International Journal on Emerging Research Areas (IJERA) *ISSN:2230-9993*

DOI: 10.5281/zenodo.14714351

A Survey on Automating Answer-Sheet Evaluation Using AI Techniques

Fabeela Ali Rawther
Dept. of Computer Science
Amal Jyothi College of Engineering,(Autonomous)
fabeelaalirawtheramaljyothi.ac.in

Raihana Rasaldeen
Dept. of Computer Science
Amal Jyothi College of Engineering,(Autonomous)
raihanarasaldeen2025@cs.ajce.in

Stefi Marshal Fernandez
Dept. of Computer Science
Amal Jyothi College of Engineering,(Autonomous)
stefimarshalfernandez2025@cs.ajce.in

Irin Rose Jaison
Dept. of Computer Science
Amal Jyothi College of Engineering,(Autonomous)
irinrosejaison2025@cs.ajce.in

Ria Mariam Mathews
Dept. of Computer Science
Amal Jyothi College of Engineering,(Autonomous)
riamariammathews2025@cs.ajce.in

***Abstract*—The evaluation of answer sheets has traditionally been a time-consuming and subjective process, posing significant challenges in terms of efficiency, scalability, and fairness. With advancements in artificial intelligence (AI) and natural language processing (NLP), automated systems have emerged as promising solutions to these challenges. This study explores two key approaches for implementing automated answer evaluation: BERT-based semantic analysis and large language models (LLMs) powered by prompt engineering. BERT offers deep contextual understanding and precision in grading responses aligned with predefined answer keys, but its reliance on these keys limits its ability to evaluate creative or non-standard answers. In contrast, LLMs such as GPT-4 extend beyond predefined rubrics, utilizing both answer keys and their reasoning capabilities to assess diverse responses accurately.**

This paper examines the strengths and limitations of these approaches, highlighting their potential for improving grading accuracy, scalability, and adaptability. By integrating advanced OCR technologies for digitizing handwritten responses, these models can provide a holistic evaluation system. The work emphasizes the need for flexible frameworks that balance precision and creativity, ensuring fair and efficient evaluation in diverse educational contexts. Through this exploration, we aim to guide the development of scalable AI-driven solutions for modern assessment challenges.

***Index Terms*—Automated grading, Natural Language Processing, AI in education, Answer-key based assessment, Real-time insights, OpenAI API.**

I. INTRODUCTION

The process of grading answer sheets in traditional educational settings has long been plagued by several challenges. Manual grading is not only time-intensive but also susceptible to human error and bias, often resulting in inconsistent evaluations. These inefficiencies become more pronounced as class sizes grow and the complexity of assessments increases. Evaluators may unconsciously be influenced by factors unrelated to the quality of answers, such as handwriting, leading to disparities in grading. Moreover, the repetitive nature of manual evaluation makes it exhausting and prone to fatigue-induced errors, further impacting fairness and accuracy. As a result, ensuring uniform and unbiased assessment has become a critical challenge for educational institutions worldwide.

The need for automation in the grading process arises from the limitations of manual evaluation and the evolving demands of modern education systems. Automated systems promise not only to address inefficiencies but also to scale effectively, accommodating the ever-increasing number of students and diverse assessment formats. Automation can standardize grading across various evaluators, reducing the subjectivity inherent in manual processes. Additionally, with the integration of advanced technologies like artificial intelligence (AI) and natural language processing (NLP), automated systems can evaluate complex, subjective answers with a degree of semantic understanding that rivals human evaluators. This shift toward automation is essential to meet the growing demands for equity, scalability, and efficiency in assessment.

Automated grading systems bring several benefits that go beyond speed and accuracy. They enable timely feedback for students, which is crucial for continuous learning and improvement. With AI-driven technologies, such systems can evaluate a wide range of question types, from multiple-choice to descriptive answers, providing flexibility and versatility in assessments. Furthermore, such systems reduce the workload on educators, allowing them to focus on pedagogy and student engagement rather than repetitive administrative tasks. By ensuring consistent and unbiased grading, automation also helps foster trust in the evaluation process, an essential aspect of modern education.

This paper explores the design and development of an innovative framework, the Comprehensive AI-Powered Answer Evaluation Framework for Automated Assessment, which aims to leverage advanced technologies to transform the grading process. Our ongoing study focuses on integrating optical character recognition (OCR) for digitizing handwritten responses with state-of-the-art evaluation methodologies, includ-

DOI: 10.5281/zenodo.14714351

ing BERT for semantic analysis and large language models (LLMs) for nuanced answer evaluation. By studying these technologies and their applications, we seek to understand how they can work in tandem to create a scalable, accurate, and equitable solution for modern educational assessment. This framework represents an aspirational step toward addressing the shortcomings of manual grading while harnessing the strengths of AI to shape the future of efficient and fair evaluations in education.

