructed-response questions, providing feedback, offering personalised learning support and redesigning assessments[19]. Among the above situations, automated scoring and feedback generation have been mentioned most frequently because they directly address the actual requirements of reducing teacher workload and improving the speed of feedback.

A LLMs have demonstrated the ability to score open-ended and constructed-response items automatically. A LLMs are better than the previous rule-based scoring system at handling diverse students' free-form answers and identifying the meaning in natural language. It is also suitable for presenting students' reasoning in many different ways in mathematics education. Recent research on math constructed-response scoring has demonstrated that LLMs can improve the scalability of scoring; however, their reliability is still influenced by the Design of the task, scoring rubrics, selection of the model and human verification[24].

A LLMs can be used to analyse the steps in a student's mathematical error detection process rather than just grading the final answer. According to the above studies, LLMs are prone to calculating errors, misunderstanding the concepts, failing to reason correctly or using improper problem-solving methods. A LLMs can still produce inconsistent results in multi-step reasoning for mathematical problems, symbolic manipulation, geometric interpretation or rigid proof logic. Therefore, mathematical assessment requires stronger domain adaptation than general educational assessment[25-27].

Generate explanatory feedback, correction suggestions and learning recommendations in the feedback generation module of LLMs. As a kind of formative assessment, this function is to provide students with feedback on their errors and help them correct these errors in their study by Dai *et al.*[25]. However, according to the above studies, the feedback provided by LLMs may be too general, too long, too difficult, or not in line with how students actually think. The quality of feedback for learning mathematics needs to ensure accuracy in content, identify errors at a certain stage and match the curriculum objectives.

Teacher assistance is also a relatively underdeveloped case. Some research shows that Large Language Models (LLMs) can be employed to summarise students' answers, generate preliminary feedback and reduce the labour involved in repeated grading[20,22]. However, most of the current studies still focus on student-oriented functions. Little research has been carried out on what teachers do with LLM-generated assessment data, such as analysing, validating or using it[22]. Therefore, teacher-AI collaboration for the purpose of mathematical intelligent educational assessment has yet to be thoroughly explored.

Core technologies used to improve LLM-based assessment: The second main result is related to the technical means used for improvement of LLM-based educational assessment. Most of the studies published to date have focused on prompt-based assessment, retrieval-augmented generation (RAG), automated scoring models, adaptive feedback generation and mathematical reasoning evaluation[21]. The above methods are to enhance the accuracy, reliability, interpretability and personalisation of LLM-based assessment systems.

Prompts are frequently used to guide LLMs in the completion of tasks, rubrics are applied to assess the quality of the answers and organised feedback is provided. Step-by-step prompting can help motivate students in mathematics to demonstrate their thoughts at different times and not just present the end result[23]. Prompt-based methods are not suitable for problems that have changed in type, student answer format or difficulty. Therefore, prompt design by itself cannot guarantee a stable performance of mathematical assessments.

A RAG is mainly used to reduce hallucination and improve the factual reliability of LLM-generated educational content. By connecting LLMs with external knowledge sources, such as curriculum materials, scoring rubrics, textbook content, or domain-specific resources, RAG-based systems can generate feedback that is more aligned with task requirements[22]. This is particularly important in mathematics education because incorrect explanations or beyond-syllabus feedback may mislead students.

Automated scoring studies show that LLMs can support the evaluation of open-ended responses, but the reliability of scoring still requires careful control. Studies comparing LLM-generated scores with human grading suggest that LLMs may achieve useful levels of agreement in some tasks, but scoring performance varies across question types, disciplines, scoring rubrics and model settings[19]. Therefore, LLM-based scoring should not be treated as a fully independent replacement for teacher judgment, but as a support tool that still requires validation and supervision.

The other is an adaptation function. Recently, some research has demonstrated that large language models (LLMs) are capable of offering tailored and detailed improvements to students' learning through self-reflection and correction[25]. The Quality Level of the feedback will be determined by how precisely the model recognizes the student's answer and what shortcomings in their studies have been identified. The demands of mathematical assessment are generally higher and thus, the feedback needs to be both logically sound and educationally meaningful for the students at that level.

