Advances in Auto-Grading with Large Language Models: A Cross-Disciplinary Survey

Fredrick Eneye Tania-Amanda Nkoyo[1], Chukwuebuka Fortunate Ijezue[1], Maaz Amjad[1], Ahmad Imam Amjad[2], Sabur Butt[3], Gerardo Castañeda-Garza[3]

[1]Texas Tech University, Texas, USA

[2]The University of Punjab, Pakistan

[3]Tecnológico de Monterrey, Mexico

tafredri@ttu.edu, cijezue@ttu.edu, maaz.amjad@ttu.edu, Ahmadimamamjad@gmail.com, saburb@tec.mx, g.castaneda@tec.mx

Abstract

With the rise and widespread adoption of Large Language Models (LLMs) in recent years, extensive research has been conducted on their applications across various domains. One such domain is education, where a key area of interest for researchers is investigating the implementation and reliability of LLMs in grading student responses. This review paper examines studies on the use of LLMs in grading across six academic sub-fields: educational assessment, essay grading, natural sciences and technology, social sciences and humanities, computer science and engineering, and mathematics. It explores how different LLMs are applied in automated grading, the prompting techniques employed, the effectiveness of LLM-based grading for both structured and open-ended responses, and the patterns observed in grading performance. Additionally, this paper discusses the challenges associated with LLM-based grading systems, such as inconsistencies and the need for human oversight. By synthesizing existing research, this paper provides insights into the current capabilities of LLMs in academic assessment and serves as a foundation for future exploration in this area.

1 Introduction

Grading has traditionally been a manual process conducted by human teachers or graders, which can be time-intensive, laborious, and subject to inconsistencies due to individual judgment (Gnanaprakasam and Lourdusamy, 2024). To circumvent some of these issues, standardized examinations and rubrics are designed. Nonetheless, these may fail to detect variations in student ability or in learning styles (Gnanaprakasam and Lourdusamy, 2024). Furthermore, traditional grading methods fail to deliver tailored feedback at scale, further decreasing the value of exams as opportunities for personalized assessment (Haque et al., 2022).

Manual grading has significant mental and physical implications both for educators and students (Skaalvik and Skaalvik, 2017) and students (Hough, 2023). Due to its repetitive and time-consuming nature, it leads to physical and mental fatigue for educators. Previous research indicates that the stress associated with manual grading can also hinder educators' ability to focus on other critical aspects of teaching, such as lesson planning and student engagement (Hakanen et al., 2006). For students, the subjective nature of manual grading can introduce biases, which may negatively impact students' academic outcomes and their trust in the evaluation process (Wigfall, 2020). Delayed feedback from manual grading can leave students in prolonged uncertainty, which may increase their anxiety levels (England et al., 2019).

On the contrary, the rapid advancements in the field of artificial intelligence (AI), and the introduction of LLMs, capable of understanding and generating human-like text, have shifted this paradigm. LLM-based grading is any grading technique that leverages powerful Large Language Models to automate the evaluation of student responses, offering potential benefits in speed, consistency, and scalability. This shift is particularly relevant in educational settings where large volumes of assessments, such as essays and short answers, need efficient processing. Research conducted by Grandel et al. (2024) showed the ability of LLM-based grading techniques to reduce grading time by 81.2%. AI-automated grading could reduce the workload on educators, allowing them to spend more time teaching. Such systems could ensure consistency and objectivity in evaluations, reducing human biases and providing fair assessments for all students.

While individual studies have demonstrated LLM applications in specific educational domains, a comprehensive cross-disciplinary analysis is essential to understand broader patterns, identify transferable methodologies, and reveal domain-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 477–498

specific challenges. Educational assessment varies significantly across disciplines—from objective STEM problem-solving to subjective humanities analysis—making it crucial to examine how LLMs perform across this spectrum. A cross-disciplinary perspective enables identification of universal best practices, domain-specific adaptations, and systematic gaps that single-domain studies cannot reveal. Moreover, such an approach allows for the synthesis of methodological insights that can inform both researchers and practitioners across diverse educational contexts.

This review paper surveys the current landscape of LLM-based assessment across six academic domains—educational assessment, essay grading, natural sciences and technology, social sciences and humanities, computer science and engineering, and mathematics. It synthesizes findings from 30 recent studies, analyzing how LLMs are applied in different assessment formats, the prompting strategies used, their alignment with human evaluators, and the contextual variables influencing performance. In doing so, this paper provides a cross-disciplinary framework for understanding the capabilities and limitations of LLM-based grading systems. It also highlights methodological trends, emerging implementation strategies, and the evolving role of human-AI collaboration in educational assessment. Overall, this paper provides a timely cross-disciplinary survey that will serve as a useful reference. It is well-scoped and captures key themes in LLM-based grading. Moreover, it brings to light current challenges and limitations in the area, such as rubric drift and LLM transparency issues.

2 Data Collection

We conducted a systematic literature review using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to identify relevant studies on the use of large language models (LLMs) in educational assessment. The search aimed to include published academic research between January 1, 2022, through January 14, 2025. The selection focused on works that addressed LLM use in grading, feedback generation, essay evaluation, short-answer marking, domain-specific assessments, and pedagogical implications.

Our review focuses on six academic domains—educational assessment, essay grading, natural sciences and technology, social sciences and humanities, computer science and engineering, and mathematics—selected to represent the breadth of educational assessment contexts where LLMs are being applied. These domains were chosen to span the spectrum from highly structured (mathematics, computer science) to open-ended assessments (humanities), include both technical and non-technical fields, and represent different cognitive complexity levels as defined by Bloom's taxonomy. This selection enables a comprehensive analysis of how LLM performance varies across assessment types, content domains, and evaluation criteria while maintaining sufficient depth within each domain.

Articles were gathered from multiple scholarly databases and repositories as detailed in Table 1, including Google Scholar, arXiv, IEEE Xplore, ACL Anthology, and ERIC (Education Resources Information Center). We also examined proceedings from key conferences, including ACL, EMNLP, EDM (Educational Data Mining), LAK (Learning Analytics and Knowledge), and AIED (Artificial Intelligence in Education). Keywords used in the searches included combinations of: "large language models," "educational assessment," "automated grading," "essay scoring," "student feedback," "ChatGPT," "GPT-4," "short answer evaluation," and "AI in education."

Studies were selected based on predefined inclusion criteria: (1) empirical studies involving LLM applications in educational assessment, (2) published between 2022-2025, (3) sufficient detail on methodology and results, and (4) focus on grading, feedback, or evaluation tasks. The PRISMA flow diagram detailing the study selection process is presented in Figure 1.

The initial search yielded 104 articles and reports. After removing duplicates, irrelevant papers (e.g., not focused on education or assessment, theoretical works without application), we filtered 48 full-text articles assessed for eligibility. We excluded the remaining articles for insufficient empirical content, lack of focus on assessment, or being out of scope (e.g., general education technology without AI involvement). This rigorous selection process yielded 30 articles for the final review. The included studies comprised 19 peer-reviewed publications and 11 preprints, reflecting the rapidly evolving nature of this research area.

3 Variables of Study

To analyze LLM applications in educational assessment, we define a set of key variables that form the basis for comparing diverse implementations. These variables are grouped into four main categories. First, the assessment types include Multiple-Choice Questions (MCQs), Short-Answer Questions, Essay Assessments, Programming Assignments, Mathematics Assessments, and Handwritten Assessments. Second, the studies span a broad range of education levels—from Early and Primary education through Secondary, Undergraduate, Graduate, to Professional Education. Third, human annotators are classified into distinct groups: Expert Evaluators, Experienced Educators, Novice Evaluators, Field Practitioners, and Unspecified Graders, a categorization that is crucial for understanding how LLM outputs compare with human judgment. Finally, evaluation metrics employed across the studies include Cohen's Kappa, Quadratic Weighted Kappa, Krippendorff's Alpha, Pearson and Spearman Correlations, Accuracy, F1 Score, and Win Rate. Collectively, this framework is essential for identifying patterns and making meaningful cross-disciplinary comparisons of LLM-assisted assessment. Detailed tables for each variable category can be found in Tables 2, 3, 4, and 5 in the Appendix.

4 LLMs in Assessment

Large language models face significant challenges in educational assessment contexts, particularly when evaluating higher-order cognitive tasks and providing nuanced feedback comparable to human experts (Kasneci et al., 2023; Gnanaprakasam and Lourdusamy, 2024). However, researchers have developed innovative approaches to address these limitations, demonstrating increasingly promising results across diverse educational settings.

