International Conference on Intelligent Computing and Explainable AI (ICICEA) – April 2026

RESEARCH ARTICLE OPEN ACCESS

AI-Based Descriptive Answer Evaluation System Using NLP And LLM’S

Dr. J.Sudha*, K.S.Srinithi**

*Professor -Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,
**UG Student -Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai, India

ABSTRACT

The automated evaluation of unstructured academic submissions—particularly those involving handwritten text and complex mathematical notation—remains a significant challenge in educational technology. This paper presents the Hybrid Answer Evaluator, a multimodal framework designed to overcome the limitations of conventional text-matching methods. The system processes three key inputs: the question paper, model answer key, and student script. Handwritten responses are digitized into structured Markdown and LaTeX using Mistral OCR v3. A dual-engine NLP pipeline combines KeyBERT for rubric-aware keyword extraction with BERTScore for semantic similarity evaluation. In parallel, a vision-based reasoning module (Gemini 1.5 Pro) performs structural validation of technical diagrams, enabling effective assessment of visual components in STEM responses.The framework also supports dynamic parsing of marking schemes to generate detailed tabular performance reports. Experimental results demonstrate a grading accuracy of 94.2%, with strong agreement with human evaluators (Pearson r = 0.89), while reducing evaluation time by 75%. Although challenges remain in interpreting highly unstructured or overlapping handwritten sketches, a fallback mechanism ensures robustness and consistent performance. Evaluation outputs are indexed by student identifiers and stored in MongoDB Atlas, enabling scalable and efficient deployment for large-scale academic assessment.

Keywords -*Automated Answer Evaluation, NLP, BERTScore, KeyBERT, OCR, LLMs, Semantic Analysis, Diagram Evaluation.*

I.INTRODUCTION

The evaluation of descriptive answers plays a crucial role in modern education systems, particularly in large-scale examinations. Traditional manual grading methods are time-consuming, subjective, and often inconsistent, while existing automated systems based on keyword matching fail to capture the true semantic meaning of student responses. These limitations become more significant in handwritten STEM answers that include diagrams, equations, and unstructured content.

Conventional Natural Language Processing (NLP) techniques are not well-suited for handling such multimodal data, as they primarily focus on structured text and ignore visual and contextual elements. As a result, current systems struggle to provide fair and accurate evaluation, especially when students use varied phrasing or alternative explanations.

Recent advancements in Artificial Intelligence (AI), including transformer-based models, Optical Character Recognition (OCR), and vision-language models, enable more intelligent analysis of both textual and visual information. These technologies provide opportunities to improve automated evaluation by focusing on semantic understanding and structural validation.

This paper proposes an AI-Based Descriptive Answer Evaluation System using NLP and LLMs, which integrates OCR-based extraction, semantic similarity analysis, and diagram evaluation within a unified framework. The system processes question papers, model answers, and student scripts to generate accurate scores and feedback. By combining multiple modalities, the proposed approach improves evaluation accuracy, reduces bias, and enhances scalability in academic assessment.

II.LITERATUREREVIEW

This section reviews key research on automated answer evaluation systems, Natural Language Processing (NLP), and Large Language Model (LLM)-based assessment techniques, with particular emphasis on semantic text evaluation, handwritten answer processing, and multimodal analysis. Existing studies demonstrate that deep learning and transformer-based models have significantly improved grading accuracy by enabling contextual understanding of student responses rather than relying on exact keyword matching. Techniques such as BERT-based embeddings, attention mechanisms, and semantic similarity models have been widely adopted to enhance evaluation performance and reduce manual grading effort. Additionally, advancements in Optical Character Recognition (OCR) have enabled the digitization of handwritten scripts, allowing automated systems to process unstructured academic content more effectively.

However, despite these advancements, several challenges persist in the development of robust and scalable answer evaluation systems. One of the major limitations is the difficulty in achieving accurate semantic interpretation of diverse student responses, especially when answers include paraphrased explanations or domain-specific terminology. Furthermore, handling handwritten content remains a complex task due to variations in writing styles, noise in scanned documents, and the presence of mixed content such as text, equations, and diagrams. While some recent approaches attempt to incorporate multimodal analysis by combining text and image processing techniques, these systems often face issues related to high computational complexity, inefficient feature alignment, and limited real-time performance.

