Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

An AI-Powered Sentence-BERT Framework for Automated Assessment of Open-Ended Responses: A Multi-Model Evaluation and Response-Length Bias Analysis

Temidayo Oluwatosin Omotehinwa [1,*], Folashade Habibat Omotehinwa [2], Adeyi Thomas Edeh [1], and Yakubu Aliyu Ibrahim [1]

[1] Department of Mathematics and Computer Science, Federal University of Health Sciences, Otukpo 972261, Nigeria;
e-mail : temidayo.omotehinwa@fuhso.edu.ng; adeyiedeh@gmail.com; yakubu.ibrahim@fuhso.edu.ng
[2] Department of Chemistry, Federal University of Health Sciences, Otukpo 972261, Nigeria;
e-mail : folashade.omotehinwa@fuhso.edu.ng
* Corresponding Author: Temidayo Oluwatosin Omotehinwa

Abstract: The automated assessment of open-ended responses remains challenging due to the complexity and subjective nature of human grading. A further methodological challenge is determining which pretrained Sentence-BERT model and scoring configuration can provide reliable agreement with human assessment while maintaining computational efficiency. This study presents an AI-powered framework based on Sentence-BERT models for automated grading of short-answer responses, focusing on model performance, reliability, response-length sensitivity, and hybrid scoring. A custom dataset of student responses to open-ended questions, annotated by three independent human raters, was developed and pre-processed. Five Sentence-BERT models, all-MiniLM-L12-v2, all-mpnet-base-v2, stsb-roberta-large, paraphrase-MiniLM-L12-v2, and distilbert-base-nli-stsb-mean-tokens, were evaluated using semantic similarity scoring and multiple metrics, including Pearson and Spearman correlations, Cohen's Kappa, MAE, RMSE, and ICC. Among the evaluated models, all-MiniLM-L12-v2 provided the most favorable balance of grading performance, reliability (ICC = 0.527), and computational efficiency. Human ratings showed stronger dependence on response length (Pearson r = 0.663) than the evaluated AI models (r = 0.338–0.474), indicating reduced sensitivity to verbosity. Weight-sensitivity analysis showed that the 25% semantic–75% rubric configuration had the lowest continuous-score MAE (1.8499) and RMSE (2.3980) among the tested configurations, supporting the complementary contribution of holistic semantic similarity and rubric-based concept coverage. The selected model and hybrid scoring mechanism were implemented in a web-based prototype providing automated scores and interpretable feedback while retaining instructor review.

Keywords: AI-Based Assessment; all-MiniLM-L12-v2; Educational Technology; Hybrid Scoring; Semantic Text Similarity; Sentence-BERT; Short Answer Grading; Transformers.

Received: July, 23rd 2026
Revised: September, 7th 2026
Accepted: September, 10th 2026
Published: September, 20th 2026
Curr. Ver.: September, 20th 2026

1. Introduction

Effective assessment provides valuable evidence of students' learning and conceptual understanding. Although multiple-choice questions are widely used because they are easy to standardize, they primarily assess recognition rather than the active recall, explanation, and knowledge synthesis required in professional practice [1], [2]. Consequently, open-ended short-answer questions (SAQs), which require students to construct and articulate responses in natural language, are increasingly important [3], [4]. However, variation in how students express their understanding makes such responses difficult to assess consistently [5]–[7].

Manual assessment of open-ended responses is labor-intensive and subjective, with scores potentially affected by fatigue, inter-rater variability, and bias [6]–[9]. These challenges

become more pronounced in large-scale settings, where the volume of submissions can hinder timely grading and feedback [10]. This has increased the need for scalable and reliable Automatic Short Answer Grading (ASAG) systems capable of capturing semantic meaning beyond surface-level word matching [4], [11].

Research on automated short-answer grading has evolved from rule-based systems and keyword matching toward deep learning and transformer-based architectures. Traditional approaches commonly relied on surface-level features, such as word frequency and n-gram overlap, which were limited in handling diverse student phrasing that deviates from a reference answer. The emergence of transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), enabled richer contextual representations through bidirectional language modelling [12], [13]. Sentence-BERT subsequently extended this capability by producing sentence-level embeddings that facilitate semantic comparison between student responses and reference answers [14], [15].