Early evaluations by Teckwani et al. (2024) in the physiological education domain revealed that LLMs, such as GPT-3.5, GPT-4o, and Gemini, achieved only moderate alignment with human graders (for example, Gemini reached 71% agreement with $r = 0.672$), whereas experienced faculty demonstrated superior consistency (80% agreement, $r = 0.936$). This divergence was especially pronounced on higher-order cognitive tasks, which has driven further research into methods for enhancing LLM assessment performance. To address these challenges, several studies have focused on structured rubrics and frameworks (see Appendix A.4.2). For example, Morjaria et al. (2024) found that when ChatGPT-4 was paired with question-specific rubrics, score inflation was reduced and the correlation with human reviewers improved to between $r = 0.6$ and 0.7. In parallel, Yuan and Hu (2024) observed that Llama-UKP models, when provided with well-defined assessment criteria (see Table 2), achieved high agreement with human evaluators (Spearman $\rho = 0.843$). These results underscore that explicit, rubric-based guidance consistently leads to more interpretable and reliable feedback in diverse educational contexts.

In addition to structured frameworks, advanced prompting strategies have emerged as critical tools for optimization (see Appendix A.1). The Reason-Act-Evaluate" (RAE) prompt introduced by Li et al. (2024) structures the assessment process into three clearly defined stages: reasoning about criteria, performing an assessment, and reviewing the outcome (see Appendix A.4.3). When applied to 1,235 student-generated texts, the RAE method achieved 76.5% accuracy and demonstrated strong alignment in dimensions such as logical reasoning ($\rho = 0.824$). This approach not only mirrors human grading practices but also significantly boosts the overall reliability of LLM outputs without necessarily relying on cutting-edge architectures.

Furthermore, the most promising results have been observed when LLMs are integrated into hybrid human-AI systems. Tools such as EvalGen, developed by Shankar et al. (2024b), combine LLM-generated assessments with human oversight to mitigate challenges like criteria drift (refer to Appendix A.4). Similar hybrid approaches proposed by Sinha et al. (2023), Khan et al. (2023), and the "Assisted RAE" method by Li et al. (2024) reinforce the idea that human-AI collaboration can enhance assessment consistency and integrity while reducing individual grader workload.

Finally, comparative model insights reveal that while newer LLMs often outperform older ones, the overall effectiveness of an LLM-based assessment system depends more on the quality of prompting and implementation strategy than simply on model recency. Open-source models like Llama-UKP, when used with robust methods, can perform comparably to proprietary systems (Yuan and Hu, 2024). Complimenting this, Li's finding—that Assisted RAE achieved 76.5% accuracy—demonstrates that strategic prompt engineering can be just as influential as acquiring the latest

model updates (Li et al., 2024).

4.1 Essay Grading

On the widely-used ASAP dataset (Automated Student Assessment Prize, a collection of 17,043 student essays across eight prompts with expert human scores). (The Hewlett Foundation, 2012), performance varies significantly across model architectures and implementation approaches. See Appendix A.2 for a detailed description. Xiao et al. (2024)'s dual-process framework using LLaMA3-8B achieved Quadratic Weighted Kappa (QWK) scores of approximately 0.7, approaching state-of-the-art models (QWK = 0.79) while maintaining over 80% score consistency. Similarly, Tang et al. (2024) found that GPT-4 achieved moderate reliability (QWK=0.5677) with criteria-referenced prompts, though still below human reliability benchmarks (QWK=0.6573). In contrast, Kundu and Barbosa (2024)'s evaluation of ChatGPT on the same dataset showed weaker correlation with human scores (r=0.21–0.23), though Llama-3 models demonstrated 130–173% improvement over baseline metrics, highlighting the rapid evolution in open-source model capabilities for educational assessment.

Among the prompting methods, Jauhiainen and Garagorry Guerra (2024)'s implementation of verification-based chain-of-thought prompting (see Appendix A.1) with the RAG framework achieved remarkable consistency, with 68.7% of ChatGPT-4 grades remaining stable across multiple evaluations and 72.2% aligning closely with human assessments. This approach parallels Xiao et al. (2024)'s dual-process framework, which distinguishes between a "Fast Module" for rapid predictions and a "Slow Module" for detailed feedback when confidence is low—a design inspired by Kahneman's dual-processing theory (Kahneman, 2011). Both studies demonstrate how thoughtful prompt design can dramatically improve performance even without requiring the most advanced models, with Xiao's open-source implementation achieving a 35% win rate (see Table 5) when compared to GPT-4 explanations despite using the smaller LLaMA3-8B model. Tang et al. (2024) further established that lower temperature settings (0.0) consistently produced better human alignment across models, highlighting how parameter tuning complements prompt engineering in optimizing assessment quality.

Supporting our observation that LLMs in human grading workflows show particular promise, Xiao et al. (2024)'s human-AI experiments revealed that novice graders improved from QWK 0.53 to 0.66 (approaching expert-level performance of 0.71) when provided with LLM-generated feedback, while experts reached QWK 0.77 with AI assistance. These findings align with Farrokhnia et al. (2024)'s assertion that AI tools can effectively reduce teacher workload while maintaining assessment quality. The complementary relationship between human and AI evaluation extends beyond efficiency gains, with Kundu and Barbosa (2024) noting that humans and LLMs employ distinctly different evaluation criteria—humans prioritizing essay length (r=0.74) while LLMs focus more on technical elements like grammar—suggesting that hybrid approaches can provide more comprehensive assessment than either alone.

Interactive assessment frameworks represent an emerging frontier, moving beyond static grading toward dynamic, dialogue-based evaluation systems. Hong et al. (2024)'s CAELF (Contestable AI Evaluation with Logic and Feedback; see Appendix A.4) introduces a multi-agent framework that enables students to challenge grades through structured debate, with Teaching-Assistant Agents discussing essay quality while a Teacher Agent resolves conflicts using principles from computational argumentation (Dung, 1995). When tested on 500 critical thinking essays (Hugging Face, 2023), this approach improved interaction accuracy by 44.6% over GPT-4o while maintaining correct evaluations in 80-90% of cases. More importantly, the system admitted mistakes 10-20% more frequently than baselines, demonstrating improved metacognitive awareness. Human evaluators particularly praised the clarity and actionable nature of the feedback, aligning with advances in LLM-driven formative assessment (Dai et al., 2023).

4.2 Natural Sciences & Technology

In the natural sciences domain, Henkel et al. (2024a) demonstrated that GPT-4 achieved near-human performance on 1,710 K-12 short-answer questions from the Carousel dataset (Cohen's $\kappa = 0.70$ compared to human $\kappa = 0.75$), with metrics of 85% accuracy, 0.87 precision, and 0.85 recall. In contrast, GPT-3.5 only reached a κ of 0.45, highlighting rapid advancements between model generations. Similarly, Tobler (2024)'s GenAI-Based Smart Grading system attained strong alignment with human evaluators (Krippendorff's $\alpha = 0.818$,

95% CI [0.689, 0.926]) in university-level assessments. Further comparisons by Latif and Zhai (2024) revealed that a fine-tuned GPT-3.5-turbo outperformed BERT across six scientific tasks, particularly excelling in multi-class (10.6% improvement) and unbalanced multi-label scenarios. Meanwhile, Wu et al. (2024)'s work on the open-source Mixtral-8x7B-instruct model showed moderate rubric alignment (F1=0.752) and a scoring accuracy of 54.58%. Their "Full-shot + Holistic Rubrics" prompting strategy outperformed both human-created rubrics (50.41%) and non-rubric baselines (33.5%), underscoring the impact of structured prompting on assessment quality.

Notably, efficiency gains in science education are compelling. GPT-4 completed evaluations of 1,710 short-answer questions in approximately 2 hours, compared to 11 hours for manual grading (Henkel et al., 2024a), and Tobler (2024)'s system also demonstrated significant time savings in university-level assessments. Overall, these findings indicate that carefully structured, rubric-based prompts and advanced LLM architectures not only enhance performance but also offer substantial efficiency improvements in scientific assessments.

4.3 Social Sciences & Humanities

Lundgren (Lundgren, 2024) and Kostic (Kostic et al., 2024) evaluated GPT-4 in advanced humanities assessments using distinct approaches. Lundgren's study of master-level political science essays showed that GPT-4's mean scores (approximately 5.03–5.60) generally aligned with human scores (around 4.95), although interrater reliability was very low (Cohen's $\kappa \leq 0.18$, $\leq 35\%$ agreement). In contrast, Kostic's assessment of German-language business transfer assignments revealed that GPT-4 produced markedly different scores from human evaluators (e.g. 52/50/60 vs. an average human score of about 26). Furthermore, Kooli and Yusuf (Kooli and Yusuf, 2024) reported moderate positive correlations between LLM and human grading (Pearson $r = 0.46$, Spearman $r = 0.518$, $p = 0.008$) for open-ended exam responses, while Pinto et al. (Pinto et al., 2023) observed strong LLM performance on structured exam grading. These results suggest that LLMs tend to evaluate well-defined, bounded responses more reliably than extended analytical writing, which requires nuanced human interpretation.

