IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

Autograder : Automated handwritten subjective answer evaluation system

Sarthak Karmalkar[1], Ansari Siraj[2], Shakti Singh[3], Mrudula kulkarni[4],

Prof. Naved Raza Q. Ali[5]

Undergraduate Survey Paper, Department of Computer Engineering, SKNCOE, Savitribai Phule Pune University, Pune 411041, India[1,2,3,4]

Department of Computer Engineering, SKNCOE, Savitribai Phule Pune University, Pune 411041 India[5]

Abstract: In today's technologically advancing academic environment, the timely and accurate evaluation of subjective answers plays a crucial role in educational assessment. While objective-type answers can be easily evaluated using automated systems, the assessment of subjective responses demands more sophisticated techniques that consider context, content relevance, structure, and grammatical accuracy. This research paper presents an AI- based assessment framework capable of evaluating both handwritten and typed subjective answers using Machine Learning (ML) and Natural Language Processing (NLP) methods. Handwritten responses are first digitized using Optical Character Recognition (OCR), which converts the input into textual data. Subsequently, the evaluation process utilizes semantic similarity measures, keyword extraction, and grammatical analysis. The framework integrates pretrained language models along with custom-trained classifiers to compare student responses against reference solutions, enabling the assessment of contextual accuracy and logical coherence. The proposed method reduces human bias, enhances consistency through algorithmic scoring, and significantly minimizes manual grading effort. The results this work demonstrates that the system achieves high accuracy (99.30%) and outperforms traditional evaluation techniques in speed and reliability. This framework offers a scalable and intelligent approach for subjective answer assessment, contributing meaningfully to the integration of AI in modern education.

Keywords: Optical Character Recognition (OCR), Convolutional Neural Networks (CNN), Machine learning(ML) Natural Language Processing (NLP), Large Language Models (LLM), Subjective Answer Assessment.

I. INTRODUCTION

Assessing private responses is an important but delicate part of educational assessment. Unlike objective questions, private answers are different in length, structure, and vocabulary, and they bear mortal observers to assess content as well as the environment and expression of the response. This process can be very time-consuming, mentally exacting, and error-prone.

At the same time, handwriting recognition capabilities have become increasingly useful in numerous sectors. The parallels stem from the capability of handwriting recognition to reduce homemade workloads, as it's a way to convert a handwritten source to a machine-readable format, thus automating data entry or other document processing that requires homemade review. Although Optical Character Recognition (OCR) was originally seen as a result of getting published textbooks into a machine-readable format, ultramodern systems include systems to fetch and interpret handwritten input, which is a much bigger ask due to the variation in lots of factors like writing style, size, and symbols generally referred to as handwriting variation [1].

To address these enterprises in handwriting recognition surroundings, there's clear substantiation in the literature to suggest that machine literacy methods, in particular Convolutional Neural Networks (CNNs), can gain state-of- the-art performance on offline handwriting recognition [2].

Still, occasionally you want to go beyond just recognition and development of alternate-order additions to the offline recognition operation, which includes but is not limited to it. It would be easy to imagine what use this type of technology could be for educationalists looking to automate handwriting answer evaluation.

This paper presents a system that isn't only suitable for comparing pupils handwritten answers against "correct" answers; it also provides a richer subcase of modelling, as it uses Natural Language Processing (NLP) paradigms and strategies

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁓ Peer-reviewed & Refereed journal ⁓ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14652

for comparing answers to destined reference answers using token recognition and semantic similarity measures (among others), to produce meaningful content. This Research paper introduces a new system that automatically checks descriptive answers using Machine Learning (ML) and Natural Language Processing (NLP) technologies. The system is designed to work similarly to a human grader. It reads and understands the response, compares it to the ideal answer, picks out important ideas, and then gives a score by looking at how accurate and relevant the answer is.

To tackle these issues, this paper presents an automated system for assessing descriptive answers utilizing Machine Learning (ML) and Natural Language Processing (NLP). The system mimics how a human grader operates—comprehending the response, comparing the response with a perfect answer, determining major concepts, and rating based on relevance and correctness.

The focus of this project is to test how effective the Mistral 7B language model is in extracting handwritten text. Comparison of the OCR output from the model to the ground truth transcriptions at word level was given in a comprehensive evaluation.

Section 1 of the paper contains Introduction, section 2 contains Literature Survey, section 3 contains Proposed Methodology, section 4 contains Result, section 5 contains Future Work, Section 6 contains Challenges, section 7 contains Conclusion and section 8 contains References.

II. LITERATURE SURVEY

This literature survey explores recent advancements in AI- driven automated answer evaluation systems, focusing on methodologies like OCR, neural networks, and large language models. Preetha S. et al. [1] presented handwriting recognition, with a special focus on CNN for handwritten character recognition. She and her team discuss about seven techniques for handwriting recognition. Namely, CNN, Incremental method, Semi incremental method, Slope and slant, Line and word segment, Part based method, Ensemble method. She also mentioned about OCRs as they lack accuracy.

Jamshed Menon et al. [2] presented comprehensive review on OCRs for handwriting recognition between 2009-2019. The paper summarized various machine algorithms used for Text recognition like, support vectors, decision tree, K-nearest neighbour. Among all machine learning algorithm, SVM is most accurate (92-98.4%). This was tested on six most widely spoken language. Further, the paper challenges research in OCR. Also, recognition in real-world settings and requires large amount of dataset size and variety.

Varun Aggarwal et al. [3] presented, the computer-assisted system to automate the assessment of student’s answers to subjective questions. This paper argues that due to traditional assessment methods, a lot of time can go into waste and some selection bias or human bias does exist. Therefore, authors recommended using similarity algorithms like “cosine similarity”, “Fuzzy Wuzzy”, or “Jaccard similarity” for comparing the answers provided by students to the reference answers. The study was carried out on the data of 100 IT students' answers concerning the APIs. The results indicate that the cosine similarity performed best, with an accuracy of 74.7% compared to that of a human grader. The paper concludes on the importance of having multiple reference answers by the testing center as one way of boosting the accuracy of the automated assessment system.

Sharad Bharadiya et al. [4] proposed the use of Machine Learning for automatic evaluation of answers. The system uses optical character recognition on handwritten answers to compare them with a database of keywords and length parameters given by teachers. This will scale down the timescale and human effort invested in grading subjective questions. Hence, results in faster evaluation and consistent than by hand grading.

Era Johri et al. [5] presented an architecture of subjective answer evaluation using Semantic Learning, Sentence Encoding, Similarity Matrix. Similarity Matrix used by them allowed the generation of feedback for the students answer but at the same time system was not capable of evaluating contradictory answers compared to the model answer.