Despite these advances, important methodological challenges remain. The growing number of pretrained Sentence-BERT models introduces a model-selection problem because the models differ in pretraining objectives, model capacity, embedding characteristics, and computational requirements. Such differences may affect their ability to capture semantic equivalence between student responses and reference answers. Consequently, performance observed for one model cannot necessarily be assumed to generalize to other architectures, while selection based on a single metric may provide an incomplete basis for practical deployment.

Previous studies have evaluated individual transformer models or compared relatively limited sets of architectures [2], [16], [17]. However, systematic comparisons of multiple Sentence-BERT models under the same assessment conditions, using complementary measures of grading error, correlation, agreement, reliability, and computational efficiency, remain limited. Evidence is also limited regarding response-length effects across human raters and different Sentence-BERT models [3], [9]. Furthermore, relatively few studies bring together comparative model evaluation, response-length analysis, hybrid semantic–rubric scoring, interpretable feedback, and practical deployment within a single assessment framework [10]. Taken together, these gaps suggest the need for an evaluation framework that considers not only predictive association with human scores, but also grading error, agreement, reliability, sensitivity to response characteristics, and computational practicality.

Accordingly, this study develops and evaluates an AI-powered framework for automated assessment of open-ended short-answer responses. The framework systematically compares pretrained Sentence-BERT models using complementary measures of performance, agreement, reliability, response-length sensitivity, and computational efficiency, and subsequently integrates the selected model with hybrid semantic–rubric scoring, interpretable feedback, and a web-based prototype. The study does not propose a new Sentence-BERT architecture or training algorithm; rather, it examines how existing pretrained models can be systematically evaluated, selected, and operationalized for open-ended response assessment. The main contributions are as follows:

- It conducts a multi-model evaluation to identify the most suitable pretrained Sentence-BERT model among those tested under the evaluation criteria adopted in this study.
- It applies a statistical validation framework incorporating correlation, quadratic weighted Cohen's Kappa, MAE, RMSE, ICC, the Friedman test, Kendall's W, and corrected post-hoc comparisons to examine model performance and agreement from complementary perspectives.
- It examines response-length sensitivity and compares the association between response length and scores assigned by the evaluated AI models and human consensus ratings.
- It identifies all-MiniLM-L12-v2 as the preferred model among those evaluated, based on the observed balance of grading error, reliability, and computational efficiency under the study conditions.
- It develops an AI-powered web-based automated short-answer grading prototype that combines semantic and rubric-based scoring to provide automated scores and interpretable feedback while retaining instructor review.

The remainder of this paper is organized as follows. Section 2 reviews relevant research on automated short-answer assessment. Section 3 describes the methodology, including the dataset, preprocessing, evaluated Sentence-BERT models, hybrid semantic–rubric scoring

framework, evaluation metrics, and statistical analyses. Section 4 presents and discusses the experimental findings, including model performance, reliability, response-length sensitivity, computational efficiency, error analysis, weighting sensitivity, component ablation, and prototype implementation. Finally, Section 5 summarizes the main findings, limitations, and directions for future research.

2. Related Work

Previous research on Automatic Short Answer Grading (ASAG) has addressed practical assessment systems, transformer-based scoring, model comparison, response-length effects, fairness, and automated feedback. Collectively, these studies demonstrate substantial progress in semantic assessment while also revealing several unresolved issues concerning model selection, evaluation reliability, and practical deployment.

Study [10] introduced the GradeAid framework to address the limited availability of practical software tools for instructors. Its findings highlighted the gap between research prototypes evaluated primarily through isolated performance measures and systems designed to support real-world grading. Similarly, [18] proposed a hybrid BERT-based approach with customized attention layers to model complex linguistic relationships in short-answer assessment. These studies illustrate advances in practical deployment and semantic modelling, but they leave open the question of how alternative Sentence-BERT models perform when evaluated under common assessment conditions.

Recent research has increasingly examined model reliability, fairness, comparative performance, and computational efficiency. An empirical analysis of GPT-4 for short-answer grading found relatively consistent scores across demographic groups, while also reporting leniency bias compared with human markers [1]. Study [11] compared BERT, RoBERTa, and all-MiniLM-L12-v2, highlighting the importance of considering both grading performance and computational efficiency. Other work has investigated the influence of response characteristics, including length and complexity, on automated scores [9], while [19] demonstrated the importance of computationally efficient AI-supported workflows for large-scale examination datasets. These findings suggest that assessment models should be evaluated not only by their agreement with human scores but also by their robustness to response characteristics and their practical computational requirements.