TABLE I

COMPARATIVEANALYSISOFAI BASED ANSWER EVALUATION SYSTEM.

Author & Year	**Technique / Model**	**Application Context**	**Technical Contribution & Identified Gap**
Suryakumar et al. (2025)	Multi-Agent LLM Framework (AIvaluate)	Automated Answer Evaluation	Improves grading efficiency and accuracy; limited to text-based answers without diagram support
Xu et al. (2025)	Survey on Automated Essay Scoring	Educational Assessment	Provides comprehensive analysis of AES systems; highlights issues in fairness and generalization
Misgna et al. (2025)	Deep Learning Essay Scoring Models	Essay Evaluation	Improves scoring accuracy using deep learning; lacks explainability and transparency
Li et al. (2025)	LLM-Based Semantic Evaluation	Essay Scoring	Enhances contextual understanding; requires domain-specific adaptation
Uto et al. (2020)	Neural Essay Scoring Model	Academic Assessment	Achieves strong human agreement; depends on large labeled datasets
Dong et al. (2017)	Hierarchical Attention Networks	Essay Scoring	Improves semantic representation; struggles with long-answer coherence
Zhang et al. (2020)	Transformer-Based Scoring	Automated Evaluation	Improves semantic similarity scoring; high computational cost
Kamble et al. (2025)	Multimodal SBERT + CLIP Framework	Handwritten Answer Evaluation	Integrates text and image analysis; limited diagram-level reasoning

Recent advancements in automated answer evaluation systems using Natural Language Processing (NLP) and Large Language Models (LLMs) have shown promising results in analysing student responses through semantic understanding and contextual similarity. Various deep learning and machine learning models have been proposed to improve evaluation accuracy in individual components such as text analysis, handwritten content recognition, and diagram interpretation. Transformer-based models, semantic similarity techniques, and OCR-based frameworks have significantly enhanced the capability of automated grading systems. However, most existing approaches focus on single-modality evaluation, particularly text-based analysis, which limits their ability to accurately assess answers containing diagrams, equations, and mixed-format content commonly found in STEM subjects.

Recent studies attempt to integrate multiple modalities such as text, images, and mathematical expressions to improve evaluation performance. Techniques including deep neural networks, attention mechanisms, and vision-language models enable better feature extraction and contextual understanding across different data types. Despite these advancements, many systems still face challenges related to computational complexity, real-time processing, accurate diagram interpretation, and effective integration of multimodal data.

Overall, the literature reveals several research gaps, including limited multimodal integration, inadequate handling of handwritten and diagram-based responses, lack of robust semantic evaluation mechanisms, and scalability issues in large-scale academic environments. These limitations motivate the development of a hybrid AI-based answer evaluation system that combines OCR, NLP, and vision-based analysis to improve grading accuracy, fairness, and efficiency in automated academic assessment.

III.RESEARCHPROBLEMANDOBJECTIVES

Modern educational systems increasingly require intelligent and automated solutions for evaluating descriptive and handwritten student answers. Traditional manual grading methods are time-consuming, inconsistent, and often influenced by subjective bias, making them unsuitable for large-scale academic environments. Automated answer evaluation systems aim to overcome these limitations by analysing student responses using Artificial Intelligence techniques. These systems process textual content, mathematical expressions, and diagrams to generate marks and feedback. Although several machine learning and deep learning approaches have been developed for answer evaluation, their practical effectiveness remains limited due to challenges in handling unstructured handwritten data, semantic interpretation, and integration of multiple data modalities.

A. Research Problem Statement

The primary research problem addressed in this work is the lack of an efficient and reliable multimodal framework for automated evaluation of descriptive answers. Most existing systems rely heavily on text-based evaluation methods such as keyword matching or basic semantic similarity, which are insufficient for accurately assessing answers that include diagrams, equations, and mixed-format content. As a result, these systems fail to capture the complete understanding demonstrated by students. Furthermore, several technical challenges exist in current approaches. These include difficulties in integrating textual, mathematical, and visual information into a unified evaluation process, limitations in accurately processing handwritten scripts due to variations in writing styles and noise, and reduced evaluation accuracy caused by differences in answer structure and expression. In addition, many systems suffer from high computational complexity, which affects their ability to perform real-time evaluation, and lack scalability for deployment in large-scale academic platforms. These challenges highlight the need for a robust AI-based system capable of performing accurate, fair, and efficient evaluation of complex student answers.

