Evaluating the Landscape of Automated Scoring with GenAI: A Systematic Review

Warren Li WARREN@LD-INSIGHTS.COM
Melanie Kurimchak MELANIE@LD-INSIGHTS.COM
Alexandra Andres ALEXIS@LD-INSIGHTS.COM
John Whitmer JOHN@LD-INSIGHTS.COM
Davis, CA

Abstract

As Generative Artificial Intelligence (GenAI) capabilities rapidly expand in educational settings, assessing the validity, reliability, and fairness of these tools is critical before widespread classroom deployment. While there has been a large amount of research, it is not yet clear what methods are being used to evaluate these foundational concepts and what the findings are. We present a work-in-progress systematic literature review focusing exclusively on automated item scoring. Based on a final coded corpus of 107 empirical studies published since January 2023, our preliminary analysis reveals that while prompt-based Large Language Models (LLMs) are achieving parity with traditional transformer-based methods, significant gaps remain regarding fairness evaluation and the representation of diverse student populations. Quadratic weighted kappa (QWK) is the dominant agreement metric, used in 41.8% of studies, and fairness is acknowledged by 55% of studies but empirically evaluated in only 11.2%. This paper offers evidence-based guidance for responsible classroom integration of AI.

Keywords: Automated Item Scoring, Generative AI, Fairness, Metareview

1. Introduction and Background

Historically, automated scoring systems have relied heavily on handcrafted linguistic features or custom-trained, fine-tuned transformer models to evaluate student writing (Haller et al., 2022). Yet, the advent of publicly available LLMs has shifted the landscape. Researchers and developers are testing whether they can use an off-the-shelf model to score complex student responses, bypassing resource-intensive training of bespoke models (Ramesh et al., 2025). In practice, this means that a single general-purpose model could score student responses from a short natural-language prompt, without being trained on labeled examples for that specific task, thereby increasing the potential use of automated scoring to new items and reducing the cost of creating these algorithms. The speed of this shift has led to rapid innovations in evaluation methods and results that are not yet well understood.

There is also a critical gap in the literature regarding the responsible evaluation of these systems (Abeysinghe and Circi, 2024). Traditional educational measurement frameworks and standards offer a principled foundation for assessment design with consideration of fairness and other components of responsible AI, but these approaches often lack explicit guidelines for the nuances of GenAI and other advanced computational methods. Fairness has long been a central requirement in educational measurement, where a score is expected to mean the same thing for students from different backgrounds, but whether LLM-based

scoring meets that standard is still largely untested. If these systems are to be safely deployed in real-world settings however, it is important to produce a measurement-informed synthesis of how these models are currently being evaluated and where they fall short.

The findings presented here are a subset of a work-in-progress, large-scale systematic review; this paper focuses specifically on automated item scoring. By analyzing the evaluation metrics, modeling techniques, and fairness considerations of 107 empirical studies, we highlight existing vulnerabilities and propose pathways to responsibly incorporate AI-based scoring.

2. Methodology

To assemble a comprehensive corpus, we conducted systematic database searches across peer-reviewed publications, preprints, and technical reports.[1] The search strategy used both keyword-based queries and citation network analysis to ensure sufficient coverage of the rapidly expanding field of GenAI in education. Selected studies were published from January 2023 onward, aligning with publicly available, post-ChatGPT 3.5 models.

The papers must be empirical studies that explicitly evaluate the quality of LLM outputs, with a focus on assessment use cases. Research methods must be described in sufficient detail to support robust analysis, and studies containing major methodological faults threatening the validity of the findings were excluded. While the broader research project analyzes multiple domains, the results in this paper are filtered for studies on automated item scoring such as essay evaluation, short constructed responses, and other tasks traditionally assessed by humans.

We developed a standardized coding schema to systematically categorize the diverse evaluation methods and model choices present across the literature. To capture how models were operationalized and assessed, a team of 8 reviewers with background in AI and education, documented up to 10 performance evaluation metrics per study, prompt engineering strategies (e.g., zero-shot, few-shot, chain-of-thought), and whether sampling fairness was empirically evaluated. Coding proceeded with pairs working from a shared codebook; disagreements were resolved through discussion to reach a target inter-rater reliability of Krippendorff's $\alpha \geq .80$. This metric was selected as it handles a variable number of raters, missing data, and different data types in a single metric that is easily interpretable. The resulting dataset is a coded corpus of 107 empirical papers. Materials and the coding codebook are openly available in an Open Science Project and the study is pre-registered.[2]

