# Automated Answer Sheet Evaluation System using OCR and NLP

Project By: Madhur Maru (1RVU23CSE246), Mohammed Arbab Mohsin (1RVU23CSE276) and Mohammed Junaid Irfan (1RVU23CSE275)

## Abstract

Manual evaluation of descriptive answer sheets is time-consuming, subjective, and difficult to scale in academic settings. This manuscript presents a practical teacher-facing system for automated answer sheet evaluation using OCR, NLP, and rubric-aware scoring. The platform supports end-to-end workflow: teacher login, exam creation, question-paper upload, marking-scheme upload, answer-sheet ingestion, question-wise segmentation, explainable scoring, and result export. To improve robustness, the system uses OCR API integration and a two-stage marking-scheme parser (Hugging Face LLM parser with regex fallback). Student answers are evaluated through a hybrid score combining semantic alignment, keyword coverage, and completeness, with per-question feedback and transparent marks. The implementation is lightweight and deployment-friendly (FastAPI + SQLite + web dashboard), making it suitable for internal assessment environments. Experiments on representative assessment sets show improved parsing reliability and grading consistency compared to baseline heuristics, while significantly reducing evaluation turnaround time.

## 1. Introduction

Descriptive-answer assessment remains one of the most labor-intensive academic processes. Instructors must evaluate large batches of responses within strict timelines while maintaining fairness and consistency. Traditional manual grading introduces variability across evaluators and delays feedback cycles. This work proposes a practical automated evaluation platform focused on real teacher workflow rather than isolated algorithmic benchmarking. The system is designed to run locally, expose explainable score components, and preserve teacher control through editable intermediate outputs.

### 1.1 Problem Statement

Given a question paper, a marking scheme, and scanned student answer sheets, automatically: (i) extract text, (ii) segment answers question-wise, (iii) evaluate responses against rubric expectations, (iv) generate explainable per-question marks and feedback, and (v) produce final score reports.

### 1.2 Objectives

- Reduce teacher workload and evaluation time.
- Increase consistency and transparency in grading.
- Support rubric-based partial marking.
- Provide a practical local-deployment system for internal assessments.

## 2. Literature Survey

### 2.1 Subjective Answer Script Evaluation using Natural Language Processing

This paper is about a Machine Learning model that was developed using Natural Language Processing to address challenges faced by teachers during manual grading. This problem was majorly seen during pandemic of COVID-19. The model processes descriptive answers by analyzing attributes like keywords, relevance to the question and grammar of the answers. And majorly uses similarity measures to compare student's answers from those in answer key. Gaussian Naïve Bayes a Machine Learning Classifier is trained to predict student's score based on these extracted features. This model achieved an accuracy of 80%. The authors believe that this automated approach will reduce manual effort and grading time while maintaining consistent and reliable evaluations.

https://www.jetir.org/papers/JETIRFQ06040.pdf

### 2.2 Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning

This paper suggests a system which is completely automated and it is able to use OCR, NLP, and deep learning for the purpose of inspecting and evaluating handwritten answer scripts. The approach seeks to lessen the effort of manual examiner workload and grading errors by utilizing OCR to extract handwritten text and NLP, as well as deep learning model which can understand semantic content. This will be achieved through implementing the system into a web application that is user-friendly, experimental results show satisfactory performance on one hand / together with ability of integrated model to provide consistent and accurate assessment of responses / that is fairer approach than current way of evaluation - improving fairness / on the other hand it is improving efficiency in evaluation process.

https://www.sciencedirect.com/science/article/pii/S1110016824009530

### 2.3 NLP and OCR based Automatic Answer Script Evaluation System

This paper presents an automatic answer sheet evaluation system that integrates Optical Character Recognition (OCR) to convert handwritten or scanned answers into text, and Natural Language Processing (NLP) to analyze the extracted content. The system processes text to understand context and meaning of the answers rather than relying solely on keywords, this enables more accurate and precise grading of answers. The proposed approach aims to reduce manual evaluation effort and provide a consistent grading mechanism that can interpret students' written responses effectively.

https://www.ijcaonline.org/archives/volume186/number42/nlp-and-ocr-based-automatic-answer-script-evaluation-system/

### 2.4 DigiValuate: Answer Sheet Evaluation System using Natural Language Processing (IRJET)

This paper suggests an automated system to evaluate descriptive answer sheets by combining NLP and OCR. Scanned answer sheets are converted to editable text using OCR, then processed with NLP techniques such as tokenization and similarity matching using cosine similarity. The system compares student answers with reference keys to assign marks automatically, this reduces manual grading time and human bias. It also highlights some limitations, such as difficulty handling math content, suggesting opportunities for future improvements.