Farrukh Bashir et al. [6] combined the various methods like NLP, Tokenization, Stop words removal, POS tagging, Lemmatization, Stemming, Case folding, Bag of Words, TF-IDF, Word2Vec, Cosine Similarity, Jaccard Similarity, WMD in order to improve the answer evaluation methods, which rose the efficiency of the system to 88%.

Kavita Shirsat et al. [7] proposed a model to evaluate subjective answers using NLP, Machine Learning, Similarity Index (Cosine Similarity) and Universal Sentence Encoding. The proposed model used 3 factors for evaluating the answers, namely similarity index, grammar and question specific parameters, but small size of dataset used to train the model

Impact Factor 8.471 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

inhibits the efficiency of the system. Universal Sentence Encoder(USE) converts sentences into 512 dimensional embedding vector, which preserves the actual meaning and context of the answers of the students. Question Specific Parameter searches for a particular keyword in the written paragraphs of the students. Assigned values 0 and 1, based upon the grammar of the answer(0 for bad grammar and 1 for good grammar).

Shreya Singh et al. [8] built a system for the evaluation of handwritten sheets of students by using OCR to convert handwritten answers into digital text, summarizing the answer to further evaluate the responses. The use of Recurrent Neural Network (RNN) improved the image recognition performance, along with algorithms like Word2Vec and TF-IDF to find the importance of every word in the answer.

Nandita Bharambe et al. [9] presented a system that combined both handwriting to digital text conversion and the evaluation of the same. They used supervised learning algorithms like Artificial Neural Network (ANN) and OCR to detect the text present in the handwriting. In ANN, the use of Back-propagation allows the neurons to improve their performance by sending errors to the previous neurons. To evaluate the answers, they used Cosine Similarity, matching the keywords of the model answer and also giving importance to the length of an answer. Marks are assigned out of 10; if the length of an answer is less than the required length, marks are given between 1 to 4. Conversely, if the length exceeds the threshold set by them, then the marks of the student will lie between 5 and 10.

Prerana M S et al. [10] introduce a framework that integrates models like BERT, GPT-3, CNN, LSTM, and SVM with Optical Character Recognition (OCR). Their goals include improving scoring accuracy and scalability, while also addressing challenges such as keyword dependency and the significant resources needed for model training.

Vijay Kumari et al. [11], leverage BERT and TF-IDF to highlight the balance between automation and consistency. Their system basically included two important modules Checker and Evaluator. Their method relies heavily on predefined correct answers, which can create difficulties when faced with ambiguous responses.

Sheik Abdullah et al. [12], concentrate on OCR, NLP, and Connectionist Recurrent Neural Networks (CRNN), addressing issues related to handwriting variability and the challenges of data collection. Their research offers valuable insights into the complexities involved in automating the grading process. In [13], they mentioned OCR, NLP, and ML are used to automate answer script grading. It helps increase accuracy and efficiency but has issues in terms of handwriting variability and the inability to evaluate the non-text components properly. Open- ended questions need more development for their proper evaluation as well as diverse formats of answering scripts.

Vaibhav Shikhar Singh et al. [14] share their findings where the system achieves an accuracy rate of 83.14% by utilizing Artificial Neural Networks (ANN) for OCR tasks, although they recognize ongoing challenges related to handwriting quality. Collectively, these studies demonstrate notable advancements in automatic answer evaluation systems while underscoring the necessity for further research to improve their effectiveness in educational contexts.

Madhavi Kulkarni et al. [15] built a system to evaluate handwritten answer sheets using OCR, NLP, ML, BERT and cosine similarity. OCR can parse handwritten content into a structured format, but not process further before giving the NLP model, which actually performs textual data analysis, helping to match student answers with model answers pre-defined in machine-readable text. So, this provides predictions of new submissions based on analysis of the data. It also creates instant feedback for students highlighting where they may have strayed, in order for the assessment process to be more efficient and precise.

Md. Afzalur Rahaman et al. [16] Implemented a system for evaluation of handwritten sheets of students by using Bidirectional LSTM Network (BiLSTM), Convolutional Neural Networks (CNN). They includes powerful tools such as: Natural language processing (NLP), support vector regression (SVR) and bayesian linear ridge regression(BLRR) to Automatically grade handwritten answer scripts. The system will save educators time, eliminate human error, and be for practice learning and then improve. But even though the system is highly practical, its limitations such as computational complexity along with contextual understanding and recognition challenges have raised several important aspects that need to be addressed. However, today the model gets only around 80% accuracy far from enough that confidence can be instilled to deploy the system into a production environment for handling complex textual and image-based content.[15]
Sangeeta Mangesh et al. [17] presented Subjective Answer Script Evaluation using Natural Language Processing, in which various techniques were combined, namely NLP, Gaussian Naive Bayes Approach, Machine learning, and Cosine Similarity. The proposed system will generate meaning from textual content using the techniques of NLP. It will classify and predict scores based on NLP features of the text and train it using previously evaluated answers. The system achieves around 80% accuracy and improves by adding some mathematical tools that will assess specific mathematical and chemical equations with precision.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁚ Peer-reviewed & Refereed journal ⁚ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

III. PROPOSED METHODOLOGY

3.1 System Architecture

The proposed system involves two primary user roles: students and teachers, each interacting with the system to perform distinct tasks. The system is designed to facilitate the automated evaluation of handwritten subjective answers using advanced machine learning and natural language processing (NLP) techniques.

The backend of the system follows a microservice architecture, where each service operates independently, ensuring scalability, modular development, and fault tolerance. All data transactions pass through a centralized MongoDB database, which stores student responses, evaluations, and other relevant metadata. The system's design ensures that key user interactions, such as answer submissions, login authentication, and evaluation retrievals, are processed efficiently and securely.

Machine learning models are integrated to provide personalized feedback, automated grading, and result analytics. MongoDB ensures efficient storage and retrieval of large volumes of educational data, making the system suitable for institutional-scale deployments.

Diagram Explanation

The architecture diagram illustrates the end-to-end flow of data and user interactions within the system. Here's how it works:The Login Module enables both students and teachers to authenticate securely. Through the Answer Submission interface, teachers upload handwritten answer sheets in image or PDF format. These are processed using OCR Conversion via the Mistral API to extract text from handwriting. The extracted text undergoes Preprocessing and Fine-Tuning to clean and align it with evaluation criteria using domain-specific knowledge.

The refined text is sent to a Large Language Model (LLM) through the Mistral Developer Platform for contextual evaluation and grading. The Result Storage and Retrieval module saves the scores in a MongoDB database, which students can access via the View Score module. A Service API Layer ensures communication between all components and manages system orchestration. This diagrammatic representation highlights the modularity and workflow of the answer evaluation system.

