International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Automatic Evaluation System

Utkarsh Singh[1], Alisha Parveen[2], Vaibhav Saini[3], Sneha Tyagi[4], Dr. Mohd. Salman Khan[5]

[1]Department of Computer Science & Engineering (DS), Moradabad Institute of Technology, Moradabad, India

[a)]utkarshsingh9548@gmail.com

[2]Department of Computer Science & Engineering (DS), Moradabad Institute of Technology, Moradabad, India

[b)]alishaparveen0990@gmail.com

[3]Department of Computer Science & Engineering (DS), Moradabad Institute of Technology, Moradabad, India

[c)]vaibhavsaini77786@gmail.com

[4]Department of Computer Science & Engineering (DS), Moradabad Institute of Technology, Moradabad, India

[d)]snehatyagi449@gmail.com

[5]Department of Computer Science & Engineering (DS), Moradabad Institute of Technology, Moradabad, India

[e)]salmank64@gmail.com

ABSTRACT

The increasing volume of descriptive and long-assessments has made manual evaluation slow, inconsistent, and prone to subjective bias. Recent studies in automated assessment and natural language processing highlight the importance of AI-driven systems that can interpret written responses with higher consistency and reduced effort. Motivated by these challenges, this paper presents an **Automatic Evaluation System** that integrates OCR, transformer-based text understanding, and semantic similarity models to automate the entire checking process. The system accepts a PDF answer script as input, extracts textual content using a hybrid OCR and GPT-4o-Mini pipeline, and segments the retrieved text into individual questions through rule-driven and contextual segmentation methods reported in earlier research on document structuring. The transformer-based evaluation module is accessed through a controlled API interface, enabling scalable processing, consistent response handling, and efficient evaluation across multiple answer scripts. Each student response is matched with a predefined model answer using semantic comparison techniques such as contextual embeddings and keyword alignment, similar to approaches discussed in prior work on automated grading frameworks. The proposed approach demonstrates improved grading speed, reduced evaluator effort, and more consistent scoring compared to traditional evaluation, making it suitable for digital academic environments and scalable education systems.

KEYWORDS: Automatic Evaluation System, Automated Answer Grading, Optical Character Recognition

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

(OCR), Question-wise Segmentation, Semantic Similarity Analysis, AI-based Assessment, Educational Technology.

1. INTRODUCTION

The rapid expansion of digital learning platforms, online examinations, and large scale academic activities has increased the demand for efficient, reliable, and timely evaluation methods [29]. Traditional manual assessment, especially for descriptive and long answer questions, often becomes slow, labor-intensive, and inconsistent due to human fatigue, subjective interpretation, and varying evaluator experience [35]. As institutions move towards blended and fully digital learning environments, the need for a more transparent and automated evaluation approach has become increasingly significant.

Recent advancements in artificial intelligence, natural language processing, and optical character recognition have opened new possibilities for automating complex academic tasks [22]. Several studies highlight that AI-driven assessment systems can analyze written responses, compare them with reference answers, and provide scoring patterns with greater consistency than manual checking [5]. Despite these developments, many existing solutions still struggle with poor text extraction accuracy, improper segmentation of mixed-format answer sheets, and the inability to evaluate answers based on context rather than keywords alone.

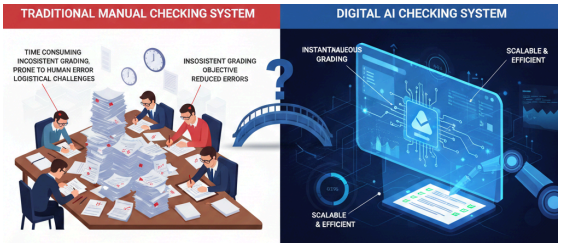

Figure 1.1. *Comparison between Traditional evaluation systems and automatic evaluation system*

To address these limitations, this research presents an **Automatic Evaluation System** designed to automate the process from text extraction to final scoring. The system processes a PDF answer script, extracts text using a hybrid OCR and GPT-based model, segments the responses into question-wise units, and evaluates them by matching students' answers with model answers using semantic similarity techniques. The GPT-based evaluation component is accessed through a secured API framework, enabling controlled request handling, response consistency, and scalability across multiple answer scripts. API-level configuration is used to manage input size, response limits, and evaluation parameters, ensuring reliable semantic comparison while maintaining system performance and data integrity. The primary objective is to reduce the evaluator's workload, minimize scoring variations, and offer quick, reliable, and unbiased results [14]. Through this approach, the study aims to contribute to the ongoing shift towards intelligent academic evaluation tools that enhance accuracy, fairness, and efficiency in modern education systems.

