ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

AI-Assisted Evaluation of Handwritten Answers with a Smart Feedback Loop

[1] Mirnalini E R*, [2] Karpagam S, [3] Nick Anto Roy A, [4] Venkata Lakshmi S

[1] [2] [3] [4] Department of Artificial Intelligence and Data Science, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India

Corresponding Author Email: [1] mirnalinierm@gmail.com, [2] karpagamsekarm@gmail.com, [3] nickanto555@gmail.com, [4] venkatalakshmis@skcet.ac.in

Abstract*— Grading messy handwritten answers takes too long. One person might mark differently than another. Some students get unfair results because of how they write or make small grammar mistakes. This paper discusses the use of artificial intelligence to help fix these problems. Initially the handwritten text is converted to digital text by using intelligent document processing. Then comes understanding meaning: does the answer match what was expected? Transformers: smart language models check how close the ideas are. Grammar gets reviewed at the same time, without rigid rules but with context awareness. Scores come out based on both sense and structure. Feedback follows, shaped around each learner's response. It points out strengths, highlights gaps, guides improvement. Fairness stays central throughout. Handwriting differences matter less now. Mistakes in syntax don't block comprehension anymore. The system adapts instead of rejecting. Learning moves forward without waiting weeks for results. When mistakes happen, teachers can step in to check and change AI grades. Their fixes go straight back into the system, helping it learn over time. Instead of just numbers, a live dashboard shows patterns - like where students struggle most or how tests are performed overall. Watching real usage reveals faster grading without losing accuracy. Feedback gets better too. As a result, this becomes a better assisted solution for large classrooms and fair evaluation is done. Learning from each decision makes the whole process tougher against errors down the line.

***Keywords*— Artificial intelligence, handwritten evaluation, semantic similarity, human-in-the-loop, personalised feedback.**

I. INTRODUCTION

Manual evaluation of handwritten, subjective answers continues to be a key part of educational assessment, especially in subjects that require detailed reasoning and conceptual explanations. However, traditional evaluation methods are time-consuming and prone to inconsistencies caused by human bias, fatigue, and differing interpretations. As class sizes increase and assessments become more frequent, teachers are under greater pressure to provide timely, fair evaluations along with valuable feedback that supports student learning. These issues emphasize the need for smart systems that can support teachers without taking over the important role of human judgment. Recent progress in artificial intelligence, natural language processing, and document analysis has made it possible to develop automated grading systems that can analyse structured responses [1], [2]. Nonetheless, evaluating handwritten answers adds more complexity because of differences in handwriting, grammatical errors, and unclear meanings. Many current solutions focus only on automation, often ignoring transparency, flexibility, and educational benefits. Therefore, a successful evaluation system must balance efficiency with clarity and relevance to learning. This paper introduces an AI-assisted framework that combines intelligent document processing, semantic similarity analysis, and grammar checks to assess handwritten responses. While fully automated grading systems operate without human involvement, the suggested method includes a human-in-the-loop mechanism that enables teachers to verify and improve AI-generated evaluations. The system enhances itself over time through an intelligent feedback loop, ensuring flexibility across various subjects and assessment formats. Moreover, an interactive analytics dashboard offers insights into learning behaviours and performance trends, supporting data-informed educational choices. The rest of this paper is structured as follows: Section II covers related research and current methods in automated assessment. Section III outlines the proposed system's architecture and approach. Section IV explains the experimental setup and results of the evaluation. Section V explores the implications and limitations, and Section VI concludes the paper by suggesting directions for future research.

II. RELATED WORK

Traditional automated evaluation systems have attempted to reduce manual grading effort through rule-based and machine learning approaches. Early rule-based evaluation systems relied on predefined keywords and rigid answer templates to assign marks [5]. These systems operate by matching words or phrases from student responses with model answers. While computationally simple, keyword-based systems fail to capture semantic meaning, ignore creativity, and cannot adapt to varied writing styles. Furthermore, they provide minimal or no constructive feedback, limiting their usefulness as learning tools.

