International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-6, November 2025
DOI: https://doi.org/10.55524/ijircst.2025.13.6.1
Article ID IRP-1685, Pages 1-7
www.ijircst.org

A Comprehensive Review on Descriptive Answer Script Evaluation Techniques

Pradeep Rao K B[1], Dr. Thyagaraju G S[2], Sahana Kumari B[3], and Prasad S R[4]

[1, 3, 4] Assistant Professor, Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India
[2] Professor, Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India

Correspondence should be addressed to First Author Name; kbpradeeprao@gmail.com

ABSTRACT- Descriptive answer script evaluation involves assessing student's long-form textual answers and it remains a critical and challenging task in educational assessment. Conventional approach of manual grading has shortcoming such as it is time consuming and subjective. To overcome the drawback of manual grading, automated and semi-automated evaluation systems are introduced. This review presents an in depth examination of techniques used for descriptive answer script evaluation. These techniques consists of rule-based methods, machine learning based methods, deep learning based methods, transformer-based and hybrid approaches. Recent advancements in transformer architectures and large language models (LLMs) have greatly improved contextual understanding, scalability, and grading accuracy. In this review, we gather findings from recent studies, and discuss model performance, computational efficiency, and strategies to reduce bias. Despite these advances, significant issues remain concerning generalizability, fairness, explainability, multimodal processing, dataset availability, and pedagogical integration. To address these limitations, there is need of domain-agnostic LLM frameworks, fairness-aware learning, multimodal evaluation, and real-time feedback systems. This review aims to provide researchers and educators with a cohesive view point on current progress, research gaps, and future directions in automated evaluation of descriptive answer script evaluation.

KEYWORDS- Automated Grading, Deep Learning, Large Language Models, Transformer Models

I. INTRODUCTION

Descriptive answer script evaluation is the process of assessing student's long-form, open-ended responses. Descriptive answer script evaluation involves subjective judgment, comprehension of context, and evaluation of conceptual understanding rather than rote recall. Unlike objective assessments such as multiple-choice or short-answer questions, descriptive answers demand interpretation of linguistic nuances, coherence, argumentation, and depth of explanation. Traditionally, such evaluations have been performed manually by human examiners, making the process time-consuming, labor-intensive, and prone to variability or bias. With the rapid growth of digital learning environments and large-scale assessments, there has been a strong motivation to develop automated or semi-automated systems capable of evaluating descriptive answers accurately, consistently, and efficiently. These systems aim to emulate human judgment while reducing subjectivity [1] and improving scalability [2] in educational assessment.

Over the years, a wide range of computational techniques have been explored for descriptive answer evaluation. These techniques evolved from early rule-based methods and statistical similarity measures to machine learning and deep neural network models [3] that capture semantic and contextual relationships. The recent rise of transformer-based language models like BERT [4], RoBERTa, and GPT has significantly advanced the ability of automated systems to understand meaning, coherence, and relevance in student responses. However, despite these advances, several challenges persist — including handling diverse writing styles, interpreting partially correct answers, ensuring fairness across linguistic variations, and providing explainable and transparent scoring criteria. Addressing these challenges remains a critical research focus in developing robust and equitable descriptive answer evaluation systems.

The specific objectives of the review are as follows:

- To summarize various descriptive answer script evaluation techniques.
- To make a comparative analysis of different studies carried out on descriptive answer script evaluation.
- To analyze the evolution of descriptive answer script evaluation techniques.
- To highlight existing limitations in research and propose future research directions

The paper is organized as follows: Section II discusses the various descriptive answer evaluation techniques. Section III discusses about various research works carried out on descriptive answer script evaluation. Section IV traces the progress of descriptive answer evaluation methodologies. Section V identifies research gaps and outlines future research directions. Finally, section VI concludes the paper by summarizing key findings and emphasizing the importance of research in this field.