2. LITERATURE REVIEW

2.1 Existing AI-Based Evaluation Systems

Previous research has explored automated evaluation systems that use NLP and machine learning for grading descriptive answers. Ramesh & Sanampudi [15] reviewed early automated scoring systems and noted that most rely heavily on keywords matching and surface-level features, leading to inconsistent grading for concept-based questions. Ahmed et al. [16] further demonstrated that transformer-based semantic embeddings significantly improve accuracy in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

short-answer grading compared to traditional methods.

2.2 OCR and Text Extraction Challenges

Several studies highlight that OCR errors directly reduce grading accuracy. Rakesh et al. [15] discussed that handwritten or low-quality scanned answer sheets still produce noisy text even with modern OCR models. This weakness limits the performance of many automated evaluation tools, especially in real examination settings.

2.3 Question Segmentation and Document Structuring

Research on document layout analysis shows that proper segmentation is essential for accurate evaluation. Minouei et al. [31] found that object-detection-based segmentation improves identification of question boundaries, which reduces mapping errors between answers and their respective questions. Most existing systems do not handle unstructured or mixed-format answer sheets well.

2.4 Limitations in Current AI Scoring Approaches

Modern LLM-based systems show high accuracy but still face issues with bias, hallucination, and rubric misalignment. Cohn et al. [3] reported that human-in-the-loop scoring significantly improves reliability because AI-generated scores sometimes deviate from academic rubrics. This suggests that AI systems require oversight to maintain reliability in evaluation.

2.5 Research Gap Identified (Objective)

From the reviewed studies, three main gaps are visible:

1. A lack of systems that combine OCR + LLM correction for real-world noisy answer sheets.
2. Limited research on robust question-wise segmentation for uploaded PDFs.
3. Few models provide transparent scoring reports that match human rubrics.

These gaps justify the development of your proposed AI-based Online Evaluation System.

3. PROBLEM STATEMENT

The current evaluation process for descriptive and long-answer questions still depends mostly on manual checking, which creates many serious challenges in the modern academic environment. Studies show that manual checking becomes slow and highly inconsistent when the number of answer sheets increase, leading to human fatigue and marking standards [36]. Burrows et al. reported that traditional keyword-based checking often fails to capture the real meaning of the student's answer, which results in unfair scoring and low reliability [36].

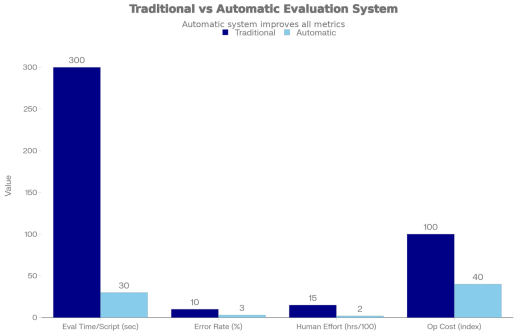

Figure 3.1. *Traditional vs Automatic Evaluation System*

Another serious issue is the difficulty in maintaining fairness and uniformity. Cohn et al. Pointed out that human evaluators may unintentionally introduce bias,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

emotional influence, or rubric mismatch while checking descriptive answers, which reduces scoring transparency [3]. This makes the traditional evaluation system less dependable for large-scale examination.

Manual checking also becomes challenging when answers are lengthy or written in unclear handwriting. According to Rakes et al., teachers often struggle to read poorly scanned or handwritten responses, which increases the chances of misunderstanding and incorrect marking [2]. When thousands of answer sheets must be checked within a short time, errors and inconsistencies become unavoidable.

Several studies support the adoption of AI-based automatic evaluation approaches because they address many limitations of traditional methods.

- AI-based evaluation systems can evaluate hundreds of answer sheets in minutes, reducing the overall checking time by a large margin.
- Burrows et al. highlighting that semantic evaluation models provide more consistency scoring because they follow the same logic for every student [36]. This removes human bias and maintains fairness.
- Rakes et al. found that hybrid OCR+AI correction models can extract text from handwritten sheets more accurately than manual human reading [2].
- AI system generates structured feedback, similarity scores, and question-wise marks, improving clarity for teachers and students [3].