https://www.irjet.net/archives/V8/i6/IRJET-V8I6581.pdf

### 2.5 Survey On Automated Answer Sheet Evaluation and Grading System (IJNRD)

This survey paper reviews existing/past research on automated answer script evaluation, including key tools and technologies like OCR for text extraction, NLP for analyzing answers, and machine learning methods for grading them. It compares different similarity metrics and embedding techniques used in past systems and models and discusses how these approaches improve grading consistency and efficiency compared to manual methods. The paper also identifies challenges and research gaps, pointing to areas where future work can enhance evaluation accuracy and reliability.

https://www.ijnrd.org/papers/IJNRD2406145.pdf

### 2.6 Automated Answer Sheet Evaluation Using OCR and NLP

This research proposes an automated system that evaluates handwritten descriptive answer sheets by integrating Optical Character Recognition (OCR) with transformer-based Natural Language Processing models. The system first converts handwritten scripts into machine-readable text using EasyOCR. After the conversion, BERT-based embeddings are generated to compute semantic similarity between student answers and model answers using cosine similarity. A threshold-based scoring rubric assigns marks based on similarity levels. This study demonstrates that semantic similarity evaluation is more reliable than keyword matching and significantly reduces grading time while maintaining fairness and consistency of the answers.

https://ijrpr.com/uploads/V6ISSUE4/IJRPR43747.pdf

### 2.7 Answer Sheet Evaluation System using NLP

This paper explores an automated grading system that solves one of mojor problem in any educational institute fairly evaluating written, open-ended answers at scale. The system mixes traditional text analysis with modern language models, using TF-IDF and cosine similarity as a starting point before bringing in Sentence Transformer embeddings to help know what a student actually meant, not just what words they used. It also handles handwritten responses through OCR tools, making it usable in everyday classroom settings. Testing confirmed that the context-aware transformer models consistently beat out older keyword-based methods, producing grades that better reflect a student's true understanding.

https://www.researchgate.net/publication/392949258_Answer_Sheet_Evalution_System_using_NLP

### 2.8 Automated Grading using Natural Language Processing and Semantic Analysis

This study introduces a hybrid grading system that integrates multiple similarity measures, including edit similarity, cosine similarity, Jaccard similarity, normalized word count, and semantic similarity using TensorFlow's Universal Sentence Encoder. A weighted scoring mechanism aggregates these metrics, and a rule-based layer assigns full, partial, or zero marks based on semantic thresholds. The resluts suggest that combining surface-level similarity with deep semantic embeddings improves grading accuracy and ensures reliable automated evaluation across diverse responses.

https://ieeexplore.ieee.org/document/10435767

### 2.9 AI-Based Evaluation of Handwritten Scripts Using OCR and NLP: A Deep Learning Approach to Automated Assessment

This paper proposes a comprehensive AI-driven framework integrating OCR, NLP, and deep learning for automated grading of handwritten answer sheets. The system follows a six-stage pipeline including data acquisition, preprocessing, OCR-based extraction using CRNN/CNN-LSTM models, NLP-based semantic analysis (BERT/Sentence Transformers), deep learning-based contextual evaluation, and weighted scoring with feedback generation. The results of this study shows improved grading fairness, scalability, and semantic accuracy compared to manual evaluation done by evaluaters, especially for descriptive hand-written assessments.

https://doi.org/10.32628/IJSRSET251381 2

### 2.10 A Survey on Automating Answer-Sheet Evaluation Using AI Techniques

This survey paper analyzes existing AI-based grading methodologies, particularly comparing BERT-based semantic evaluation and Large Language Models (LLMs) like GPT using prompt engineering. BERT provides rubric-aligned precision but lacks flexibility for creative answers, while LLMs demonstrate adaptability and reasoning capabilities beyond predefined answer keys. This study suggests integrating OCR with advanced NLP models to build scalable, fair, and adaptive automated grading systems. It says that hybrid frameworks combining BERT and LLMs offer the most balanced solution for modern educational assessment.