II. DESCRIPTIVE ANSWER EVALUATION TECHNIQUES

Descriptive answer evaluation has evolved through several

methodological paradigms—each building upon the limitations of its predecessors. The major categories of techniques include Rule-Based, Machine Learning-Based [5], Deep Learning-Based [6], Transformer and Pretrained Model-Based [7] and Hybrid and Ensemble Techniques [8]. These techniques offer unique advantages in terms of accuracy, interpretability, and scalability, but also introduce specific challenges that affect their adaptability to real-world assessment environments.

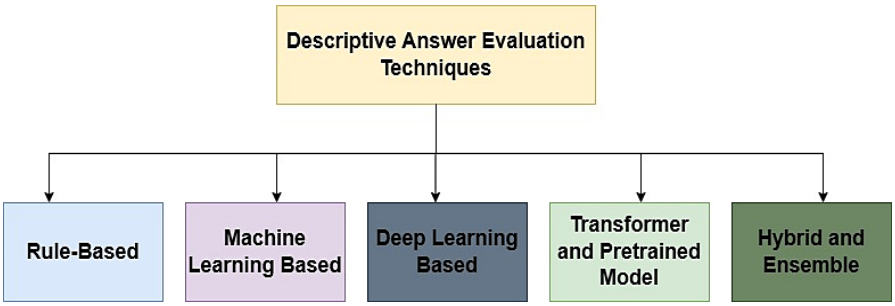

Figure 1: Descriptive Answer Evaluation Techniques

A. Rule-Based Techniques

Rule-based techniques were among the first methods developed for automating the evaluation of descriptive answers. These systems operate by applying predefined linguistic rules and matching of key terms, phrases, or structures between student responses and model answers. Different rule-based approaches include keyword and n-gram matching, Latent Semantic Analysis (LSA) [9], and cosine similarity based on term frequency–inverse document frequency (TF–IDF) representations. The main strength of rule-based methods is their simplicity, transparency, and low computational requirements. Rule-based techniques are easy to interpret and suitable for short, fact-based responses. However, they are limited in their ability to capture semantic meaning and struggle when response involves paraphrasing, varied sentence structures, or conceptual explanations. Additionally, manual rule construction restricts scalability and adaptability across domains, making such techniques less effective for large-scale educational settings.

B. Machine Learning-Based Techniques

Machine learning-based approaches made a significant transition from fixed rule frameworks to data-driven evaluation. These models employ supervised learning algorithms such as Support Vector Machines (SVM) [10], Logistic Regression, and Random Forests trained on features extracted from text—covering lexical richness, syntactic patterns, semantic similarity [11], and discourse indicators. Their primary strength lies in the ability to learn scoring patterns from data, which enables more flexible and context-aware evaluation. By incorporating multiple linguistic features, they provide a more comprehensive evaluation of answer quality. Machine learning-based techniques suffer from certain limitations.

C. Deep Learning-Based Techniques

The use of deep learning has brought major improvements in evaluating descriptive answers. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks [12], can automatically learn useful patterns from text without manual feature design. These systems recognize word order and contextual meaning, allowing them to interpret the flow and coherence of a student's written response more effectively than earlier methods. The primary strength of deep learning techniques lies in their ability to perform representation learning, allowing them to capture subtle linguistic and semantic patterns, which enhances scoring accuracy and robustness. These models demand large amount of training data and significant computational resources. They also function as "black boxes", offering limited interpretability, which can hinder transparency and fairness in educational assessment. Their performance may also degrade when faced with domain shifts or small datasets.

D. Transformer and Pretrained Model-Based Techniques

The rise of transformer models and pretrained language models like BERT, RoBERTa, T5, and GPT [13] has changed how descriptive answers are automatically checked. These models use attention mechanisms that help them look at how words relate across a sentence, which lets them understand meaning at a deeper level. When these models are trained for grading, they often reach very high accuracy on most test datasets. Their strength comes from their strong sense of context and their ability to transfer what they learned from one task to another, even when little data is available. They do especially well when judging how relevant, clear, and well-connected a student's answer is to the question.

Still, there are some clear issues. These models need expensive hardware and a lot of fine-tuning to work properly. Since they are trained on large collections of online data, they sometimes carry unwanted bias. Their design is also very complex, so it is not always easy to explain how or why a particular score was given. This lack of clarity can be a problem in education, where fairness and transparency matter.

