STM JOURNALS

Emerging Trends in
Languages

ISSN: 3049-0693
Volume 3, Issue 2, 2026
July–December
DOI (Journal): 10.37591/ETL

https://journals.stmjournals.com/etl

Research ETL

Automated Evaluation of Descriptive Answers

Subhakanta Panda[1]*, Harendra Kumar[2], Dhawaleswar Rao[3], Gitanjali Mishra[4]

Abstract

This paper offers a smarter system of automatic evaluation of descriptive responses in educational tests. The time-consuming aspect of testing with traditional manual marking, the subjectivity of that process, and the impossibility of scaling it makes it inapplicable, particularly to large academic environments. To resolve the issues, the proposed solution combines the methods of Natural Language Processing (NLP) and hybrid image-text processing on the responses of handwriting and typed answers. The system follows five key steps, namely, text preprocessing, feature extraction/embedding generation, similarity computation, post-processing, and evaluation. It uses the lexical, syntactic, semantic, and discourse-level characteristics to quantify the quality of answers. The common techniques that are applied to compare student responses to gold-standard answers created by experts include Bag of Words (BoW), TF-IDF, N-grams, cosine similarity, Jaccard similarity, LSA, and LDA. Additionally, deep learning-based language models and contextual embeddings are incorporated to improve semantic understanding, enabling the system to recognize paraphrased responses and conceptually equivalent answers more accurately. The framework also supports automated feedback generation, helping students identify areas for improvement while assisting educators in reducing evaluation workload and maintaining consistency across assessments. The resulting score is established by the proximity of prediction of the system compared to the assessment of the experts. The system should serve to offer gradable, objective, and uniform grading while still being in line with academic rubrics and assessment criteria.

Keywords: Natural language processing (NLP), machine learning (ML), text similarity analysis, semantic analysis, feature extraction, automated scoring system, artificial intelligence (AI).

INTRODUCTION

Assessment is a significant issue in the education system as it determines performance of the students, their academic growth and their future. Since the growth of online learning space has been rapid, especially, post-COVID-19, online learning and computerbased testing has become popular in schools and universities. In as much as automated systems are effective in the evaluation of objective questions like Multiple Choice Questions (MCQs), descriptive responses like essays and short responses have remained one of the major gaps in automated systems. At this stage, the evaluation of descriptive answers is mostly done through teachers in a manual assessment. This is a traditional approach that contains several challenges. It is also time consuming, tedious and it is also prone to human subjectivity or changeable scoring standards. The higher the number of students to teacher, the higher the pressure of manual evaluation. Moreover, these issues as sided scoring, poor division of time, and lack of compatibility of marking criteria can also influence manual grading [1].

The Automated Descriptive Answer Scoring has become a significant field of research in both Natural Language Processing (NLP) and Educational Technology. These systems study

***Author for Correspondence**
Subhakanta Panda
E-mail: subhakantapanda034@gmail.com

[1]Teaching Associate, [3]Associate Professor, Department of CSE, Centurion University of Technology and Management, R. Sitapur, Odisha, India
[2]Project Associate, [4]Assistant Professor, Department of CSE, Gandhi Institute of Engineering and Technology, Gunupur, Odisha, India

Received Date: June 29, 2026
Accepted Date: July 14, 2026
Published Date: July 15, 2026

Citation: Subhakanta Panda, Harendra Kumar, Dhawaleswar Rao, Gitanjali Mishra. Automated Evaluation of Descriptive Answers. Emerging Trends in Languages. 2026; 3(2): 1–9p.

student responses based on a variety of linguistic characteristics such as lexical (word usage), syntactic (sentence structure) semantic (meaning and relevance) and discourse (coherence and organization) features. Nevertheless, numerous of the available machine learning-based systems are strongly dependent on large training datasets that are domain-specific, costly, and hard to acquire. Also, models trained in a particular area of study might not be effective in a different area and hence their generalizability becomes scarce. This project therefore suggests a hybrid approach that combines both image processing and text processing approaches in order to overcome these challenges [2]. As a number of the examinations are still done on paper and pencil, the answer scripts are initially scanned and processed on paper script. The system transcribes handwritten text into machine-readable text and matches student responses with a standard (gold-standard) response that was prepared by the domain experts. Different methods of computing similarity are used in testing the proximity between the student response and the standard response. This similarity is used to generate the final score which is checked against expert evaluation. The proposed system is to offer just, impartial, constant, and effective evaluation of descriptive responses. It helps create a real computerized assessment system of modern education, as it prevents the use of manual work and will enhance the accuracy of evaluation [3].

