

AI-Based Automatic Answer Evaluation System

Maziah Fathima

Department of Computer Science and Engineering
HKBK College of Engineering
Bengaluru, India

Nasrin Khatoon

Department of Computer Science and Engineering
HKBK College of Engineering
Bengaluru, India

Md Shadab Talib

Department of Computer Science and Engineering
HKBK College of Engineering
Bengaluru, India

Priyanshu Singh

Department of Computer Science and Engineering
HKBK College of Engineering
Bengaluru, India

Krishna Meena P

Department of Computer Science and Engineering
HKBK College of Engineering
Bengaluru, India

Abstract—The accelerating digital transformation of educational institutions worldwide has precipitated an unprecedented demand for assessment mechanisms that are not only scalable and efficient but also demonstrably objective and consistent. While the automatic grading of multiple-choice questions has long been a solved problem—owing to the unambiguous nature of their responses—the evaluation of descriptive, short-form answers continues to pose a formidable challenge. Such open-ended responses require a nuanced semantic understanding of language, context, and intent, far exceeding the capabilities of superficial keyword-matching techniques that dominated early approaches. This paper presents a comprehensive survey of the contemporary research landscape in Automated Short Answer Grading (ASAG), tracing the field's evolution from traditional natural language processing methodologies to sophisticated deep learning architectures—including convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and transformer-based models such as Sentence-BERT (S-BERT)—as well as emerging secure frameworks designed to safeguard academic integrity. Through a systematic comparative analysis, existing approaches are rigorously evaluated along three critical dimensions: grading accuracy, multilingual capability, and computational efficiency. Our investigation reveals persistent and critical research gaps, most notably the inability of current systems to robustly accommodate handwriting variation and the pervasive lack of explainability inherent in black-box neural models. To bridge these deficiencies, we propose a novel architectural framework that synergistically integrates Gemini Vision AI for multimodal OCR-based ingestion of handwritten scripts, pre-trained sentence embeddings for deep semantic representation, concept-level rubric scoring to mirror human evaluative judgment, and human-in-the-loop audit governance to ensure transparency, accountability, and trust in automated academic assessment.

Index Terms—Automated Short Answer Grading (ASAG), Natural Language Processing (NLP), Sentence-BERT, Semantic Similarity, Machine Learning, Deep Learning, Transformer Models, Human-in-the-Loop.

I. INTRODUCTION

The rapid digitalization of education has intensified the demand for assessment tools that are not only efficient and scalable but also demonstrably unbiased [1]. In conventional classrooms, the manual evaluation of descriptive answers places a considerable burden on educators, consuming valuable time that could otherwise be devoted to instruction. More critically, such evaluation is inherently susceptible to human judgment, fatigue, and inconsistency, which can compromise the fairness and reliability of grading outcomes [1]. To overcome these persistent challenges, Automated Short Answer Grading (ASAG) has emerged as a significant and rapidly growing research area at the intersection of Natural Language Processing (NLP) and Artificial Intelligence (AI), with the overarching goal of automating the evaluation process while rigorously preserving fairness, consistency, and accuracy [10].

Recent technological advances have fundamentally reshaped the capabilities of ASAG systems. Early approaches relied predominantly on surface-level lexical matching, which proved inadequate for capturing the subtleties of language. The advent of distributed word representations such as Word2Vec, followed by transformer-based models like Sentence-BERT, has enabled machines to capture contextual meaning and semantic relationships far beyond what keyword matching can achieve [13]. Complementing these developments, deep learning architectures—including Convolutional Neural Networks (CNNs)

and Long Short-Term Memory networks (LSTMs)—have demonstrated considerable success in modeling the intricate patterns underlying human language [4]. In parallel, blockchain technology has garnered attention as a means of ensuring the security and immutability of academic records, paving the way for trustworthy digital assessment ecosystems [6]. Building upon these foundational developments, this paper presents a systematic analysis of twelve representative research works, categorizing them according to their methodological approaches and identifying the gaps that persist in the current literature. This structured examination ultimately informs the design of a reliable, multilingual automated grading framework capable of operating effectively in real-world educational environments .