E. Hybrid and Ensemble Techniques

Hybrid and ensemble methods mix the strengths of several systems to make grading more stable and flexible. A hybrid model might use simple rule-based tools like TF-IDF or BLEU along with modern embeddings from BERT. Ensemble systems combine predictions from many algorithms or networks to get a more balanced result. This mix helps correct the weaknesses of single models and usually improves accuracy for many types of answers. It also

handles messy or mixed data better. But using many models together makes the whole system more complicated. Training, tuning, and understanding the results all become harder, and the setup needs more computing power to stay consistent and fair.

III. RELATED WORK

Recent studies on descriptive answer evaluation have made strong use of progress in natural language processing (NLP) and machine learning (ML). The goal has been to make grading faster, more consistent, and closer to how humans think. Researchers have tested different techniques, such as checking meaning similarity, using neural networks, and trying out large language models to better capture what students are trying to say. These methods usually include cleaning and preparing the text, finding useful features, and using embeddings to compare student answers with ideal responses. The results not only help produce fairer scores but also save teachers' time and give students quicker, more meaningful feedback.

Ashoka et al. [14] introduced hybrid architecture by integrating advanced Deep Learning (DL) and Natural Language Processing (NLP) techniques. The system incorporated a Cosine similarity network to achieve accurate similarity scoring. An OCR (Optical Character Recognition) model was utilized for the transformation of handwritten text into a digital format. For generating embeddings, the research employed a Universal Sentence Encoder. LLM was integrated for comprehensive contextual analysis of the answers. A Deep Columnar CNN was included to effectively handle complex answer formats and symbols. The integrated technique demonstrated significant improvements over conventional methods in terms of evaluation criteria. It achieved an accuracy of 93.8%, a precision of 94.1%, a recall of 92.7%, and an F1-score of 93.4%

Manikandan et al. [15] integrated natural language processing (NLP) and machine learning (ML) methodologies to accurately score descriptive answers. Semantic techniques like word2vec were used to represent words in which they preserve meaning. Word Mover's Distance (WMD) was used for semantic similarity measurement between answers. Without using any machine learning model, the proposed techniques achieved 85% accuracy in automatically scoring answers, and 86.3% accuracy when adding a machine learning model.

Singh et al. [16] developed a machine learning-assisted model to automate the evaluation of subjective answer sheets in the education sector. An artificial neural network with three layers was constructed. Rectified Linear Activation Unit (ReLU) and Sigmoid activation functions were used in the ANN. The ANN was trained to predict scores based on extracted features. The trained model was integrated into a user-friendly web application using the Streamlit library. A dataset was specifically constructed through research surveys for training and testing the model. The study achieved an accuracy of 83.14% after employing techniques like text cleaning, preprocessing, and feature extraction. The model designed enhanced grading efficiency and accuracy while also providing valuable feedback to students.

C Wangwiwattana et al. [17] introduced a method that leverages Large Language Models (LLMs) for automatically assessing student's short-answer responses. The approach demonstrated effectiveness across several use-cases such as answer matching, keyword extraction, and clustering. The proposed approach achieved 99.03% accuracy rate. The system was capable of generating tailored, real-time feedback for students.

Metan et al. [18] proposed an automated subjective answer evaluation system using machine learning and natural language processing. System utilized cosine similarity algorithm to get the amount of similarity between the answers and grade them accordingly. Cosine similarity algorithm achieved 87% accuracy and outperformed existing systems in terms of time complexity.

Konade et al [19] considered various approaches for assessment and ultimately utilized the most suitable model that meets its requirements. The system specifically implemented the SBERT (Sentence-BERT) architecture to generate vector embeddings of sentences. The utilization of sentence transformers and sentence embedding was highlighted as a cornerstone for achieving efficient and robust semantic analysis of textual responses. The proposed system's results were compared against actual human moderator results across various domains of examination subjects. It achieved a significant accuracy of 95.1% when assessing theoretical answers.