The focus of the math-reasoning assessment by LLMs is also on mathematical reasoning. Recently, some studies have been conducted using mathematical reasoning benchmarks and survey-based methods to explore the problem-solving capabilities of LLMs in mathematics[26]. The above studies have shown that LLMs have improved in mathematical reasoning;

Table 1: Thematic synthesis of reviewed studies

Research theme	Main focus	Synthesized finding	Remaining limitation
LLM-based educational assessment	Assessment redesign, intelligent tutoring, learning support	LLMs can support assessment innovation and personalized learning	Reliability, fairness, privacy and over-reliance remain concerns
Automated scoring	Open-ended responses, constructed-response assessment, formative assessment	LLMs can improve scoring scalability and reduce repetitive grading work	Scoring reliability still depends on rubrics, task design and human validation
Mathematical reasoning	Problem solving, reasoning evaluation, mathematical assessment	LLMs show potential in mathematical reasoning and problem-solving support	Complex reasoning, symbolic representation and proof-based tasks remain challenging
Feedback generation	Explanatory feedback, correction suggestions, adaptive feedback	LLMs can generate detailed and process-oriented feedback	Feedback may be inaccurate, generic, or misaligned with students' actual needs
Technical optimization	Prompt design, RAG, adaptive feedback, reasoning evaluation	These methods can improve reliability, contextual alignment and learning support	Integration across technologies remains insufficient
Dataset and benchmark support	Mathematical reasoning benchmarks, open-ended response datasets, feedback-oriented data	Different resources support different assessment functions	No single dataset covers all needs of mathematical intelligent educational assessment

however, their performance is still affected by the complexity of the problem, symbolic representation, reasoning depth and evaluation design. Therefore, in the future, mathematical assessments need to be adjusted to evaluate students' thinking process, errors made and the reason for those errors as well.

Dataset and benchmark support in mathematical assessment studies: The third main result is the Dataset and Benchmark Support. Existing studies have used various kinds of datasets and benchmarks to evaluate the performance of LLMs, such as mathematical reasoning benchmarks, open-ended response datasets, feedback-oriented assessment data and multimodal or context-rich mathematical tasks[25]. The above resources provide support for evaluating the ability to reason, scoring performance, feedback quality and learning assistance.

Mathematical reasoning benchmarks can be used to test the problem-solving and logical-step-following capabilities of LLMs in mathematics. Reasoning benchmarks are not always designed to diagnose learning problems in the classroom. Some benchmarks focus on the correctness of the final answer or high-level problem-solving skills and offer little information about students' error patterns, knowledge-point labels or feedback quality.

Open-ended response datasets are closer to educational assessments because they contain students' free-form answers and are frequently graded with rubrics. The above data can be used to examine the level at which LLMs have understood student responses, how grades are assigned and support for formative assessment provided. However, there may still be deficiencies in the subject coverage, annotation quality and detailed process-level error information.

Feedback-oriented assessment data can be used to find out whether LLMs are helping students learn effectively in actual teaching situations, rather than just being correct[26]. Based on the above research, LLMs can generate rich data for practice; however, it is not known whether this data is valid. All of the above aspects will be used as indicators to measure the quality of feedback.

For mathematical intelligent educational assessment, the reviewed literature indicates that no single dataset or benchmark can fully support all requirements[27]. Some resources are strong in mathematical reasoning evaluation, but weak in feedback generation. Some are useful for open-ended assessment, but lack mathematical subject specificity. Some support feedback analysis, but do not include long-term learning traces. Therefore, dataset selection should be aligned with the specific assessment task, such as automated scoring, error diagnosis, feedback generation, multimodal understanding, or adaptive learning support.