More advanced approaches employ natural language processing techniques such as TF-IDF similarity scoring and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

transformer-based models like BERT to measure semantic alignment between student answers and reference content [1], [3]. These methods improve grading accuracy by capturing contextual meaning beyond surface-level keyword matching. However, many existing systems function as black-box models with limited explainability, making it difficult for educators and students to understand how scores are generated. In addition, most similarity-based frameworks focus solely on scoring and lack mechanisms for personalized feedback, topic-wise improvement guidance, or integration with broader learning analytics.

Recent research highlights the importance of combining automation with interpretability and educator involvement [6]. Fully automated systems risk reinforcing hidden biases and reducing pedagogical transparency. Therefore, emerging trends emphasize hybrid human-AI evaluation frameworks that support scalable assessment while maintaining human oversight. Despite these advances, there remains a need for an integrated system that combines handwriting recognition, semantic understanding, grammar analysis, explainable scoring, and adaptive feedback within a unified architecture [4], [5].

III. PROPOSED METHODOLOGY

A. Input Acquisition and OCR

The evaluation pipeline begins with the acquisition of handwritten answer sheets in the form of scanned images or PDF documents uploaded through a file interface or mobile capture module. The system is designed to support diverse input formats and varying image quality conditions commonly encountered in real classroom environments. Preprocessing at the image level includes resolution normalization, contrast enhancement, skew correction, and noise reduction to improve OCR accuracy. These enhancements ensure consistent text extraction even when answer sheets contain uneven lighting, background artifacts, or overlapping handwriting. Transformer-based Optical Character Recognition models such as TR-OCR are employed to convert handwritten content into machine-readable text with high robustness across varied writing styles [4]. Since the reliability of downstream semantic analysis depends heavily on accurate transcription, this stage is optimized to minimize recognition errors and maintain fairness across students with different handwriting characteristics. The overall architecture of the proposed framework is illustrated in Fig. 1.

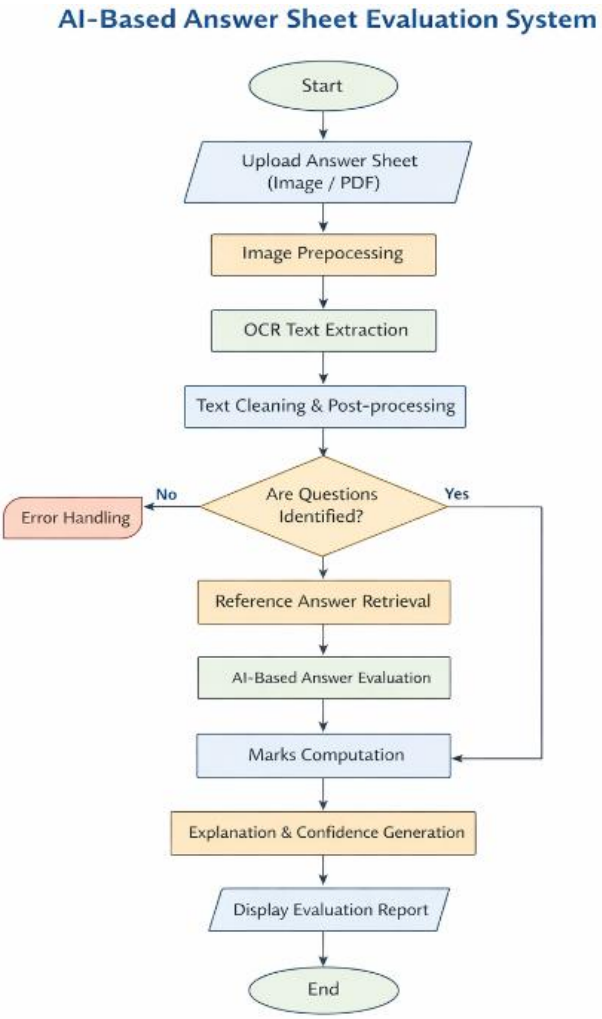

Figure 1. Overall workflow of the AI-assisted evaluation system

B. Text Preprocessing

Text obtained from OCR often includes noise, spelling inconsistencies, and formatting issues due to variations in handwriting. To address this, a preprocessing stage utilizing natural language processing techniques is implemented to standardize and organize the text. Tools like NLTK are employed to eliminate unnecessary symbols, perform tokenization, apply lemmatization, and correct spelling mistakes while maintaining the original meaning [5]. Filtering out stop words and identifying sentence boundaries help maintain a consistent linguistic structure without sacrificing the depth of the content. For longer texts, an automatic summarization module based on transformer models such as GPT or T5 is used to condense lengthy content while preserving key points. Additional error correction methods help minimize the effects of OCR inaccuracies. This well-structured preprocessing process creates a clear and consistent text format, enhancing the accuracy and efficiency of later semantic analysis and scoring.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