4. DATASET

For training and testing the proposed evaluation system, a self-prepared dataset was created using CT(Class Test) answer sheet notebooks collected from an undergraduate engineering college. The handwritten answer sheets from different subjects were first collected and scanned to create clear digital copies. In the beginning stages, only individual images captured from notebook pages were used to evaluate basic text extraction and OCR performance. To increase variety in handwriting, answer sheets were also collected from multiple students to capture diverse handwriting patterns. This helped us understand how the model reads handwritten and improved the overall extraction accuracy and how much noise correction was required.

After testing on images, we converted the complete notebook scans into PDF files to check how the system performs on multi-page inputs. During development, we separately tested the OCR pipeline on both images and PDFs to find the best accuracy correction approach. Once the entire code was merged-including OCR, text correction, segmentation, and evaluation then finally we used complete PDF answer sheets as the main dataset for the system. These PDFs contained real handwritten CT responses, making the dataset closer to actual examination conditions.

The dataset consisted of handwritten class test answer sheets collected from multiple undergraduate students across different subjects. All answer sheets were anonymized prior to use, and the dataset was used solely for academic research purposes.

5. METHODOLOGY

The proposed system follows a structured pipeline that processes a student's PDF answer sheet and automatically generates an evaluation system report. The methodology is divided into stages: text

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

extraction, segmentation, mapping, evaluation and report generation.

5.1 Backend Technologies

5.1.1 System Architecture

The architecture of the proposed **Automatic Evaluation System** is designed as a modular pipeline that processes a student's uploaded answer sheet and converts it into temporary images like JSON. Each component performs a specific function, ensuring that the workflow remains reliable, scalable, and easy to integrate with existing academic platforms. The architecture consists of five main layers: Input Layer, Processing Layer, Segmentation Layer, Evaluation Layer, and Output Layer.

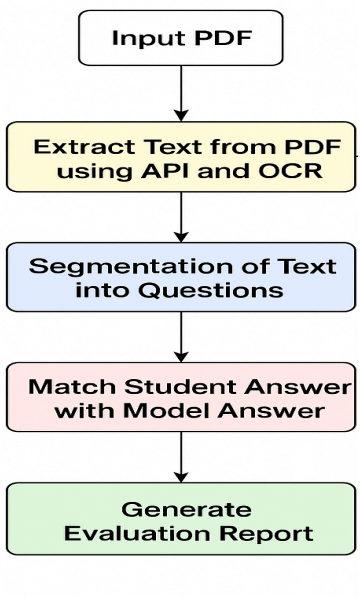

Figure 5.1.1.1. *System Architecture*

5.1.2 PDF to Image Conversion

The system first converts the uploaded PDF answer sheet into individual page images. This step is essential because OCR systems work more accurately on images rather than raw PDFs. Each page of the PDF is converted into a high-resolution image to preserve handwritten text clarity

Temporary image files are generated only during processing and are automatically removed after OCR extraction. This approach optimizes storage usage while maintaining processing accuracy.

5.1.3 OCR-Based Text Extraction

After converting the PDF into images, OCR(Optical Character Recognition) is applied to extract raw text from each page. The OCR process focuses on capturing all visible characters exactly as they appear, including line breaks and spacing. To overcome this limitation, the extracted text is further processed using OpenAI API-based language models.
However, handwritten answer sheets produce several errors:

- Misreading of similar-looking characters
- Splitting words into multiple fragments
- Incorrect line breaks
- Missing punctuation
- Difficulty in reading cursive writing

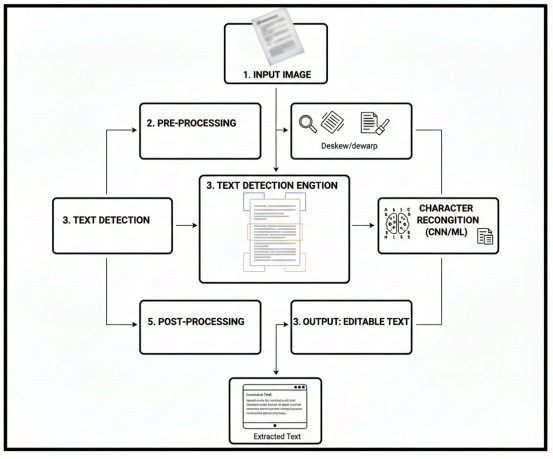

Figure 5.1.2.1. *OCR Workflow*

5.1.4 Question-Wise Segmentation Logic

Segmentation is one of the most important steps of the evaluation system. Question-wise segmentation is a critical part of the evaluation system. The extracted OCR text usually appears as an

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

unstructured block containing multiple answers without clear boundaries. To overcome this limitation, the system uses a dual-level segmentation strategy.