Reliability and ethical considerations have also emerged as important concerns in automated grading. Low reliability and systematic bias have been identified as barriers to wider adoption of AI-based assessment [3]. In clinical dental education, [20] documented a tendency for AI models to over-score incorrect or partially correct responses, describing this phenomenon as "optimism bias" and highlighting the risks of uncontrolled use in high-stakes assessment. Study [21] therefore advocated instructor-in-the-loop systems in which AI supports, rather than replaces, human judgement. In addition, [22] found that lexical similarity measures such as Cosine and Jaccard similarity are insufficient to capture the nuanced semantic meaning required for higher education assessment. Together, these studies indicate that reliable automated assessment requires semantic modelling alongside appropriate validation and human oversight.

Research has also increasingly considered the pedagogical value of automated feedback. Study [23] developed a system for the Algeomath platform that generated personalized feedback using submission history and teacher-authored examples. Similarly, [22] used GPT-4 to generate rubric-informed feedback that identified conceptual gaps and suggested improvements to student responses. These studies demonstrate the potential of automated feedback to extend ASAG beyond score prediction toward more informative learning support. However, they primarily focus on feedback generation rather than integrating feedback with systematic multi-model evaluation, response-length analysis, and comparative assessment reliability.

Overall, the literature indicates progress across several complementary dimensions of ASAG, including semantic modelling, model comparison, fairness analysis, automated feedback, and practical system development. Nevertheless, these dimensions have generally been investigated separately. Comparative studies often consider individual models or limited architecture sets, while evaluation frequently relies on a restricted number of metrics. Evidence also remains limited regarding response-length effects across human raters and multiple Sentence-BERT models. Moreover, relatively few studies integrate model comparison, reliability

analysis, response-length sensitivity, hybrid scoring, interpretable feedback, and practical deployment within a unified framework [2], [3], [9], [16], [17]. This leaves a methodological gap in determining how pretrained Sentence-BERT models should be compared and selected when grading error, agreement, reliability, response-length sensitivity, and computational efficiency are considered jointly.

To address this gap, the present study evaluates five pretrained Sentence-BERT models under common assessment conditions using complementary measures of grading error, correlation, agreement, reliability, response-length sensitivity, and computational efficiency. It further examines semantic–rubric weighting and component ablation to assess the contribution of the scoring components before implementing the selected configuration in a web-based prototype with interpretable feedback and instructor review. This design provides a controlled basis for examining model selection while connecting comparative evaluation with the practical requirements of automated open-ended response assessment.

3. Methodology

3.1. Study Design

This study adopts a quantitative experimental design to develop and evaluate an AI-powered system for automated grading of open-ended short-answer responses. The methodology combines natural language processing with statistical evaluation to assess model performance, reliability, response-length sensitivity, and computational efficiency. A multi-model comparative framework is used to evaluate pretrained Sentence-BERT architectures, with human consensus scores serving as the reference standard. The overall methodological workflow is presented in Fig. 1.