C. Semantic and Grammar Evaluation

Following preprocessing, the cleaned text undergoes semantic similarity analysis to evaluate conceptual alignment with reference answers. Sentence Transformer embeddings are used to capture deep contextual meaning beyond surface-level keyword matching [3]. This embedding-based approach allows the system to recognize paraphrased or creatively expressed answers that still convey correct understanding. Grammar and linguistic quality are assessed in parallel using automated tools such as LanguageTool or Grammarly APIs [5]. These tools analyze sentence structure, punctuation, clarity, and grammatical correctness to produce quantitative quality indicators. The dual evaluation strategy ensures that both conceptual accuracy and expressive clarity contribute to the final assessment, creating a balanced evaluation framework suitable for subjective responses.

D. AI Scoring Engine

The AI scoring engine integrates semantic relevance, grammar quality, and keyword coverage into a unified evaluation model. A hybrid scoring strategy combining rule-based weighting and adaptive machine learning techniques is used to generate preliminary marks [6]. This modular design enables flexible weighting policies that can be customized for subject-specific assessment criteria. For example, conceptual disciplines may emphasize semantic alignment, while language-based subjects may prioritize grammar and expression. Transparency is maintained by exposing scoring components so educators can understand how each metric contributes to the result. This interpretability builds trust in AI-assisted grading and supports its adoption in real educational environments.

To ensure fairness and consistency, the scoring engine applies normalization techniques that account for variations in answer length, vocabulary richness, and writing style. Rather than rewarding verbosity, the system evaluates informational density and conceptual clarity. Weighted aggregation functions combine multiple evaluation signals into a balanced score, preventing any single metric from disproportionately influencing the outcome. Confidence estimation mechanisms further indicate the reliability of automated scoring, allowing uncertain cases to be flagged for human review. This risk-aware design strengthens the integrity of the evaluation process. The scoring engine is also designed to support continuous improvement through adaptive learning. Corrections provided by educators are incorporated into model refinement cycles, enabling the system to align more closely with pedagogical expectations over time. Logging and audit trails preserve scoring decisions, ensuring accountability and traceability in institutional settings. By integrating adaptability, explainability, and reliability, the AI scoring engine functions not merely as an automated grader but as a decision-support system that enhances educational assessment while maintaining human oversight.

E. Feedback Generation and Human-in-the-Loop

Constructive feedback is created using prompts from large language models that convert assessment criteria into individualized learning recommendations. The system emphasizes students' strengths, pinpoints areas of misunderstanding, and offers specific suggestions for improvement, all expressed in encouraging and easy-to-understand language. Instead of just being a grading tool, the feedback system functions as a supportive learning assistant that promotes reflection and the development of skills. A human-in-the-loop interface lets teachers review, adjust, and explain AI-generated scores, ensuring that the feedback is educationally sound and fair. Teacher corrections are saved and used in an ongoing learning process that constantly improves the system's performance [6]. This collaborative approach combines the efficiency of automation with the expertise of human educators.

F. Analytics Dashboard and Student Portal

The framework also features an interactive analytics dashboard that collects performance data from various assessments to aid in making informed educational decisions. Visualization tools show trends in understanding, common grammar issues, and differences in performance across topics. Teachers can use these insights to modify their teaching methods and create focused interventions. A special self-evaluation portal for students allows them to submit their work on their own, track their progress, and get immediate feedback. Personalized progress tracking turns assessments into a continuous improvement process rather than a one-time grading event. By combining analytics, automation, and human judgment, the system enhances the link between evaluation and ongoing learning. The implemented analytics interface is shown in Fig. 2 and Fig. 3.