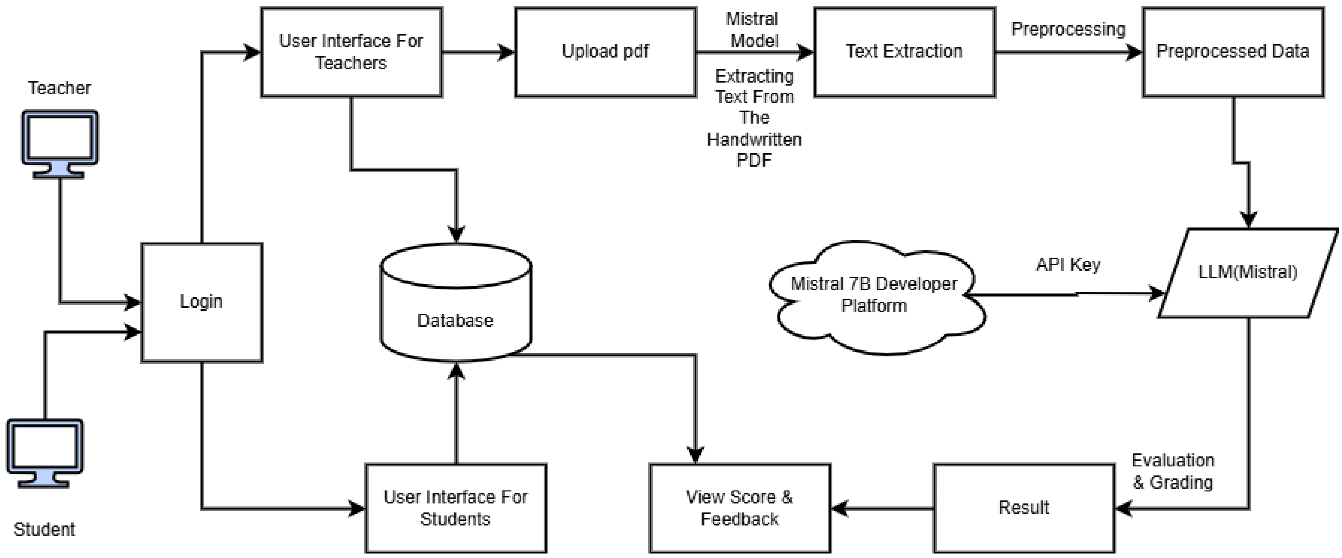

Fig. 3.1 System Architecture

The architecture's modular design ensures flexibility, ease of maintenance, and future extensibility making the system not only reliable but also adaptable to evolving educational needs. This comprehensive workflow demonstrates the integration of AI, OCR, and modern web technologies to automate and enhance handwritten answer evaluation at scale.

3.2 Proposed Algorithm for Evaluation

Algorithm : Automated Handwritten Answer Evaluation System INPUT: Uploaded handwritten scripts (PDF/Image)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

OUTPUT: Structured evaluation results including scores and feedback.

1. for each uploaded file f do
2. convert f to a standardized image format if needed
3. use Mistral API to extract OCR text from each page/image
4. aggregate extracted text into unified document doc_f
5. apply preprocessing on doc_f:
6. – remove noise, extra spaces, line breaks
7. – correct OCR misreads (e.g., “0” ↔ “O”, “1” ↔ “I”)
8. – normalize structure (bullets, paragraphs)
9. structure doc_f for compatibility with LLM
10. load corresponding question paper Q_f
11. segment doc_f into answer chunks based on question markers
12. for each answer chunk a_i in doc_f do
13. formulate prompt by pairing a_i with q_i from Q_f
14. send prompt to fine-tuned LLM
15. receive JSON {score, justification}
16. store {q_i, a_i, score, justification} in evaluation record E_f
17. end for
18. store E_f, doc_f, and file metadata in database
19. associate E_f with user ID and file ID
20. end for
21. for each E_f do
22. display original file, OCR text, and evaluation in Streamlit UI
23. allow manual score adjustment by teacher
24. end for
25. collect user feedback on LLM accuracy
26. validate and add feedback to training dataset
27. retrain and redeploy LLM periodically
28. return all evaluation results

3.3 Methods of Evaluation

Several crucial steps are involved in evaluating student responses:

1. Length and Completeness of the Response Length and Completeness of the Response

The system evaluates whether the student’s answer satisfies the specified length requirement (for example, 300 words) to ensure it aligns with the expected structure and depth necessary for a thorough assessment.

2. Pertinence

Using NLP techniques, the semantic similarity between the student’s response and the predetermined correct answer is evaluated to determine the relevance of the response. This step assesses how closely the concepts and context of the student’s response match the intended solution.

3. Logical and Contextual Accuracy

The system evaluates the reasoning and context of the student’s response using the Mistral AI model. This ensures that the answer is logical and accurately addresses every aspect of the query, providing a robust measure of accuracy and completeness. 4. Correctness of Grammar Grammar-checking algorithms are used by the system to evaluate the syntax and grammar of the student’s response. This ensures that the evaluation considers the clarity and quality of the answer.

3.4 Accuracy and Performance Metrics

The system's evaluation performance is measured using several key metrics to ensure it aligns with human evaluation standards:

1. Cosine Similarity Score: This score quantifies the similarity between a student’s response and the reference solution, helping assess the alignment of the response’s meaning with the correct answer.
2. Precision, Recall, and F1-Score: These metrics evaluate the system's ability to correctly identify relevant answers (precision) and retrieve correct answers (recall). The F1-score balances these metrics to provide an overall performance measure.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

3. Human Evaluation Benchmarking: To ensure reliability, the system's scores are compared with those given by human evaluators. The average deviation between system-generated and human scores verifies the consistency and accuracy of the evaluation.
4. Mistral AI Model Accuracy: The Mistral AI model, utilized for language processing and answer evaluation, is tested against a set of student responses. It achieves a high accuracy rate of 99.30%, confirming its ability to evaluate answers based on context, grammar, and relevance.

3.5 Visualization and Student Feedback

Once the evaluation is complete, the system provides the students with comprehensive feedback:

1. Marks: The final score is calculated based on the above evaluation criteria.
2. Feedback: Constructive comments on grammar, relevance, context, and logic.
3. Visual Representation: The system provides visual feedback through bar charts and graphs, allowing students to track:

a. Overall performance on the question.
b. Performance analysis by criteria, such as grammar, content relevance, and logical accuracy.
c. Areas for improvement based on the evaluation.

This feedback helps students better understand their strengths and areas for growth, making the learning process more transparent and actionable.