PROBLEM STATEMENT

Assessment is an important principle in the contemporary education system to gauge the performances and comprehensions of the students. Whereas questions that belong to the objective form like Multiple Choice Questions (MCQs) can be automatically graded through computer-based assessment, descriptive answers such as essays and short answers remain mostly to manual scoring by the instructors. This conventional assessment process is cumbersome, tedious and in most cases inaccurate [4].

As the number of students increases, the teachers have had to work more and thus making use of manual grading harder. Human assessment can also create problems of biased judgment, absence of consistency, inadequate time allocation, and disparities in marking criterion among the assessors. In addition, students can give correct answers in various forms, and it is hard to make sure that they are graded fairly and consistently. Whereas there are a few automated systems of grading, most of them are based on massive training datasets in the domain, which are costly and time-intensive to produce [5]. Furthermore, models learnt on one topic area do not work effectively on another topic area. The second big issue is that in paper-based examinations, handwritten answer scripts must be assessed through proper text recognition, which is followed by correct assessment [6].

Hence, an automated system that is based on reliable, efficient, and domain-adaptable assessment and can appraise descriptive answers across the board and in a consistent manner is required. The system is supposed to cut human labor and reduce biases, process handwritten responses, and score students correctly by matching their answers to the answers set by the experts through sophisticated text analysis tools [7].

OBJECTIVE

- Creation of computerized scoring of answer books in handwriting.
- Reading paper-based examination handwritten answer scripts.
- Application of scale-space method in segmentation of words in handwritten document.
- Training a handwritten answer recognizer model in the form of a neural network.
- Using a post-processing component in form of a neural word embedding to refine the output provided by the system.
- Comparing the answers of students and expert-prepared standard (gold-standard) answers.
- Similarity between student and standard answers: computing with syntactic and semantic similarity measures.
- Production of system-predictive scores, which are based on similarity measures.
- Checking the functioning of the system against the assumption by comparing the predicted scores to the mean scores of two experts (gold-standard score).

- Creating an effective computerized scoring system of descriptive responses.
- Taking into account various assessment criteria grammar, vocabulary, coherence, structure, relevance, and the usage of examples.
- The input data to the proposed system is a social science textbook, and its results are tested [8].

METHODOLOGY

This proposed system is based on Hybrid Methodology, which combines the two-image processing and text processing techniques in order to automatically assess descriptive answers which are written in pen-and-paper exams. As most institutions continue to use traditional handwritten exams, the system initially takes scanned answer scripts and then does a text-based analysis [9].

First, two or more domain experts make a standard answer to every question. These are the evaluators who give marks to the student answers independently. The inter-rater agreement between the two experts must be good to be reliable and fair. The last score of the reference, which is referred to as the gold-standard score, is the mean of the scores provided by the two assessors. This gold-standard measure is the measure whereby the performance of the system is assessed [10].

The system scans the handwritten answer scripts and converts them into machine-readable text, as a result of the handwritten recognition techniques. The main part of the system at that point, compares the Student Answer to the standard Answer which is prepared. This is compared by application of several natural language processing methods at word and sentence levels [11].

The system analyzes responses based on syntactic and semantic similarity (meaning and relevance) as well as grammar and structure. There are various methods of computing similarity used in determining the similarity between the student answer and the standard answer based on its content, form and context (Figure 1).

Lastly, the system comes up with a predicted score, which is determined by the similarity calculated. The model is effective when assessment of the distance between the system generated score and the gold standard score determined by experts is measured. The less the disparity the greater the accuracy of the system (Figure 2).