Matos et al. [20] incorporated several machine learning and NLP models to evaluate answer script. Techniques included WordNet, latent semantic analysis, word to vector (Word2Vec), universal sentence encoder, convolution neural network, and Siamese network. Some of these models produced distance and similarity scores, while others predicted whether sentences were similar or not. The system demonstrated an average training time of 92 minutes and achieved an accuracy of 76.2%.

Table 1 summarizes recent studies on automated descriptive answer evaluation, highlighting model performance, efficiency gains, and bias mitigation effectiveness.

Table 1: Comparative Analysis of Different Studies

Study	Model Performance Comparison	Efficiency Gains	Bias Mitigation Effectiveness
Dada et al. [21]	Outperformed baseline models on multiple benchmark datasets	Achieved scalable and consistent grading with reduced manual effort	Introduced semantic alignment to reduce bias in concept matching
(Selvam & Vallejo [22]	Combined transformer models with human validation for balanced performance	Reduced grading time by 40% through AI-human collaboration	Human oversight addressed algorithmic bias and contextual insensitivity
Suryakumar et al. [23]	Compared LLM-based scoring with semantic similarity benchmarks	Reduced evaluation time by 70-80% compared to manual grading	System designed for fairness in technical and theory-based tests
Zhu et al. [24]	Combined LLMs with knowledge	Multimodal framework enhanced	Bias correction for handwriting

	graphs for conceptual evaluation	grading consistency	neatness and answer length
Sura et al. [25]	Compared keyword matching, WMD, and BERT-based models	Automated grading reduced workload and errors	Used multiple NLP and parsing techniques to reduce bias

Figure 2 presents the accuracy obtained by various models discussed in previous studies. The graph shows that transformer-based and hybrid methods generally perform better than traditional machine learning and statistical approaches. This trend indicates a shift toward more context-sensitive and meaning-focused evaluation methods in descriptive answer assessment.

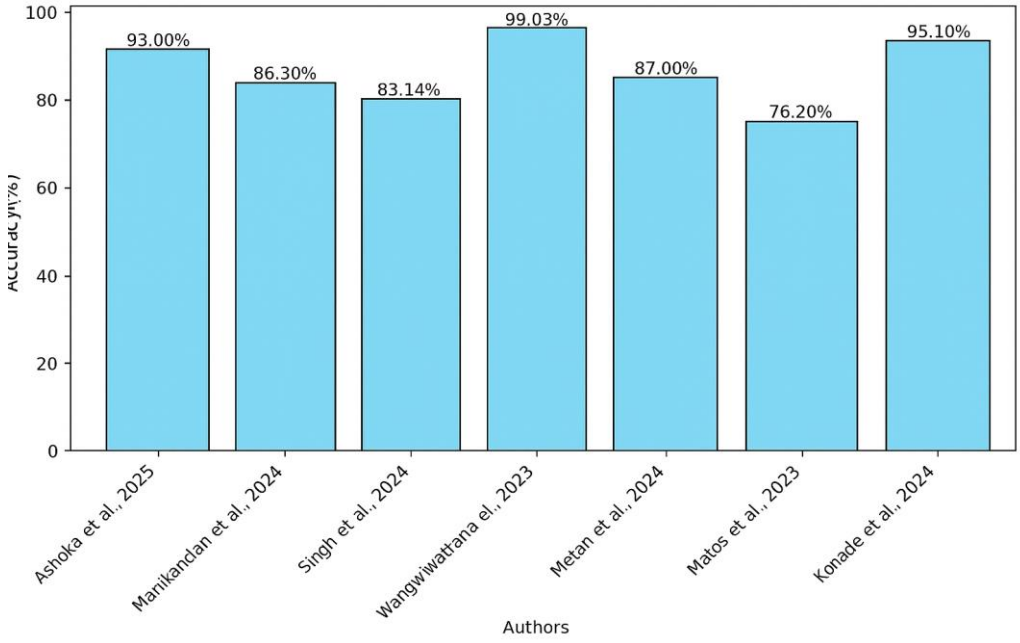

Figure 2: Accuracy reported in surveyed research works

III. EVOLUTION OF DESCRIPTIVE ANSWER EVALUATION TECHNIQUES

Work on checking descriptive answers has changed a lot over the past two decades. In the early days, most researchers were trying simple things like matching important words or counting how many keywords appeared in a student's answer. The goal was to make grading faster and more consistent. As better language-processing tools came along, researchers started using models that could look deeper into meaning and grammar. This helped reduce personal bias and made grading fairer. In recent years, bigger and more advanced systems have made it possible to focus on ethics and transparency as well, so that automated evaluation feels more trustworthy.