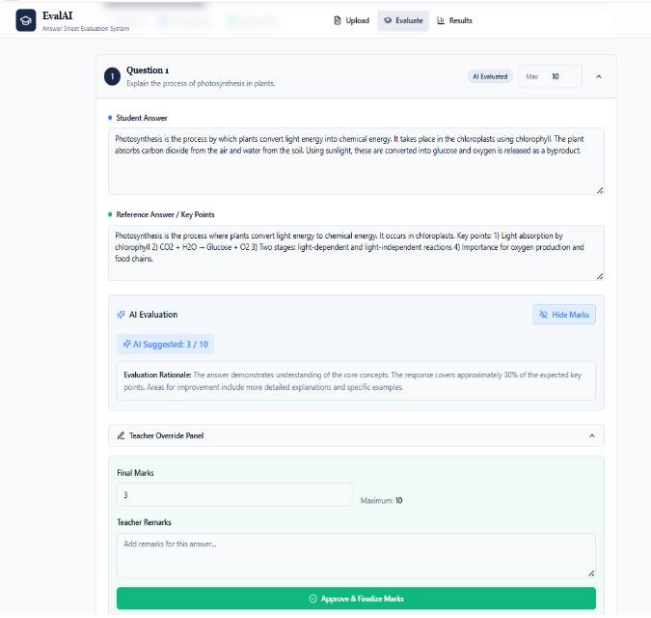

Figure 2. Evaluation dashboard

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

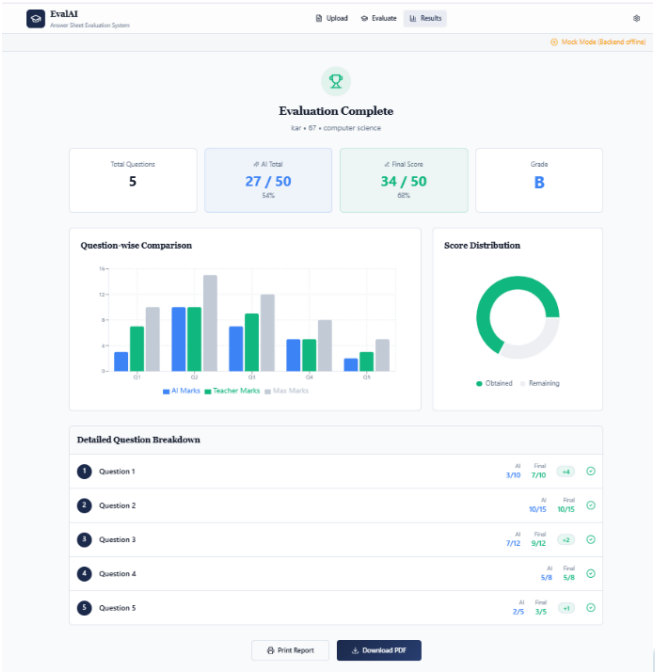

Figure 3. Results analysis

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

The proposed framework was implemented as a working prototype and evaluated using a dataset of handwritten student answer sheets collected to represent diverse writing styles, response lengths, and subject domains. The dataset included subjective descriptive answers to simulate realistic classroom evaluation scenarios. Images were captured using both scanning devices and mobile cameras to reflect practical deployment conditions. The evaluation focused on the accuracy of OCR transcription, semantic scoring reliability, grammar assessment performance, and the effectiveness of automated feedback generation.

B. Performance Analysis

Experimental testing demonstrated that the AI-assisted pipeline significantly reduced grading time compared to traditional manual evaluation. Automated preprocessing and semantic scoring enabled rapid initial assessment, while the human-in-the-loop validation preserved fairness and pedagogical oversight. High agreement was observed between AI-generated scores and educator evaluations, indicating consistent performance. Semantic similarity models successfully recognized paraphrased answers and contextually correct variations, overcoming limitations of keyword-based scoring. Grammar analysis contributed additional quality metrics that improved overall evaluation depth.

C. Observations and Impact

Educators reported that AI-generated feedback was constructive and pedagogically meaningful, helping students understand conceptual gaps and writing weaknesses. Analytics dashboard results revealed recurring learning patterns and class-wide performance trends, enabling targeted instructional interventions. Students using the self-evaluation portal showed increased engagement due to immediate feedback and progress tracking. These observations suggest that the hybrid framework enhances both grading efficiency and learning effectiveness, transforming assessment into a continuous improvement process rather than a one-time scoring event.