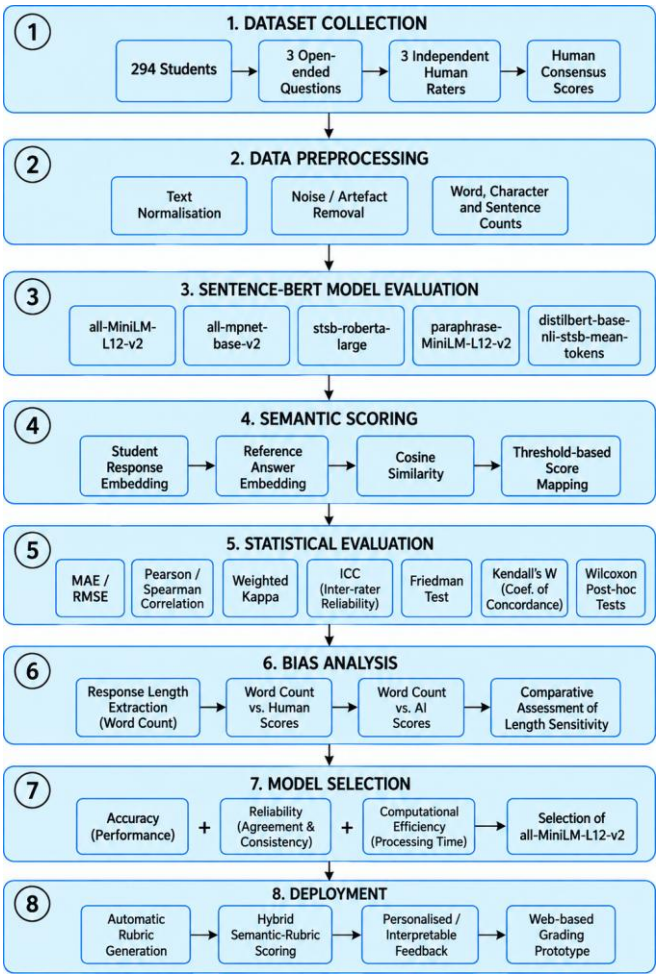

Figure 1. Methodological workflow of the proposed AI-powered Sentence-BERT framework for automated assessment of open-ended responses.

The experimental procedure comprises dataset preparation, text preprocessing, comparative evaluation of five pretrained Sentence-BERT models, statistical validation against human assessment, response-length sensitivity analysis, and computational-efficiency assessment. The model selected through this comparative evaluation is subsequently incorporated into the hybrid semantic–rubric scoring framework and implemented in a web-based prototype. This design therefore links model evaluation with the subsequent scoring and deployment stages while maintaining a common evaluation setting across all tested models.

3.2. Dataset Collection and Structure

The dataset used in this study was derived from an authentic continuous assessment administered to 294 undergraduate students at the Federal University of Health Sciences, Otukpo, Benue State, Nigeria, who were enrolled in COS101, an introductory computer course. Each student responded to three open-ended short-answer questions (Q1, Q2, and Q3), resulting in 882 question-level responses. Three independent human raters assessed each response using the same marking guide and scoring rubric following a calibration exercise to establish a common understanding of the assessment criteria. The arithmetic mean of the three ratings was used as the consensus reference score.

The dataset was selected because open-ended responses introduce substantial variation in wording, response length, and sentence structure, making automated short-answer grading more challenging than fixed-response assessment. As shown in Table 1, responses ranged from 1 to 245 words and from 1 to 24 sentences across the three questions. This variation provides an appropriate setting for evaluating whether Sentence-BERT models capture semantic content rather than relying primarily on superficial characteristics such as response length. The availability of three independent ratings for each response also enables the assessment of inter-rater variability and agreement between automated scores and human judgments.

The locally collected dataset also provided a consistent assessment context, including common questions, reference answers, marking criteria, and three independent human ratings. Public ASAG datasets provide valuable benchmarks, but they differ in domains, question sets, grading schemes, and annotation protocols. Therefore, the present study prioritized a controlled, multi-rater dataset to evaluate model differences under identical conditions, while recognizing that validation on established public ASAG benchmarks would be valuable for assessing broader generalizability.

The models were evaluated in a zero-shot setting without fine-tuning on the study dataset. This design enabled a controlled comparison of pretrained Sentence-BERT models without introducing dataset-specific fine-tuning effects. All models therefore received the same student responses, reference answers, scoring procedure, and evaluation conditions. Classical machine-learning baselines, fine-tuned transformer models, and LLM-based approaches were outside the scope of this study, which focuses on comparative model selection within the pretrained Sentence-BERT family. The descriptive characteristics of the student responses are presented in Table 1.

Table 1. Descriptive characteristics of the student responses.

Feature	Q1	Q2	Q3
Responses (n)	294	294	294
Mean word count	31.82	35.87	32.57
Std. dev. (word count)	29.64	31.96	29.75
Median word count	25	28	25
Minimum word count	1	1	1
Maximum word count	234	243	245
Mean character count	229.09	251.17	223.75
Mean sentence count	1.58	1.49	1.60
Median sentence count	1	1	1
Maximum sentence count	24	21	17
Very short responses (<3 words)	8	3	3
Missing responses	0	0	0

As shown in Table 1, each question contains 294 responses with no missing entries. Mean response length ranges from 31.82 to 35.87 words, while the standard deviations and wide ranges indicate considerable variation in student expression. Most responses contain a single sentence (median = 1), although some extend beyond 20 sentences. These results show substantial variation in response length and structure across the three questions, providing the basis for the subsequent response-length sensitivity analysis.