Text Preprocessing

The first and the most crucial stage of the system is Text Preprocessing. In this phase, the raw text in which handwritten answer scripts have been converted to (and then transformed to a digital text) is purified and made standardized to analyze. Natural language has changes, redundant words and inconsistency; therefore, preprocessing is used to provide uniformity. It involves tokenization (individual words or sentences are extracted), lowercasing (the entire text is changed to lowercase to prevent occurrence of word similarity in different cases), removal of stop word (words such as is, and others) and lemmatization (the conversion of words to their root forms). This stage eliminates noise and prepares a structured and consistent data on which to proceed with more computational analysis [12].

Feature Extraction

Bag of Words (BoW): Represents text by the frequency of words. Simple but ignores word order and context TF-IDF (Term Frequency-Inverse Document Frequency):

Enhances BoW by weighing terms inversely by their frequency across all documents, giving more importance to rare but significant words

The term frequency TF (t, d) measures how frequently a term t occurs in a document d.

It is given by:

$$TF(t,d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

The inverse document frequency IDF (t, D) measures how important a term is within the entire corpus D. It is given by

$$(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

N-grams: Extends BoW and TF-IDF to include contiguous sequences of n items (words, characters)

Given a text with T tokens (words or characters), the number of possible n-grams is calculated as:

Number of n-grams = $T - n + 1$

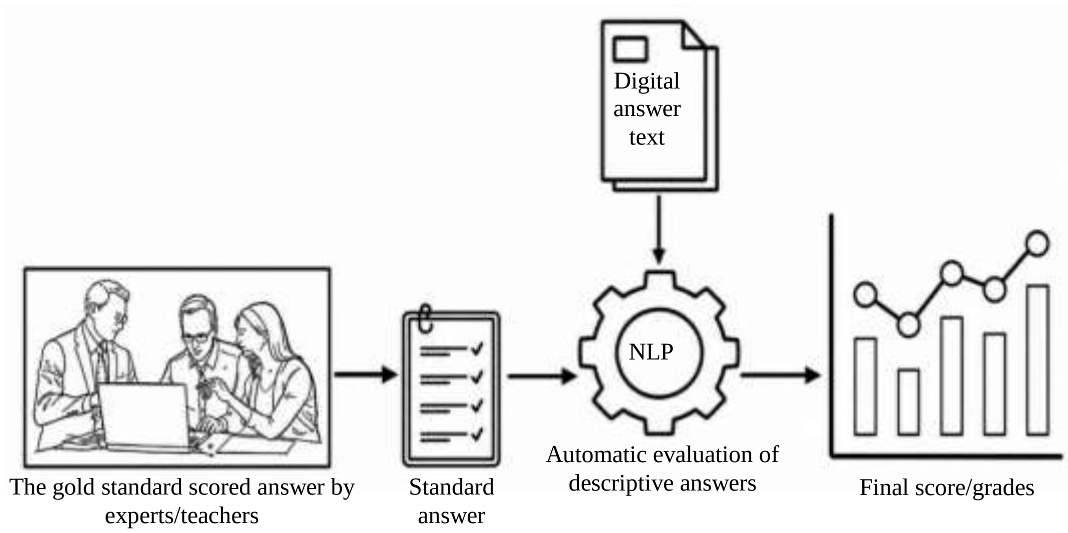

Figure 1. Methodology of the proposed automated descriptive answer evaluation system**.**

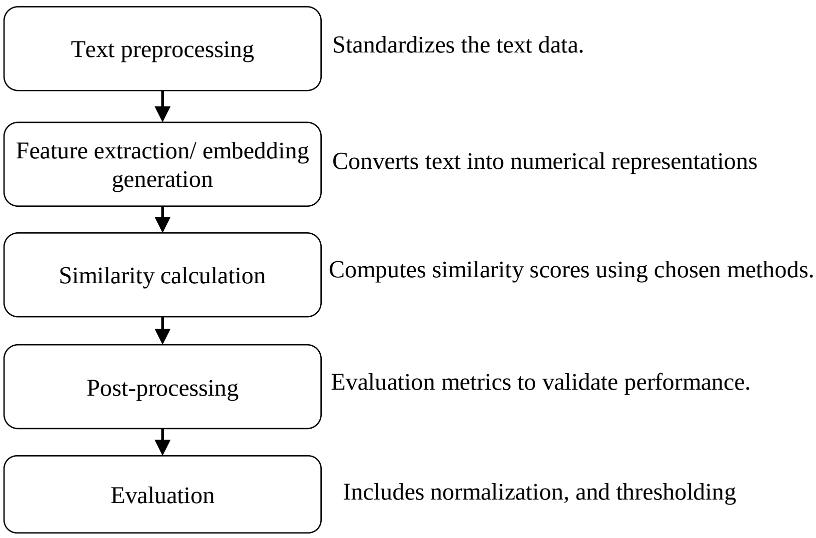

Figure 2. Phases of the proposed project workflow.