Thematic synthesis of reviewed studies: Table 1 summarizes the main research themes identified from the reviewed literature. The synthesis shows that LLMs have strong potential in mathematical intelligent educational assessment, but current studies still have limitations in reliability, subject adaptation, feedback quality, teacher-AI collaboration and dataset suitability.

Synthesized research gaps: Based on the reviewed studies, five major research gaps can be identified.

First, the mathematical subject adaptability of LLMs remains insufficient. Although LLMs perform well in general language understanding, they may still make errors in symbolic reasoning, step-by-step verification, geometric interpretation and subject-specific error diagnosis[26]. Existing studies have not fully solved the problem of adapting general-purpose LLMs to the strict logical requirements of mathematics.

Second, the quality of LLM-generated feedback still needs improvement. Many systems can judge whether an answer is correct, but they cannot always provide accurate, targeted and learner-friendly explanations[25,26]. In particular, feedback may be too general, too difficult, beyond the syllabus, or disconnected from the student's actual reasoning process. This shows that feedback quality should be evaluated not only by fluency, but also by pedagogical usefulness and mathematical correctness.

Third, the adaptive learning closed loop is not yet complete. Although some studies discuss personalized learning support, many of them stop at one-time assessment or feedback generation[27]. There is still limited research on the continuous cycle of assessment, feedback, relearning, reassessment and learning status updating. This limits the ability of LLM-based systems to support long-term learning improvement.

Fourth, teacher support is still relatively weak. At present, more attention has been paid to student-oriented functions and the role of teachers in interpreting and applying LLM-generated assessment results has not been fully explored. More effort needs to be made to use LLMs for error analysis and trend monitoring of student learning to inform targeted remedial measures.

Fifth, Dataset Integration is also lacking. Existing datasets and benchmarks offer some support for the three sub-tasks of supporting mathematical reasoning, evaluating open-ended responses and providing feedback; however, they are often designed for different purposes. There is a deficiency in an all-in-one dataset that can simultaneously support mathematical error diagnosis, feedback generation, multi-modal input and educational assessment validation.

In short, based on the above results, LLMs show good prospects for use in intelligent educational assessment for mathematics, particularly in automated scoring, error analysis and feedback generation and personalized learning support. However, the existing studies are still limited in mathematical domain adaptation, feedback reliability, adaptive learning continuity, teacher-AI collaboration and dataset suitability. The above-mentioned synthetic results provide the foundation for the following discussions on implications and directions for future research.

Interpretation of the main findings: According to the results of this review, large language models should not be regarded solely as aids to improve the efficiency of grading. The first problem of mathematical intelligent educational assessment is whether LLMs can provide good support for learning by accurately identifying difficulties in learning, giving prompt and relevant feedback and adjusting teaching plans. Good feedback can help learners know what they need to improve in relation to their learning goals and more importantly, it should show them how to improve rather than just tell them the correct answer[28].

Therefore, this will also apply to teaching of mathematics. Mathematical assessment can verify that a student has solved a problem correctly and observe how well they can employ logic, concepts, symbols, procedures, problem-solving strategies, etc., in this process. Therefore, the Design of LLM-based assessment should support process-level interpretation. If only a binary indicator of "correctness" is provided by an LLM, its educational value will be limited. A better assessment system should be able to explain why an error occurred, what concept is being addressed and how the learner can correct the solution.

Based on the above research, LLM-based mathematics assessment needs to be linked with students' self-regulated learning. Formative assessment and self-regulated learning are related in that feedback can help learners know how well they have learned and adjust their own study plans[29]. Therefore, the feedback generated by LLMs should not be regarded as final rectification. It should help learners think independently and independently choose to learn further.

Another is that the design of LLM-based assessment should be a human-AI collaboration. Intelligent tutoring systems and AI-assisted learning tools have been proposed to support students' studies for a long time, but they are not intended to replace teachers' teaching. Research on hybrid human-AI learning in recent years has also proposed that AI systems should support teachers and not take over their roles[29,30]. Teachers are still necessary at the level of mathematical assessment to validate the results produced by LLMs, address any misunderstandings students have and adjust teaching plans accordingly.