A. Foundational Automated Grading Techniques (1998-2010)

At the beginning, grading systems were mostly rule-based. They used fixed sets of words and fuzzy logic to check how close an answer was to the reference solution. These methods were quite basic, but they helped teachers grade large batches of papers in a consistent way. Even though the technology was limited, it gave researchers the first practical tools for automated scoring and showed what could be done with early computing power.

B. Emergence of NLP and Machine Learning Approaches (2011–2015)

After 2011, things moved quickly. Researchers began applying natural language processing methods such as Latent Semantic Analysis, n-gram models, and vector space approaches. Machine learning was added to pick up on sentence structure and meaning. Many studies at this stage tried to make the system learn how teachers think when they grade. There was also more interest in connecting grading methods to specific subject areas so that the results matched the expectations of different fields.

C. Integration of Deep Learning and Clustering Methods (2016–2018)

By 2016, deeper networks like LSTM and CNN started to appear in grading work. They helped the systems understand context, not just single words. Some projects tried clustering and active learning so the software could point out answers that still needed human review. Hybrid systems, which combined traditional rules with newer models, became popular because they gave more balanced and explainable results.

D. Expansion of Transformer Models and Explainability (2019–2021)

When transformer models such as BERT arrived, the field changed again. These systems could read full sentences in context, which made grading more accurate. Still, researchers noticed that people didn't always trust the results. Because of that, several papers looked at ways to explain how the system reached a grade. A few even tested memory-based or multi-agent setups to deal with longer or more complex student answers.

E. Multimodal and Ethical Frameworks in Automated Grading (2022–2023)

Around 2022, new ideas appeared. Instead of focusing only on typed text, researchers started to include images, diagrams, and handwritten work. They used OCR tools along with language models to bring all that information together. Ethical concerns also became more important. Many teams talked about fairness and bias, making sure grading did not favor certain language styles or groups of students. Human-supported review systems and rubric-based methods were used to keep results consistent and fair.

F. Large Language Processing and Human-Centric Grading Innovations (2024–2025)

The latest research has been testing very large language systems such as GPT-4 for descriptive answer grading. Some studies used reasoning step by step or learning from feedback to make the grades clearer and more transparent. The current idea is to let automation handle the heavy work while teachers keep the final control. This approach aims to save time, protect academic honesty, and still keep grading fair across different subjects.

IV. GAPS AND FUTURE RESEARCH DIRECTIONS

Even with all this progress, automated grading is far from perfect. Systems still struggle when the question style or topic changes. They sometimes misread creative or unusual answers. Bias and transparency also remain ongoing concerns. The next steps should focus on making systems more adaptable, more open about how they work, and easier for teachers to understand and trust.

A. Generalizability of LLM-Based Grading Systems

Most large language-driven grading tools work well only for a few subjects or writing styles [26], [27]. When the topic changes, performance often drops. Future work should look at making these systems more flexible so that they can adapt to new areas without retraining from scratch. Transfer-learning techniques and smaller domain-specific adjustments could make grading more reliable and consistent across different disciplines.

B. Bias and Fairness in Automated Grading

Language and cultural background still influence how automated graders score answers [28]. Even with some human checking, small biases remain. The best way forward is to build wider and more varied datasets, include samples from different cultures, and keep checking grading patterns for hidden bias. Continuous review and fairness testing can help ensure that every student, regardless of language or region, receives equal treatment.

C. Explainability and interpretability of AI grading

High-performing AI models, especially deep learning and LLMs, often operate as black boxes, limiting educator's understanding of how scores are determined [29]. Future approaches should combine interpretable models with advanced AI systems and incorporate techniques that clarify model decisions.