In addition, GPT-4 appears to prioritize evaluation criteria differently from human graders by favoring middle-range grades and language quality over the extremes preferred by human evaluators who emphasize analytical depth. Kostic also noted that human evaluators vary widely due to factors such as fatigue and subjectivity, a variability not observed in LLM scoring. Such complementary characteristics indicate the potential for hybrid approaches that integrate human expertise with the computational consistency of LLMs (Williamson et al., 2012).

4.4 Computer Science

Xie et al. (2024) developed a framework that employs LLMs for rubric generation, initial grading, and post-grading review. Their system iteratively refines rubrics using sampled student responses from the OS and Mohler datasets, and employs group comparisons to enhance assessment consistency. This approach parallels Grandel et al. (2024)'s GreAIter system, which achieved a grading accuracy of 98.21% while reducing grading time by 81.2% for programming assignments.

Performance evaluations across different LLM architectures show that both proprietary and open-source models can yield competitive outcomes. Yousef et al. (2025)'s BeGrading system, based on fine-tuned open-source LLMs, demonstrated only a 19% absolute difference relative to the benchmark Codestral model when grading programming assignments. Similarly, Koutcheme et al. (2024) and Smolić et al. (2024) found that models such as CodeLlama, Zephyr, GPT-3.5, and Gemini offer useful insights and perform comparably in providing feedback on programming assignments.

Targeted prompt engineering, like all other domains has also significantly impacted computer science. Tian et al. (2024)'s systematic evaluation of four prompting strategies revealed that few-shot-rubric prompting consistently outperformed zero-shot approaches, with strong agreement observed for criteria such as Greet Intent (QWK = 0.698) and Default Fallback Intent (QWK = 0.797). These findings are supported by Duong and Meng (2024), who demonstrated that combining GPT-4 with few-shot prompting and Retrieval Augmented Generation (RAG)[1] achieved the highest performance (Pearson correlation of 0.844).

[1] See Appendix A.4 for details on RAG implementation.

4.5 Mathematics

In Mathematics, multi-agent systems represent a particularly promising direction, exemplified by (Chu et al., 2024)'s GradeOpt framework, which employs three specialized LLM-based agents—Grader, Reflector, and Refiner—working in concert to optimize mathematics assessment. When evaluated on a dataset of 1,218 teacher responses to five mathematics questions, the GPT-4o-powered system achieved impressive performance metrics (0.85 accuracy and 0.73 Kappa). Further testing on an expanded dataset of 6,541 responses demonstrated significant improvement on specific questions, enhancing accuracy from 0.70 to 0.78 and Kappa scores from 0.52 to 0.64. We also see prompting strategies significantly influencing LLM performance in mathematics assessment, with chain-of-thought approaches demonstrating particularly strong results. Henkel et al. (2024b)'s comprehensive evaluation using the AMMORE dataset (53,000 question-answer pairs from African middle school students) compared six different grading methods ranging from simple string matching to sophisticated chain-of-thought prompting with GPT-4. The results showed that chain-of-thought prompting excelled particularly on challenging edge cases, achieving 92% accuracy where other methods struggled and boosting overall accuracy from 98.7% to 99.9%. This approach yielded impressive precision (0.97), recall (0.98), and F1 scores (0.98), demonstrating how well-designed prompting strategies can substantially enhance mathematics assessment quality. When implemented within a Bayesian Knowledge Tracing framework ($P(L_0) = 0.4$, $P(T) = 0.05$, $P(S) = 0.299$, $P(G) = 0.299$), these improvements translate to more accurate student mastery estimation, highlighting the practical educational value of such advancements. In mathematics assessment, GPT-4 in particular and its variants demonstrate particularly strong performance across multiple studies and assessment contexts, from GradeOpt's 0.85 accuracy on teacher responses to Henkel's 99.9% accuracy with chain-of-thought prompting on middle school mathematics. These results consistently outperform traditional NLP approaches like SBERT and RoBERTa as demonstrated in Chu's comparative evaluation. The performance advantage appears most pronounced when LLMs are implemented with sophisticated prompting strategies or multi-agent architectures, suggesting that continued advances in implementation methods may yield further improvements even with existing model architectures.

5 Discussion and Analysis

5.1 The Explainability Imperative

A critical consideration for the widespread adoption of LLM-based assessment is the fundamental need for explainable decisions in educational contexts. Unlike other AI applications, educational assessment directly impacts student learning, progression, and opportunities, making transparency not just desirable but essential. Students require clear explanations of their grades to understand learning gaps and improve performance, while educators need interpretable feedback to guide instructional decisions. The current "black box" nature of leading LLMs presents a significant barrier to educational adoption, as stakeholders cannot adequately justify or contest assessment decisions.

5.2 Patterns in LLM Assessment Performance

The reviewed studies reveal substantial variations in LLM assessment performance across academic disciplines. Figure 2 shows that mathematics and general education yield high human–LLM agreement rates (0.74 and 0.72, respectively), whereas humanities assessments exhibit notably lower alignment (0.46)—a pattern that mirrors our observation that structured formats (see Table 2) offer clearer evaluation criteria than open-ended tasks.

GPT-4 consistently outperforms earlier models in well-structured contexts (Henkel et al., 2024a; Chu et al., 2024; Henkel et al., 2024b), yet its reliability diminishes on complex, subjective tasks. For example, in political science essays, Lundgren (2024) observed that despite similar mean scores, GPT-4 showed very low interrater reliability (Cohen's $\kappa \leq 0.18$, $\leq 35\%$ agreement). Likewise, Kostic et al. (2024) reported that GPT-4 produced scores that differed dramatically from human evaluators in business administration assessments.

Human–LLM agreement studies consistently report moderate alignment: Jauhiainen and Garagorry Guerra (2024) found that 72.2% of GPT-4 grades differ by at most one grade from human scores, while Teckwani et al. (2024) noted 71% agreement between LLMs and human graders compared to 80% among humans. Tobler et al. (2024) achieved strong alignment (Krippendorff's $\alpha = 0.818$), though with notable qualitative differences

in rubric interpretation. Comparative evaluations further reveal that, while GPT-4 attains high reliability with criteria-referenced prompts (QWK = 0.5677; Tang et al. (2024)) it still falls slightly short of human benchmarks (QWK = 0.6573). Similarly, Xiao et al. (2024) demonstrated that LLaMA3-8B achieved QWK scores near 0.7 with 80% score consistency, and Morjaria et al. (2024) reported moderate to good correlations (r = 0.6–0.7) in medical education, despite discrepancies in 65–80% of cases.

Figure 3 reveals that single LLM approaches dominate current research (50%), while emerging alternatives such as multi-agent frameworks (10%) and chain-of-thought implementations (6.7%) show superior performance. Overall, these findings suggest that although the gap between human and LLM assessment is narrowing in structured domains, significant differences persist in evaluating complex, open-ended tasks due to varying evaluation approaches and priorities.

5.3 Methodological Approaches and Their Effectiveness

The literature reveals an evolution in prompting techniques, with more sophisticated approaches consistently outperforming simpler implementations across diverse educational contexts. As shown in Figure 4, semi-automated (0.90) and chain-of-thought approaches (0.81) demonstrate the highest human-LLM agreement rates, substantially outperforming single LLM implementations (0.53). These findings align with our categorization of prompting strategies in A.1, where we distinguish between simple zero-shot implementations and more advanced approaches like chain-of-thought. Chain-of-thought and few-shot prompting strategies have proven significantly more effective than zero-shot implementations Wu et al. (2024); Tian et al. (2024); Henkel et al. (2024b) across multiple disciplines as explained in Section 4. Multi-agent frameworks Hong et al. (2024); Chu et al. (2024); Xie et al. (2024) represent another promising methodological direction, allowing for more sophisticated assessment processes that mimic human evaluation workflows, as described in A.4.

Similarly, context-aware approaches that incorporate domain-specific knowledge show particular promise for enhancing assessment quality. Retrieval-augmented generation (RAG) Duong and Meng (2024); Jauhiainen and Garagorry Guerra (2024), as detailed in A.4 has emerged as an effective technique for contextualizing assessments with relevant educational materials. While many people have totally relied on AI to score, we also see many Hybrid human-AI approaches. These approaches yield optimal results in educational assessment by leveraging the complementary strengths of both human evaluators and LLMs. As noted by Xiao et al. (2024), positioning LLMs as assistants rather than replacements enhances overall evaluation quality and efficiency. Kundu and Barbosa (2024) observed that humans and LLMs apply different evaluation criteria—humans prioritizing essay length (with $r = 0.74$) while LLMs focus on technical elements like grammar—suggesting that a combined approach offers a more comprehensive assessment. This complementarity is evident across disciplines; for instance, Morjaria et al. (2024) reported that GPT-4 showed moderate to good correlation with human assessors ($r = 0.6$–0.7) in medical education, yet discrepancies persisted, and Teckwani et al. (2024) further reinforced the importance of human oversight by finding that human graders demonstrated 80% agreement compared to 71% for LLMs, particularly on higher-order cognitive tasks.