Dataset and benchmark suitability for mathematical assessment: Selection of the Dataset and Benchmark is a Problem in LLM-based Mathematical Intelligent Educational Assessment. As shown by the above studies, the various datasets and benchmarks have different purposes. Some are more suitable for the assessment of mathematical reasoning, some for scoring open-ended answers, some for evaluating the quality of feedback and some for adaptive learning or classroom decision support. Therefore, when selecting a dataset, we need to know why we plan to use it for training and what the scale of the data is.

Benchmarks of mathematical reasoning can be used to determine the problem-solving ability and difficulty handling of LLMs in mathematics. They can be used to test reasoning ability, symbolic understanding and problem-solving skills.

Table 2: Comparative analysis of dataset and benchmark suitability

Dataset or benchmark type	Main scope	Suitable assessment task	Main limitation	Suitability for mathematical intelligent assessment
Mathematical reasoning benchmarks	Mathematical problem solving and reasoning performance	Testing LLM reasoning ability and problem-solving accuracy	Often lack student error patterns and feedback labels	Suitable for evaluating reasoning capacity, but insufficient for student learning diagnosis
Open-ended response datasets	Student written answers and rubric-based scoring	Automated scoring and constructed-response assessment	May lack reasoning-step annotations and misconception labels	Suitable for scoring tasks, but should be combined with error diagnosis data
Feedback-oriented data	Feedback generation and learning support	Evaluating explanation quality and correction suggestions	Feedback quality is difficult to measure automatically	Suitable for evaluating personalized feedback when combined with expert review
Adaptive learning data	Learning traces, progress records, repeated assessment results	Supporting assessment-feedback-relearning cycles	Requires longitudinal data collection	Suitable for validating whether LLM feedback improves learning over time
Multimodal mathematical data	Problems involving diagrams, tables, graphs, or visual context	Evaluating mathematical understanding beyond text-only input	May lack process-level diagnosis labels	Useful for multimodal assessment, but not sufficient alone for feedback evaluation

However, their weakness is that they often focus on model-level performance rather than student-level learning diagnosis. A model may be able to solve a benchmark problem correctly, but it is not necessarily indicated that this has identified a student's misconception or provided effective learning feedback. Recent studies on LLMs in mathematics education have also shown that current LLMs still have deficiencies in complex reasoning, proof-based tasks and reliable explanation generation[31].

Open-ended response datasets are more directly related to educational assessment[18]. They are suitable for determining whether LLMs can interpret varied answers from students, apply scoring rubrics and support formative assessment. However, the open-ended response datasets may not always include detailed reasoning-step annotations or misconception labels. Therefore, they are less suitable for analysing errors in mathematics. Scoring labels are not enough for the mathematical intelligent assessment; we also need process-level labels, error categories and knowledge-point information.

Feedback-oriented data can be employed to identify the degree to which an LLM's responses assist teaching[24,25]. The Quality of Feedback does not depend on Fluency or Length. Assess the correctness, specificity, clarity, suitability for learners and practicality of the problem. Feedback in mathematics education should be precise in terms of mathematics and aligned with how students have been thinking. If the feedback is accurate but too difficult, too general, or beyond what the learner can currently handle, it will have little teaching effect.

Adaptive learning data are needed when the goal is to build a continuous assessment-feedback-relearning cycle. Self-regulated learning research suggests that learning support should help students monitor progress and adjust strategies over time[32]. Therefore, if an LLM-based system aims to support adaptive learning, it should not rely only on single-task datasets. It also needs longitudinal learning data, knowledge-state records and repeated assessment results. Without such data, it is difficult to verify whether feedback actually improves learning over time.

Multimodal or context-rich mathematical data are also important for future mathematical assessment. Mathematics problems often include diagrams, tables, graphs, symbolic expressions, or real-world contexts. However, multimodal data alone are not enough. To support educational assessment, such data should also include student responses, reasoning traces, error labels and feedback evaluation criteria. Otherwise, they can test visual or reasoning ability, but cannot fully support learning diagnosis shown in Table 2.