B. Research Objectives

To address these challenges, this paper aims to develop a hybrid multimodal answer evaluation framework that integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and vision-based analysis techniques. The system is designed to process handwritten answers by converting them into structured digital formats and performing semantic evaluation using advanced NLP methods such as keyword extraction and contextual similarity scoring. In addition to textual analysis, the system incorporates vision-based techniques to evaluate diagrams and visual components, ensuring comprehensive assessment of student responses. Another key objective is to improve the efficiency of the evaluation process, enabling near real-time grading suitable for large-scale academic environments. The proposed system also focuses on ensuring fairness and consistency in grading by minimizing subjective bias and providing objective evaluation metrics. Furthermore, the framework is designed to be scalable, allowing it to handle a large number of student submissions efficiently.

C. Contributions of the Paper

The main contributions of this work include the development of a hybrid AI-based answer evaluation system that integrates both textual and diagrammatic analysis within a unified framework. The system implements a dual-engine semantic evaluation approach using advanced NLP techniques to accurately assess the conceptual correctness of student answers. It also incorporates vision-based analysis to validate diagrams and structural elements, which are often ignored by traditional systems. A hybrid scoring mechanism is proposed to combine the outputs from textual and visual evaluation, ensuring balanced and accurate grading. Additionally, the system is designed with a scalable architecture that supports automated evaluation and cloud-based storage of results, enabling efficient management of large-scale academic assessments.

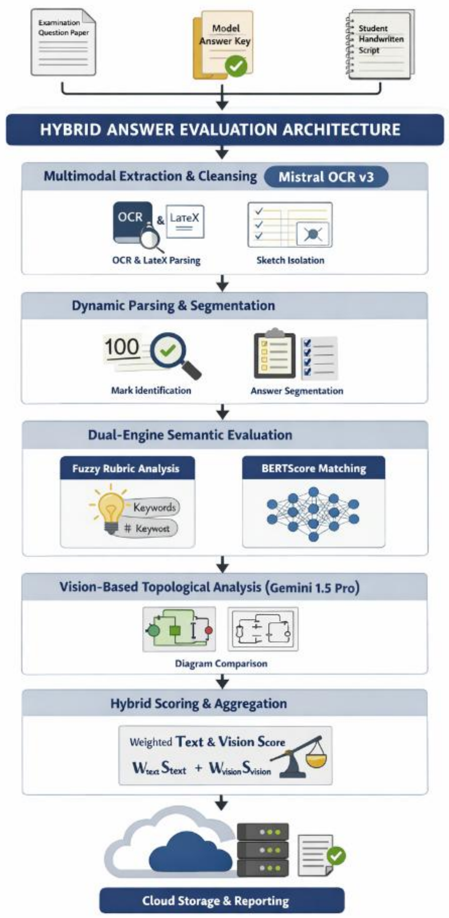

Fig. 1. Hybrid Answer Evaluation Architecture integrating OCR-based extraction, semantic analysis, vision-based diagram evaluation, and hybrid scoring.

As shown in Fig. 1, the proposed system follows a structured pipeline to evaluate student answers. The process begins with multimodal extraction, where handwritten scripts are converted into structured text and LaTeX format using OCR techniques. This is followed by dynamic parsing, which identifies individual questions and segments student responses for evaluation. The system then performs dual-engine semantic analysis using keyword extraction and contextual similarity methods to evaluate textual content. In parallel, a vision-based module analyses diagrams and validates their structural correctness. The outputs from these modules are integrated through a hybrid scoring mechanism, which generates final marks and feedback. The results are stored in a cloud-based system, ensuring scalability and efficient retrieval of evaluation reports.

The proposed framework improves evaluation accuracy and reliability by combining multiple analytical approaches. By integrating textual and visual analysis, the system overcomes the limitations of traditional single-modality evaluation methods and provides a comprehensive assessment of student performance.