Similarity Calculation

The basic step is the Similarity Calculation, during which, the answer of the student is evaluated against the standard answer prepared by the expert. Similarity measures based on the numerical vectors of the generated number in the earlier stage like Cosine Similarity, Jaccard Similarity and other semantic similarity methods are implemented. Such techniques establish the degree of similarity between the response of the student and the content, structure and meaning of the standard answer. What is obtained is the similarity score indicating the extent to which the student was right and relevant [13].

Cosine Similarity

Measures the cosine of the angle between two vectors in a multi-dimensional space. Commonly applied to BoW, TF-IDF, and embedding vectors. The cosine similarity value ranges between -1 (completely opposite) and 1 (exactly the same).

$$\text{Similarity} = (A \cdot B) / (\|A\| \, \|B\|) \quad (1)$$

Where:

- A and B are vectors representing word frequencies.
- $A \cdot B$ is the dot product of vectors A and B.
- $\|A\|$ and $\|B\|$ are the Euclidean norms (magnitudes) of vectors A and B.

Jaccard Similarity

Measures the size of the intersection divided by the size of the union of two sets of words. Often used for binary or multi-set data. Jaccard Similarity

$$(A, B) = |A \cap B| / |A \cup B| \quad (2)$$

Where:

- $|A \cap B|$ denotes the number of elements in the intersection of sets A and B.
- $|A \cup B|$ denotes the number of elements in the union of sets A and B.

Post-Processing

The calculated similarity score is narrowed down in the postprocessing stage to have equal grading. Normalization is used to bring the similarity value to an appropriate marking range (e.g. 0-10 or 0-100). Minimum or maximum score limits can also be set with the help of thresholding techniques. This measure protects consistency, the equitableness, and consistency of the scoring system used and adherence to the criteria of academic assessment [14].

Evaluation

The final phase is Evaluation in which the success of the system is assessed. This is contrasted to the score that is predicted by the system and that of the gold-standard score that is the mean score of two expert evaluators. Some of the evaluation measures are accuracy, reliability and consistency which are used to determine the similarity between the human judgment and the automated grading. This step determines overall performance and validity of the proposed system [15].

Advanced Text Representation and Similarity Methods

- *Co-occurrence Vector:* identifies the frequency of coincidence of words in a window. Similarity is obtained using co-occurrence matrices of words.
- *Statistical & Syntactical Methods:* Uses a variety of statistical representations (e.g. mean, variance) and syntactic structure analysis (e.g. parse trees) to represent text structure.
- *Latent Semantic Analysis (LSA):* Applies Singular Value Decomposition (SVD) to come up with a smaller dimensionality of the TF-IDF matrix and reveal latent semantic structures.
- *Latent Dirichlet Allocation (LDA):* A probabilistic model that represents documents as mixtures of topics, each characterized by a distribution of words (Table 1).

Table 1. Techniques with their output feature groups.

Technique / Feature	Lexical	Syntactic	Semantic	Discourse
Cosine similarity	Yes	Yes	Yes	–

Jaccard similarity	Yes	Yes	–	–
Co-occurrence vector similarity	Yes	Yes	Yes	–
TF-IDF similarity	Yes	–	–	–
LDA similarity	–	–	Yes	–
LSA similarity	–	–	Yes	–
BERT similarity	–	–	Yes	–
SBERT similarity	–	–	Yes	–
Number of sentences	–	Yes	–	–
Length in words	–	Yes	–	–
Average sentence length	–	Yes	–	–
Number of unique words	Yes	–	–	–
Spelling errors	Yes	–	–	–
Number of nouns, verbs, adjectives, prepositions, discourse connectives	Yes	Yes	–	Yes
Average clauses per sentence	–	Yes	–	–
Sum tree depth	–	Yes	–	–
Average tree depth	–	Yes	–	–