IV. RESULT

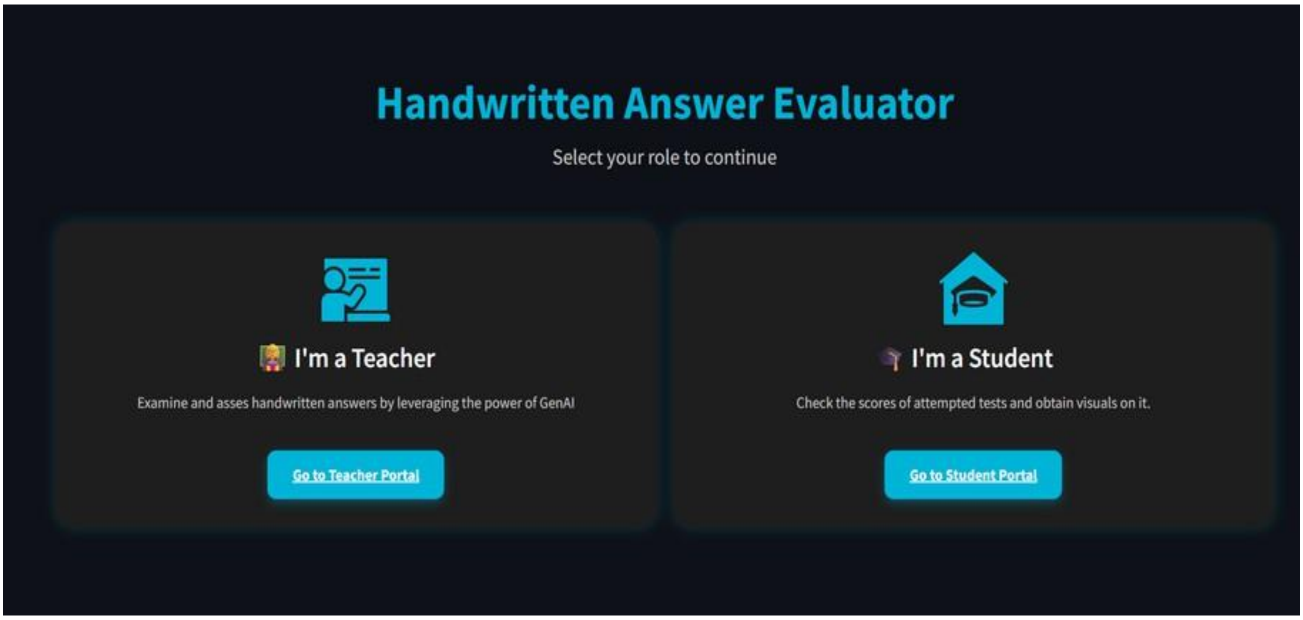

Fig. 4.1

The role selection page is the initial interface of our Autograder system, designed to streamline user access by allowing individuals to choose their role as either a teacher or a student. This intuitive landing page ensures users are directed to the appropriate workflow tailored to their needs. Teachers are guided toward creating and managing assessments, while students proceed to view their respective scores. The clean and minimal design promotes ease of use and quick decision-making. By establishing clear user roles from the beginning, the system enhances user experience and maintains organized functionality across different types of users within the platform.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

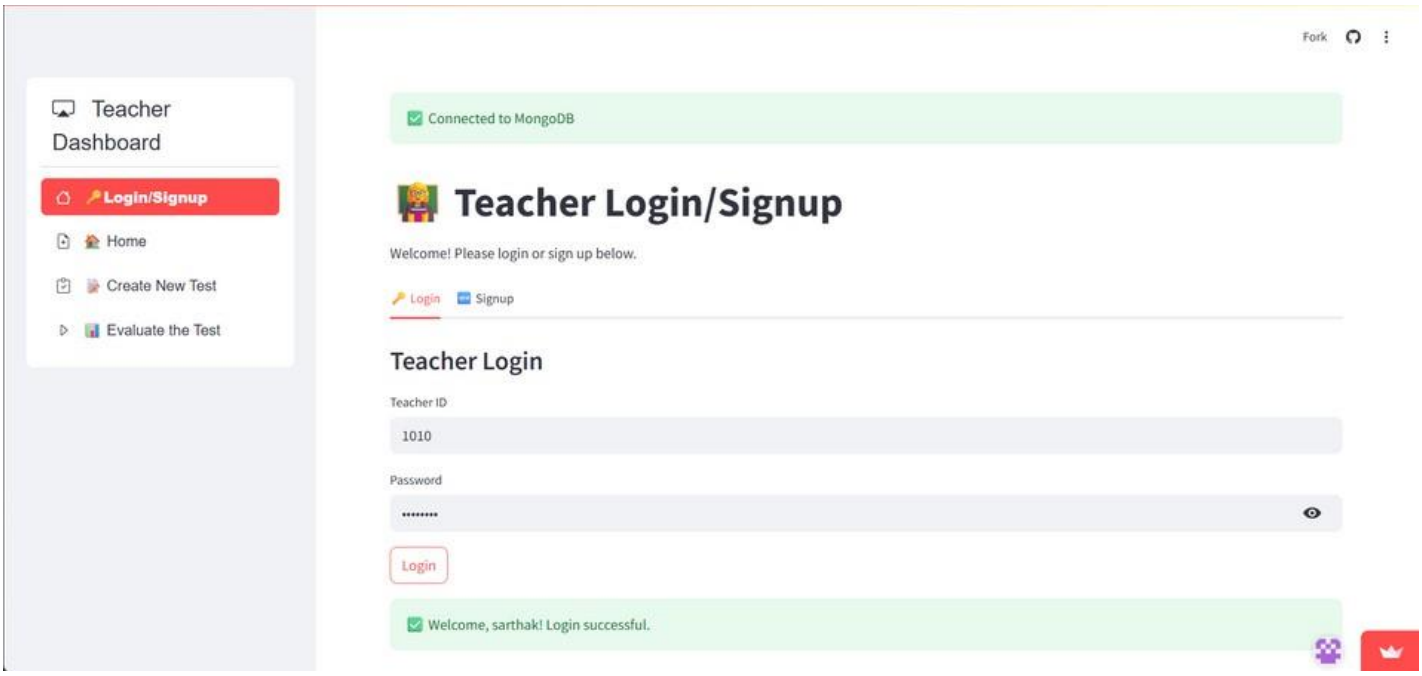

Fig. 4.2

The teacher login page serves as a secure gateway for educators to access their accounts within the Autograder system. It requires users to enter their registered email and password, ensuring authorized access to sensitive academic data. With a clean and intuitive design, the page facilitates quick and hassle-free login, allowing teachers to seamlessly continue their tasks. Upon successful authentication, educators are directed to their dashboard where they can create assignments, manage questions, and review student performance. This page plays a vital role in maintaining the system's security while providing teachers with smooth access to their personalized teaching environment.