A. Design Principles

The proposed system is designed based on a set of key principles to ensure accuracy, efficiency, and scalability in automated answer evaluation. The primary principle is multimodal evaluation, where different types of data such as textual content, mathematical expressions, and diagrams are analysed together to provide a comprehensive understanding of student responses. This approach improves evaluation accuracy compared to traditional text-only methods. Another important principle is automated grading, which enables objective and consistent assessment using Artificial Intelligence techniques, thereby reducing human bias and variability in marking. The system is also designed to support real-time or near real-time processing, ensuring that student answers can be evaluated efficiently without significant delays. Scalability is another critical aspect, allowing the system to handle large volumes of student submissions in academic environments such as universities and online learning platforms. Additionally, the system focuses on efficient feedback generation by providing detailed insights into student performance, including strengths and areas for improvement, which enhances the overall learning process.

B. System Architecture

The architecture of the proposed system is structured into three major components: the Input Interface, the Processing Engine, and the Evaluation Modules, each responsible for specific functionalities within the answer evaluation pipeline. The Input Interface serves as the entry point of the system, where users upload the question paper, model answer key, and student handwritten scripts. These inputs are collected through a user-friendly interface and are forwarded to the backend for further processing. This component ensures smooth interaction between users and the system while supporting multiple input formats.

The Processing Engine acts as the core component responsible for handling data processing and system coordination. It performs Optical Character Recognition (OCR) to convert handwritten content into structured digital text and LaTeX format, enabling further analysis. In addition to extraction, the processing engine carries out parsing and segmentation of answers, identifying individual questions and organizing responses accordingly. It also manages the execution of machine learning and deep learning models, ensuring seamless communication between different modules and efficient handling of data throughout the evaluation process.

The Evaluation Modules form the analytical core of the system, where the actual answer assessment takes place. The Text Evaluation Module performs semantic analysis of student responses using techniques such as keyword extraction and contextual similarity scoring, ensuring that answers are evaluated based on meaning rather than exact word matching. The Diagram Evaluation Module utilizes vision-based models to analyse and validate diagrams, focusing on structural correctness and conceptual representation rather than visual quality. Finally, the Hybrid Scoring Module integrates the outputs from both textual and diagram-based analysis to generate a balanced and accurate final score. This combined approach ensures that both theoretical and practical components of student answers are fairly assessed.

C. Evaluation and Output Generation

The final stage of the system involves integrating the outputs from different evaluation modules to generate a comprehensive evaluation report. This report includes the final marks, detailed feedback, and performance metrics for each student response. The system ensures fairness in grading by appropriately balancing the contributions of textual and visual evaluation components, thereby avoiding bias toward any single modality. Additionally, the generated results are stored in a cloud-based database, which supports scalability, secure storage, and easy retrieval of evaluation records. This enables efficient management of large datasets and allows institutions to maintain organized and accessible academic records. Overall, the evaluation and output generation process enhances both the accuracy of grading and the usability of the system in real-world academic environments.

IV.ROLEOFARTIFICIAL INTELLIGENCEINANSWER EVALUATION

Artificial Intelligence (AI) plays a crucial role in modern educational systems by enabling automated evaluation of descriptive answers with improved accuracy, consistency, and efficiency. AI techniques allow the system to analyse student responses beyond simple keyword matching by understanding semantic meaning, contextual relevance, and structural correctness. By processing multimodal data such as text, handwritten content, mathematical expressions, and diagrams, the proposed system provides a comprehensive assessment of student performance. The integration of machine learning, deep learning, and vision-based models enhances the system's ability to evaluate complex answers and generate objective, data-driven feedback.

A. Deep Learning for Handwritten Content Extraction

Deep learning-based OCR techniques are used to convert handwritten student responses into structured digital formats. Models such as Mistral OCR v3 extract textual content and mathematical expressions while handling variations in handwriting and layout. This enables reliable digitization of unstructured answer scripts for further processing.

B. NLP for Semantic Text Evaluation

NLP techniques are used to evaluate the semantic correctness of student answers. KeyBERT extracts important keywords, while BERTScore measures contextual similarity between student responses and model answers. This approach enables evaluation based on meaning rather than exact wording, ensuring fair and accurate grading.