D. Handling multimodal and handwritten responses

Many current systems cannot fully handle responses that include drawings or diagrams. leading to reduced accuracy and bias [30]. Future Research should develop methods that integrate OCR and multimodal reasoning to reliably assess such responses.

E. Efficiency and computational cost of multi-agent LLM systems

Multi-agent LLM frameworks improve scoring accuracy, but require significant computational resources, limiting real-time scalability [31]. Future research should explore lightweight LLM architectures and efficient prompt engineering to reduce cost and latency.

F. Dataset Scarcity and Annotation Quality

At present, there are very few large and reliable datasets that include a wide range of languages or question types. This shortage limits both the training and testing of automated grading models [32]. In the future, researchers should work on developing multilingual and diverse datasets. Using semi-supervised or active learning methods can also help reduce the time and expense needed for manual annotation.

G. Feedback Quality and Pedagogical Integration

In many cases, automated systems provide feedback that is too general and does not reflect deeper thinking or creativity [33]. Future systems should aim to give more useful and detailed comments that support the actual learning goals. They should also fit easily into existing teaching platforms so that learners and teachers can receive real-time guidance and feedback during study sessions.

VI. CONCLUSION

The way descriptive answers are checked has changed slowly over the years. In the beginning, people used simple steps like rule checking or counting words to decide the marks. Later, some researchers tried using machine learning, and after that, deeper models became common. Now, a mix of systems is used. The newer ones, which include large language models and multimodal ideas, can understand the meaning of an answer better and handle more kinds of responses. When two or more models work together, they often give steadier and more balanced results.

Even with these improvements, many gaps still remain. Most systems perform well in one subject but not in others. Some are unfair or hard to explain. Reading handwriting or combining text with other input is still a challenge. Another issue is the shortage of big, clean, and well-marked datasets. Many systems also fail to give useful comments that help students learn from their mistakes.

To move ahead, researchers should try to build models that can work across subjects and languages. The systems need to be fair and clear, and they should be able to read both typed and handwritten work. Collecting larger and more varied data will help a lot. Most of all, feedback should be simple, clear, and actually useful for learning. If these things improve, automated grading will not only save time but also support real learning in classrooms.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] F. Moers, "Discretion and bias in performance evaluation: the impact of diversity and subjectivity," Accounting,

Organizations and Society, vol. 30, no. 1, pp. 67–80, 2005. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=491023

[2] P. Tayal, P. P. Shetty, R. Harshita, and Sreenath. M. V., "Evaluation of handwritten descriptive responses using machine learning – A survey," in Proc. 2nd Int. Conf. Futuristic Technologies (INCOFT), Nov. 2023, pp. 1–5. IEEE. Available from: https://doi.org/10.1109/incoft60753.2023.10425768

[3] S. K. A., N. George, and S. M. Varghese, "Descriptive answer script grading system using CNN-BiLSTM network," Int. J. Innovative Research in Computer Science and Technology, vol. 9, no. 5, pp. 139–144, 2021. Available from: https://doi.org/10.35940/IJRTE.E5212.019521

[4] Aftan, Sulaiman, and Habib Shah., "A survey on BERT and its applications," in Proc. 20th Learning and Technology Conf. (L&T), Jan. 2023, pp. 161–166. IEEE. Available from: https://doi.org/10.1109/LT58159.2023.10092289

[5] S. Jagannathan, K. A. Sriram, and P. Vasuki, "Automatic evaluation of answer scripts," AIP Conf. Proc., vol. 2966, no. 1, p. 020008, Mar. 2024. AIP Publishing LLC. Available from: https://doi.org/10.1063/5.0189777

[6] Rahaman, Md Afzalur, and Hasan Mahmud., "Automated evaluation of handwritten answer script using deep learning approach," Transactions on Machine Learning and Artificial Intelligence, vol. 10, no. 4, pp. 1–16, 2022. Available from: https://doi.org/10.14738/tmlai.104.12831