The dataset therefore captures both concise and more elaborate response styles, providing a heterogeneous basis for evaluating semantic grading and examining the relationship between response characteristics and automated scores. For each response, let Q_i denote the response to question i, where $i \in \{1,2,3\}$, and let $HR1_i$, $HR2_i$, and $HR3_i$ denote the scores assigned by the three human raters. The consensus human score for question i was calculated as the arithmetic mean of the three ratings:

$$\bar{H}_i = \frac{HR1_i + HR2_i + HR3_i}{3}, \quad i \in \{1,2,3\} \tag{1}$$

The total human score for each response was then calculated by summing the consensus scores across the three questions:

$$H_{\text{total}} = \sum_{i=1}^{3} \bar{H}_i \tag{2}$$

This aggregation reduces the influence of individual rater variation and provides a more stable reference score for evaluating automated grading performance.

3.3. Data Preprocessing and Feature Extraction

A text preprocessing pipeline was applied to improve data quality and consistency while preserving information relevant to semantic assessment. Text normalization was first performed using Unicode standardization to correct encoding anomalies, including malformed apostrophes and quotation marks. Noise removal included the elimination of irrelevant non-alphanumeric symbols, Markdown artefacts such as asterisks, and redundant whitespace. Regular expressions were used to remove irrelevant characters while preserving essential punctuation. For the subsequent response-length sensitivity analysis, auxiliary textual features were extracted from the cleaned responses, including word count, character count, and sentence count. These features were used to examine whether automated scores were systematically associated with response length.

3.4. Sentence-BERT Models

Five pretrained Sentence-BERT models were purposively selected to provide variation in transformer architecture, model capacity, embedding dimensionality, and sentence-embedding training objectives while maintaining a common sentence-embedding evaluation paradigm. The selected models include lightweight MiniLM-based models, a higher-capacity MPNet model, a RoBERTa-large model, and a DistilBERT-based model, providing variation in computational cost and representational capacity. They also differ in training configurations, including general-purpose sentence embedding, paraphrase-oriented training, and NLI/semantic-textual-similarity-oriented training. This diversity enables the study to examine whether differences in architecture, capacity, and training objective are associated with differences in grading accuracy, agreement, reliability, and computational efficiency.

The comparison was restricted to the Sentence-BERT family to maintain a controlled evaluation of pretrained models designed to generate semantically meaningful sentence-level embeddings. Because the proposed grading approach relies on semantic similarity between student responses and reference answers, Sentence-BERT provides a common representation and scoring paradigm across the evaluated models. Restricting the comparison to this family also allows performance differences to be examined primarily in relation to model architecture, capacity, embedding characteristics, and training objective rather than differences between fundamentally different assessment paradigms. Accordingly, this study does not seek to establish the superiority of Sentence-BERT over other NLP approaches; instead, it focuses on model selection within the Sentence-BERT family for the specific task and evaluation

setting considered. The architectural characteristics and common experimental settings of the evaluated models are summarized in Table 2.

Table 2. Configuration and architectural characteristics of the evaluated Sentence-BERT models

Model	Architecture	Parameters	Embedding Dimension	Pretraining / Training Objective	Evaluation Mode	Batch Size	Similarity	Score Mapping
all-MiniLM-L12-v2	MiniLM-L12	~33M	384	General-purpose sentence embedding	Zero-shot	32	Cosine	0–5 threshold mapping
all-mpnet-base-v2	MPNet-base	~109M	768	General-purpose sentence embedding	Zero-shot	32	Cosine	0–5 threshold mapping
stsb-roberta-large	RoBERTa-large	~355M	1024	STS-oriented sentence embedding	Zero-shot	32	Cosine	0–5 threshold mapping
paraphrase-MiniLM-L12-v2	MiniLM-L12	~33M	384	Paraphrase/sentence-embedding training	Zero-shot	32	Cosine	0–5 threshold mapping
distilbert-base-nli-stsb-mean-tokens	DistilBERT-base	~66M	768	NLI + STS sentence embedding	Zero-shot	32	Cosine	0–5 threshold mapping

All models were evaluated under identical zero-shot conditions using the same student responses, question-specific reference answers, batch size, cosine-similarity measure, and score-conversion thresholds. The dataset served as the evaluation dataset for comparative assessment rather than as an independent hold-out benchmark. The comparative results were subsequently used to select the Sentence-BERT model and inform the semantic–rubric weighting configuration. Because the dataset contributed to model and scoring-configuration selection, it should not be interpreted as an independent test set. Nevertheless, the common examination context enabled evaluation under consistent and realistic assessment conditions.