3. Preliminary Findings and Observations

3.1. Model Usage and Performance Trends

Most studies in our corpus compare multiple models and vary their fine-tuning and prompting strategies, testing a median of 5 models per study. The OpenAI family (particularly GPT-4 and GPT-3.5) is the most frequently tested (62 papers, 56%), though there is also a strong showing from the BERT family of models, including RoBERTa and DeBERTa (31 papers, 28%). Some of this emphasis on OpenAI models may reveal early studies in which

1. The underlying reference corpus is maintained as a public library at https://aievidencehub.org/lib.
2. The codebook and its iterations are available at https://osf.io/fd4up/overview.

this was the dominant LLM; it is also one of the most readily available. Open-source models, such as the LLaMA family (18 papers, 16%), appear less frequently. Some studies utilize proprietary models that do not provide enough methodological transparency for replication; these studies were filtered out from the review database during selection.

When comparing evaluation approaches in Figure 1, more than 50% of the sources used zero-shot prompting as one of their tested approaches. While custom-trained, fine-tuned transformer models currently remain the most accurate for operational scoring, prompt-based models are achieving performance parity, particularly in English writing tasks, with QWK accuracy gaps ranging from 0.01 to 0.13.

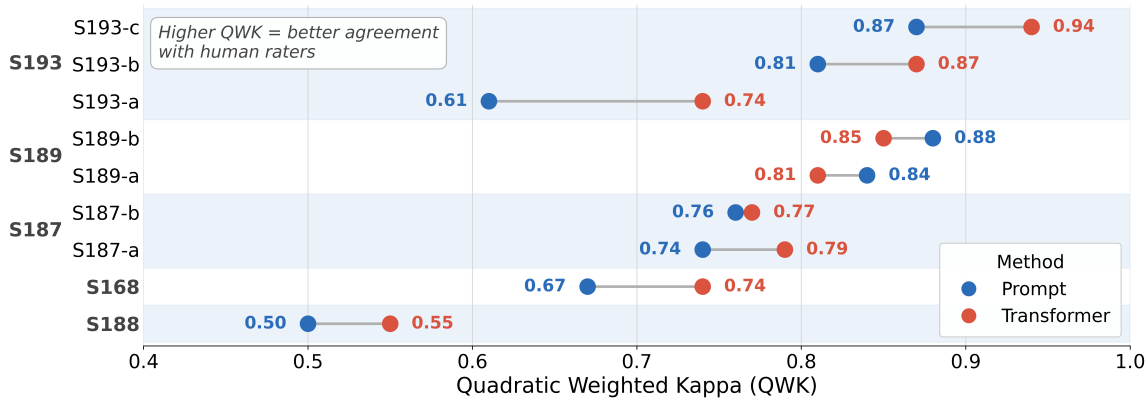

Figure 1: This visualization compares 9 research questions across 5 studies that compare prompt-based models to transformer-based models in writing assessments.

3.2. Evaluation Task Contexts and Methodological Gaps

As detailed in Table 1, the vast majority of automated scoring research focuses heavily on writing and language arts contexts. General short answers and essays dominate the literature, comprising 51.7% of the reviewed studies. There are fewer studies that expand into other domains, particularly social and emotional learning (SEL) or durable skills (2.9%).

Table 1: Papers by content area of focus. One paper may fit multiple categories.

Content Area	Percentage	Content Area	Percentage
Standardized Assessment	0.6%	Language Assessment	4.6%
SEL/Durable Skills	2.9%	Other	12.6%
Formative Assessment	3.4%	Writing (STEM Explanations)	20.1%
Math Problems	4.0%	Writing (Short Answer, Essays)	51.7%

Many papers rely on multiple metrics to assess scoring quality. QWK is the primary metric, utilized by 41.8% of the sources to measure agreement between a human "gold standard" and LLM-generated scores. QWK adjusts results for unbalanced data, which

accounts for uneven distributions often found in graded work. Additionally, 28.2% of the studies used some format of raw agreement. There is a noticeable influence of "traditional" machine learning classification metrics in the literature, with several studies employing F1 scores (15.3%), and AUROC (4.0%) to evaluate model performance. The dominance of these metrics is understandable given their roots in traditional automated scoring research, but agreement with human raters is not equivalent to validity. A model can track human scores closely and still be systematically unfair to particular groups of students.