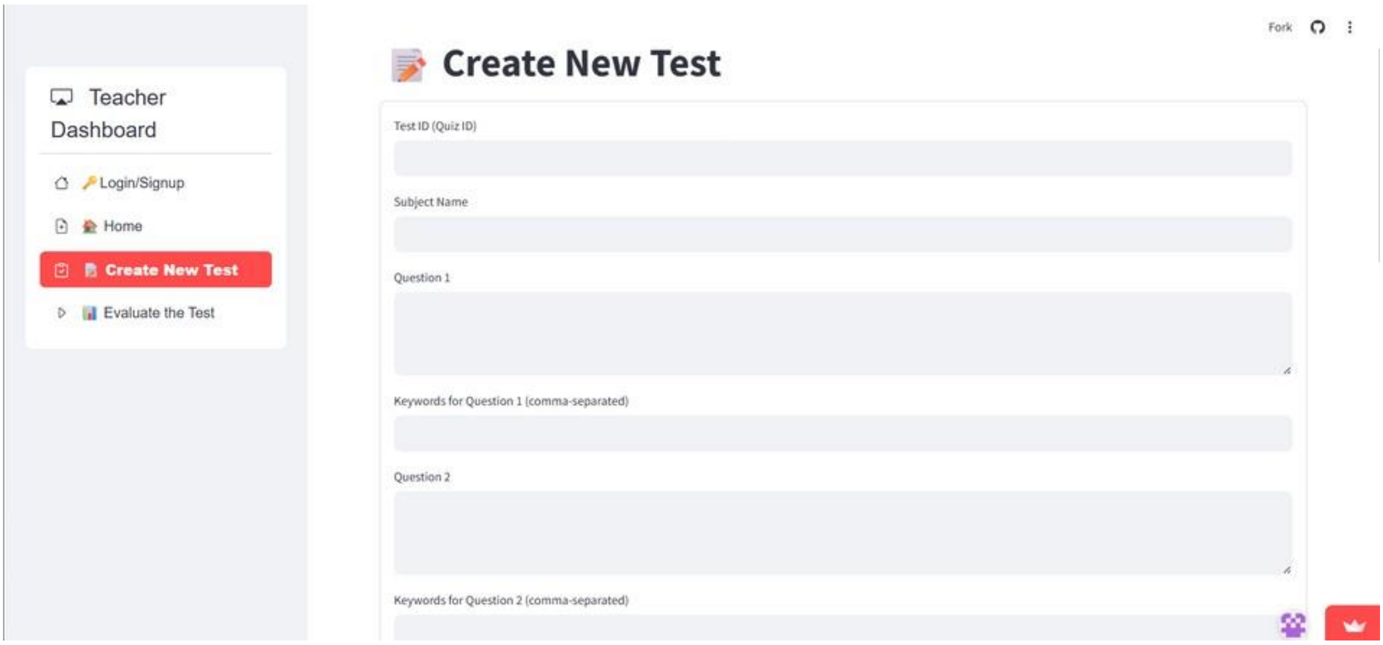

Fig. 4.3

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14652

The create test page is a key feature of the Autograder system that enables teachers to design and publish assessments with ease. On this page, educators can input a unique test ID, assign a test name, and enter up to five custom questions. The interface is simple, organized, and optimized for quick test creation, allowing teachers to focus on crafting meaningful assessments. Once the test is created, it is saved and made accessible to students for submission and evaluation.

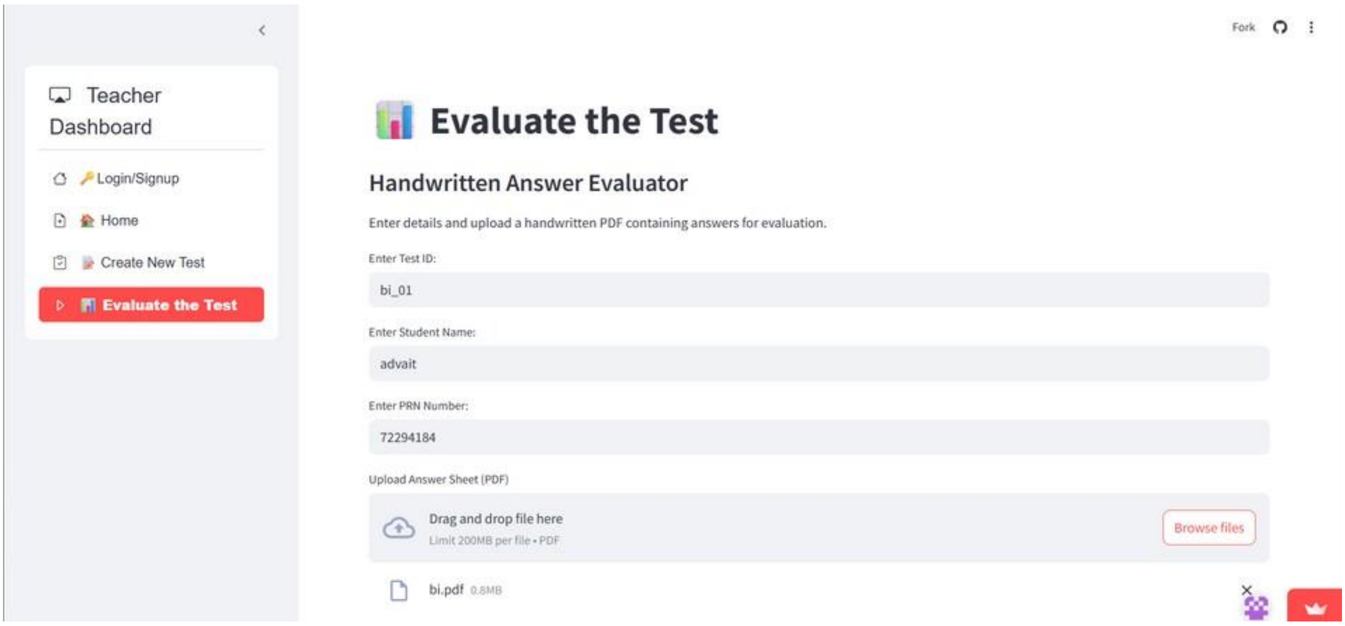

Fig. 4.4

The handwritten answer evaluation page is a powerful tool in the Autograder system that allows teachers to upload and evaluate students' handwritten answer sheets. On this page, educators can select and upload a PDF file of the student's answer sheet, enter the corresponding test ID, and provide the student's name. The system then processes the uploaded document using advanced evaluation algorithms to assess the responses. With a user-friendly interface and seamless upload functionality, this page simplifies the grading process for handwritten submissions. It enhances efficiency, reduces manual effort, and ensures accurate evaluation while maintaining a smooth and structured workflow for teachers.