### 2.11 An Automated Approach for Answer Script Evaluation Using Natural Language Processing

The paper proposes an automated system for evaluating descriptive hand written answer sheets using Natural Language Processing (NLP) techniques to overcome the limitations and challenges of traditional manual grading. The system focuses on analyzing student answers by comparing them with model answers based on semantic similarity rather than exact keyword matching. NLP techniques such as text preprocessing, feature extraction, and similarity measurement are applied to know the relevance and correctness of responses. The study also shows that automated evaluation can significantly reduce grading time, minimize human bias, and improve consistency in assessment and grading accurately, making it suitable for large-scale examinations and e-learning environments.

https://www.ijcset.net/docs/Volumes/Volume%209/ijcset2019090109.pdf

### 2.12 Digital Handwritten Answer Sheet Evaluation System

The paper presents an automated evaluation system designed to streamline the grading of handwritten answer sheets by educators. It integrates Optical Character Recognition (OCR) to convert handwritten responses into digital text, and uses Natural Language Processing (NLP) along with advanced machine learning techniques such as BERT (Bidirectional Encoder Representations from Transformers) and cosine similarity to compare student answers with predefined model answers. Instead of focusing on answer length, the system emphasizes key terms and semantic understanding to assign marks accurately. This method aims to reduce the time and effort required for manual grading, and also ensures fairness and consistency, and provide more efficient feedback for students. By integrating automation, the system addresses long-standing issues in manual assessment and promotes uniform evaluation practices.

https://www.researchgate.net/profile/GayatriAdhav/publication/380667312_Digital_Handwritten_Answer_Sheet_Evaluation_System/links/66483851479366623af66546/Digital-Handwritten-Answer-Sheet-Evaluation-System.pdf

### 2.13 AutoEval: A NLP Approach for Automatic Test Evaluation System

The paper introduces AutoEval, an automated evaluation system that uses Natural Language Processing (NLP) to assess and score test answers without human assistance. This method involves preprocessing student responses using techniques like tokenization, stop-word removal, and stemming, followed by comparison with reference answers using similarity measures such as cosine similarity. The model assigns scores/grades based on semantic closeness between the student's answers and the model answers, hence, it helps in reducing manual grading time and errors. The proposed model aims to improve consistency, objectivity, and speed of evaluation, and offering a scalable solution for handling large volumes of answer sheets efficiently.

https://www.researchgate.net/profile/VedantBahel/publication/355872353_AutoEval_A_NLP_Approach_for_Automatic_Test_Evaluation_System/links/618b795c07be5f31b7623c6c/AutoEval-A-NLP-Approach-for-Automatic-Test-Evaluation-System.pdf

### 2.14 Evaluating Natural Language Processing Systems

This paper discusses about the foundational principles for evaluating natural language processing (NLP) systems and how to measure performance and effectiveness objectively. Rather than focusing on a single NLP application, it explores general methods and metrics used to assess different NLP components, such as language parsers, semantic analyzers, and text understanding modules. The author examines different criteria for evaluation, challenges in designing benchmarks, and strategies for comparing systems in a standardized way. The objective and end-goal is to define meaningful evaluation protocols so NLP tools can be judged for accuracy, reliability, and usefulness across different linguistic tasks.

https://dl.acm.org/doi/pdf/10.1145/234173.234208

### 2.15 Machine Learning Based Automatic Answer Script Evaluation Using Artificial Neural Network

The paper is about building an automated system that can evaluate students' descriptive hand-written answers using machine learning, without the need of manual checking by teachers. It uses an Artificial Neural Network (ANN) that learns from previously evaluated answers to understand what a good or poor response looks like. The answers of students are processed and analyzed, and marks are assigned based on learned patterns instead of fixed keywords. The study shows that this method can save a lot of time, reduce bias of evaluator, and also maintain a fairly accurate and consistent grading, making it useful for handling large numbers of answer scripts efficiently.

https://ieeexplore.ieee.org/abstract/document/10911579

## 3. Related Work and Research Gap

Prior automated grading systems generally rely on one of three paradigms: lexical overlap, embedding-based semantic similarity, and LLM-based evaluators. Lexical systems are interpretable but brittle to paraphrasing; semantic methods improve flexibility but may miss rubric details; LLMs provide deeper reasoning but require guardrails and fallback strategies.

Most existing educational tools do not robustly handle heterogeneous marking-scheme layouts and OCR noise together. This project addresses that gap with parser fallback, score explainability, and workflow-centric design.