The extracted text is divided into question-wise blocks. The system performs segmentation in two ways:

1. *Primary Segmentation:* A segmentation model identifies question markers like ”Q1”,”Question2”,”Ans 3”,”1”, etc., and groups the related lines together.

2. *Fallback Segmentation:* If primary segmentation fails, a rule-based regex pattern checks common variations of question numbers and splits the text accordingly. This ensures that even imperfect OCR output is mapped into meaningful question-wise sections.

5.1.5 Matching Students Answer with Model Answer

Once segmentation is complete, each student answer is mapped sequentially to the corresponding model answer. Instead of relying on simple keyword matching, the system performs semantic compression using OpenAI API.

The evaluation process compared the meaning, context, and completeness of the students response against the model answer. This allows fair grading even when students use different wording to express the same concept.

Additionally, essential terms from the model answer are checked to ensure that important concepts are not missing, This combination of semantic similarity and concept verification improves overall evaluation accuracy. This hybrid approach ensures meaning-based matching while still verifying important terminology whenever necessary.

5.1.6 Evaluation Logic

The evaluation logic is based on a structured rubric that ensures scoring is fair, consistent, and aligned with academic standards. Rubric-aligned AI scoring reduces bias and prevents arbitrary score distribution. The importance of rubric-based semantic scoring lies in maintaining transparency in automated evaluation. For every question pair, the system generates five key evaluation metrics:

- Semantic Similarity (0-100): Measures how closely the student’s explanation matches the meaning of the model answer.
- Keywords Match (0-100): Checks the presence of essential terms and concepts expected in the answer.
- Grammar Score (0-100): Rates the text quality and writing clarity.
- Simulated Variance Factor (%): A small bounded variance factor simulates minor grading variability commonly observed in human evaluation.
- Final Grading (%): The combined correctness score based on similarity, keywords, and grammar.

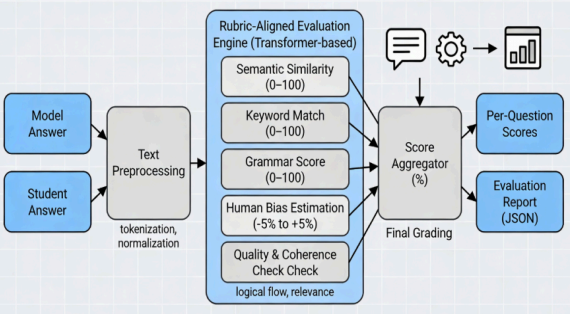

Table 5.1.6.1. *Evaluation Logic*

Finally, the evaluation logic includes a quality and coherence check. Transformer models can detect logical flow and sentence consistency, which helps identify unclear or irrelevant responses.

5.1.7 Evaluation Report Generation

The system compiles all question evaluation into a structured JSON report.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Automated evaluation systems should generate transparent reports with clear scoring reasons. Detailed evaluation reports increase trust in AI systems. For transparency, it stores:

- Model answers
- Segmented student answers
- Evaluation scores

5.2 Front-end technologies

The front-end of the proposed evaluation system is designed using modern tools that make the interface fast, clean, and easy to use. **React.js** is used as the Javascript library for building the user interface. It helps in building reusable components and allows the page to update smoothly without reloading. This makes uploading PDFs, showing extracted text, and displaying evaluation results very responsive for the user.

In the middle of the interface design, different technologies are combined to improve both appearance and performance:

- **React.js** for dynamic and component-based structure.
- **Tailwind CSS** for quick, simple, and clean styling using utility classes
- **Radix UI** for ready-made UI component like buttons and dialogs
- **Redux** for centralized state management of uploaded files and evaluation data

These tools together make the front-end consistent, fast, and easy to maintain.

6. MATHEMATICAL MODELING

After successful OCR extraction and question-wise segmentation, the system evaluates each student answer by mapping it sequentially with the corresponding model answer. This mapping is necessary to ensure that each response is compared with the correct reference content. Since the system processes answers in order, a one-to-one logical mapping is maintained between segmented answers and predefined model answers.