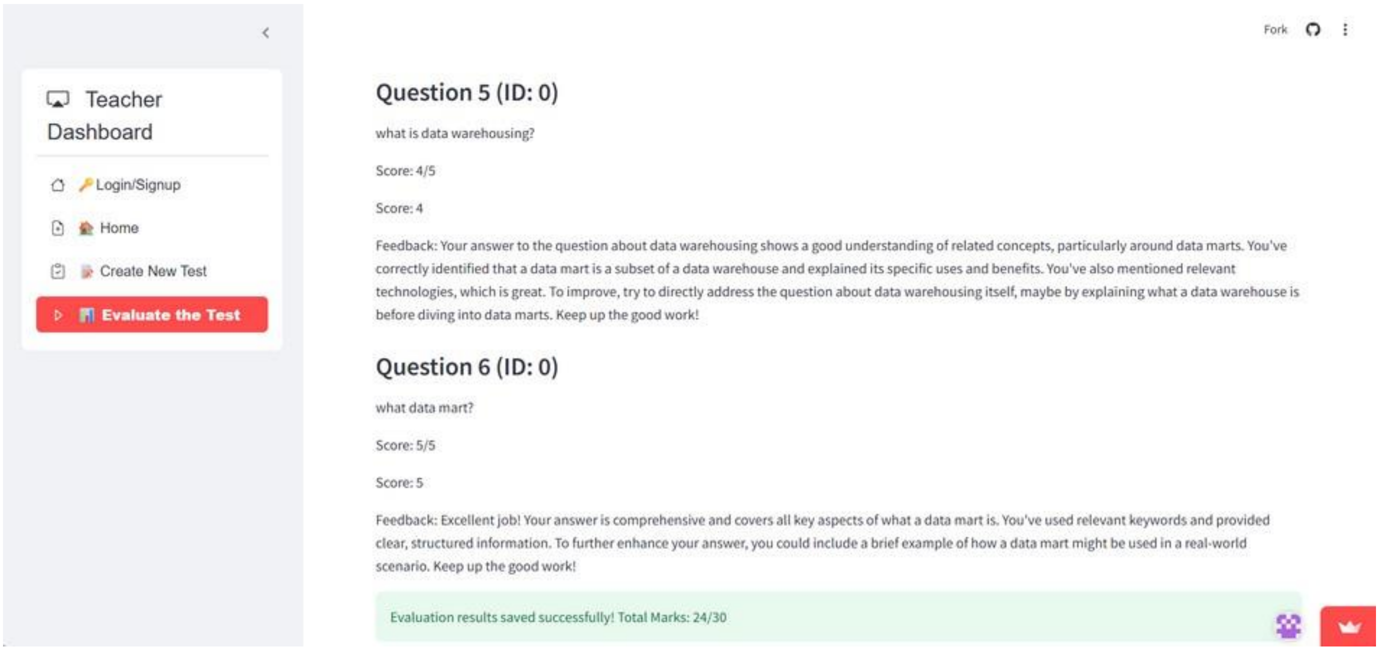

Fig. 4.5

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

The Evaluation page(Fig. 4.5) shows how LLMs evaluate each and every answer along with feedback for every answer. At the end the total score is displayed

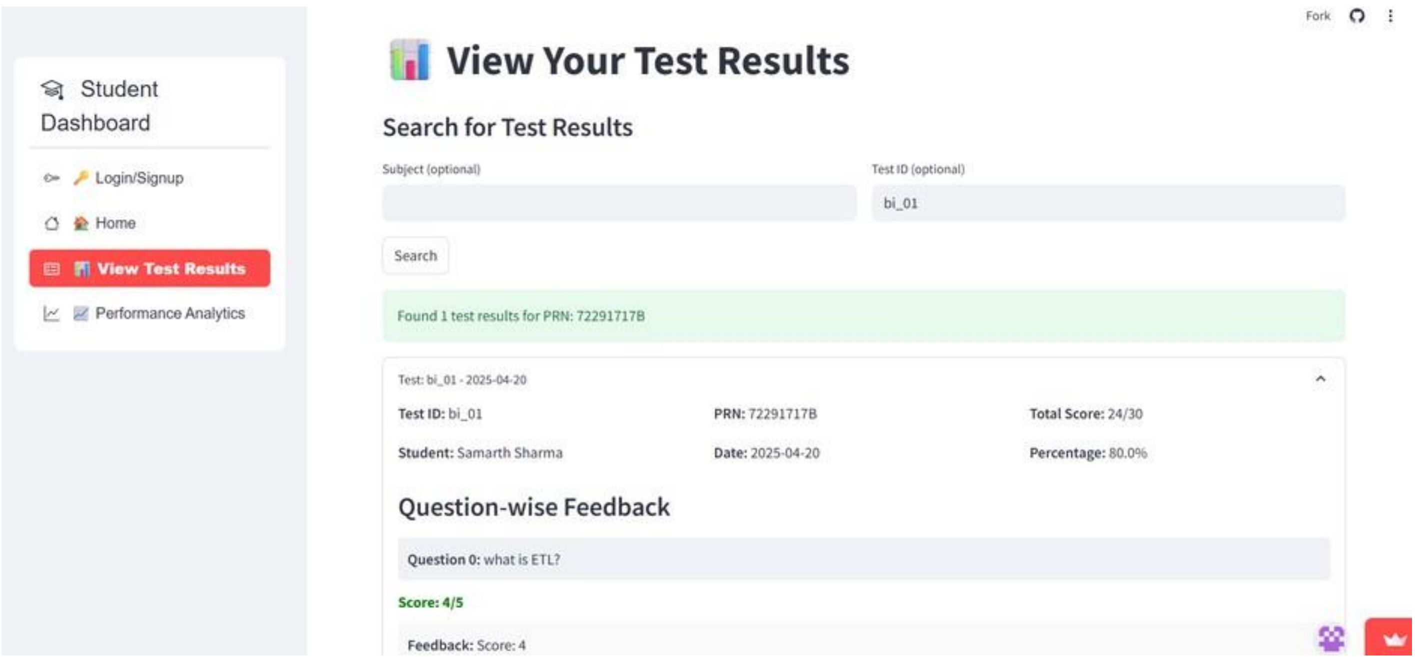

Fig. 4.6

The view score page is a student-focused feature in the Autograder system that allows learners to check their performance across different subjects. By entering the subject ID and test ID, students can instantly access their evaluated scores for specific assessments. This page offers a clean and responsive design, ensuring students can easily retrieve their results without confusion. It provides clear visibility into individual test scores, helping students track their academic progress and identify areas for improvement. Serving as a transparent feedback channel, the view score page enhances the learning experience by delivering quick, reliable access to performance insights.