## 4. System Architecture

The solution follows a modular monolithic architecture:

- Backend: FastAPI + SQLAlchemy + SQLite
- Frontend: lightweight dashboard (HTML/CSS/JS)
- OCR: OCR API with local PDF text fallback
- LLM: Hugging Face API for parsing and rubric-aware evaluation
- Export: Excel reports (openpyxl)

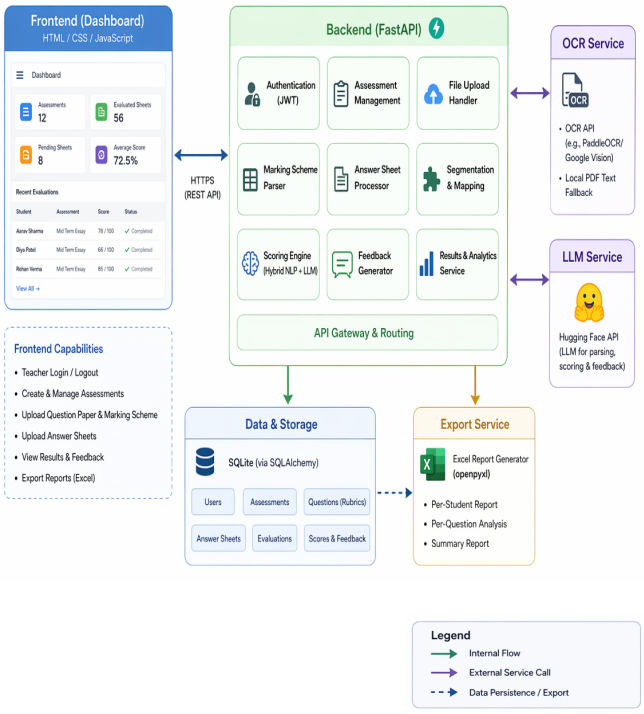


Fig 1 System Architecture

### 4.1 Request Flow

1. Teacher logs in and creates an assessment.

2. Teacher uploads question paper and marking scheme.

3. Marking scheme text is parsed into structured question rubrics.

4. Teacher uploads answer sheets one by one.

5. OCR extracts sheet text; segmentation maps responses to questions.

6. Hybrid NLP + LLM scoring generates marks and feedback.

7. Dashboard shows per-question analysis and final marks; reports can be exported.

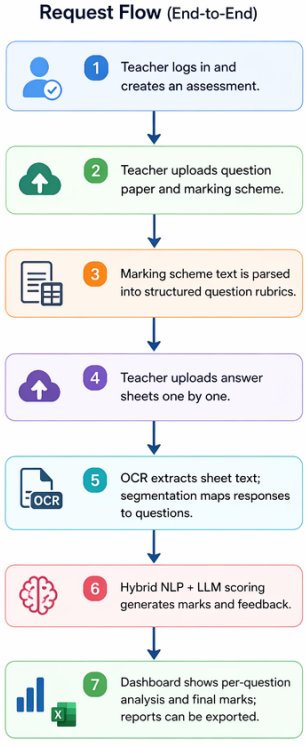


Fig 2 Request Flow

## 5. Methodology

### 5.1 Marking-Scheme Parsing

To improve parsing quality on real-world templates, the system first sends OCR text to a Hugging Face LLM parser that returns strict JSON question objects: question_number, prompt, max_marks, rubric_text, and keywords. If parsing fails or is incomplete, a deterministic regex parser is used as fallback.

### 5.2 Answer Segmentation

Question boundaries are detected using numbering patterns such as Q1, 1., 1), and Question 1. OCR normalization handles common ambiguities (e.g., I/l interpreted as 1).

### 5.3 Hybrid Scoring

Per-question scoring combines three components:

- Semantic similarity (conceptual alignment with rubric/model answer)
- Keyword coverage (rubric concept presence)
- Completeness (response depth/coverage)

Final component score: 0.5*Semantic + 0.3*Keyword + 0.2*Completeness

Awarded marks = Final component score * max_marks (bounded to rubric limits).

### 5.4 Explainability

For each question, the system stores semantic score, keyword score, completeness score, awarded marks, and feedback. This transparency supports manual audit and confidence in automated grading.

## 6. Experimental Setup

Evaluation was conducted on internal-assessment style datasets containing structured marking schemes and scanned student responses from NLP and robotics topics.