Figure 5 reveals important relationships between assessment types and frameworks, with certain combinations demonstrating particular prevalence. Single LLM approaches dominate essay (3 studies) and short-answer assessment (3 studies), while more specialized frameworks like chain-of-thought appear primarily with mathematical problem solving. These patterns suggest domain-specific optimization of LLM implementation strategies, aligned with our categorization of assessment formats in 2.

6 Conclusion

This review shows that LLM applications in educational assessment are advancing rapidly across various disciplines. Our analysis of 30 studies suggests that these models can help reduce the grading workload while still maintaining quality, particularly in structured contexts—GPT-4, for instance, is already nearing human-level performance in mathematics and science assessments. Innovations like chain-of-thought prompting, multi-agent frameworks, and retrieval-augmented generation are proving to be game changers for improving assessment accuracy.

However, challenges remain. LLMs continue

to struggle with nuanced, subjective evaluations in the humanities and social sciences, and rubric adherence is inconsistent. Technical hurdles, such as processing handwritten responses and handling complex programming tasks, further complicate the picture. Overall, the evidence supports a hybrid human-AI approach: LLMs are most effective when they serve as helpful assistants—automating routine tasks and generating detailed feedback—while human experts handle the more complex evaluations.

While this review focuses specifically on educational assessment, the use of LLMs as evaluators ("LLM as a judge") is a rapidly growing area across multiple domains including legal document review, content moderation, and research evaluation. Our findings regarding prompting strategies, human-AI collaboration, and reliability challenges likely have broader applicability beyond education, suggesting opportunities for cross-domain learning and methodological transfer.

Looking forward, future research should emphasize domain-specific fine-tuning, standardize prompt engineering practices, and explore multimodal assessment strategies. Moreover, more classroom-based validation studies are needed to assess the long-term impact on learning outcomes. Despite the rapid progress in LLM technology, it is clear that human oversight remains essential for achieving high-quality educational assessment.

7 Limitations

7.1 Methodological Limitations

This review, while comprehensive within its scope, has several methodological limitations that should be acknowledged. Our analysis is based on *limited sample size and generalizability concerns*, as the review includes 30 studies, which may limit the generalizability of our findings, particularly regarding framework performance comparisons shown in Figure 3. Some framework categories are represented by only 1-2 studies, making it difficult to draw robust conclusions about their relative effectiveness. The small sample size is partly due to the nascent nature of LLM applications in educational assessment, with most research emerging only after 2022. Future reviews with larger sample sizes will be needed to validate these preliminary patterns and provide more statistically robust comparisons across framework types.

A significant issue affecting our analysis is *methodological heterogeneity* across the reviewed studies. The studies exhibit significant methodological diversity, using different datasets, evaluation metrics, experimental protocols, and LLM configurations. This heterogeneity limits direct comparability and complicates the generalization of findings across studies. For instance, studies within the same discipline often use different datasets (e.g., some essay grading studies use ASAP while others use proprietary datasets), making it challenging to attribute performance differences to framework choices versus dataset characteristics. Additionally, generative AI systems employ various decoding strategies (beam search, temperature settings, top-p sampling) that can significantly impact output quality and consistency, yet these technical parameters are inconsistently reported across studies.

Our organizational approach presents another methodological consideration. While our discipline-based organization provides domain-specific insights valuable for understanding how LLMs perform across different educational contexts, an alternative methodological organization (e.g., by prompting strategies, assessment types, or hybrid architectures) might have enabled different analytical perspectives and cross-cutting insights. This organizational choice may limit the visibility of methodological patterns that transcend disciplinary boundaries. Future reviews could explore cross-cutting methodological themes to complement the domain-specific patterns we identify.

The *under-representation of K-12 studies* in our review likely reflects both limited published research in this educational level and potential search strategy limitations. K-12 educational technology adoption often faces greater institutional barriers, ethical considerations, and regulatory requirements than higher education, potentially slowing research publication in this area. Additionally, our keyword strategy may have inadvertently favored higher education terminology, though we attempted to include broad terms like "educational assessment" and "K-12." This limitation suggests that our findings may be more applicable to higher education contexts, with K-12 applications requiring additional targeted research.

Another concern affecting the quality of our analysis is *publication quality variability*. Over one-third of the reviewed studies (11 out of 30, or 37%) are preprints that have not undergone formal peer review. While preprints provide valuable insights into cutting-edge research and emerging trends in

LLM-based educational assessment, their inclusion introduces potential quality variability to our analysis. Preprints may contain methodological limitations, incomplete evaluations, or preliminary findings that could change during the peer review process. This limitation is particularly relevant given the rapidly evolving nature of LLM technology, where researchers often share findings quickly through preprint servers to keep pace with technological advances.

7.2 Challenges from the Literature

Several recurring challenges emerge from the literature that must be addressed before widespread educational adoption of LLM assessment systems can occur. A prominent issue is Rubric adherence problems. While Kostic et al. (2024) report poor adherence to assessment criteria in business evaluations despite explicit rubrics, Tobler (2024) observed that AI sometimes adheres more strictly to rubrics than humans, indicating divergent interpretations (see Appendix A.4.2).

Another critical limitation is the *inadequate description of human grader characteristics*. Approximately 40% of studies classify human evaluators as "Unspecified Graders" (see Table 4), making it difficult to contextualize performance metrics and understand the influence of grader expertise—as exemplified by Xiao et al. (2024)'s finding that novice graders scored significantly lower (QWK of 0.53) compared to experts (QWK of 0.71).

A further challenge is the persistence of *grading inconsistencies* across domains. For example, Lundgren (2024) found that GPT-4 exhibits a central tendency bias (favoring middle grades) in political science essays, while Kooli and Yusuf (2024) reported that ChatGPT is more conservative in social science assessments. Similarly, Smolić et al. (2024) noted discrepancies between LLM-provided numerical grades and human standards in programming, highlighting challenges in aligning qualitative feedback with quantitative accuracy.

Significant *technical limitations* also remain for processing specialized content and complex assessment scenarios. Liu et al. (2024) encountered OCR issues in handwritten mathematics, with false positives averaging 27%, and model architecture continues to affect reliability—larger models like GPT-4 consistently outperform smaller ones, although improvements in open-source models and fine-tuning are narrowing this gap. Another concern is the *"black box" nature of commercial LLMs*, which raises issues of transparency and explainability in educational assessment. The proprietary models (e.g., GPT-3.5/4) offer little insight into their internal decision-making processes, complicating the justification of evaluation decisions. This also leads to *critical governance questions*, as institutions risk disruptions if vendor-controlled systems are modified or discontinued. While promising open-source alternatives (Yousef et al., 2025; Koutcheme et al., 2024) offer more transparent solutions, they require substantial technical capacity to implement and maintain.

There is also a notable limitation in the *scarcity of real-world, classroom-based implementations*, especially in K-12 contexts. Most studies are controlled experiments, raising concerns about ecological validity and practical challenges. Moreover, there is an *imbalanced focus on educational levels*, with over 60% of studies focusing on higher education while early childhood and primary education remain underexplored. This is particularly problematic given the distinct developmental, pedagogical, and ethical requirements for younger learners, where *ethical and privacy considerations* are especially pronounced. Collectively, these challenges call for further research into standardized methods, transparent AI, and real-world strategies to bridge the gap between experimental promise and practical assessment.

References

Carousel Learning. 2024. Carousel short answer dataset. Carousel Learning Platform.

Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. 2024. A llm-powered automatic grading framework with human-level guidelines optimization. *arXiv preprint arXiv:2410.02165*.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pages 323–325. IEEE.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Ta Nguyen Binh Duong and Chai Yi Meng. 2024. Automatic grading of short answers using large language models in software engineering courses. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10. IEEE.

B. J. England, J. R. Brigati, E. E. Schussler, and M. M. Chen. 2019. Student anxiety and perception of difficulty impact performance and persistence in introductory biology courses. *CBE—Life Sciences Education*, 18(2):ar21.

Mohammadreza Farrokhnia, Seyyed Kazem Banihashem, Omid Noroozi, and Arjen Wals. 2024. A swot analysis of chatgpt: Implications for educational practice and research. *Innovations in education and teaching international*, 61(3):460–474.

Johnbenetic Gnanaprakasam and Ravi Lourdusamy. 2024. The role of ai in automating grading: Enhancing feedback and efficiency. In *Artificial Intelligence and Education-Shaping the Future of Learning*. IntechOpen.

Skyler Grandel, Douglas C Schmidt, and Kevin Leach. 2024. Applying large language models to enhance the assessment of parallel functional programming assignments. In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 102–110.

Jari J Hakanen, Arnold B Bakker, and Wilmar B Schaufeli. 2006. Burnout and work engagement among teachers. *Journal of school psychology*, 43(6):495–513.