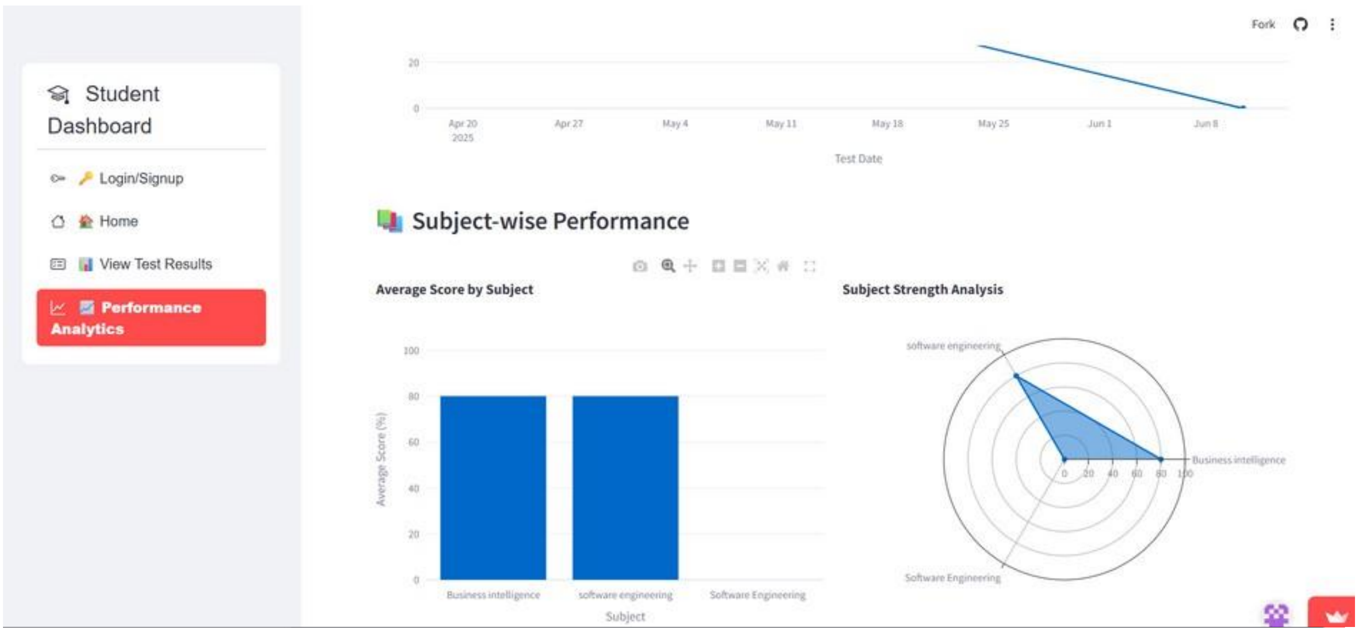

Fig. 4.7

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14652

The performance evaluation page shows visual representation for previously attempted tests for individual subject as well as all subject

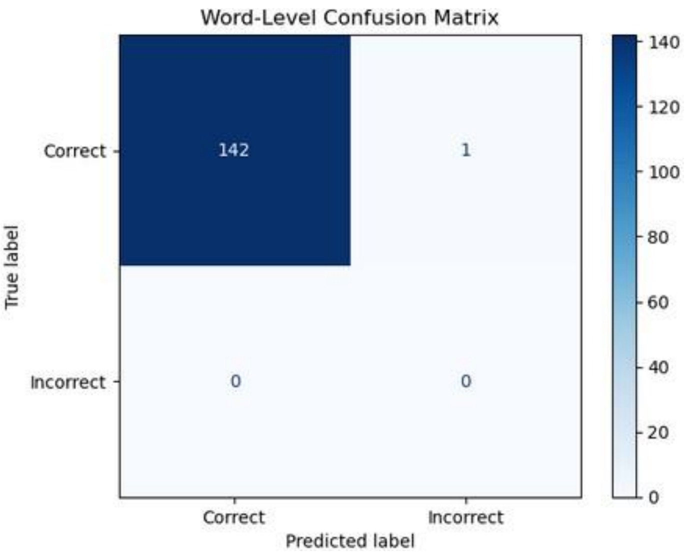

Fig. 4.8

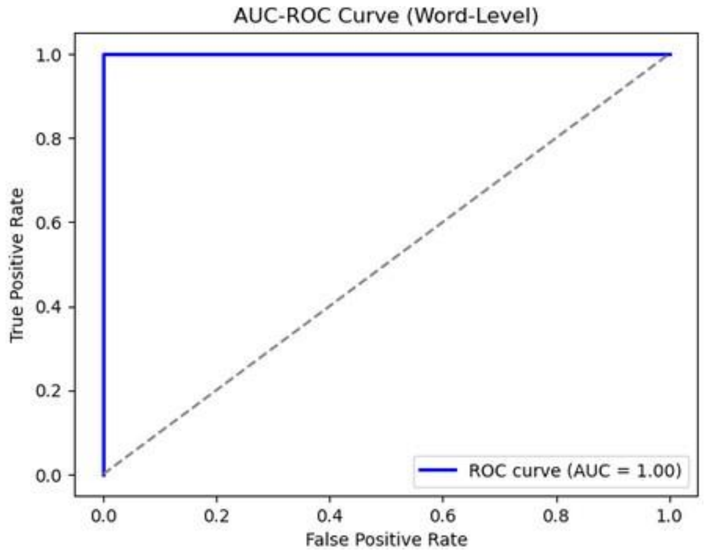

Fig. 4.9

To rigorously assess the performance of the Mistral 7B language model in handwritten text extraction tasks, we conducted a comprehensive evaluation using both confusion matrix analysis and the AUC-ROC (Area Under the Receiver Operating Characteristic) curve. The evaluation was carried out at the word level, comparing the model-generated OCR output against a ground truth transcription. Using a custom evaluation function based on Levenshtein similarity, we quantified the correctness of each word prediction. The confusion matrix highlighted the distribution of correct versus incorrect recognitions, yielding a word-level accuracy of 99.30%. This high accuracy score indicates that the model was able to correctly recognize the vast majority of handwritten words with minimal error.

To further investigate the model’s discriminatory power between correct and incorrect word predictions, we plotted the ROC curve based on similarity scores. The resulting AUC (Area Under the Curve) score reinforced the model's effectiveness, demonstrating its strong capability in distinguishing between accurately and inaccurately predicted words. The ROC curve maintained a favorable balance between sensitivity and specificity, suggesting that the model performs reliably across varied handwriting styles and complexities. Together, these metrics substantiate the robustness and precision of the Mistral 7B model in real-world handwritten text extraction scenarios, establishing it as a viable solution for document digitization and OCR applications.

In addition to accuracy, the evaluation highlighted a substantial improvement in processing efficiency. While manual assessment of a single handwritten answer sheet typically takes approximately 3 minutes (180 seconds), our automated system accomplished the same task in just 20 seconds. This translates to an 88.89% reduction in evaluation time, significantly enhancing throughput and enabling large-scale deployment in academic or administrative settings where time is a critical factor. The combined accuracy and speed advantages position the Mistral 7B model as a highly effective solution for real-time handwritten document evaluation and digitization workflows.