EVALUATION METRICS

Pearson Correlation

Tests the linear association or rank association between two groups of similarity scores. Pearson correlation coefficient is a measure of the linear relationship between two variables that are continuous. When the value of a linear connection is 1 then it is said to be perfect, when -1 it is said to be per-fect and when zero it is said to be nonexistent [16].

$$\mathrm{r} = [\mathrm{n}(\Sigma \mathrm{xy}) - \Sigma \mathrm{x} \Sigma \mathrm{y}] / \sqrt{\mathrm{n}(\Sigma \mathrm{x2}) - (\Sigma \mathrm{x})2[\mathrm{n}(\Sigma \mathrm{y2}) - (\Sigma \mathrm{y})2]}$$

Spearman’s Rank Correlation

It is unlike Pearson because it can be applied when the data is non-normally distributed, and there need not always be a linear relationship among the data.

$$\rho = 1 - \mathrm{n} \ \frac{6\Sigma \mathrm{di2}}{(\mathrm{n2} - 1)}$$

Root Mean Square Error (RMSE)

Calculates the mean size of the difference between anticipated and real similarity ratings. Making a comparison between ground truth scores and predicted similarity scores. The smaller RMSE value implies the enhanced prediction quality (Figure 3).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RESULT ANALYSIS

The suggested hybrid system is effective and achieves automatic analysis of descriptive responses through image processing and text processing. The answer scripts in handwritten formats are correctly

Emerging Trends in Languages
Volume 3, Issue 2
ISSN: 3049-0693

translated into machine-readable format and matched against the standard answers prepared by experts [16]. The scoring mechanism which is based on similarity yields similar results which are close to the gold-standard scores of human assessors (Figure 4). It has a good accuracy, consistency, and reliability in grading. It enhances less manual work, less bias, and standardized evaluation criteria [17]. In general, the model is an effective and pragmatic approach to use in the computerized evaluation of descriptive responses within learning institutions (Figure 5).

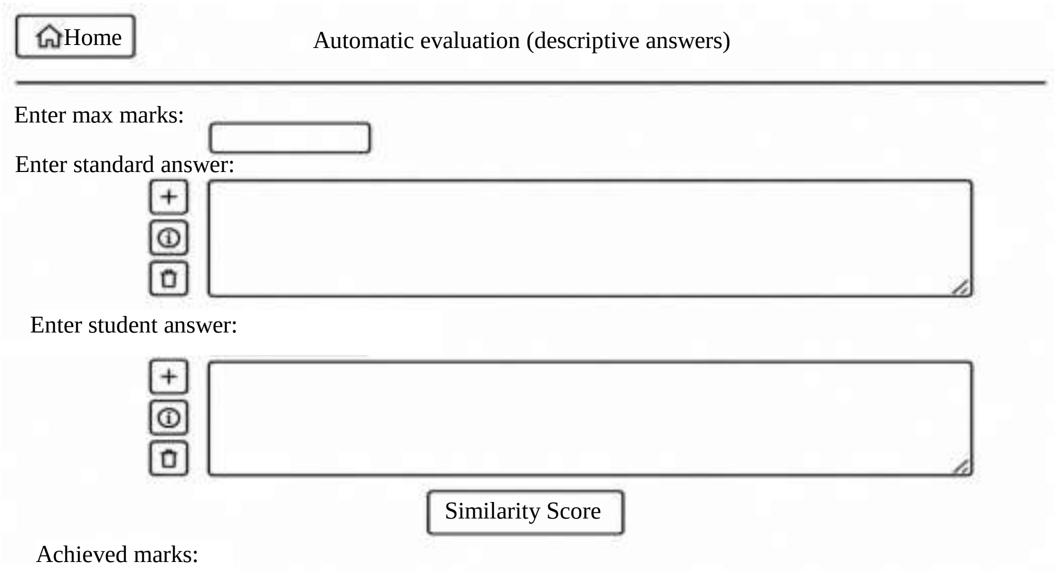

Figure 3. Result analysis interface of the automated evaluation system.