3.3. The Fairness Gap

The most striking finding to date is the absence of fairness analysis. While over half of the papers (55%; 59 papers) acknowledge fairness within their literature reviews, they generally do not conduct analysis. When one is conducted, which occurred in 12 papers (11.2%) it is more commonly focused on the sampling and representation of the training dataset rather than the potential bias or other analyses of scoring results.

Only three papers in our corpus (Morris et al., 2025; Huang et al., 2025; Perkoff, 2025) measured scoring bias through post-hoc statistical measures (e.g., Cohen's d or SMD) between demographic groups. One study that tested directly for bias found that GPT-4o changed its essay scores when the same writing was rephrased in a different linguistic style, suggesting the model reacts to surface features rather than to the quality of the response (Meng et al., 2025). Yet, the literature lacks sufficient research on automated scoring for students with disabilities and multilingual learners, who together represent nearly one in six U.S. public school students (NCES, 2024; OSEP, 2022). There are currently no evaluations regarding intersectionality or twice-exceptional students in our sample.

4. Discussions and Implications

There are several actionable takeaways for practitioners and researchers. Prompt design and rubric quality are levers that are both accessible and consequential. Across the studies we reviewed, improving the scoring task itself with clearer rubrics, retrieval, or fine-tuning often improved results much more than switching to a larger model. Scoring design deserves attention before model scale. Investing in elaborated, example-rich rubrics is one of the highest-yield steps before deployment. Second, task type should govern expectations: structured reasoning or curriculum-specific tasks warrant additional validation, human review, or hybrid scoring designs.

With only 12 papers evaluating fairness and only three conducting post-hoc bias checks on marginalized student populations, these innovative approaches to conducting automated scoring are largely untested for equitable classroom use. The literature reviewed here skews heavily toward English-language writing assessment using frontier models such as GPT-4, meaning practitioners working with multilingual learners, students with disabilities, or math-heavy content should treat existing results as directionally informative but not directly applicable without local validation. Until robust fairness benchmarks are established, practitioners should exercise caution in their use of LLMs to classify student work.

This work in progress has several limitations. The model landscape changes quickly and specific models named here should be read as a recent snapshot rather than a fixed ranking. Much of the corpus draws on older, human-scored response datasets, which pushes coverage

toward English short-answer and essay writing and K-12 settings. The review also covers a single scoring domain and is still being coded, thus the figures reported here will shift as more studies are included.

5. Conclusion and Next Steps

Future steps will include expanding our analysis beyond automated scoring to other assessment domains, specifically item generation and formative feedback. Ultimately, this research will culminate in the release of a public, open-access dataset and a comprehensive evidence rubric designed to bridge the gap between rapid AI deployment and psychometric rigor. LLM-based scoring is best framed as a starting point requiring context-aware calibration, not a plug-and-play solution. By aligning our findings with professional measurement standards, we aim to provide researchers, developers, and educators with actionable, theory-driven guidance for evaluating GenAI tools in real-world classrooms.

References

Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches. *arXiv preprint arXiv:2406.03339*, 2024.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*, 2022.

Yue Huang, Corey Palermo, Ruitao Liu, and Yong He. An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation. *Chinese/English Journal of Educational Measurement and Evaluation*, 6(2):5, 2025.

Lionel Meng, Shamya Karumbaiah, Vivek Saravanan, and Daniel Bolt. Identifying biases in large language model assessment of linguistically diverse texts. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, pages 204–210, Pittsburgh, PA, October 2025. National Council on Measurement in Education (NCME). URL https://aclanthology.org/2025.aimecon-wip.25/.

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. Automated scoring of constructed response items in math assessment using large language models. *International journal of artificial intelligence in education*, 35(2):559–586, 2025.

NCES. Students with disabilities, 2024. URL https://nces.ed.gov/programs/coe/indicator/cgg/students-with-disabilities.

OSEP. OSEP fast facts: Students with disabilities who are english learners, 2022. URL https://sites.ed.gov/idea/new-osep-fast-facts-students-with-disabilities-english-learners/.

Elana Margaret Perkoff. Leveraging large language models for automated scoring of constructed response mathematics items. *Annual Meeting of the National Council on Measurement in Education*, 2025.

G Ramesh, Mahammad Sahil, Shashank A Palan, Darshan Bhandary, Teli Abhishek Ashok, J Shreyas, and N Sowjanya. A review on nlp zero-shot and few-shot learning: methods and applications. *Discover Applied Sciences*, 7(9):966, 2025.