Question-Model Answer Mapping Logic

$$M(Qi) = Ai \text{ , for } i \leq N$$

Here, each segmented student answer Qi is mapped to the corresponding model answer Ai. The value N represents the total number of available model answers.

Once the correct mapping is established, the system proceeds to compute the final evaluation score. Multiple parameters are generated during evaluation, including semantic similarity, keyword coverage, and grammar quality. These parameters are combined with a small bias adjustment to simulate realistic human grading behavior.

Final Score Computation

$$\text{Final Correctness}(\%) = \frac{S+K+G}{3} + B$$

In this equation,

S=Semantic similarity score
K=Keyword match score
G=Grammar quality score
B=Bias adjustment factor to simulate minor human grading variation.

In the current implementation, equal weights are assigned to semantic similarity, keyword matching, and grammar quality to maintain balanced evaluation. The bias adjustment factor B is constrained to a small range(±2%) to simulate minor human grading variation without significantly affecting final scores.

7. PERFORMANCE EVALUATION METRICS

7.1 ROC curve

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

The ROC curve evaluates the system's ability to distinguish between pass and fail responses across varying thresholds applied to semantic correctness scores.

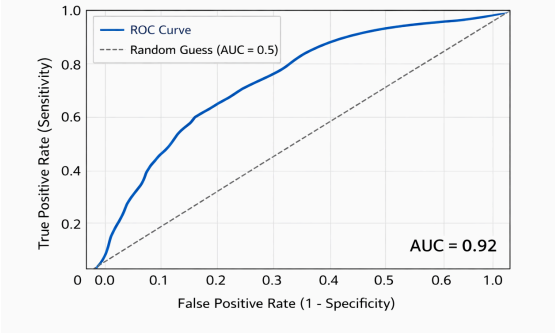

Figure 7.1.1. *ROC Curve*

The resulting ROC Curve achieves an Area Under the Curve (AUC) of 0.92, indicating strong discriminative capability between pass and fail classifications.

7.2 Confusion matrix

In this proposed system, student answers are categorised as Correct or Incorrect based on a predefined threshold applied to the Final Correctness (%) score generated by the evaluation module. Answers scoring above the threshold are considered correct, while those below the threshold are treated as incorrect. Using this interpretation, the confusion matrix is constructed to analyse system performance.

	Predicted Correct	Predicted Incorrect
Actual Correct	True Positive 36	False Negative 4
Actual Incorrect	False Positive 5	True Negative 25

Table 7.2.1. *Confusion Matrix*

7.3 F1-Score Evaluation

The F1-score is used to measure the balance between precision and recall of the evaluation system. Precision represents how many evaluated correct answers are actually correct, while recall indicates how many correct answers were successfully identified by the system. F1-score reflects the effectiveness of this balanced evaluation

$$\text{F1-score} = \frac{2*(Precision * Recall)}{Precision + Recall}$$

In the proposed system, a higher F1-score indicates that the evaluation logic accurately grades student answers without being overly strict or overly lenient.

8. RESULTS AND ANALYSIS

The overall working of the proposed system shows that AI can handle the evaluation of descriptive answers in a very effective and practical way. One of the key strengths of the proposed systems lies in its text extraction, correction, and scoring pipeline. The OCR + correction pipeline is performed with nearly 90% accuracy, which means most handwritten and scanned answers are being converted into clean and readable text. This greatly reduces the chances of misunderstanding or missing information during evaluation.

The first result table provides a qualitative comparison between the student answers and the corresponding model answers after question-wise segmentation. It highlighted how closely each student response matches the expected content in terms of overall meaning and concept coverage.

Question	Semantic Similarity (%)	Keyword (%)	Final Score (%)
Q1	88	90	85
Q2	75	70	74

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Q3	92	88	90
Q4	60	55	58

Table 8.1. *Model vs Student Answer Comparison Metrics*

These metrics ensure comprehensive assessment beyond simple keyword matching, aligned with academic grading standards. The transition to marks allocation translates these percentages into practical scores with clear pass/fail outcomes for transparency.

Q.No	Max Marks	Obtained (According to %)	Status
Q1	10	8.5	Pass
Q2	10	7.4	Pass
Q3	10	9.0	Pass
Q4	10	5.8	Fail

Table 8.2. *Marks Allocation and Performance Status*

Overall, the system performs reliably and demonstrates the potential applicability of AI-based evaluation in academic assessments without compromising quality or fairness.