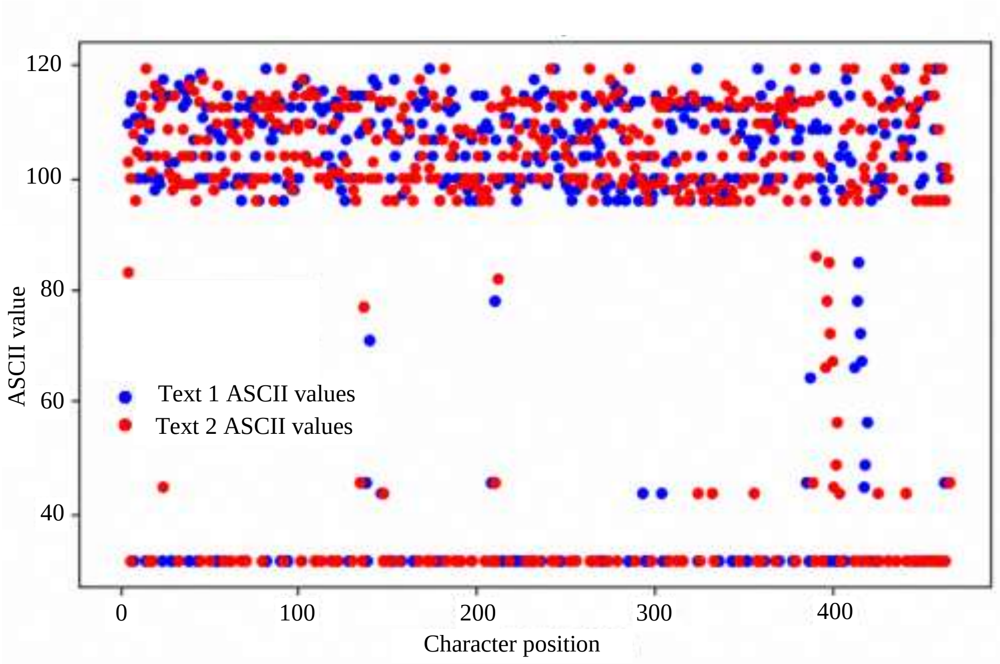

Figure 4. Scatter plot between two descriptive text inputs.

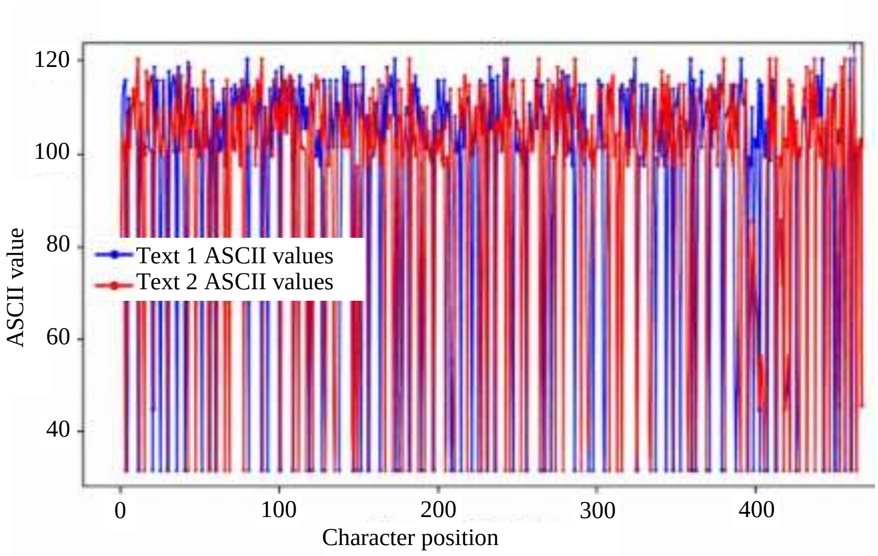

Figure 5. Line chart between two descriptive text input.

CONCLUSION

This paper introduced a mixed approach to the automatic analysis of the descriptive responses through the combination of text processing and image processing tools. The system deals with impairments of manual evaluation especially in pen-and-paper based assessments. Neural network-based recognition methods are used to scan, divide, and translate handwritten answer scripts into machine readable text. The similarity between the syntactic and semantic similarity of the processed student answers and the expert prepared standard answers are tested.