Based on the above comparison, none of the existing datasets or benchmarks can meet all the demands of mathematical intelligent educational assessment. A reasonable way to do this is to combine all the above data in accordance with different purposes. Reasoning benchmarks can evaluate problem-solving ability; open-ended response datasets support automated scoring; feedback-oriented data assess feedback quality; adaptive learning data support learning-loop validation; and multimodal data help build visual mathematical understanding. The above combined set of data is more in line with the features of mathematics assessment.

Implications for LLM-based mathematical assessment design: Based on the discussion above, the three focuses of design for LLM-based mathematical assessment systems are as follows:

Strengthen the Mathematical Domain Adaptation in the first place. General-purpose LLMs are good at generating fluent explanations but may not be accurate in mathematics. Symbolic-mathematics assessment can be employed to identify the entire process of a student's thinking and correct any errors. Although recent studies on the application of LLMs in mathematics education have shown that they can provide prompt explanations and learning support, they are still weak in dealing with difficult problems and offering reasonable justifications[33]. Therefore, domain-specific prompts, mathematical knowledge bases, curriculum-aligned rubrics and expert verification are needed.

Secondly, add a few types of feedback evaluation. A good LLM feedback system should inform students that the answer is incorrect and, at the same time, guide them to learn why they made a mistake and what they should study next. Based on the research results of feedback, it should be more concrete, feasible and closely aligned with learning goals[34]. Therefore, the future study should examine various aspects of LLM-generated feedback, such as its mathematical accuracy, ease of understanding, specificity, suitability for learners and teaching applications.

Thirdly, foster collaboration between teachers and artificial intelligence for assessment. Research on human-AI cooperation suggests that AI should assist teachers in dealing with routine or initial analysis; teachers will still be responsible for pedagogical interpretation and instruction[35]. Therefore, in the field of mathematics assessment, LLMs can be employed to assist with the initial grading and collection of errors, providing teachers with prompt feedback, while the final decision on intervention strategies should still be made by teachers.

FUTURE RESEARCH DIRECTIONS

First, a new system for mathematical assessment will be introduced in the future. Such frameworks should also assess the accuracy of the final answer, the reasoning process, types of errors, quality of feedback and learning improvements. Learning also includes developing procedural skills, building concepts and mastering symbols at the same time.

Second, in the future, more integrated datasets for mathematical intelligent educational assessment will be created. Most of the existing datasets support only one part of the evaluation, such as reasoning assessment or scoring. In the future, the datasets will also include student answers, reasoning paths, error annotations, knowledge points, feedback annotations and learning progress data. Thus, we will be able to assess the quality of education that LLMs can offer better.

Thirdly, more experiments are needed to be carried out in actual classrooms. Many of the current studies focus on benchmark evaluation or short-term experimental tasks. Mathematics' intelligent assessment is also to help improve teaching and learning. A long-term study of the classroom should be carried out to determine whether LLM-generated feedback improves students' learning outcomes, how teachers can use LLM-generated summaries effectively and whether students show improved self-regulation in repeated rounds of evaluation and feedback.

Finally, future studies will also address the problems of ethics and teaching. LLM-based assessment may have the following risks: Inaccurate feedback, student over-reliance, teacher deskilling, privacy problems and unfair evaluation. Therefore, the following systems will include human verification, open scoring rules, data protection measures and clear boundaries for the use of AI in educational assessment.

CONCLUSION

Review of recent studies on the application of Large Language Models in mathematical intelligent educational assessment. Based on the above results, it is found that LLMs can help build automated scoring and reasoning analysis systems for teacher-aided evaluation. Moreover, this paper proposes that rather than scoring students' math exercises, good uses of LLMs can help us locate their mistakes in the work and offer prompt corrections and suggestions for improvement. By organising the literature from the perspectives of assessment function, technical support and dataset suitability, this paper offers a targeted reference for subsequent studies and practical system design of mathematical intelligent educational assessment.