Sakib Haque, Zachary Eberhart, Aakash Bansal, and Collin McMillan. 2022. Semantic similarity metrics for evaluating source code summarization. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, pages 36–47.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024a. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 300–304.

Owen Henkel, Hannah Horne-Robinson, Maria Dyshel, Nabil Ch, Baptiste Moreau-Pernet, and Ralph Abood. 2024b. Learning to love edge cases in formative math assessment: Using the ammore dataset and chain-of-thought prompting to improve grading accuracy. *arXiv preprint arXiv:2409.17904*.

Shengxin Hong, Chang Cai, Sixuan Du, Haiyue Feng, Siyuan Liu, and Xiuyi Fan. 2024. " my grade is wrong!": A contestable ai framework for interactive feedback in evaluating student essays. *arXiv preprint arXiv:2409.07453*.

Lory Hough. 2023. The problem with grading. *Ed. Magazine*. Accessed: 2025-03-16.

Hugging Face. 2023. Critical thinking essays dataset. Hugging Face Datasets Hub.

Jussi S Jauhiainen and Agustín Garagorry Guerra. 2024. Generative ai in education: Chatgpt-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, pages 1–18.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, USA.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Rehan Ahmed Khan, Masood Jawaid, Aymen Rehan Khan, and Madiha Sajjad. 2023. Chatgpt-reshaping medical education and clinical management. *Pakistan journal of medical sciences*, 39(2):605.

Chokri Kooli and Nadia Yusuf. 2024. Transforming educational assessment: Insights into the use of chatgpt and large language models in grading. *International Journal of Human–Computer Interaction*, pages 1–12.

Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. 2024. Llms in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147.

Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. 2024. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 52–58.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders? *arXiv preprint arXiv:2409.13120*.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Kangkang Li, Chengyang Qian, and Xianmin Yang. 2024. Evaluating the quality of student-generated content in learnersourcing: A large language model based approach. *Education and Information Technologies*, 30:2331–2360.

Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. 2024. Ai-assisted automated short answer grading of handwritten university level mathematics exams. *arXiv preprint arXiv:2408.11728*.

Magnus Lundgren. 2024. Large language models in student assessment: Comparing chatgpt and human graders. *arXiv preprint arXiv:2406.16510*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Leo Morjaria, Levi Burns, Keyna Bracken, Anthony J Levinson, Quang N Ngo, Mark Lee, and Matthew Sibbald. 2024. Examining the efficacy of chatgpt in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, 3(1):32–43.

Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. In *Proceedings of the XXXVII brazilian symposium on software engineering*, pages 293–302.

Shreya Shankar, Haotian Li, Parth Asawa, Madelon Hulsebos, Yiming Lin, JD Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G Parameswaran, and Eugene Wu. 2024a. Spade: Synthesizing data quality assertions for large language model pipelines. *arXiv preprint arXiv:2401.03038*.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024b. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Ranwir K Sinha, Asitava Deb Roy, Nikhil Kumar, Himel Mondal, and Ranwir Sinha. 2023. Applicability of chatgpt in assisting to solve higher order problems in pathology. *Cureus*, 15(2).

Einar M Skaalvik and Sidsel Skaalvik. 2017. Dimensions of teacher burnout: Relations with potential stressors at school. *Social Psychology of Education*, 20:775–790.

Ema Smolić, Marko Pavelić, Bartol Boras, Igor Mekterović, and Tomislav Jagušt. 2024. Llm generative ai and students' exam code evaluation: Qualitative and quantitative analysis. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 1261–1266. IEEE.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14).

Swapna Haresh Teckwani, Amanda Huee-Ping Wong, Nathasha Vihangi Luke, and Ivan Cherh Chiet Low. 2024. Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. *Advances in Physiology Education*, 48(4):904–914.

The Hewlett Foundation. 2012. Automated student assessment prize (asap) dataset. Kaggle.

Xiaoyi Tian, Amogh Mannekote, Carly E Solomon, Yukyeong Song, Christine Fry Wise, Tom Mcklin, Joanne Barrett, Kristy Elizabeth Boyer, and Maya Israel. 2024. Examining llm prompting strategies for automatic evaluation of learner-created computational artifacts. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 698–706.

Samuel Tobler. 2024. Smart grading: A generative ai-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX*, 12:102531.

Catrin Wigfall. 2020. Grading standards do impact student achievement. Accessed: 2025-03-16.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Xuansheng Wu, Padmaja Pravin Saraf, Gyeong-Geon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2024. Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *arXiv preprint arXiv:2407.18328*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. Human-ai collaborative essay scoring: A dual-process framework with llms. *arXiv preprint arXiv:2401.06431*.

Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2024. Grade like a human: Rethinking automated assessment with large language models. *arXiv preprint arXiv:2405.19694*.

Mina Yousef, Kareem Mohamed, Walaa Medhat, Ensaf Hussein Mohamed, Ghada Khoriba, and Tamer Arafa. 2025. Begrading: large language models for enhanced feedback in programming education. *Neural Computing and Applications*, 37(2):1027–1040.

Bo Yuan and Jiazi Hu. 2024. An exploration of higher education course evaluation by large language models. *arXiv preprint arXiv:2411.02455*.

A Terminology and Definitions

This appendix provides comprehensive definitions for key terms, methodologies, and frameworks referenced throughout this review, offering detailed context beyond the condensed definitions in the main text.

A.1 Prompting Strategies

A.1.1 Zero-shot prompting

Direct instruction to the LLM to perform an assessment task without providing examples or demonstrations. The model relies entirely on its pretraining knowledge to understand the assessment criteria and generate appropriate evaluations. Tang et al. (2024) examined this approach for essay evaluation, finding it achieved lower reliability

(QWK=0.4321) compared to few-shot approaches. Zero-shot prompting represents the simplest implementation but typically under-performs more sophisticated strategies, particularly for complex or domain-specific assessments.

A.1.2 Few-shot prompting

Providing the LLM with a limited number (typically 3-6) of example question-answer pairs and their corresponding evaluations before asking it to assess new responses. This approach establishes a pattern for the model to follow when generating its own assessments. Tian et al. (2024) demonstrated that few-shot-rubric prompting consistently outperformed zero-shot approaches when assessing chatbot projects, with particularly strong performance in structured dimensions like Greet Intent (QWK=0.698) and Default Fallback Intent (QWK=0.797). Duong and Meng (2024) found that GPT-4 with 6 examples achieved a Pearson correlation of 0.694 with human graders, substantially outperforming simpler implementations.

A.1.3 Chain-of-thought prompting

Guiding the LLM to articulate step-by-step reasoning before providing a final assessment, mimicking human cognitive processes in evaluation. This approach is particularly effective for mathematical and logical evaluations requiring multi-step reasoning. Henkel et al. (2024b) used this method on the AMMORE dataset of 53,000 question-answer pairs from African middle school students, showing it increased mathematics assessment accuracy from 98.7% to 99.9% and was especially effective for complex edge cases (92% accuracy where other methods struggled). While more computationally intensive than simpler prompting methods, chain-of-thought approaches consistently demonstrate superior performance for complex assessment tasks requiring logical reasoning.

A.1.4 Reason-Act-Evaluate (RAE) prompting

A structured three-stage process where the LLM first reasons about assessment criteria (contemplating evaluation dimensions and standards), then performs the actual assessment (applying these criteria to the student response), and finally reviews its own assessment for accuracy, consistency, and adherence to rubrics. Li et al. (2024) developed this approach for evaluating student-generated content, achieving 76.5% accuracy across 1,235 articles with particularly strong performance in structured dimensions like logical reasoning ($\rho = 0.824$). This technique incorporates meta-cognitive awareness into the assessment process, enabling self-correction and improved reliability.

A.1.5 Rubric-guided prompting

Explicitly incorporating detailed assessment rubrics into LLM prompts, providing structured evaluation criteria that guide the model's judgment. This approach improves alignment with human evaluation standards by making assessment criteria explicit rather than implied. Morjaria et al. (2024) found this approach significantly reduced score inflation tendencies when using ChatGPT-4 to evaluate medical students' short-answer assessments, achieving moderate to good correlation (r=0.6-0.7) with human assessors. Similarly, Yuan and Hu (2024) demonstrated that rubric incorporation enabled Llama-UKP models to achieve remarkable correlation with human evaluators (Spearman: 0.843) when assessing higher education courses.

A.2 Dataset

A.2.1 ASAP dataset

The Automated Student Assessment Prize dataset, released by the Hewlett Foundation in 2012 (The Hewlett Foundation, 2012), containing 17,043 student essays across eight distinct prompts with expert human scores. Each prompt represents a different essay type (e.g., persuasive, source-based, narrative) and grade level (ranging from grade 7 to 10), with varying length requirements and scoring scales. This comprehensive collection has become the standard benchmark for automated essay scoring systems, enabling direct comparison of different approaches. Studies by Xiao et al. (2024), Tang et al. (2024), and Kundu and Barbosa (2024) used this dataset to evaluate LLM essay assessment capabilities, with Xiao et al. (2024)'s implementation achieving QWK scores of approximately 0.7, approaching state-of-the-art performance (QWK 0.79).