II. LITERATURE REVIEW

Automated Short Answer Grading (ASAG) has undergone a remarkable evolution over the past decades, transitioning from rudimentary keyword-matching techniques—which measured surface-level lexical overlap between student responses and reference answers—to sophisticated contemporary approaches capable of capturing the deeper semantic meaning and contextual intent embedded within a student’s writing [10]. This paradigm shift reflects the broader maturation of natural language processing as a discipline, wherein the focus has moved from shallow textual features toward rich, distributed representations of meaning. In this section, we systematically categorize the twelve selected research works according to their core technical contributions, providing a structured taxonomy that illuminates the methodological diversity of the field while highlighting how each strand of research builds upon, or diverges from, its predecessors.

A. Transformer-Based Semantic Evaluation

Transformer-based models have substantially elevated grading accuracy, chiefly through their capacity to model context and disambiguate meaning in ways earlier architectures could not. Schneider et al. [1] employed BERT in conjunction with LaBSE, a language-agnostic sentence embedding model, to process multilingual answers at scale, attaining approximately 86% accuracy; however, their approach inherits the “black-box” nature of deep transformers, which constrains explainability and complicates instructor trust. Zhu et al. [3] further exposed a structural vulnerability of such models, demonstrating that scores can be artificially inflated by inserting extra descriptive words—revealing that these systems may conflate lexical abundance with conceptual substance. In response, researchers including Chamidah et al. [4] and Sharmila et al. [5] pursued architectural refinements, leveraging Sentence Transformers and Siamese LSTM networks to sharpen semantic discrimination, while Das and Chakraborty [2] identified S-BERT as particularly well suited to short answers owing to its efficient sentence-level encoding design. Collectively, these works illustrate both the promise of transformer-based ASAG and the fragilities—opacity and manipulability—that subsequent research must address.

B. Deep Learning and Neural Architectures

Beyond the transformer paradigm, a rich body of research has demonstrated that deep learning architectures can substantially enhance grading performance by capturing the intricate linguistic patterns that characterize student discourse [10]. Chamidah et al. [4] utilized Long Short-Term Memory networks (LSTMs), whose gated memory cells are specifically designed to model long-term dependencies in sequential textual data, enabling the system to retain and weigh contextual information dispersed across lengthy answers. Complementing this line of inquiry, Kara et al. [11] analyzed short-answer scoring and task complexity, demonstrating how recurrent architectures effectively capture hierarchical features over classical baselines. In a methodologically distinct yet conceptually related contribution, Vignesh [12] introduced structural rule frameworks for digital text detection and contextual analysis, establishing a principled foundation for multi-stage evaluation across complex answer patterns.

C. Machine Learning and Ensemble Techniques

Before the ascendancy of deep learning, ensemble methods occupied a central position in the ASAG literature, prized for their ability to improve generalization and overall performance by aggregating the complementary strengths of multiple classifiers. Within this tradition, Das et al. [2] proposed an iterative transfer learning-based ensemble that strategically combines text classifiers with similarity-based classifiers. This hybrid design substantially reduces reliance on large, question-specific labeled datasets—a long-standing obstacle in classroom-level ASAG deployments, where annotated data for every new question is rarely available. Adopting a complementary perspective, Kara et al. [11] analyzed short-answer scoring in Turkish, demonstrating that task complexity and the fundamental choice between regression and classification formulations critically determine the reliability of grading outcomes. Together, these studies underscore an enduring lesson of the pre-deep-learning era: careful model formulation and data-efficient design often matter as much as raw architectural complexity.