SIGNIFICANCE STATEMENT

This study is significant as it systematically synthesizes current research on Large Language Models in mathematical intelligent educational assessment, highlighting their potential to enhance automated scoring, personalized feedback and learning diagnosis. It identifies key limitations in domain adaptation, feedback quality and dataset integration, providing a structured foundation for improving AI-driven assessment systems and guiding future research in mathematics education.

REFERENCES

1. Gallifant, J., A. Fiske, Y.A.L. Strekalova, J.S. Osorio-Valencia and R. Parke *et al*., 2024. Peer review of GPT-4 technical report and systems card. PLOS Digital Health, Vol. 3. 10.1371/journal.pdig.0000417.
2. Imran, M. and N. Almusharraf, 2024. Google Gemini as a next generation AI educational tool: A review of emerging educational technology. Smart Learn. Environ., Vol. 11. 10.1186/s40561-024-00310-z.
3. Abbasi, M., P. Váz, J. Silva, F. Cardoso, F. Sá and P. Martins, 2026. Comparative performance analysis of large language models for structured data processing: An evaluation framework applied to bibliometric analysis. Appl. Sci., Vol. 16. 10.3390/app16020669.
4. Baker, R.S. and P.S. Inventado, 2014. Educational Data Mining and Learning Analytics. In: Learning Analytics: From Research to Practice, Larusson, J. and B. White (Eds.). Springer, New York, USA, ISBN: 978-1-4614-3305-7, pp: 61-75.
5. Kochmar, E., D.D. Vu, R. Belfer, V. Gupta, I.V. Serban and J. Pineau, 2022. Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. Int. J. Artif. Intell. Educ., 32: 323-349.
6. Koedinger, K.R. and A. Corbett, 2005. Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In: The Cambridge Handbook of the Learning Sciences, Sawyer, R.K. (Ed.), Cambridge University Press, Cambridge, United Kingdom, ISBN: 9780511816833, pp: 61-78.
7. VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. Educ. Psychologist, 46: 197-221.
8. Hattie, J., 2012. Visible Learning for Teachers: Maximizing Impact on Learning. 1st Edn., Routledge, London, United Kingdom, ISBN: 9780203181522, Pages: 296.
9. Black, P. and D. Wiliam, 1998. Assessment and classroom learning. Assess. Educ. Princ. Policy Pract., 5: 7-74.
10. Shute, V.J., 2008. Focus on formative feedback. Rev. Educ. Res., 78: 153-189.
11. García-Méndez, S., F. de Arriba-Pérez and M. del Carmen Somoza-López, 2025. A review on the use of large language models as virtual tutors. Sci. Educ., 34: 877-892.
12. Williamson, B., R. Eynon, J. Knox and H. Davies, 2023. Critical Perspectives on AI in Education: Political Economy, Discrimination, Commercialization, Governance and Ethics. In: Handbook of Artificial Intelligence in Education, du Boulay, B., A. Mitrovic and K. Yacef (Eds.), Edward Elgar Publishing, Cheltenham, United Kingdom, ISBN: 9781800375413, pp: 553-570.
13. Brusilovsky, P., 2024. AI in education, learner control, and human-AI collaboration. Int. J. Artif. Intell. Educ., 34: 122-135.
14. Snyder, H., 2019. Literature review as a research methodology: An overview and guidelines. J. Bus. Res., 104: 333-339.
15. Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron and T.C. Hoffmann *et al*., 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, Vol. 372. 10.1136/bmj.n71.
16. Kasneci, E., K. Sessler, S. Küchemann, M. Bannert and D. Dementieva *et al*., 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individual Differ., Vol. 103. 10.1016/j.lindif.2023.102274.
17. Shi, Y., K. Yu, Y. Dong and F. Chen, 2026. Large language models in education: A systematic review of empirical applications, benefits, and challenges. Comput. Educ.: Artif. Intell., Vol. 10. 10.1016/j.caeai.2025.100529.
18. Gao, R., H.E. Merzdorf, S. Anwar, M.C. Hipwell and A.R. Srinivasa, 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. Comput. Educ.: Artif. Intell., Vol. 6. 10.1016/j.caeai.2024.100206.
19. Foung, D., L. Lin and J. Chen, 2024. Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. Comput. Educ.: Artif. Intell., Vol. 6. 10.1016/j.caeai.2024.100250.
20. Morris, W., L. Holmes, J.S. Choi and S. Crossley, 2025. Automated scoring of constructed response items in math assessment using large language models. Int. J. Artif. Intell. Educ., 35: 559-586.
21. Mendonça, P.C., F. Quintal and F. Mendonça, 2025. Evaluating LLMs for automated scoring in formative assessments. Appl. Sci., Vol. 15. 10.3390/app15052787.
22. Pecuchova, J., Ľ. Benko and M. Drlik, 2025. Automated grading of open-ended questions in higher education using GenAI models. Int. J. Artif. Intell. Educ., 35: 3813-3846.
23. Li, Z., Z. Wang, W. Wang, K. Hung, H. Xie and F.L. Wang, 2025. Retrieval-augmented generation for educational application: A systematic survey. Comput. Educ.: Artif. Intell., Vol. 8. 10.1016/j.caeai.2025.100417.
24. Kinder, A., F.J. Briese, M. Jacobs, N. Dern, N. Glodny, S. Jacobs and S. Leßmann, 2024. Effects of adaptive feedback generated by a large language model: A case study in teacher education. Comput. Educ.: Artif. Intell., Vol. 8. 10.1016/j.caeai.2024.100349.
25. Dai, W., Y.S. Tsai, J. Lin, A. Aldino and H. Jin *et al*., 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. Comput. Educ.: Artif. Intell., Vol. 7. 10.1016/j.caeai.2024.100299.
26. Liu, T., Z. Chen, Z. Fang, W. Luo, M. Tian and Z. Liu, 2025. MathEval: A comprehensive benchmark for evaluating large language models on mathematical reasoning capabilities. Front. Digital Educ., Vol. 2. 10.1007/s44366-025-0053-z.
27. He, F., H. Lai, J. Liu, B. Wang, H. Chen, H. Liu and C. Zhang, 2025. Solving mathematical problems using large language models: A survey. Data Intell., 7: 907-946.