A.2.2 ASAP++ dataset

An extension of the original ASAP dataset developed by Mathias and Bhattacharyya (2018) that enriches the essays with additional attribute scores beyond the holistic ratings in the original dataset. These attributes include content, organization, word choice, sentence fluency, conventions, and prompt adherence. Kundu and Barbosa (2024) used this enhanced dataset to evaluate LLM assessment capabilities across multiple dimensions of

writing quality, providing more nuanced analysis of model performance on different aspects of essay evaluation.

A.2.3 Carousel dataset

A collection of 1,710 K-12 short-answer questions from science and history subjects developed by Carousel Learning (Carousel Learning, 2024). The dataset includes multiple student responses to each question along with expert human evaluations based on detailed rubrics. Questions span multiple grade levels and subject areas, providing a diverse testbed for short-answer assessment capabilities. Henkel et al. (2024a) used this dataset to evaluate GPT-4's performance on K-12 short-answer grading, finding near-human performance (Cohen's $\kappa = 0.70$ compared to human $\kappa = 0.75$).

A.2.4 AMMORE dataset

The African Middle-school Math Open Response Evaluation dataset contains 53,000 question-answer pairs from African middle school students across multiple mathematical topics. This comprehensive collection includes diverse response formats and challenging edge cases that test the limits of automated assessment capabilities, including unconventional solution methods and partial understanding demonstrations. Henkel et al. (2024b) used this dataset to evaluate various assessment approaches, finding that chain-of-thought prompting achieved 99.9% overall accuracy and 92% accuracy on challenging edge cases where simpler methods struggled.

A.2.5 Mohler dataset

A computer science short-answer dataset containing 2,273 student responses to technical questions with expert human grades. This dataset features specialized computer science content requiring domain-specific knowledge for accurate assessment, including algorithm descriptions, theoretical explanations, and applied problem-solving. Xie et al. (2024) and Duong and Meng (2024) used this dataset to evaluate LLM performance on computer science assessment, with Duong and Meng (2024) achieving a Pearson correlation of 0.694 using GPT-4 with few-shot prompting.

A.2.6 OS dataset

A dataset of operating systems concept questions and student responses used by Xie et al. (2024) to evaluate their multi-agent assessment system. This specialized collection focuses on technical computer science concepts and includes varied response types requiring domain-specific knowledge for accurate evaluation. The dataset exemplifies the challenges of assessing technical subject matter where specialized terminology and conceptual precision are essential for accurate evaluation.

A.3 Framework Definitions

- **Mixed-initiative**: Systems combining human and AI decision-making with dynamic role allocation
- **OCR+LLM**: Optical Character Recognition integrated with Large Language Models for handwritten content
- **Semi-automated**: Human-AI collaborative systems where AI provides initial assessment subject to human review
- **Multi-agent**: Multiple LLM instances with specialized roles working collaboratively

A.4 Specialized Concepts

A.4.1 Criteria drift

The phenomenon is where evaluation standards evolve or shift during the assessment process, potentially compromising consistency and fairness. This can occur with both human and LLM evaluators and represents a significant challenge for maintaining reliable assessment standards. Shankar et al. (2024a) identified this as a fundamental challenge in LLM assessment, where initial evaluation criteria may be applied differently to later responses. Criteria drift manifests in several forms:

- **Standard inflation/deflation**: Gradual shifting of grading standards to become more lenient or strict over time.
- **Criteria reinterpretation**: Subtle changes in how specific rubric elements are interpreted across different responses.
- **Priority shifting**: Changes in the relative importance assigned to different evaluation criteria during the assessment process.
- **Context effects**: Earlier responses influencing the evaluation of later responses through comparative judgment rather than fixed standards.

Addressing criteria drift requires explicit metacognitive awareness and structured review processes, which multi-agent LLM frameworks like Chu et al. (2024)'s GradeOpt implement through specialized roles such as the "Reflector" agent dedicated to consistency monitoring.

A.4.2 Rubric-based approach

Assessment methodologies that employ structured evaluation frameworks with explicitly defined criteria and performance levels to ensure consistent, transparent evaluation. In LLM assessment, rubric-based approaches involve providing models with these structured frameworks to guide evaluation. Key elements include:

- **Dimension specification**: Clearly identified aspects of performance to be evaluated (e.g., content coverage, organizational structure, technical accuracy, language use).
- **Performance descriptors**: Explicit descriptions of what constitutes different quality levels for each dimension, typically ranging from excellent to unsatisfactory.
- **Weighting schemes**: Optional specifications regarding the relative importance of different dimensions in the overall assessment.
- **Scoring mechanics**: Clear instructions on how to convert qualitative judgments into numerical scores, ensuring consistent quantification of performance.

Studies by Morjaria et al. (2024), Wu et al. (2024), and Yuan and Hu (2024) demonstrated that incorporating detailed rubrics significantly improved LLM assessment alignment with human evaluation, particularly for complex responses requiring multi-dimensional evaluation. Morjaria et al. (2024) specifically found that rubric incorporation reduced ChatGPT-4's tendency toward score inflation in medical education contexts.

A.4.3 Assisted RAE approach

An enhancement to the basic Reason-Act-Evaluate framework developed by Li et al. (2024) that incorporates metadata analysis and additional contextual information to improve assessment quality. This approach augments the three-stage RAE process (reasoning about criteria, performing assessment, evaluating quality) with supplementary information about the assessment context, student characteristics, or relevant educational standards. The assisted version achieved 76.5% accuracy when evaluating student-generated content across 1,235 articles, with particularly strong performance in structured dimensions like logical reasoning ($\rho = 0.824$).

A.4.4 CAELF framework

Contestable AI Evaluation with Logic and Feedback, a multi-agent framework developed by Hong et al. (2024) that enables students to challenge AI-generated grades through structured debate. The system employs teaching assistant agents for initial evaluation and discussion of contested grades, while a teacher agent resolves conflicts using principles from computational argumentation theory (Dung, 1995). When tested on 500 critical thinking essays, this approach improved interaction accuracy by 44.6% over GPT-4o alone, maintained correct evaluations in 80-90% of cases, and admitted mistakes 10-20% more frequently than baselines, demonstrating improved metacognitive awareness.

A.4.5 Retrieval-Augmented Generation (RAG)

Enhancing LLM evaluation by retrieving and incorporating relevant reference materials from external sources to contextualize the assessment. This approach integrates domain-specific knowledge beyond the model's training data, improving performance on specialized subjects. Duong and Meng (2024) applied this method to software engineering course assessment, dramatically improving Pearson correlation from 0.694 to 0.844 by incorporating course materials into the evaluation process. RAG implementations are particularly valuable for domain-specific assessments where specialized knowledge or context is essential for accurate evaluation.

A.4.6 Multi-agent frameworks

Using multiple specialized LLM instances that perform different aspects of the assessment process in collaboration, mimicking human evaluation workflows with distinct roles. These frameworks typically include components like initial graders, reviewers, and arbitrators that communicate to produce a refined assessment. Hong et al. (2024)'s CAELF framework exemplifies this approach, employing teaching assistant agents for initial evaluation and a teacher agent to resolve conflicts, improving interaction accuracy by 44.6% over single-agent approaches. Similarly, Chu et al. (2024)'s

GradeOpt employed three distinct agents—grader, reflector, and refiner—working collaboratively to achieve 0.85 accuracy and 0.73 Kappa in mathematics assessment. While more complex to implement, multi-agent frameworks consistently demonstrate superior performance, particularly for nuanced assessment tasks requiring multiple perspectives.

A.4.7 Automated Short Answer Grading (ASAG)

A field focused on using computational methods to automatically evaluate student responses to short-answer questions. ASAG systems typically analyze the semantic content of responses against reference answers or rubrics to determine correctness, completeness, and relevance. LLM-based ASAG frameworks like GradeOpt (Chu et al., 2024) represent advanced approaches that can evaluate nuanced understanding beyond simple keyword matching.

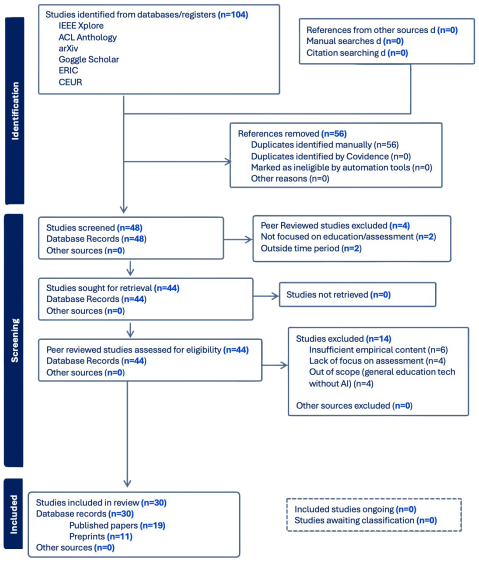

Figure 1: PRISMA flow diagram showing the study selection process.