V. DISCUSSION

A. Educational Impact

The results demonstrate that AI-assisted evaluation can significantly transform how subjective assessment functions in educational environments. By reducing grading delays and providing immediate personalized feedback, the system shifts assessment from a purely summative exercise to a formative learning experience. Students receive actionable insights into their conceptual understanding and writing quality, enabling continuous improvement. For educators, the system reduces administrative workload while enhancing visibility into class-wide performance trends. This dual benefit strengthens the connection between assessment and pedagogy, promoting a more learner-centred evaluation culture.

B. Technical Strengths

A key strength of the proposed framework lies in its modular architecture. Each layer—OCR processing, linguistic normalization, semantic evaluation, scoring logic, and feedback generation—operates independently, allowing flexible upgrades and subject-specific customization. This extensibility ensures long-term scalability and compatibility with future AI advancements. The emphasis on explainable scoring further distinguishes the system. Transparent evaluation metrics allow educators to understand how scores are derived, increasing trust and encouraging adoption. The hybrid human-in-the-loop mechanism reinforces reliability by combining machine efficiency with expert oversight.

C. Limitations and Ethical Considerations

Although the system shows strong performance, there are still several limitations. OCR accuracy can decrease when dealing with very messy handwriting or poor image quality, which might impact the reliability of the evaluation. Semantic models may need specific adjustments to manage specialized academic terminology. Ethical issues also need attention, such as algorithmic bias, fairness among different language groups, and the proper handling of student data. Ensuring inclusiveness and transparency is crucial for ethical implementation. AI-based evaluation systems should be governed by clear frameworks that safeguard privacy and ensure accountability.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 3, March 2026

D. Future Directions

Future research could expand the framework in various ways. Developing multilingual handwriting recognition would increase accessibility in international education. Semantic models that adapt to specific domains could improve scoring accuracy in technical subjects. Real-time feedback systems might help students while they are writing their answers, turning evaluation into an interactive learning tool. Combining the system with adaptive learning platforms could link assessment results with personalized learning materials. These improvements would bring AI-assisted evaluation closer to modern educational goals focused on personalization and ongoing development.

VI. CONCLUSION

This paper presents an AI-supported framework for assessing handwritten subjective responses using a hybrid architecture that blends automation with human supervision. By combining handwriting recognition, natural language processing, semantic similarity modelling, grammar assessment, and explainable scoring methods, the system tackles ongoing issues in conventional evaluation, such as inefficiency, inconsistent grading, and lack of personalized feedback. The framework redefines assessment as an interactive learning experience rather than a purely administrative task. Experimental results show that the proposed system significantly decreases the manual effort required while preserving scoring accuracy and educational value. The inclusion of a human-in-the-loop mechanism ensures fairness and responsibility, while the feedback module supports student learning by offering practical insights. The analytics dashboard broadens the impact of evaluation by helping teachers identify overall learning patterns and develop focused teaching strategies. Collectively, these elements create a scalable assessment system that promotes both efficiency and educational quality.

One of the main strengths of the framework is its modular and scalable structure. Each component of the system can develop separately, making it possible to incorporate enhanced OCR technologies, more sophisticated language models, or specialized scoring methods without having to rebuild the whole system. This flexibility helps maintain its relevance in the quickly evolving field of AI. While there are still issues to address, such as variations in handwriting, customization for specific domains, and ethical implementation, the framework offers a solid base for responsibly integrating AI into education. Looking ahead, future studies might focus on evaluating handwriting in multiple languages, personalizing scoring to individual learners, and developing real-time feedback systems that support students while they are writing their answers. Enhancing the analytics component to provide predictive insights could turn assessments into a more proactive learning aid. In the end, the combination of artificial intelligence and educational analytics indicates a move toward more open and student-focused evaluation methods. Rather than replacing teachers, AI acts as a smart partner that improves teaching quality and supports ongoing student growth.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019.

[2] Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[3] T. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[4] R. Smith, “An overview of the Tesseract OCR engine,” in Proceedings of the International Conference on Document Analysis and Recognition, 2007.

[5] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2020.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.