II. LITERATURE SURVEY

The growing demand for automated systems in various fields has led to significant advancements in natural language processing (NLP), optical character recognition (OCR), and related technologies. These advancements provide a foundation for developing AI-driven frameworks for automated answer evaluation, as they address challenges in digitizing, analyzing, and grading textual responses. This section surveys relevant studies that have contributed to these domains, highlighting their relevance to the proposed framework.

A. NLP and Text Analysis

NLP techniques have proven effective in analyzing textual data for tasks requiring deep semantic understanding, such as text classification and contextual evaluation. Advanced architectures like BERT and BiLSTM models have shown success in detecting nuanced relationships between words and phrases, making them suitable for grading paraphrased or contextually similar responses. These models, leveraging embeddings and attention mechanisms, ensure evaluations focus on meaning rather than surface-level text similarity, offering significant potential for scalable and consistent automated grading [15][16][19].

Additionally, NLP models integrated with predefined rubrics enable subjective grading, standardizing evaluations across large datasets. This approach supports the development of robust systems adaptable to diverse educational contexts while maintaining fairness and scalability [3][11][15][16].

B. Handwriting Recognition and OCR

OCR systems play a critical role in digitizing handwritten responses, enabling downstream processing. Recent developments in deep learning-based handwriting recognition, such as CNNs, have significantly improved accuracy in extracting structured data from diverse handwriting styles. These advancements ensure precision in recognizing even complex handwritten content, which is essential for creating reliable grading systems [4][17][18].

Moreover, feature extraction techniques within OCR systems have been enhanced to identify fine-grained details in handwritten text, such as annotations or mathematical symbols, which are often present in academic responses. Enhanced OCR methodologies also support the identification of mathematical symbols and annotations, which are common in academic responses. These features allow OCR to provide a strong foundation for processing both textual and graphical content accurately [8][17][18].

C. CNN Frameworks for Text Line Rcognition

The Two-Step CNN Framework for Text Line Recognition offers a robust solution for extracting text from noisy and irregularly formatted camera-captured images. By employing a two-phase approach—segmentation for isolating text lines and character recognition using convolutional neural networks (CNNs)—this method significantly enhances recognition accuracy. It is particularly effective in handling diverse handwriting styles, variable lighting, and text density, making it ideal for educational contexts such as digitizing handwritten answer sheets. The segmentation step isolates text lines even in cluttered backgrounds, while the character recognition phase ensures high precision, addressing common challenges in OCR for handwritten responses [6][17].

When integrated with NLP models like BERT, the OCR outputs from frameworks such as the Two-Step CNN can support nuanced semantic evaluation of textual data, enhancing the overall grading process. This approach complements other advancements, such as content-based image retrieval (CBIR), which enables recognition of diagrams and graphical elements in answer sheets. Together, these technologies provide a comprehensive solution for automated answer evaluation, ensuring scalability, accuracy, and fairness in educational assessments [6][15][18].

D. Implications for the Proposed Framework

The studies highlight the need for integrating advanced NLP and OCR methodologies into a unified framework for automated grading. NLP models such as BERT enhance semantic understanding, while robust OCR systems ensure accurate digitization of handwritten responses. The inclusion of CBIR-inspired methods further extends the framework's capabilities to evaluate multimodal content, addressing diverse academic tasks. These insights provide a strong basis for building scalable and reliable evaluation systems, aligning with the dynamic needs of modern education [15][16][17][18].

III. IMAGE PREPROCESSING AND OCR TECHNIQUES

A. The Role of OCR in Automated Evaluation

OCR serves as the backbone for digitizing handwritten responses, ensuring seamless processing of textual data from physical answer sheets. OCR algorithms identify and extract textual regions, transforming them into structured data for downstream processing. Despite its significance, OCR poses several challenges, including variability in handwriting styles, noise in scanned documents, and the need for high accuracy in extracting semantic content [4][6].

B. Prominent OCR Techniques

Recent advancements have led to a range of OCR techniques suited for educational applications. These include traditional feature-based models, neural network architectures, and hybrid approaches that combine the strengths of multiple methods [5][6].