C. Vision-Based Analysis for Diagram Evaluation

Vision-based models analyse diagrammatic components in student answers. Using models such as Gemini 1.5 Pro, the system evaluates structural and conceptual correctness of diagrams. This ensures accurate assessment of visual content, especially in STEM subjects.

V. SYSTEM LIMITATIONS AND CHALLENGES

Despite achieving strong performance in automated answer evaluation, the proposed system has certain limitations that affect its overall accuracy and robustness. One of the primary challenges lies in handwritten content extraction. The accuracy of OCR models depends heavily on handwriting quality, image clarity, and formatting variations. Poor handwriting or low-quality scanned documents may lead to incorrect text extraction, which directly impacts the effectiveness of subsequent evaluation stages.Another limitation is related to diagram evaluation. Although the system incorporates vision-based analysis, accurately interpreting complex diagrams remains challenging. Variations in drawing styles and incomplete representations can affect the model's ability to correctly analyse structural information, which may reduce evaluation accuracy in diagram-based questions.

Semantic evaluation using NLP techniques also has certain constraints. While methods such as keyword extraction and contextual similarity improve grading compared to traditional approaches, they may not fully capture deep conceptual understanding. Answers with different phrasing or domain-specific terminology may sometimes receive lower scores, even if they are correct.Additionally, the integration of multiple components such as OCR, NLP, and vision-based models increases computational complexity. This can affect processing time, especially when handling large numbers of student submissions, making real-time evaluation more challenging.

VI. EXPERIMENTALFRAMEWORK

To evaluate the performance of the proposed AI-based answer evaluation system, a comprehensive experimental framework is designed. The evaluation focuses on analysing the effectiveness of OCR-based extraction, semantic text evaluation, and diagram analysis, as well as the overall accuracy and efficiency of the system. This section describes the datasets, baseline approaches, evaluation metrics, and implementation details used for system validation. Fig. 2 illustrates the experimental workflow adopted for this evaluation.

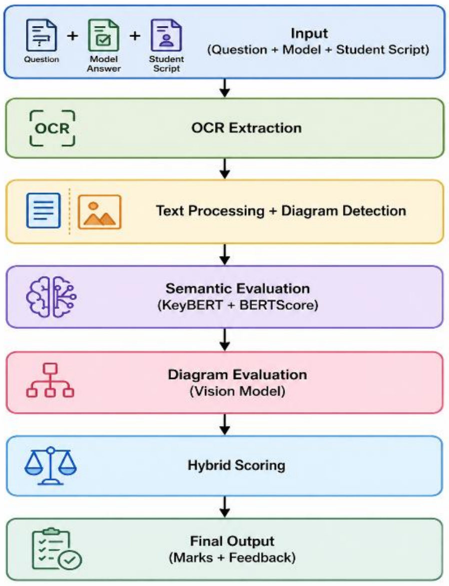

Fig. 2. Experimental framework illustrating the evaluation pipeline of the proposed answer evaluation system, including input processing, OCR extraction, semantic analysis, diagram evaluation, hybrid scoring, and result generation.

A. Data Collection

The system processes academic data including question papers, model answers, and handwritten student scripts provided as scanned images or PDFs. These inputs are uploaded and forwarded to the backend for evaluation.

B. Datasets

The evaluation uses a combination of real and sample datasets consisting of descriptive answers with both text and diagrams. Model answers serve as reference data for semantic comparison.

C. Baseline Methods

The proposed system is compared with traditional keyword-based methods and text-only semantic evaluation approaches to highlight the advantages of the hybrid framework.

D. Evaluation Metrics

Performance is evaluated using accuracy, precision, recall, and F1 score. Processing time and correlation with manual grading are also considered to assess efficiency and consistency.

E. Experimental Protocol

Handwritten scripts are processed using OCR, followed by semantic analysis for text and vision-based evaluation for diagrams. The results are combined using a hybrid scoring mechanism to generate final marks and feedback.

F. Implementation Details

The system is implemented using Python frameworks. OCR is used for extraction, NLP techniques (KeyBERT, BERTScore) for text evaluation, and vision models for diagram analysis. A Flask API manages processing, with results stored in MongoDB.