V. FUTURE WORK

While the current system performs efficiently for evaluating handwritten answers in English, future improvements aim to increase its versatility. Expanding support to multiple languages would make it more accessible in diverse educational contexts. Adding capabilities to assess visual elements like diagrams and flowcharts is also a priority, especially for science and technical subjects. Enhancing open-ended answer interpretation, integrating automated feedback, and enabling LMS connectivity are other key goals. These upgrades will help evolve the system into a more comprehensive and adaptable assessment tool

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⅔ Peer-reviewed & Refereed journal ⅔ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

VI. CHALLENGES

The system effectively automates handwritten answer evaluation using image processing and a fine-tuned Mistral 7B model, but certain limitations remain. Currently, it supports only English responses, restricting its use in multilingual settings. It also lacks the ability to interpret visual content like diagrams and flowcharts, which are crucial in many subjects. Additionally, the system struggles with extremely poor handwriting, affecting text extraction accuracy and evaluation fairness. Addressing these issues—such as by adding multilingual OCR, diagram recognition, and handwriting tolerance—will be essential for broader and more reliable application.

VII. CONCLUSION

The developed system provides a powerful and efficient alternative to manual answer sheet evaluation by combining deep learning and natural language processing. Leveraging a fine-tuned Mistral 7B model for semantic understanding, along with real-time image preprocessing and text extraction, the system achieves an impressive 99.30% word-level accuracy and reduces total evaluation time by 88.89%. Its seamless integration with a user-friendly Streamlit interface ensures smooth and accessible interaction for both teachers and students.

With future enhancements such as multilingual OCR, diagram interpretation, and improved handling of diverse handwriting styles, the system is well-positioned to become a comprehensive and scalable solution for academic evaluation. This implementation lays a strong foundation for advancing intelligent, AI-powered assessment in education.

REFERENCES

[1]. Preetha S, Afrid I, Karthik Hebbar, Nishchay S, "Machine Learning for Handwriting Recognition," *International Journal of Computer (IJC)* Volume 38, No 1, pp 93-101, 2020.

[2]. JAMSHED MEMON, MAIRA SAMI, RIZWAN AHMED KHAN, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE ACCESS r10.1109/ACCESS.2020.3012542 VOLUME 8, Aug. 2020.*

[3]. Varun Aggarwal, Rutuja Sutar, Shweta Tiwari, Pankaj Choudhary "Automated Assessment of Students Responses to the Questions using Various Similarity Techniques," *International Journal of Engineering Research & Technology (IJERT)* ISSN: 2278-0181 Published by, www.ijert.org ICSITS - Volume 8, Issue 05, Dec. 2020.

[4]. Sharad Bharadia, Prince Sinha, Ayush Kaul, Dr. Sheetal Rathi "Answer Evaluation Using Machine Learning,"*Researchgatehttps://www.researchgate.net/publication/333856264Answer_Evaluation_Using_Machine_Learn ing, 2020.*

[5]. Era Johri, Nidhi Dedhia, Kunal Bohra, Prem Chandak, Hunain Adhikari "ASSESS-Automated Subjective Answer Evaluation using Semantic Learning," *Social Science Research Network (SSRN),* 2021

[6]. Farrukh Bashir, Hamza Arshad, Abdul Rehman Javed, Natalia Kryvinska, Shahab S. Band "Subjective Answer Evaluation" *Institute of Electrical and Electronics Engineers(IEEE)* Using Machine Learning and Natural Language Processing," Institute DOI:10.1109/ACCESS.2017.DOI, 2021

[7]. Kavita Shirsat, Akash Hamare, Sourabh Pardeshi, Vaishnavi Patil "Subjective Answer Evaluation System," *Internation Research Journal of Engineering and Technology(IRJET),* volume 8, 2021

[8]. Shreya Singh, Omkar Manchekar, Ambar Patwardhan, Prof. Uday Rote, Prof. Sheetal Jagtap, Dr. Hariram Chavan "Computer Application for Assessing Subjective Answers using AI," *Social Science Research Network(SSRN), 2021*

[9]. Nandita Bharambe, Pooja Barhate, Prachi Dhannawat "Automatic Answer Evaluation Using Machine Learning," *International Journal of Information Technology(IJIT)* volume 7, 2021

[10]. [Prerana M. S., Sagarika M. Chavan, Ramit Bathula, Sreenath Saikumar, and Geetha Dayalan, "Eval - Automatic Evaluation of Answer Scripts using Deep Learning and Natural Language Processing," *International Journal of Innovative Science and Engineering (IJISAE*)

[11]. V. Kumari, P. Godbole, and Y. Sharma, "Automatic Subjective Answer Evaluation," *International Conference on Pattern Recognition Applications and Methods (ICPRAM)* 2023

[12]. A.Sheik Abdullah, S. Geetha, A.B. Abdul Aziz, Utkarsh Mishra School of Computer Science and Engineering, Vellore Institute of Technology Chennai "A pedagogical approach with NLP and Deep Learning," *Alexandria Engineering Journal (AEJ),* 2024

[13]. S. K. Virdi, H. Wadhwani, D. Hegde, A. Bhardwaj, and S. N," Automated Answer Sheet Evaluation And Grading System". *Department of Artificial Intelligence and Machine Learning, (IJNRD)* | Volume 9,Issue 06, June 2024

[14]. V. S. Singh, A. Verma, G. Srivastava, and S. Kumar, "An Automated System for Efficient Answer Sheet Evaluation," *Asian Journal of Research in Computer Science (AJRCOS),* vol. 17, no. 6, April 2024.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⩉ Peer-reviewed & Refereed journal ⩉ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14652

[15]. Madhavi Kulkarni, Gayatri Adhav, Kunal Wadile, Vrushali Deshmukh, Rohini Chavan "Digital Handwritten Answer Sheet Evaluation System" *International Journal of Computer Applications* (*IJCA*) (0975 – 8887) Volume 186 – No.16, 2024

[16]. Md. Afzalur Rahaman, Hasan Mahmud "Automated Evaluation of Handwritten Answer Script Using Deep Learning Approach" *Research Gate*, 2022

[17]. Sangeeta Mangesh, Prateek Maheshwari, Aditi Upadhyaya "Subjective Answer Script Evaluation using Natural Language Processing" *JETIR* Volume 9, Issue 8 (JETIR) August 2022

[18]. Prof. Sharayu Lokhande, Prof Shilpa Pawar, Udit Chaudhary,Akash Singh, Pranay Gaikwad, Himanshu Guleria "Automated Subjective Answer Evaluation System", *Journal Of Algebraic Statistics* Volume 13, No. 3, 2022

[19]. Ishika Aggarwal, Pallav Gautam, Gaurav Parashar, "Automated Subjective Answer Evaluation Using Machine Learning" (SSRN), 2024

[20]. António J. R. Neves, Daniel Lopes: A practical study about the Google Vision API. *Research gate* Conference: 22nd Portuguese Conference on Pattern Recognition, RECPAD 2016