DOI: 10.5281/zenodo.14714351

Two-Step CNN Frameworks: CNNs remain pivotal in OCR for their ability to learn hierarchical representations. Two-step frameworks refine text-line recognition by incorporating preprocessing stages to eliminate noise and segmentation errors. These approaches have demonstrated superior performance in processing complex layouts and non-standard handwriting styles [6][8].

Latent Feature Vector Analysis: Using latent feature vectors extracted from neural networks enhances OCR's ability to distinguish subtle handwriting nuances. Such methods allow for improved recognition rates, particularly for handwritten digits, equations, and diagrams, as demonstrated in recent research [5].

Hand-Drawn Diagram Recognition: Handwritten responses often include diagrams, charts, or mathematical notations. Techniques integrating diagram recognition algorithms with text extraction enable holistic evaluation of structured and unstructured content. These approaches employ deep learning for object recognition and contextual analysis of diagram components [4][8].

C. Challenges in OCR Implementation

The effectiveness of OCR in answer sheet evaluation systems is limited by; Variations in handwriting and formatting present significant challenges for OCR systems, as they must accommodate diverse styles and inconsistencies across different answer sheets. Additionally, the presence of non-textual artifacts, such as smudges, stains, or stray marks, can further complicate the digitization process, potentially leading to inaccuracies in text recognition. Recognizing complex characters, especially those found in diagrams, mathematical notations, or specialized symbols, adds another layer of difficulty, requiring advanced feature extraction techniques and robust algorithms to ensure accurate processing [4][6][8].

To overcome these issues, hybrid approaches combining OCR with manual verification or post-processing steps are proposed as future solutions.

IV. METHODS FOR ANSWER EVALUATION

A. Semantic Analysis with BERT

BERT, a state-of-the-art transformer-based NLP model, provides a powerful mechanism for semantic analysis in automated answer evaluation. Its bidirectional encoder processes text by analyzing both preceding and succeeding words, enabling a deep contextual understanding of student responses. This capability allows BERT to compare responses with an answer key meaningfully, identifying paraphrased or contextually similar answers with precision. Furthermore, BERT's ability to embed textual data into a shared semantic space ensures that responses closely aligned with the expected answer receive appropriate credit. This precision makes it particularly suitable for evaluations requiring strict adherence to predefined rubrics.

One of BERT's significant advantages is its ability to ensure consistency in grading, as it adheres to a predefined answer key. This minimizes discrepancies caused by subjective interpretations. BERT can also be fine-tuned with domain-specific datasets, making it adaptable to technical or discipline-specific evaluations. Its speed and scalability make it an excellent choice for processing large volumes of data, which is particularly beneficial in academic settings with a high number of responses. These strengths establish BERT as a robust choice for structured evaluation scenarios.

Context Sensitivity: BERT's high sensitivity to word placement ensures precision but can result in overly rigid evaluations. Minor changes in phrasing or structure may lead to misinterpretations, making it less effective for creative or flexible responses. This rigidity is beneficial for structured tasks but limits adaptability in subjective contexts.

Preprocessing Needs: Effective use of BERT requires preprocessing tasks like normalizing punctuation, casing, and text formatting. Without proper preprocessing, inconsistencies in input data may reduce the model's accuracy. Automating preprocessing pipelines can help streamline this process for large-scale evaluations.

Potential Applications: BERT is highly effective in tasks requiring technical precision, such as programming or mathematical evaluations. Its ability to focus on key terminology ensures consistent grading aligned with strict rubrics. This makes it ideal for structured and objective assessment scenarios.

However, BERT's dependency on an answer key is a notable limitation. It cannot evaluate responses that, while correct, deviate from the predefined answers. This restricts its ability to recognize creative or non-standard solutions. Additionally, training BERT for specific tasks requires substantial labeled data and computational resources, making it less practical for applications where datasets are scarce or diverse.

B. Large Language Models and Prompt Engineering

Large language models (LLMs) like OpenAI's GPT-4 offer a highly flexible alternative for automated answer evaluation. Unlike BERT, which relies heavily on a predefined answer key, LLMs can interpret and evaluate responses beyond the constraints of the key. This allows LLMs to recognize valid and creative answers that may not be explicitly defined, making them highly adaptable to diverse evaluation scenarios. LLMs leverage advanced prompting techniques to align their outputs with specific grading requirements, ensuring relevance and accuracy.

One of the key advantages of LLMs is their ability to handle both structured and unstructured evaluation tasks. By incorporating well-designed prompts that include questions, rubrics, and example answers, LLMs can perform nuanced evaluations and provide justifications for their decisions. This transparency enhances trust in the automated grading process. LLMs are also versatile, capable of managing open-ended questions and subjective responses effectively. Their scalability and adaptability make them suitable for a wide range of educational contexts.