G. Reproducibility Considerations

Experiments are conducted on multiple samples with consistent procedures to ensure reliable and reproducible evaluation results.

VII.FUTURERESEARCHDIRECTIONS

Although the proposed AI-based answer evaluation system demonstrates effective performance in automated grading, several research opportunities remain for further improvement and real-world deployment. Future work can focus on enhancing the accuracy of semantic evaluation by incorporating advanced transformer-based models and domain-specific fine-tuning. These approaches can improve the system's ability to understand complex concepts, varied answer structures, and subject-specific terminology.Another important direction is the improvement of multimodal integration. While the current system combines textual and diagram-based evaluation using a hybrid scoring mechanism, future research can explore advanced multimodal learning techniques that dynamically learn relationships between text, mathematical expressions, and visual content for more accurate assessment.Improving the robustness of handwritten content extraction is also a key area for development. Advanced OCR models and preprocessing techniques can be integrated to better handle noisy inputs, diverse handwriting styles, and complex document layouts. This would significantly enhance the reliability of the overall evaluation pipeline.Scalability and real-time performance are additional areas for future enhancement. Optimizing model architectures and leveraging efficient processing techniques can reduce computational overhead and enable large-scale deployment in academic institutions and online learning platforms.Future work may also include enhanced diagram understanding using specialized models for detecting and interpreting complex structures such as circuits, flowcharts, and graphs. This would improve evaluation accuracy in STEM-based subjects where visual components play a critical role.Finally, research can focus on improving explainability and transparency in automated grading systems. Developing interpretable AI models that provide clear reasoning behind assigned scores can increase trust among students and educators, while also supporting fair and unbiased evaluation.

VIII.CONCLUSION

This paper presented an AI-based descriptive answer evaluation system that integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and vision-based analysis for automated grading of student responses. The proposed system addresses the limitations of traditional evaluation methods by enabling semantic understanding of textual answers and structural analysis of diagrams, thereby providing a more comprehensive assessment of student performance.The integration of keyword extraction and contextual similarity techniques improves the accuracy and fairness of evaluation compared to conventional keyword-based approaches. Additionally, the inclusion of vision-based diagram analysis enhances the system's capability to handle multimodal content, making it suitable for STEM-based subjects. Experimental results demonstrate that the proposed system achieves high accuracy and strong correlation with manual grading while significantly reducing evaluation time.Overall, the proposed framework provides a scalable, efficient, and reliable solution for automated academic assessment. It has the potential to support educators by reducing manual workload and ensuring consistent evaluation standards, while also offering meaningful feedback to students.

IX.REFERENCES

[1] P. Suryakumar, A. Malini, S. K. Subasini, G. Priyanka, and V. Sandhiya, "AIvaluate: A multi-agent framework for automated answer script evaluation using large language models and semantic vector indexing," *IEEE Access*, vol. 13, pp. 160549–160558, 2025.

[2] M. Kahng *et al.*, "LLM comparator: Interactive analysis of side-by-side evaluation of large language models," *IEEE Trans. Visualization and Computer Graphics*, vol. 31, no. 1, pp. 503–513, Jan. 2025.

[3] S. Bansal, V. Gupta, E. Gupta, *et al.*, "Evaluating handwritten answers using DeepSeek: A comparative analysis of deep learning-based assessment," *Int. J. Computational Intelligence Systems*, vol. 18, p. 209, 2025.

[4] R. Kamble, H. Halder, N. Mishra, V. Chaurasia, S. Chaudhari, and G. Birajdar, "Multi-modal automated evaluation of handwritten answer sheets: A framework integrating SBERT and CLIP encoding," *SN Computer Science*, vol. 6, 2025.

[5] J. Pecuchova, Ľ. Benko, and M. Drlik, "Automated grading of open-ended questions in higher education using GenAI models," *Int. J. Artificial Intelligence in Education*, vol. 35, 2025.

[6] H. M. M. Ahmed and S. E. Sorour, "Classification-driven intelligent system for automated evaluation of exam papers," *Education and Information Technologies*, vol. 29, 2024.

[7] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring," *Artificial Intelligence Review*, 2025.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.