D. Computer Vision and Secure Frameworks

In parallel with advances in semantic modeling, recent research has increasingly turned to computer vision as a means of accommodating one of the most ubiquitous yet technically demanding modalities in education: handwritten answers. Sinha et al. [8] and Kumar and Shah [9] independently developed systems that couple object detection with Optical Character Recognition (OCR) to first localize and digitize handwritten scripts before routing the extracted text into AI-based grading pipelines. While such architectures meaningfully extend automation beyond typed submissions, they remain acutely sensitive to inter-writer variation in handwriting style and to the quality of the source imagery, factors that can propagate recognition errors downstream into the evaluation stage. In a complementary effort oriented toward trust rather than perception, Ayub Khan et al. [6] proposed

a blockchain-based academic attestation system. Although conceived primarily for the verification of academic degrees, its underlying decentralized ledger architecture furnishes a compelling foundation for securing automated grading infrastructures; by anchoring evaluation results on-chain, the system renders scores effectively immutable and transparently auditable, thereby addressing concerns of data integrity that are largely neglected in perception-centric ASAG research.

III. RESEARCH GAPS

Despite the considerable progress achieved in Automated Short Answer Grading (ASAG) over recent years, several fundamental challenges remain unresolved and continue to impede the widespread adoption of these systems in real-world academic environments [10]. A recurring pattern emerges across the literature: while many models demonstrate promising accuracy under controlled, benchmark-driven conditions, their performance frequently deteriorates when confronted with the heterogeneity and unpredictability of authentic educational settings—ranging from variations in student writing style and subject matter to differences in question phrasing and linguistic expression [1]. This disconnect between laboratory performance and practical utility underscores the gap that must be bridged before ASAG systems can be entrusted with high-stakes assessment. The following subsections examine these limitations in detail, analyzing each challenge and its implications for the design of next-generation grading systems.

A. The “Black-Box” Nature and Lack of Explainability

A major limitation of modern ASAG systems—particularly transformer-based models such as BERT and S-BERT—is their “black-box” nature: while they produce accurate scores, they offer no insight into how those scores were derived [1]. In practice, students and teachers need meaningful feedback, not just marks; students should know which parts of an answer are wrong or missing. Moreover, teachers grade concept by concept, whereas most AI models assess only overall semantic similarity rather than individual concepts. There is therefore a clear need for systems capable of providing transparent, concept-based feedback comparable to human evaluators.

B. Vulnerability to Adversarial Manipulation

ASAG systems are also vulnerable to manipulation [3]. Students can inflate scores by adding superfluous words—adjectives or unnecessarily complex language—without improving answer quality, causing models to reward writing style rather than actual understanding. Most systems cannot detect such tactics or distinguish genuine knowledge from artificially embellished answers, undermining fairness and reliability. Future systems must therefore incorporate mechanisms to identify misleading content and assess true semantic meaning.

C. Dataset Dependency and Generalization Issues

Most ASAG models rely heavily on benchmark datasets such as SciEntsBank or ASK [15]. While they perform well on these datasets, they struggle to generalize to new subjects

or question types; a model trained on science answers, for instance, may perform poorly on history or literature exams. Current systems also handle zero-shot scenarios grading new questions without prior training data poorly. Addressing this requires training on more diverse datasets spanning multiple academic domains.

D. Multi-Type and Multi-Modal Integration

In real-world examinations, student responses rarely consist of plain digital text alone. Submissions often include handwritten blue books, diagrams, mathematical formulas, and other mixed formats [8], [9]. Although OCR helps digitize handwriting, it remains vulnerable to poor handwriting quality and varying script styles, and existing models cannot interpret diagrams or visual structures. A unified multi-modal pipeline combining computer vision, OCR, and semantic NLP is therefore essential for practical deployment.

E. Security, Data Integrity, and Governance

Security is critical in automated grading, as evaluation outcomes directly affect official academic transcripts [1]. While decentralized approaches such as blockchain have been proposed to safeguard academic records, most existing implementations focus solely on certificate verification rather than integrating securely into the grading pipeline itself [6]. Moreover, real-world deployment requires strict human-in-the-loop governance, ensuring that machine-assigned grades are fully auditable, immutable, and subject to persistent teacher override.