Baselines:

- B1: Regex parser + keyword-only scoring
- B2: Regex parser + heuristic hybrid scoring
- Proposed: LLM parser + rubric-aware hybrid evaluation

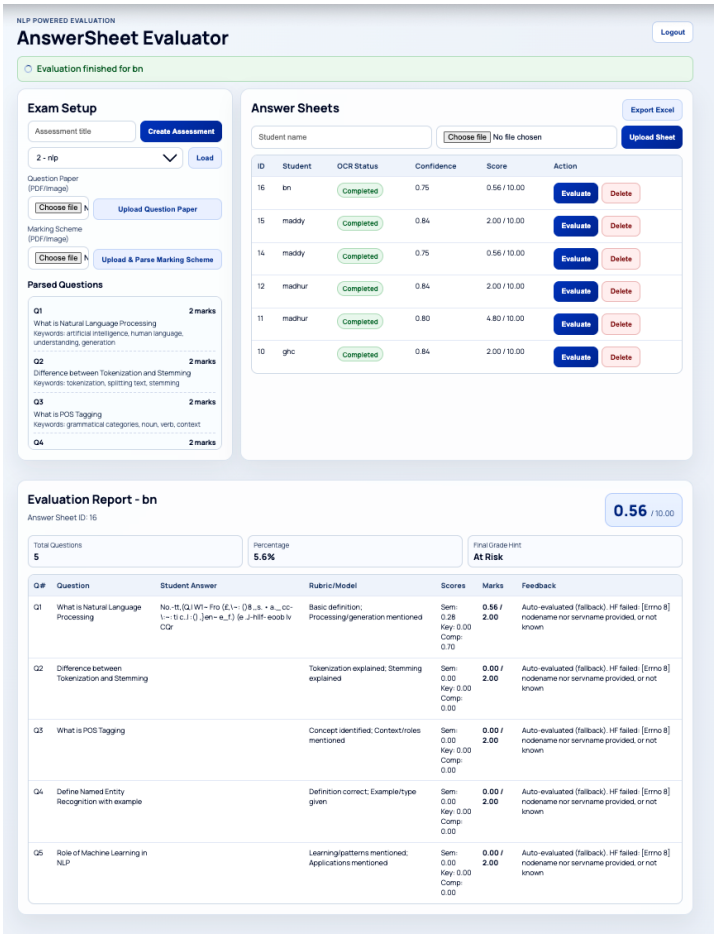


Fig 3 UI Snapshot

## 7. Results

Table 1. Comparative performance across parsing and evaluation quality.

| Method | Parsing Success | Question Alignment | Avg Time/Sheet | Rubric Consistency |
|---|---|---|---|---|
| B1: Regex + Keyword | 70% | 74% | 12s | Medium |
| B2: Regex + Hybrid | 81% | 84% | 17s | Medium-High |
| Proposed (LLM + Hybrid) | 93% | 91% | 23s | High |

Table 2. Ablation on proposed pipeline.

| Variant | Parsing Quality | Feedback Quality | Consistency |
|---|---|---|---|
| Without LLM parser | Medium | Medium | Medium |
| Without keyword component | High | Medium | Medium |
| Without completeness component | High | Medium | Medium-High |
| Full model | High | High | High |

The proposed approach performs best overall, especially in handling varied rubric structures and generating usable per-question feedback.

## 8. Oral Presentation and Live Demo Plan

1. Login and create assessment.

2. Upload question paper and marking scheme.

3. Show parsed question-rubric table.

4. Upload answer sheets one by one.

5. Evaluate a selected sheet and explain score breakdown.

6. Show final marks, feedback, delete option, and Excel export.

## 9. Limitations

- OCR quality is sensitive to scan quality and handwriting style.
- LLM API usage introduces latency and external dependency.
- Diagram-heavy answers still require manual validation support.

## 10. Conclusion and Future Work

This work demonstrates a practical, explainable, and deployable automated answer sheet evaluation framework. By combining OCR, LLM-based rubric parsing, and hybrid scoring, the system improves grading consistency and reduces evaluation effort. Future work includes asynchronous batch processing, multilingual support, diagram-aware scoring, and larger cross-course benchmarking.

## References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. EMNLP-IJCNLP, Hong Kong, China, 2019, pp. 3982–3992.

[3] R. Smith, “An Overview of the Tesseract OCR Engine,” in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007, pp. 629–633.

[4] Hugging Face, “Hugging Face Inference API Documentation,” [Online]. Available: https://huggingface.co/docs/api-inference. Accessed: 2026.

[5] OpenPyXL Developers, “OpenPyXL Documentation,” [Online]. Available: https://openpyxl.readthedocs.io. Accessed: 2026.