Reasoning Transparency: LLMs have the ability to provide detailed explanations for their grading decisions, offering insights into how specific criteria were met. This enhances the

DOI: 10.5281/zenodo.14714351

interpretability of the grading process, building trust among educators and students.

Training Independence: LLMs rely on extensive pretraining across diverse datasets, enabling them to perform well without requiring significant task-specific fine-tuning. This reduces the need for custom training resources, making them efficient for rapid deployment in various contexts.

Creative Applications: LLMs excel in evaluating tasks that involve subjective or open-ended responses, such as essays or research proposals. Their ability to analyze and interpret nuanced answers makes them ideal for grading creative and exploratory work.

Risk Management: LLMs occasionally generate plausible but incorrect responses, known as hallucinations. To mitigate this, strategies like comparing outputs from multiple models or implementing post-evaluation verification processes can be employed to ensure reliability.

Despite their strengths, LLMs face certain limitations. Their reliance on effective prompt engineering can result in inconsistencies if prompts are poorly designed. Additionally, the computational cost of running large-scale LLMs like GPT-4 can be prohibitive for widespread adoption. LLMs alsocarry the risk of generating plausible but incorrect responses, necessitating oversight to ensure reliability.

Both BERT and LLMs bring unique advantages to automated answer evaluation. While BERT ensures consistent and rubric-aligned grading in structured contexts, LLMs offer flexibility and creativity in assessing diverse and open-ended responses. However, the choice between these methods depends on the specific requirements of the evaluation scenario, including the need for scalability, adaptability, and cost-efficiency. Together, these approaches highlight the potential of AI to transform educational assessment, with future advancements likely to bridge their respective limitations.

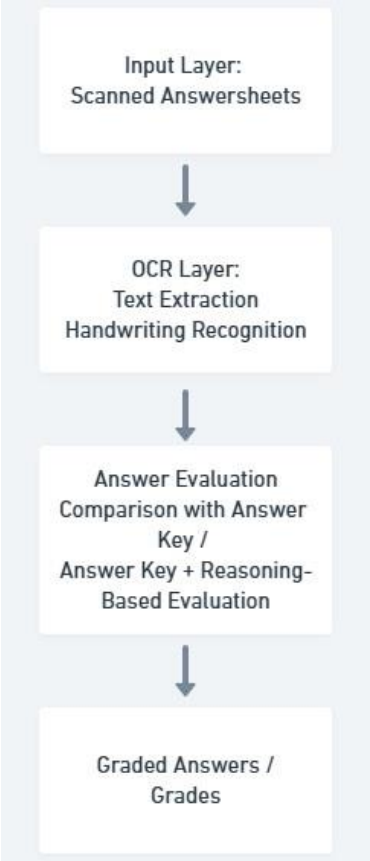

Fig. 1. Workflow of the Proposed Solution

V. CHALLENGES AND FUTURE WORK

Despite advancements in AI and NLP, several challenges remain in implementing effective automated answer evaluation systems. Variations in handwriting, non-standard formatting, and the presence of artifacts like smudges or stray marks continue to hinder the accuracy of OCR technologies. Recognizing complex characters such as equations, diagrams, and special symbols further complicates the digitization process, necessitating more sophisticated algorithms and preprocessing techniques. Additionally, while BERT and LLMs have demonstrated strong potential for semantic analysis, they face limitations. BERT's dependence on predefined answer keys restricts its flexibility, while LLMs may occasionally generate inconsistent or incorrect responses due to their probabilistic nature.

Future work must focus on enhancing the adaptability of AI models to diverse evaluation contexts. This includes developing hybrid frameworks that combine the precision of BERT with the flexibility of LLMs to address the varied needs of educational assessments. Improvements in OCR for handwriting recognition, particularly for multilingual and complex content, are also critical. Research into efficient and transparent prompt engineering techniques can further optimize LLM performance, ensuring consistent and reliable evaluations. Moreover, robust evaluation metrics and fairness-driven frameworks are essential to mitigate biases and ensure equitable grading across diverse datasets. Exploring scalable solutions that balance computational efficiency and model accuracy will be crucial for large-scale deployment in real-world educational settings.

VI. CONCLUSION

The integration of AI technologies such as BERT and LLMs into answer evaluation systems offers transformative potential to revolutionize the grading process. These methodologies address long-standing challenges in traditional assessment, providing scalability, consistency, and efficiency. While BERT excels in structured, rubric-driven evaluations, LLMs offer flexibility and creativity, enabling the recognition of diverse and unconventional responses. However, limitations such as BERT's dependency on answer keys and LLMs' susceptibility to inconsistencies highlight the need for continued research and innovation.