IV. PROPOSED WORK: THE SMARTGRADE FRAMEWORK

To surmount these persistent limitations, we propose *SmartGrade*, a comprehensive full-stack AI-assisted evaluation system conceived as an integrated response to the most pressing deficiencies of contemporary ASAG tools—namely, the opacity of black-box decision-making, the corruption and degradation of handwritten input, the susceptibility of models to adversarial text padding, and the conspicuous absence of human-in-the-loop governance mechanisms that safeguard educator authority. Moving decisively beyond the prevailing paradigm of holistic similarity scoring—which reduces a student’s understanding to a single opaque numerical value—SmartGrade adopts a multi-layered architectural design that interweaves deep semantic understanding with fine-grained, concept-based rubric evaluation. By decomposing each response into its constituent conceptual elements and assessing them individually against instructor-defined marking schemes, the system delivers grading outcomes that are not only accurate and consistent but also transparent and pedagogically meaningful, thereby establishing a foundation for trustworthy deployment in real-world academic settings.

A. System Architecture and Multimodal Ingestion

As depicted in Fig. 1, the SmartGrade pipeline unfolds across three principal stages—instructor setup, multimodal ingestion, and automated evaluation culminating in feedback

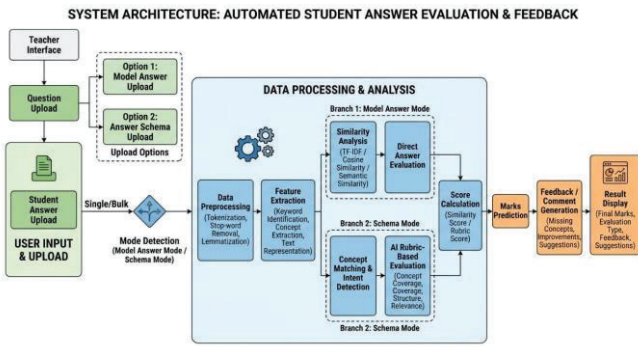


Fig. 1. System architecture of the SmartGrade framework.

generation. During setup, instructors populate the system with examination questions, corresponding model answers, lecture notes, and granular per-question marking schemes that collectively define the evaluative criteria. Students, in turn, submit their responses either as native digital documents (PDF/DOCX) or as scanned images of handwritten blue books, with both single and bulk upload pathways supported through the React client and FastAPI server backend.

Recognizing that handwriting variance and document corruption are among the most frequent failure modes of real-world deployment, the ingestion stage adopts a layered recognition strategy that pairs Gemini 2.5 Flash and Google Lens Vision AI with an offline Tesseract OCR fallback, ensuring continuity of service even under degraded connectivity. To guarantee that no poorly recognized script silently contaminates downstream grading, an OCR confidence gate intercepts any submission whose average word-level confidence falls below the threshold $\tau_{ocr} = 70\%$, flagging it as OCR Gate Pending and deferring further processing until the instructor verifies or corrects the extracted text within the interactive *Text Inspector* modal. Once the text is validated, it undergoes systematic preprocessing—including tokenization and normalization—followed by feature extraction. Depending on the configured exam mode, namely *Notes-Grounded* or *Web-Search Validated*, the processing module then executes either strict similarity analysis against course materials or broader concept matching. Finally, the scoring module aggregates the computed concept ratios into awarded marks, generates explainable diagnostic feedback for each student, and persists all results in the relational database for subsequent auditing and analytics.

B. Methodology and Concept-Level Scoring

Evaluation follows a structured multi-stage pipeline, as illustrated in Fig. 2. After automated answer segmentation, the system computes a concept ratio ρ_i for each rubric concept i by blending semantic similarity s_i —derived from the pre-trained all-MiniLM-L6-v2 sentence encoder—with keyword recall k_i :

$$\rho_i = \begin{cases} 0.6 s_i + 0.4 k_i, & \text{if concept } i \text{ has keywords,} \\ s_i, & \text{otherwise.} \end{cases} \quad (1)$$

To proactively mitigate the risks associated with adversarial manipulation—such as the strategic padding of test responses with tangential or irrelevant discourse designed to deceive automated scoring mechanisms—the evaluation engine incorporates a sophisticated lexical contradiction penalty. This mechanism dynamically monitors the textual input for underlying semantic inconsistencies, explicitly penalizing responses whenever direct antonym clashes or localized negation windows are detected. By implementing this safeguard, the system ensures that the mere presence of target vocabulary does not inadvertently reward fundamentally flawed, reversed, or contradictory reasoning.