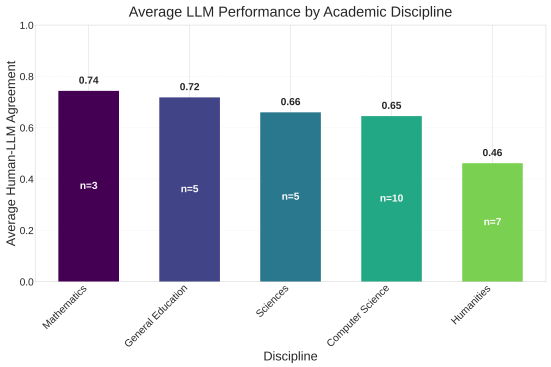

Figure 2: Average LLM Performance by Academic Discipline.

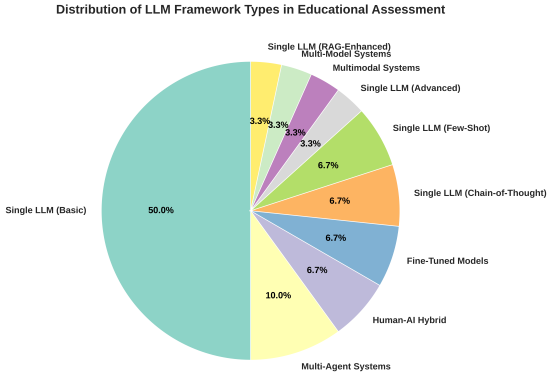

Figure 3: Distribution of LLM Framework Types in Educational Assessment (see Appendix A.3).

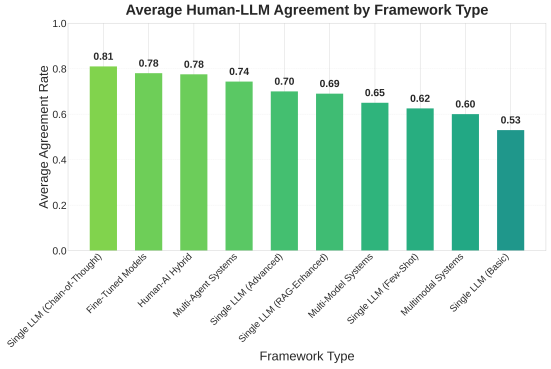

Figure 4: Average Human-LLM Agreement by Framework Type.

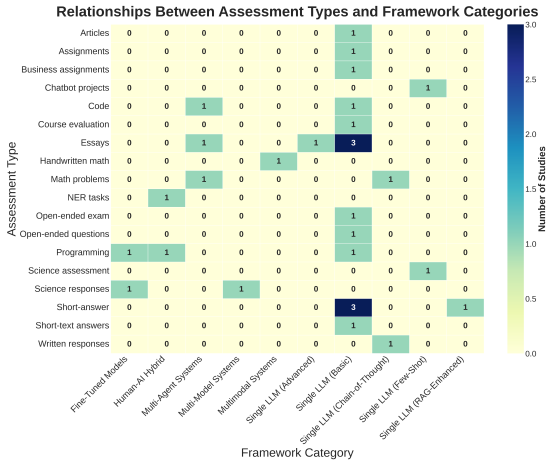

Figure 5: Relationships Between Assessment Types and Frameworks. Cell values represent the number of studies using each assessment type-framework combination (0 = no studies, 1 = one study, 2 = two studies, etc.).

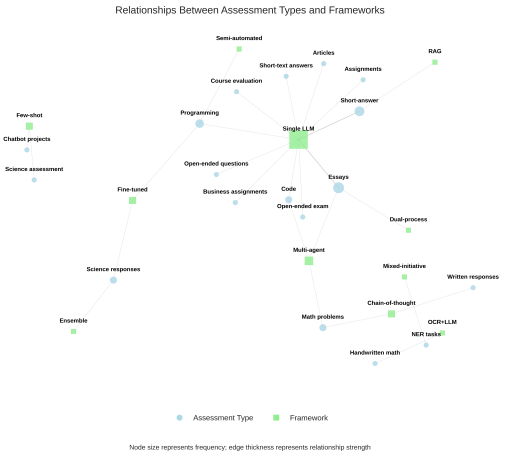

Figure 6: Network Visualization of Assessment Types and Frameworks Relationships.

Table 1: Search sources and terms used to extract peer-reviewed scientific literature on large language models in educational assessment (January 1, 2022 – January 14, 2025).

Source	Date of Search	Search Terms
Google Scholar	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")
arXiv	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")
IEEE Xplore	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")
ACL Anthology	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")
ERIC (Education Resources Info Center)	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")
CEUR	Jan 14, 2025	("large language models" OR "educational assessment" OR "automated grading" OR "essay scoring" OR "student feedback" OR "ChatGPT" OR "GPT-4" OR "short answer evaluation" OR "AI in education")

Table 2: Assessment Types in LLM Evaluation Studies.

Assessment Type	Description	Typical Evaluation Method
Multiple-Choice Questions (MCQs)	Structured questions with predefined answer options where students select from available choices.	Binary correctness evaluation; typically automated using answer keys
Short-Answer Questions	Brief text responses (typically 1-5 sentences) addressing specific, bounded questions with relatively constrained correct answers.	Rubric-based evaluation against expected key concepts or knowledge points
Essay Assessments	Extended written responses (typically >300 words) requiring development of arguments, analysis, or synthesis of information.	Multi-dimensional rubrics evaluating content quality, structure, argumentation, and language use
Programming Assignments	Code writing tasks requiring functional implementation of algorithms or solutions to computational problems.	Evaluation of correctness, efficiency, readability, and adherence to programming standards
Mathematics Assessments	Problems requiring mathematical reasoning, calculation, and demonstration of procedural or conceptual understanding.	Step-by-step evaluation of solution process, correctness, and mathematical reasoning
Handwritten Assessments	Written responses composed by hand rather than digitally, requiring OCR processing before LLM evaluation.	Content evaluation following digitization; may involve image processing and character recognition

Table 3: Education Levels in LLM Evaluation Studies.

Category	**Typical Age Range**	**Description**	**Examples in Studies**
Early Education	Ages 3-5	Pre-primary education including kindergarten and preparatory programs	Limited representation in current studies
Primary Education	Ages 6-10	Elementary school (grades 1-5 in many systems)	Wu et al. (2024), Henkel et al. (2024a)
Secondary Education	Ages 11-18	Middle and high school (grades 6-12 in many systems)	Henkel et al. (2024a), Latif and Zhai (2024)
Undergraduate Education	Ages 18-22	Bachelor's degree programs and equivalent tertiary education	Yuan and Hu (2024), Tobler (2024), Kooli and Yusuf (2024)
Graduate Education	Ages 22+	Master's and doctoral programs	Lundgren (2024), Morjaria et al. (2024)
Professional Education	Various (typically 18+)	Specialized training for specific professions (medical, engineering, etc.)	Morjaria et al. (2024), Sinha et al. (2023)

Table 4: Human Annotator Categories in LLM Evaluation Studies.

Category	**Definition**	**Typical Characteristics**
Expert Evaluators	Individuals with advanced qualifications and substantial experience in the subject matter and assessment context	PhD or equivalent qualification; 5+ years teaching/evaluation experience; specialized domain knowledge
Experienced Educators	Teachers or instructors with formal teaching qualifications and moderate experience	Master's degree or equivalent; 2-5 years teaching experience; formal pedagogical training
Novice Evaluators	Individuals with basic subject knowledge but limited assessment experience	Bachelor's degree or equivalent; <2 years assessment experience; may include teaching assistants or student peers
Field Practitioners	Domain experts who may lack formal education qualifications but possess practical expertise	Industry experience; professional certifications; variable teaching experience
Unspecified Graders	Studies where human grader qualifications are not explicitly described	Unknown qualifications and experience levels; represents a methodological limitation in some studies

Table 5: Evaluation Metrics in LLM Assessment Studies.