By combining these technologies with advancements in OCR, a holistic framework for automated grading can be

DOI: 10.5281/zenodo.14714351

developed, capable of handling both textual and graphical content with high precision. This study underscores the importance of leveraging AI to foster fairness and transparency in education, paving the way for equitable and efficient assessment systems that can adapt to the evolving needs of modern learning environments. Future advancements will continue to refine these systems, ensuring they meet the demands of educators and students alike.

REFERENCES

[1] Mohanraj G ,Nadesh R.K, Marimuthu M, and Sathiyapriya V, "An enhanced framework for smart automated evaluations of answer scripts using NLP and deep learning methods ," Multimedia Tools and Applications, Received: 26 June 2023.

[2] J. Clerk Maxwell,"Utilizing BERT for Information Retrieval, Survey, Applications, Resources, and Challenges" , ACM Comput. Surv., Vol. 56, No. 7, Article 185. Publication date: April 2024.

[3] I. S. Jacobs and C. P. Bean, "AutoEval: A NLP Approach for Automatic Test Evaluation System" ,2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON).

[4] Vanita Agrawal, Jayant Jagtap, MVV Prasad Kantipudi , "An overview of Hand Drawn Diagram Recognition Methods and its Applications," IEEE ACCESS 2024:24 January 2024. .

[5] Juhyeon Kim, Soyoung Park, Alicia Carriquiry , "A deep learning approach for the comparison of handwritten documents using latent feature vectors", The ASA Data Science Journal (17) :February 2024.

[6] Yulia S. Chernyshova And Vladimir V. Arlazarov ,Alexander V. Sheshkus , "Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images ", IEEE ACCESS 2020: Published on February 2020.

[7] "GradeAid: a frameworkfor automatic short answers grading in educational contexts—design, implementation and evaluation University Science", Received: 21 July 2022 / Revised: 24 March 2023 / Accepted: 22 April 2023 / Published online: 19 May 2023.

[8] A Nikitha; J Geetha; D.S JayaLakshmi - "Handwritten Text Recognition Using Deep Learning.", Published: 13 January 2021

[9] Shreya Singh, Omkar Manchekar, Ambar Patwardhan, Uday Rote and Sheetal Jagtap,Hariram Chavan - "Tool for Evaluating Subjective Answers using AI(TESA)",Published: 27 June 2021

[10] Angelina Patience Mulia, Pirelli Rahelya Piri,Cuk Tho - "Usability Analysis of Text Generation by ChatGPT OpenAI Using System Usability Scale Method", Published on 25 November 2023.

[11] Harsha R. Gaikwad and Arvind W. Kiwelekar - "A Generative AI-Based Assistant to Evaluate Short and Long Answer Questions", Published: 10 June 2024.

[12] Philipp Mondorf, Barbara Plank - "Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey ",Published: 6 Aug 2024.

[13] Jung X.Lee; Yeong-tae Song -"College Exam Grader using LLM AI models", IEEE Explore 2024

[14] Othmane Friha, Mohamed Amine Ferrag, Burak Kantarcı, Burak Cakmak, Arda Ozgun, and Nassira Ghoualmi-Zine. - "LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness ", Published: 9 September 2024

[15] Alan Joseph, Abhinay A K, Dr. Gee Varghese Titus, Anagha Tess B, Adham Saheer, Fabeela Ali Rawther, "Comparative Analysis of Text Classification Models for Offensive Language Detection on Social Media Platforms," International Journal on Emerging Research Areas, vol.04, issue 01, 2024, doi: 10.5281/zenodo.12515626

[16] Anu Rose Joy, "An Overview of Fake News Detection using Bidirectional Long Short-Term Memory (BiLSTM) Models," International Journal on Emerging Research Areas, vol.03, issue 01, 2023, doi: 10.5281/zenodo.8009811

[17] Arun Robin, Tijo Thomas Titus, Minu Cherian, "Improved Handwritten Digit Recognition Using Deep Learning Technique," International Journal on Emerging Research Areas, vol.03, issue 02, 2023, doi: 10.5281/zenodo.10686005

[18] Niya Joseph, Tintu Alphonsa Thomas, "A Systematic Review of Content-Based Image Retrieval Techniques", International Journal on Emerging Research Areas (ISSN:2230-9993), vol.03, issue 01, 2023 doi: 10.5281/zenodo.8019364