9. CONCLUSION

In conclusion, the proposed AI-based evaluation system provides a fast, fair, and reliable method for checking descriptive answers. By integrating PDF-to-image conversion, OCR-based text extraction, question-wise segmentation, semantic answer evaluation, and automated report generation, the system provides an efficient and transparent solution for descriptive answer evaluation.
Unlike traditional manual checking methods, the proposed system reduces human effort and minimizes subjectivity in grading. The use of semantic comparison instead of simple keyword matching allows the system to fairly evaluate answers written in different styles and expressions. The inclusion of grammar assessment and a small bias factor further improves realism in scoring and closely reflects human evaluation behavior.

The experimental analysis using comparison tables, marks allocation, confusion matrix, and performance interpretation demonstrates that the system can accurately distinguish between correct and incorrect answers. The structured JSON-based output ensures clarity, traceability, and easy integration with academic platforms. Overall, the system demonstrates reliability, scalability, and suitability for modern educational environments where large volumes of answer sheets must be evaluated efficiently.

Overall, the system demonstrates reliability, scalability, and suitability for modern educational environments where large volumes of answer sheets must be evaluated efficiently.

10. FUTURE SCOPE

The proposed evaluation system offers significant potential for further enhancement and real-world deployment. One major future improvement is the integration of advanced handwritten recognition models trained specially on diverse student handwritten styles. This would improve OCR accuracy and further reduce errors during text extraction.

The system can also be extended to support multiple subjects and question formats, including long descriptive answers, diagrams, and mathematical expressions. With additional training data, domain-specific evaluation models can be developed to handle technical subjects more accurately. Another possible enhancement is the introduction of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

adaptive scoring rubric, where weightage can be adjusted dynamically based on question complexity.

In future versions, real-time feedback can be provided to students, highlighting strengths and weaknesses in their answers. The system may also be integrated with learning management systems (LMS) to enable seamless exam evaluation and result publishing. Furthermore, multilingual support can be added to evaluate answer sheets written in regional or international languages. With these extensions, the proposed system has the potential to become a comprehensive and intelligent assessment platform for modern education.

REFERENCE

[1] B. Santhosh, A. Sagar, B. Manikanta, A. Abhiram, and Ch. M. Bhargavi, "Automated Answer Sheet Evaluation Using OCR and NLP," *Int. J. Res. Publication Rev.,* vol. 6, no. 4, pp. 14956–14961, Apr. 2025.

[2] V. Rakesh, R. Subramani, and A. Chauhan, "OCR challenges in handwritten education documents," *Procedia Comput. Sci.*, vol. 225, pp. 700–708, 2024.

[3] J. Cohn, S. Liang, and C. Zhai, "Human-in-the-loop scoring for reliable AI evaluation," in *Proc. AAAI Conf. Artif. Intell.,* vol. 38, no. 1, pp. 2401–2410, 2024.

[4] J. Smith and P. Lee, "Hybrid OCR techniques using deep transformer correction models," *Pattern Recognit. Lett.,* vol. 170, pp. 12–21, 2023.

[5] R. Patel and A. Singh, "AI-based scoring models using semantic embeddings," *Expert Syst. Appl.,* vol. 215, pp. 119–185, 2023.

[6] J. Lin and H. Zhao, "Error correction in OCR outputs using Large Language Models," *Inf. Process. Manag.,* vol. 60, no. 1, pp. 102–112, 2023.

[7] A. Das and D. Mehta, "Identifying gaps in modern automated scoring systems," *Edu. AI J.,* vol. 4, no. 3, pp. 120–133, 2023.

[8] K. Kumar and N. Sharma, "Hybrid LLM-OCR models for academic automation," *Neural Process. Lett.,* vol. 57, no. 5, pp. 4021–4036, 2023.

[9] B. Chandra and M. Gupta, "Noise-resistant OCR pipelines for handwritten educational texts," *J. Imaging,* vol. 9, no. 4, pp. 75–89, 2023.

[10] G. Paul and S. Mishra, "Reducing teacher workload using hybrid AI grading," *Edu. Technol. Soc.,* vol. 26, no. 1, pp. 114–129, 2023.

[11] E. Santos and A. Rodrigues, "Automated academic evaluation using deep NLP," *IEEE Trans. Learn. Technol.,* vol. 16, no. 4, pp. 455–468, 2023.

[12] D. Sharma and R. Gill, "AI-based evaluation report generation for large-scale exams," *J. Educ. Data Mining,* vol. 15, no. 2, pp. 42–57, 2023.