Each model was used to encode the student response and reference answer into dense vector representations in a shared semantic space. Let Q_i denote the student response to question i, and let R_i denote the corresponding reference answer. For a given Sentence-BERT model, the embeddings are obtained as:

$$\mathbf{e}_{Q_i} = f(Q_i), \quad \mathbf{e}_{R_i} = f(R_i), \tag{3}$$

where $f(\cdot)$ denotes the embedding function of the selected Sentence-BERT model.

The objective was not to compare all available large language models, but to identify a suitable lightweight Sentence-BERT encoder for automated grading. Sentence-BERT models were selected because they generate semantically meaningful sentence embeddings while requiring substantially fewer computational resources than autoregressive large language models, making them appropriate for scalable institutional deployment and real-time assessment [24].

3.5. Semantic Similarity Computation

Cosine similarity was used to measure the semantic similarity between each student response and its corresponding reference answer. Given the embeddings defined in Eq. (3), the semantic similarity is calculated as:

$$S_i = \frac{\mathbf{e}_{Q_i} \cdot \mathbf{e}_{R_i}}{\|\mathbf{e}_{Q_i}\| \|\mathbf{e}_{R_i}\|} \tag{4}$$

where S_i denotes the cosine similarity between the student response and reference answer for question i, $\mathbf{e}_{Q_i}$ is the embedding of the student response, and $\mathbf{e}_{R_i}$ is the embedding of the corresponding reference answer.

Cosine similarity was selected because it measures angular similarity in high-dimensional embedding spaces and has been widely applied to semantic textual similarity tasks [25]–[30]. This similarity score provides the continuous semantic signal used in the subsequent score-mapping and hybrid semantic–rubric scoring stages.

3.6. Scoring Mechanism and Rubric Alignment

The cosine-similarity thresholds used for automated score mapping were specified a priori as rubric-informed operational cutoffs, as shown in Eq. (5). They were established before implementation of the web-based grading system and were not empirically optimized using the study responses. Their purpose was to provide a consistent and interpretable mapping between the continuous cosine-similarity scores produced by Sentence-BERT and the discrete performance levels defined in the human scoring rubric.

$$\text{Score}(x) = \begin{cases} 5, & \text{if } \text{Sim}(x,r) \geq 0.8 \\ 3, & \text{if } 0.6 \leq \text{Sim}(x,r) < 0.8 \\ 1, & \text{if } 0.4 \leq \text{Sim}(x,r) < 0.6 \\ 0, & \text{if } \text{Sim}(x,r) < 0.4 \end{cases} \tag{5}$$

The human scoring rubric used by the raters assigns 5 marks when most or all key concepts are covered and adequately explained, 3 marks when some key concepts are covered with partial explanation, 1 mark when limited understanding of the concept is demonstrated, and 0 marks when the response is incorrect or irrelevant. Accordingly, cosine-similarity values of 0.80 and above were mapped to the highest performance level, values between 0.60 and below 0.80 to the intermediate level, values between 0.40 and below 0.60 to the lower partial-credit level, and values below 0.40 to zero credit. These thresholds were therefore intended to operationalize the qualitative distinctions in the human marking scheme rather than identify statistically optimal cutoff points. The thresholds should consequently be interpreted as rubric-informed scoring specifications rather than empirically validated semantic benchmarks. Although the same thresholding procedure was applied consistently across all evaluated models, the present study did not perform a dedicated threshold-sensitivity analysis or empirical optimization of these cutoffs.

3.7. Model Evaluation and Statistical Analysis

3.7.1. Model Comparison

The non-parametric Friedman test was applied to determine whether significant differences existed among the scores assigned by the multiple raters (three human raters or five models) to the same set of student responses across the three questions. The analysis was conducted using both total scores and per-question scores. The per-question analysis enabled a more fine-grained examination of scoring behavior while accounting for variation across student responses.