The system performance is evaluated using the gold-standard score that is the mean of the scores given by two experts in the domain. The similarity of the system-generated score and gold-standard score is a measure of the accuracy of the proposed model. The frame of the experiment shows that the system may give a stable, objective, and effective analysis of the descriptive responses and minimize the work of the human. The proposed model answer-based model takes into account grammar, vocabulary, coherence, structure, and relevance, which guarantees the all-encompassing process of assessment of the work.

REFERENCES

1. Tan H, Wang C, Duan Q, Lu Y, Zhang H, Li R. Automatic short answer grading by encoding student responses via a graph convolutional network. Interactive Learning Environments. 2023 Apr 3;31(3):1636-50.
2. Singla YK, Gupta A, Bagga S, Chen C, Krishnamurthy B, Shah RR. Speaker-conditioned hierarchical modeling for automated speech scoring. In Proceedings of the 30th ACM international conference on information & knowledge management 2021 Oct 26 (pp. 1681-1691).
3. Kaushal A, Mahowald K. What do tokens know about their characters and how do they know it?. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2022 Jul (pp. 2487-2507).
4. Xu X, Wang Y, Xu C, Ding Z, Jiang J, Ding Z, Karlsson BF. A survey on game playing agents and large models: Methods, applications, and challenges. arXiv preprint arXiv:2403.10249. 2024 Mar 15.
5. Zhu X, Wu H, Zhang L. Automatic short-answer grading via BERT-based deep neural networks. IEEE Transactions on Learning Technologies. 2022 May 19;15(3):364-75.
6. Song Y, Zhu Q, Wang H, Zheng Q. Automated essay scoring and revising based on open-source large language models. IEEE Transactions on Learning Technologies. 2024 May 6; 17:1880-90.

7. Sahu A, Bhowmick PK. Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. IEEE Transactions on Learning Technologies. 2019 Feb 7;13(1):77-90.
8. Oswald AJ, Proto E, Sgroi D. Happiness and productivity. Journal of labor economics. 2015 Oct 1;33(4):789-822.
9. Alqurashi A, Alharbi B, Sabbeh S. An Automatic Grading System for Arabic Language Short-Answer Questions Using Deep Learning. Engineering, Technology & Applied Science Research. 2025 Oct 6;15(5):26665-75.
10. Kale S. The Future of IP Protection: Harnessing the Power of AI Language and Vision Models. Technology (JARET). 2024;3(1):268-76.
11. Chavva RK, Muthyam SR, Seelam MS, Nalliboina N. A Transformer-Based Approach for Enhancing Automated Essay Scoring. In2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET) 2024 Aug 23 (pp. 1-6). IEEE.
12. Nuthalapati SB. Transforming agriculture with deep learning approaches to plant health monitoring. Remittances Review. 2022 Aug;7(1):227-38.
13. Bauer E, Sailer M, Niklas F, Greiff S, Sarbu-Rothsching S, Zottmann JM, Kiesewetter J, Stadler M, Fischer MR, Seidel T, Urhahne D. AI-based adaptive feedback in simulations for teacher education: an experimental replication in the field. Journal of Computer Assisted Learning. 2025 Feb;41(1): e13123.
14. Yu X, Lyu X, Peng R, Shen J. Solving arithmetic word problems by synergizing syntax-semantics extractor for explicit relations and neural network miner for implicit relations. Complex & Intelligent Systems. 2023 Feb;9(1):697-717.
15. Ch DR, Saha SK. Generation of multiple-choice questions from textbook contents of school-level subjects. IEEE Transactions on Learning Technologies. 2022 Nov 25;16(1):40-52.
16. Ren P, Yang L, Luo F. Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis. Education and information technologies. 2023 Jan;28(1):797-814.
17. Nguyen HT, Nguyen CT, Oka H, Ishioka T, Nakagawa M. Fully automatic scoring of handwritten descriptive answers in Japanese language tests. IEICE Technical Report; IEICE Tech. Rep.. 2021 Dec 9;121(304):45-50.