Metric	**Description**	**Typical Interpretation**
Cohen's Kappa (κ)	Measures interrater reliability between two raters, accounting for agreement occurring by chance. Scale from -1 to 1, with 1 representing perfect agreement.	< 0.40: Poor agreement $0.40 - 0.75$: Fair to good > 0.75: Excellent agreement
Quadratic Weighted Kappa (QWK)	Extension of Cohen's Kappa that assigns different weights to disagreements based on their severity. Common in essay scoring evaluation.	Similar to Cohen's Kappa, but with increased sensitivity to disagreement magnitude
Krippendorff's Alpha (α)	Reliability coefficient suitable for multiple raters and various measurement levels. Ranges from 0 to 1.	< 0.67: Insufficient $0.67 - 0.80$: Tentative > 0.80: Reliable
Pearson Correlation (r)	Measures linear correlation between two variables. Ranges from -1 to 1.	< 0.40: Weak correlation $0.40 - 0.70$: Moderate correlation > 0.70: Strong correlation
Spearman Correlation (ρ)	Measures monotonic relationships between ranked variables. Useful for ordinal data like grades.	Similar to Pearson, but for ranked data
Accuracy	Percentage of correctly identified instances. Simple measure for classification tasks.	Context-dependent; higher is better
F1 Score	Harmonic mean of precision and recall. Balances false positives and false negatives.	0 to 1 scale; higher is better
Win Rate	Percentage of instances where the LLM's assessment is preferred over alternatives in comparative evaluations.	Context-dependent; used primarily in comparative studies

Table 6: Summary of LLM Educational Assessment Research.

Reference	Discipline / Subject	Data	Data Availability	Techniques	Results
Teckwani et al. (2024)	General Education	117 assignments aligned with Bloom's taxonomy	Not mentioned	LLM evaluation (GPT-3.5, GPT-4o, Gemini)	LLMs: moderate consistency (Gemini: 71%, $r = 0.672$); Human: superior reliability (80% agreement, $r = 0.936$). It was found that LLMs struggled with higher-order tasks; poor human alignment ($\leq$44%)
Morjaria et al. (2024)	Medical Education	Medical students' short-answer assessments	Not mentioned	ChatGPT-4 as grading assistant	Moderate to good correlation with humans ($r = 0.6$–0.7); Score discrepancies in 65–80% of cases. Including rubrics reduced ChatGPT's score inflation tendency
Yuan and Hu (2024)	Higher Education	100 Chinese university courses	Not mentioned	GPT-4o, Kimi, and Llama models	Llama-UKP had strong correlation with human evaluations (Spearman: 0.843)
Li et al. (2024)	Educational Content	1,235 student articles	Not mentioned	"Reason-Act-Evaluate" prompt with metadata analysis	76.5% accuracy. Strong correlation with expert evaluations in structured dimensions (logic: $\rho = 0.824$)
Shankar et al. (2024b)	General (NLP)	Medical transcripts and product descriptions	Not mentioned	EvalGen tool with GPT-4	Criteria drift identified. Furthermore, revealed interdependence of criteria and outputs
Xiao et al. (2024)	Essay Grading	ASAP dataset and private Chinese dataset	ASAP: Publicly available	Dual-process framework with LLaMA3-8B	QWK scores ($\sim$0.7) close to SOTA (QWK 0.79); $>$80% score consistency. Novices improved from QWK 0.53 to 0.66 with AI assistance
Hong et al. (2024)	Essay Grading	500 critical thinking essays	Publicly available (Hugging Face, 2023)	CAELF multi-agent framework	Improved interaction accuracy by 44.6% over GPT-4o. Maintained correct evaluations in 80–90% of cases
Kundu and Barbosa (2024)	Essay Grading	ASAP and ASAP++ datasets	Publicly available	ChatGPT and Llama models	Weak correlation with human scores (ChatGPT: $r = 0.21$–0.23). It was found that LLMs excel in error detection but prioritize different criteria than humans
Jauhiainen and Garagorry Guerra (2024)	General Education	54 student responses	Not mentioned	ChatGPT-4 with verification-based chain-of-thought	68.7% grade consistency; 72.2% alignment with humans. Discrepancies in the model are addressable through prompt refinement
Tang et al. (2024)	Essay Grading	ASAP dataset (1,730 essays)	Publicly available	GPT-3.5, GPT-4, Claude 2	GPT-4: highest reliability (QWK = 0.5677). Lower temperature settings (0.0) produced better human alignment

Table 6: Summary of LLM Educational Assessment Research (continued).

Reference	Discipline/Subject	Data	Data Availability	Techniques	Results
Henkel et al. (2024a)	K-12 Science/History	1,710 short-answer questions (Carousel dataset)	Publicly available	GPT-4 and GPT-3.5	GPT-4: near-human performance (Cohen's $\kappa = 0.70$ vs. human $\kappa = 0.75$). 85% precision, 0.87 precision, 0.85 recall; automated grading required 2 hours vs. 11 hours manually
Wu et al. (2024)	Physics (Middle School)	12 physics science assessment items	Not mentioned	Mixtral-8x7B-instruct with few-shot prompting	Best configuration achieved 54.58% scoring accuracy. Strong correlation between human-aligned rubrics and accurate grading
Tobler (2024)	General Education	29 university students' responses	Consent required	GenAI-Based Smart Grading with GPT-4	Strong alignment with human grading ($\alpha = 0.818$). Based on results from the study, AI exhibited stricter adherence to rubrics
Latif and Zhai (2024)	Science Education	2,600 middle/high school responses	Not mentioned	Fine-tuned GPT-3.5-turbo vs. BERT	GPT-3.5: mean precision of 0.915 vs. BERT: 0.838. GPT-3.5 showed strength in multi-class tasks (10.6% improvement)
Lundgren (2024)	Political Science	60 master-level essays	Not mentioned	GPT-4 with four prompt types	Low interrater reliability (Cohen's $\kappa \leq 0.18$). GPT-4 favored middle grades; detailed prompts didn't improve accuracy
Kostic et al. (2024)	Business Administration	German-language business assignments	Not mentioned	GPT-4 with three prompt variations	Unreliable grades (e.g., overscoring). This study revealed that the automated system displayed poor rubric adherence, and is inadequate for nuanced assessment
Kooli and Yusuf (2024)	Social Science	25 open-ended exam responses	Not mentioned	ChatGPT vs. human grader	Moderate positive correlation (Pearson $r = 0.46$). ChatGPT found to be more conservative and variable than humans
Xie et al. (2024)	Computer Science	OS and Mohler datasets	Not mentioned	Multi-agent system for rubric generation	Improved grading consistency. However, challenges in achieving complete fairness and rubric precision
Yousef et al. (2025)	Programming Education	Python and Java assignments	Not mentioned	BeGrading system with fine-tuned LLMs	19% absolute difference rate. Fine-tuning small models improved performance
Koutcheme et al. (2024)	Programming Education	Programming assignments	Not mentioned	CodeLlama and Zephyr	Zephyr models performed similarly to proprietary models. Open-source LLMs can offer meaningful student feedback

Table 6: Summary of LLM Educational Assessment Research (continued).

Reference	Discipline / Subject	Data	Data Availability	Techniques	Results
Smolić et al. (2024)	Programming Education	Student code submissions	Not mentioned	GPT-3.5 and Gemini	Useful insights for code review; numerical grades inconsistent with human standards
Schneider et al. (2023)	Computer Science	Short-text answers from university courses	Not mentioned	ChatGPT-3.5	Inconsistent grading; struggled with contextual understanding and course-specific knowledge
Duong and Meng (2024)	Software Engineering	Mohler Dataset (2,273 answers) and SE Dataset (421 answers)	Not mentioned	Embedding-based and completion-based methods	GPT-4 with 6 examples: Pearson correlation of 0.694. GPT-4 superior to GPT-3.5 but at higher cost
Grandel et al. (2024)	Programming Education	Programming assignments	Not mentioned	GreAIter semi-automated system with ChatGPT-4	98.21% grading accuracy; reduced grading time by 81.2%
Tian et al. (2024)	AI Education	75 chatbot projects	Not mentioned	GPT-4 with four prompting strategies	Good performance in some dimensions (QWK=0.698). Few-shot-rubric prompting outperformed zero-shot
Pinto et al. (2023)	Software Engineering	Responses to open-ended questions	Not mentioned	ChatGPT	Aligned with expert evaluations; good at identifying misunderstandings
Gao et al. (2023)	Mechanical Engineering	Quiz dataset (70 students) and Activity dataset (85–95 students)	Not mentioned	7 NLP models (BERT, T5, etc.)	PromCSE excelled in binary tasks. NLP models struggle with precision and complex questions
Chu et al. (2024)	Mathematics	1,218 teacher responses and 6,541 teacher responses	Not mentioned	GradeOpt multi-agent framework with GPT-4o	0.85 accuracy and 0.73 Kappa. More effective than traditional methods
Liu et al. (2024)	University Mathematics	Handwritten calculus exam (54 students)	Consent required	GPT-4 with Mathpix and GPT-4V for OCR	Accuracy: 0.59 to 0.62. Whole-page OCR outperformed answer-box methods
Henkel et al. (2024b)	Middle School Mathematics	AMMORE dataset (53,000 question-answer pairs)	Publicly available	Chain-of-thought prompting and LLMs	92% accuracy on edge cases; 99.9% overall accuracy. Chain-of-thought prompting excelled but required more processing time