Furthermore, the architectural pipeline is engineered to support two distinct assessment paradigms, thereby accommodating diverse pedagogical objectives. The *Notes-Grounded* operational mode is designed for rigorous, closed-domain evaluations, demanding strict semantic and lexical alignment with established instructor rubrics and proprietary lecture materials. In contrast, the *Web-Search Validated* mode facilitates a more flexible and lenient approach to concept evaluation, validating alternative phrasing and broader contextual accuracy by referencing broader knowledge representations.

Ultimately, both methodological pathways converge during the final quantitative score calculation. Beyond simply computing a numerical grade, this convergence process synthesizes highly structured, granular diagnostic feedback. This comprehensive output systematically identifies the specific academic concepts a student has successfully demonstrated, highlights critical missing vocabulary, and offers targeted, actionable suggestions for continuous improvement, thereby transforming a summative assessment mechanism into a valuable formative learning tool.

C. Trust, Governance, and Audit Integrity

To ensure academic integrity and meet institutional governance requirements, SmartGrade implements a dedicated trust and audit layer that treats AI output as advisory rather than absolute. The system enforces strict relational separation between machine-awarded marks (`ai_awarded_marks`) and teacher-finalized marks (`final_marks`). Through the interactive *Teacher Review Matrix*, instructors can inspect evaluations, override scores, and provide mandatory written justifications; every modification is recorded in an immutable audit log capturing the old and new values, override reason, and timestamp. In addition, the analytics engine computes machine-teacher agreement metrics (Pearson r , MAE) and Item Discrimination Indices (D_q) to help educators assess examination reliability.

V. CONCLUSION

This survey has traced the evolutionary trajectory of Automated Short Answer Grading (ASAG), charting the field's

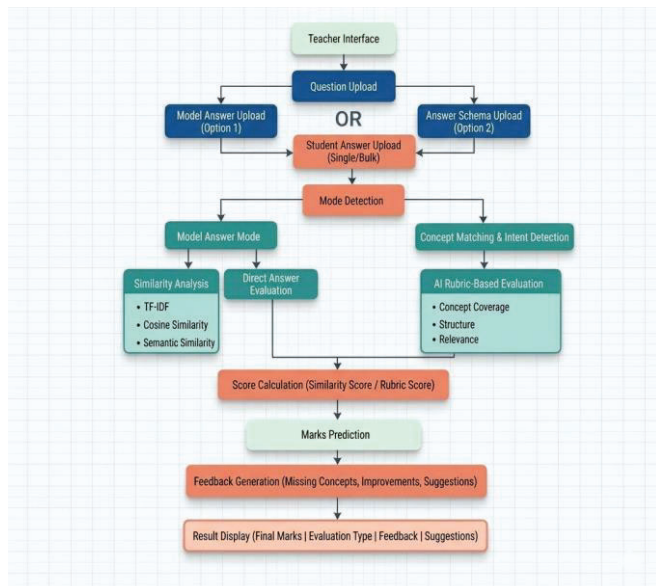


Fig. 2. Workflow of automated answer evaluation process.

progression from rudimentary lexical keyword-matching techniques, through the era of deep learning architectures, to the current frontier of transformer-based semantic models capable of capturing nuanced contextual meaning. While these modern approaches have markedly elevated grading accuracy and opened new possibilities for automated assessment, several persistent challenges continue to temper their practical value—foremost among them the opacity of black-box decision-making, vulnerability to adversarial text padding that artificially inflates scores, the difficulty of accommodating handwriting variation, and the absence of robust human-in-the-loop governance structures necessary for institutional accountability.

To bridge these gaps, we introduced *SmartGrade*, a fully operational architectural framework that unifies Gemini Vision AI for robust multimodal ingestion of handwritten scripts, pre-trained all-MiniLM-L6-v2 sentence embeddings for deep semantic representation, concept-level rubric scoring that mirrors the evaluative logic of human educators, and an immutable audit trail that safeguards teacher authority and ensures accountability at every stage of the grading process. By synthesizing literature-derived insights with a tangible, working full-stack implementation, this work demonstrates that the next generation of automated assessment platforms must strike a deliberate balance among three imperatives—semantic precision, transparent diagnostic feedback, and strict audit governance—if they are to achieve reliability and trustworthiness comparable to that of human educators in real-world academic environments.