The Friedman test compares multiple raters under repeated-measures conditions and does not assume normally distributed scores. It accounts for within-subject dependencies by comparing the scores assigned to the same student responses, thereby reducing the influence of response-specific variability on the observed differences. The test statistic was calculated as

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{6}$$

where n is the number of samples, k is the number of models or human raters, and R_j is the rank sum for model or human rater j. When the Friedman test indicated a significant difference, Kendall's coefficient of concordance was used to quantify the effect size, as given in Eq. (7):

$$W = \frac{\chi_F^2}{n(k-1)} \tag{7}$$

Post-hoc pairwise comparisons were then conducted using the Wilcoxon signed-rank test with Bonferroni correction to identify the specific pairs for which significant differences occurred.

3.7.2 Evaluation Metrics: Agreement and Reliability

A multi-metric evaluation was used to characterize agreement, predictive error, and inter-rater reliability between human raters and the evaluated models. Pearson and Spearman correlations were used to assess linear association and rank-order consistency, respectively.

Quadratic weighted Cohen's Kappa was used to assess agreement between human raters and between human and AI scores while accounting for agreement occurring by chance. Predictive error was evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), calculated as in Eq. (8).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

where y_i denotes the human consensus score and $\bar{y}_i$ denotes the corresponding automated score.

The consistency and absolute agreement among the three human raters and each AI model were further evaluated using the Intraclass Correlation Coefficient (ICC). Specifically, ICC(2,1), a two-way random-effects model for absolute agreement based on single measurements, was used to assess overall inter-rater reliability. The use of multiple complementary metrics provides a more comprehensive assessment of model performance because correlation, agreement, absolute error, and reliability capture different aspects of grading consistency.

3.7.3. Computational Efficiency

Processing time for each model was recorded to assess computational efficiency, scalability, and practical feasibility. This measure is particularly relevant to real-world deployment, where computational cost must be considered alongside grading accuracy.

3.7.4. Response-Length Bias Analysis

The bias analysis was specifically limited to response-length sensitivity as a potential source of construct-irrelevant influence on grading. Pearson and Spearman correlations between response length and scores were examined for both human raters and the evaluated AI models. This analysis was intended to determine whether scoring behavior was systematically associated with response length. It does not constitute a comprehensive assessment of algorithmic fairness.

3.7.5 Error Analysis, Weight Sensitivity, and Component Ablation

Three supplementary analyses were conducted to further examine the behavior and robustness of the automated scoring framework. First, an error analysis compared automated scores with human consensus scores to quantify the magnitude and direction of scoring discrepancies and identify representative cases of substantial over- or under-scoring. This analysis used the selected all-MiniLM-L12-v2 model under semantic-only scoring.

Second, a weight-sensitivity analysis examined the effect of different relative contributions of holistic semantic similarity and adjusted rubric scoring. Eight semantic-to-rubric weighting configurations (0:100, 25:75, 40:60, 50:50, 55:45, 60:40, 75:25, and 100:0) were evaluated using MAE, RMSE, Pearson correlation, Spearman correlation, and weighted Cohen's Kappa. The configuration yielding the lowest observed MAE and RMSE was subsequently selected for implementation. Third, a component ablation analysis compared semantic-only, adjusted-rubric-only, and hybrid 25:75 scoring to examine the contribution of each component. The same evaluation metrics were applied, with performance assessed against the human consensus scores.

3.8. Web-Based Automated Grading System

3.8.1. Web-Based Deployment

The selected Sentence-BERT model was integrated into a web-based proof-of-concept system to demonstrate the practical application of the proposed grading framework. The system provides an interface through which instructors can define assessment questions and reference answers, upload student responses, and obtain automated scores and feedback. A client–server architecture was adopted, with the user interface communicating with a backend grading service responsible for preprocessing responses, generating embeddings, applying the hybrid scoring mechanism, and returning grading results. The web-based system serves as the deployment layer for the proposed methodology, while automatic rubric generation and hybrid semantic–rubric scoring constitutes the principal methodological components of the framework.

3.8.2. Automated Rubric Generation

To improve the interpretability of semantic grading, the proposed framework automatically derives a structured rubric from the reference answer. Holistic semantic similarity provides an overall measure of correspondence between a student response and the reference answer but does not explicitly indicate which concepts have been demonstrated or omitted. The automatic rubric generation component therefore converts the reference answer into a set of semantically meaningful concepts that serve as interpretable scoring criteria.