[7] Ait Khayi, Nisrine, Vasile Rus, and Lasang Tamang, "Towards improving open student answer assessment using pretrained transformers," in the international flairs conference proceedings, vol. 34, Apr. 2021. Available from: https://doi.org/10.32473/FLAIRS.V34I1.128483

[8] N. Zhou, "Evaluating human and machine assessment: Introducing a hybrid approach for enhanced educational evaluation," Lect. Notes Educ. Psychol. Publ. Med., vol. 58, no. 1, pp. 118–124, 2024. Available from: https://doi.org/10.54254/2753-7048/58/20241716

[9] J. M. Wheeler, A. S. Cohen, and S. Wang, "A comparison of latent semantic analysis and latent Dirichlet allocation in educational measurement," Journal of Educational and Behavioral Statistics , vol. 49, no. 5, pp. 848–874, 2024. Available from: https://doi.org/10.3102/10769986231209446

[10] H. Ahmed, S. Hina, and R. Asif, "Evaluation of descriptive answers of open-ended questions using NLP techniques," In 2021 4th International Conference on Computing & Information Sciences(ICCIS), Nov. 2021, pp. 1–7. IEEE. Available from: https://doi.org/10.1109/iccis54243.2021.9676405

[11] Rozeva, Anna, and Silvia Zerkova, "Assessing semantic similarity of texts – methods and algorithms," AIP Conference Proceedings ,vol. 1910, no. 1, Dec. 2017. Available from: https://doi.org/10.1063/1.5014006

[12] N. Prabakaran, R. Kannadasan, and A. Krishnamoorthy, "A bidirectional LSTM approach for written script auto evaluation using keywords-based pattern matching," Natural Language Processing Journal, vol. 5, p. 100033, 2023. Available from: https://doi.org/10.1016/j.nlp.2023.100033

[13] G. Kortemeyer, "Performance of the pre-trained large language model GPT-4 on automated short answer grading," Discover Artificial Intelligence, vol. 4, no. 1, p. 47, 2024. Available from: https://doi.org/10.1007/s44163-024-00147-y

[14] S. B. Ashoka, K. Deep, G. Goutham, S. M. M. Ganihar, P. K. Udayaprasad, G. C. Lakshmikantha, and P. Dayananda, "Efficient automated evaluation of answer scripts using LLMs, NLP, and deep learning," In Data Science & Exploration in Artificial Intelligence, pp. 484–491, 2025. Available from: https://doi.org/10.1201/9781003589273-72

[15] Manikandan, R., RN Yakshith Sai, G. Vignavi, G. Santhosh Kumar Reddy, and N. Sri Dharshani Reddy, "Evaluating subjective answers using RNN and NLP," In Challenges in Information, Communication and Computing Technology, pp. 681–686, 2024. Available from: https://doi.org/10.1201/9781003559085-117

[16] V. S. Singh, A. Verma, G. Srivastava, and S. Kumar, "Exam assessor tool: An automated system for efficient answer sheet evaluation," Asian Journal of Research in Computer Science, vol. 17,no. 6,pp. 36-57, 2024. Available from: https://doi.org/10.9734/ajrcos/2024/v17i6455

[17] C. Wangwiwattana and Y. Tongvivat, "Automating academic assessment: A large language model approach," In 2023 7th International Conference on Information Technology (InCIT), pp. 330–334, 2023. Available from: https://doi.org/10.1109/incit60207.2023.10412991

[18] J. Metan, D. Kumar, and H. Kumar, "An automated approach to subjective answer evaluation using ML and NLP," in Proc. 2nd Int. Conf. Advances in Information Technology (ICAIT), vol. 1, July 2024, pp. 1–7. IEEE. Available from: https://doi.org/10.1109/icait61638.2024.10690635

[19] S. Konade, Y. Hirgude, A. Kulkarni, A. Jadhav, and L. Sonar, "Implementation of an automated answer evaluation system," in Proc. IEEE Int. Conf. Computing, Power and Communication Technologies (IC2PCT), vol. 5, Feb. 2024, pp. 457–462. IEEE. Available from: https://doi.org/10.1109/ic2pct60090.2024.10486693