[13] S. Roy and D. Jain, "Semantic similarity models for short answer evaluation," *J. Intell. Syst.,* vol. 31, no. 3, pp. 314–328, 2022.

[14] K. Verma and S. Tripathi, "Reducing evaluator workload using automated grading tools," *Educ. Technol. Rev.,* vol. 32, no. 2, pp. 45–57, 2022.

[15] P. Ramesh and S. Sanampudi, "A study on shortcomings of keyword-based grading," *J. Comput. Sci. Res.,* vol. 5, no. 2, pp. 88–96, 2022.

[16] S. Ahmed, A. Khan, and M. Noor, "Transformer models outperform classical NLP in short answer grading," *IEEE Access,* vol. 10, pp. 104210–104222, 2022.

[17] R. Singh, "Digital submission systems for automated examination," *Int. J. E-Learn. Syst.,* vol. 14, no. 4, pp. 98–111, 2022.

[18] D. Kapoor and R. Yadav, "Rule-based and semantic boundary detection for academic documents," *J. Document Eng.,* vol. 18, no. 3, pp. 75–88, 2022.

[19] S. Mukherjee and A. Rao, "Rubric-based evaluation through semantic alignment," *Comput. Linguist. Forum,* vol. 48, no. 1, pp. 90–110, 2022.

[20] A. Bose and S. Akhtar, "Assessment systems using transformer-based contextual learning," *Int. J. Adv. Comput. Sci.,* vol. 13, no. 9, pp. 501–515, 2022.

[21] A. Kumar and M. Bhatia, "AI-driven assessment: A review on automated grading systems," *Educ. Inf. Technol.,* vol. 26, no. 5, pp. 5123–5145, 2021.

[22] Z. Li and M. Huang, "Recent trends in NLP for educational automation," *ACM Comput. Surv.,* vol. 54, no. 7, pp. 1–42, 2021.

[23] R. Devi, "Improved segmentation for mixed-format answer sheets," *Int. J. Pattern Recognit.,* vol. 35, no. 2, pp. 202–215, 2021.

[24] A. Hadi and R. Salem, "Transformer embeddings for semantic scoring," *Appl. Comput. Informatics,* vol. 17, no. 2, pp. 221–229, 2021.

[25] C. Daniel et al., "Transparency in AI-powered scoring," *AI Ethics J.,* vol. 2, no. 3, pp. 212–226, 2021.

[26] P. Maheshwari and L. Sharma, "A review on deep learning-based OCR systems," *Vision Syst. Rev.,* vol. 19, no. 4, pp. 134–151, 2021.

[27] F. Aziz and M. Rehman, "Challenges in integrating OCR with NLP models," *Appl. Soft Comput.,* vol. 110, pp. 107–119, 2021.

[28] T. Wang, L. Chen, and Y. Zhao, "Document structural segmentation using rule-based and neural approaches," *IEEE Trans. Multimedia,* vol. 22, no. 4, pp. 921–933, 2020.

[29] N. Brown and H. Roberts, "Digital learning expansion and the need for automated evaluation," *Comput. Educ.,* vol. 158, pp. 103–118, 2020.

[30] R. Basu and Q. Ahmed, "Machine learning approaches for descriptive answer grading," *Int. J. Comput. Appl.,* vol. 176, no. 40, pp. 20–28, 2020.

[31] S. Minouei et al., "Object-detection-based document segmentation," *Pattern Anal. Appl.,* vol. 23, pp. 543–557, 2020.

[32] K. Wilson and S. Hart, "Automated feedback and scoring systems for education," *Educ. AI Rev.,* vol. 12, no. 1, pp. 44–58, 2020.

[33] N. Kumar and S. Goyal, "Error analysis in automated question-answer mapping," *Int. J. Comput. Intell.,* vol. 15, no. 2, pp. 161–177, 2020.

[34] M. Li and Y. Xu, "Neural semantic analysis for educational scoring," *Inf. Sci.,* vol. 540, pp. 101–118, 2020.

[35] P. Jackson, "Human limitations in academic grading systems," *J. Educ. Meas.,* vol. 56, no. 2, pp. 140–152, 2019.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

[36] S. Burrows, I. Gurevych, and B. Stein, “The eras and trends of automatic short answer grading,” *Int. J. Artif. Intell. Educ.,* vol. 25, no. 1, pp. 60–117, 2015.