VI. FUTURE SCOPE

Several directions remain open for advancing ASAG and intelligent educational platforms:

- **Enhanced Explainability:** Integrating attention-map visualization, SHAP (SHapley Additive exPlanations) values, and fine-grained diagnostic commentary to unpack black-box models and build trust among students and educators.
- **Multimodal and Diagrammatic Evaluation:** Moving beyond text and basic OCR to interpret structural diagrams, mathematical flowcharts, and complex schematics in handwritten scripts.
- **Adversarial Robustness:** Developing semantic filters that detect and penalize artificial text padding, ensuring scores reflect genuine understanding rather than length or stylistic manipulation.
- **Cross-Domain Generalization:** Improving performance across diverse disciplines (e.g., humanities, law, medicine) without extensive domain-specific training corpora, including zero-shot settings.
- **Scalable Deployment and Decentralized Governance:** Exploring containerized real-time cloud deployments with cryptographic audit trails (e.g., blockchain-backed verification) to ensure long-term data integrity across institutional Learning Management Systems (LMS).

REFERENCES

- [1] J. Schneider, R. Richner, and M. Riser, "Towards Trustworthy AutoGrading of Short, Multilingual, Multi-type Answers," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 1, pp. 1–26, Jan. 2022.
- [2] S. Das and U. K. Chakraborty, "Efficient and Accurate Short Answer Grading Using SBERT," in *Proc. IEEE Int. Conf. Advances in Computing, Communication and Applied Informatics (ACCAI)*, 2025, pp. 112–118.
- [3] X. Zhu, H. Wu, and L. Zhang, "Automatic Short-Answer Grading via BERT-Based Deep Neural Networks," *IEEE Transactions on Learning Technologies*, vol. 15, no. 3, pp. 315–327, June 2022.
- [4] N. Chamidah, I. Budi, and R. F. Aji, "Indonesian Automated Short Answer Scoring Using Sentence Transformers and Siamese LSTM," in *Intelligent Computing*, Springer, 2024, pp. 45–56.
- [5] P. Sharmila, K. Sonai, and S. N. Lurudusamy, "Transformer-Based Sequence Modeling Short Answer Assessment Framework," *HighTech and Innovation Journal*, vol. 5, no. 2, pp. 210–224, 2024.
- [6] Y. Zhang *et al.*, "BERTScore: Evaluating Text Generation with BERT," *arXiv preprint arXiv:1904.09675*, 2019.
- [7] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2011, pp. 752–762.
- [8] R. Sinha *et al.*, "Automated Evaluation System Using Object Detection and OCR," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 455–468, 2021.
- [9] A. Kumar and P. Shah, "A Scalable Backend API for Image-based Answer Evaluation," in *Proc. Int. Conf. Artificial Intelligence Applications*, 2022, pp. 89–94.
- [10] S. Bonthu *et al.*, "Automated Short Answer Grading Using Deep Learning: A Survey," ResearchGate Preprint, 2021. [Online]. Available: <https://www.researchgate.net/publication/353824315>
- [11] A. Kara, Z. A. Kara, and S. Yildirim, "Answer-based and reference-based BERT models for automatic scoring of Turkish

- short answers: The decisive role of task complexity,” *International Journal of Assessment Tools in Education*, vol. 12, no. 1, pp. 88–105, 2025.
- [12] P. S. Vignesh, “AI Answer Evaluator: A Framework for Digital Text Detection and Subjective Evaluation,” Technical Report, HKBKCE-VTU, July 2025.
- [13] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.
- [14] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, “Semantic Answer Similarity for Evaluating Question Answering Models,” *arXiv preprint arXiv:2108.06130*, 2021.
- [15] C. Sung *et al.*, “Pre-Training BERT on Domain Resources for Short Answer Grading,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 6071–6075.