[20] A. R. Matos, "Semantic similarity based automated answer script evaluation system using machine learning pipeline and natural language processing," Advances in Intelligent Systems and Computing, pp. 495–509, 2023. Available from: https://doi.org/10.1007/978-981-19-9819-5_36

[21] I. D. Dada, A. T. Akinwale, I. A. Osinuga, H. N. Ogbu, and T. Tunde-Adeleke, "IAttention transformer: An inter-sentence attention mechanism for automated grading," Mathematics, vol. 13, no. 18, p. 2991, 2025. Available from: https://doi.org/10.3390/math13182991

[22] M. Selvam and R. G. Vallejo, "Human-in-the-loop models for ethical AI grading: Combining AI speed with human ethical oversight," EthAIca: J. Ethics, AI and Critical Analysis, vol. 4, p. 413, 2025. Available from: https://doi.org/10.56294/ai2025413

[23] P. Suryakumar, A. Malini, S. K. Subasini, G. Priyanka, and V. Sandhiya, "Aivaluate: A multi-agent framework for automated answer script evaluation using large language models and semantic vector indexing," IEEE Access, 2025. Available from: https://doi.org/10.1109/access.2025.3608158

[24] H. Zhu, T. Li, P. He, and J. Zhou, "Enhancing automated grading in science education through LLM-driven causal reasoning and multimodal analysis," 2025. Available from: https://doi.org/10.24963/ijcai.2025/1150

[25] M. A. Sura, M. Rai, S. Khetarpaul, and S. Mishra, "AASE: AI-driven automated answer script evaluation," Authorea Preprints, 2025. Available from: https://doi.org/10.21203/rs.3.rs-6895375/v1

[26] Y. Chu, H. Li, K. Yang, H. Shomer, H. Liu, Y. Copur-Gencturk, and J. Tang, "A LLM-powered automatic grading framework with human-level guidelines optimization," arXiv preprint arXiv:2410.02165, 2024. Available from: https://doi.org/10.48550/arxiv.2410.02165

[27] X. Wu, P. P. Saraf, G. Lee, E. Latif, N. Liu, and X. Zhai, "Unveiling scoring processes: Dissecting the differences between LLMs and human graders in automatic scoring," Technology, Knowledge and Learning, pp. 1–16, 2025. Available from: https://doi.org/10.48550/arxiv.2407.18328

[28] J. H. Christyodetaputri and N. Marwa, "Realizing ethical and equitable assessment in global education through artificial intelligence," Sinergi Int. J. Education, vol. 2, no. 3, pp. 170–186, 2024. Available from: https://doi.org/10.61194/education.v2i3.590

[29] Y. Asazuma, H. Funayama, Y. Matsubayashi, T. Mizumoto, P. Reisert, and K. Inui, "Take no shortcuts! Stick to the rubric: A method for building trustworthy short answer scoring models," Communications in Computer and Information Science, pp.

337–358, 2024. Available from: https://doi.org/10.1007/978-3-031-67351-1_23

[30] S. R. Boreddy, "AI-driven handwritten assignment analysis and evaluation," International Journal for Science Technology and Engineering, vol. 13, no. 4, pp. 6810–6815, 2025. Available from: https://doi.org/10.22214/ijraset.2025.70012

[31] W. Xie, J. Niu, C. J. Xue, and N. Guan, "Grade like a human: Rethinking automated assessment with large language models," arXiv preprint arXiv:2405.19694, 2024. Available from: https://doi.org/10.48550/arxiv.2405.19694

[32] S. A. Mahmood and M. A. Abdul samad, "Automatic assessment of short answer questions: Review," Edelweiss Applied Science and Technology, vol. 8, no. 6, pp. 9158–9176, 2024. Available from: https://doi.org/10.55214/25768484.v8i6.3956

[33] D. C. Gabon, A. A. Vinluan, and J. T. Carpio, "Evaluating student communication skills and user acceptability of an NLP-based automated essay grading system," European Journal of Teaching and Education, vol. 7, no. 3, pp. 30–40, 2025. Available from: https://doi.org/10.33422/ejte.v7i3.1531