28. Pepin, B., N. Buchholtz and U. Salinas-Hernández, 2025. A scoping survey of ChatGPT in mathematics education. Digital Exper. Math. Educ., 11: 9-41.
29. Molenaar, I., 2022. Towards hybrid human-AI learning technologies. Eur. J. Educ., 57: 632-645.
30. Molenaar, I., 2022. The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. Comput. Educ.: Artif. Intell., Vol. 3. 10.1016/j.caeai.2022.100070.
31. Rizos, I. and N. Gkrekas, 2025. The impact of LLMs on mathematics education and research at the university. Social Sci. Humanit. Open, Vol. 12. 10.1016/j.ssaho.2025.101969.
32. Opesemowo, O.A.G. and H.O. Adewuyi, 2024. A systematic review of artificial intelligence in mathematics education: The emergence of 4IR. Eurasia J. Math. Sci. Technol. Educ., Vol. 20. 10.29333/ejmste/14762.
33. Turmuzi, M., S. Azmi and N.M.I. Kertiyani, 2025. ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. Teach. Teach. Educ., Vol. 170. 10.1016/j.tate.2025.105286.
34. Wang, S., F. Wang, Z. Zhu, J. Wang, T. Tran and Z. Du, 2024. Artificial intelligence in education: A systematic literature review. Expert Syst. Appl., Vol. 252. 10.1016/j.eswa.2024.124167.
35. Giannakos, M., R. Azevedo, P. Brusilovsky, M. Cukurova and Y. Dimitriadis *et al*., 2025. The promise and challenges of generative AI in education. Behav. Inf. Technol., 44: 2518-2544.