The process begins by generating candidate concepts from the reference answer using unigram, bigram, and trigram extraction. These candidates are represented using transformer-based embeddings and semantically filtered to remove candidates with weak relevance to the overall meaning of the reference answer. The remaining candidates are compared using pairwise cosine similarity, and highly similar candidates are merged to reduce redundancy and prevent multiple expressions of the same concept from receiving separate scoring contributions.

The refined concepts are then grouped using K-means clustering to identify representative conceptual units within the reference answer. The resulting units are categorized into three pedagogical dimensions: definition, explanation, and example. This categorization allows the generated rubric to represent not only whether a concept is mentioned but also the type of understanding expected from the student. Definition and explanation components are assigned weights of 40% each, while examples contribute 20% to the rubric score.

For each question, the resulting rubric consists of a set of weighted concepts derived from the reference answer rather than a fixed list of manually specified keywords. This approach allows semantically equivalent expressions to contribute to the same conceptual criterion and reduces dependence on exact lexical matching. The generated rubric also provides an interpretable basis for identifying concepts that are adequately covered and those that are missing from a student's response, thereby supporting the generation of targeted feedback. The automatically generated rubric is intended as a methodological aid rather than a replacement for instructor expertise. Instructors can inspect and refine the extracted concepts before use, allowing pedagogical judgment to be retained while reducing the effort required to construct detailed marking criteria.

3.8.3. Hybrid Semantic–Rubric Scoring Mechanism

The proposed framework combines concept-level rubric scoring with holistic semantic similarity to address the limitations of relying on either approach independently. Holistic semantic similarity captures the overall conceptual correspondence between a student response and the reference answer, allowing semantically equivalent responses to receive appropriate credit even when their wording differs. However, a single similarity score provides limited information about which aspects of the expected answer have been demonstrated. Conversely, rubric-based scoring provides interpretable evidence of concept coverage but may be sensitive to how individual concepts are expressed. The hybrid mechanism therefore combines the complementary strengths of both approaches by integrating holistic semantic understanding with explicit concept-level assessment.

- Rubric-Based Scoring:

Rubric-based scoring measures the degree to which the student response corresponds to each concept extracted from the reference answer. For each rubric concept, cosine similarity is calculated between the student-response embedding and the concept embedding. Let s_i denote the similarity between the student response and concept i and let w_i denote the weight assigned to that concept. The rubric score is computed as

$$S_r = \sum_{i=1}^{k} w_i, s_i \tag{9}$$

where k is the number of rubric concepts.

Partial credit is awarded proportionally based on the similarity thresholds defined in Section 3.6, allowing the framework to recognize varying degrees of conceptual correspondence rather than treating each concept as simply present or absent. The weighted aggregation of concept-level scores produces an overall rubric-based score that reflects the extent to which the key concepts represented in the reference answer are covered by the student response.

- Semantic Similarity Scoring

In addition to concept-level assessment, holistic semantic scoring is used to capture the overall meaning of the student response. The student-response embedding is compared with the reference-answer embedding using cosine similarity. The resulting similarity is converted to a score according to the maximum score allocated to the question, as shown in Eq. (10):

$$S_s = cosine(E_{\text{student}}, E_{\text{model}}) \tag{10}$$

where M is the maximum score allocated to the question.

The holistic semantic score complements the rubric-based score by allowing responses that express the expected knowledge using alternative wording, sentence structures, or semantically equivalent expressions to receive appropriate credit.

- Final Score Aggregation

The final score is obtained by combining holistic semantic similarity (25%) and adjusted rubric-based concept coverage (75%), as shown in Eq. (11):

$$S_{\text{final}} = 0.75 \cdot S_r + 0.25\, S_s \tag{11}$$

The greater weight assigned to the rubric component reflects its role in explicitly assessing concept-level coverage, while the semantic component captures the overall conceptual correspondence between the response and the reference answer. To prevent the rubric component from becoming disproportionately low when the response demonstrates strong overall semantic correspondence, the rubric score is adjusted using a lower bound of 50% of the holistic semantic score. The final score is then calculated using 25% of the semantic score and 75% of the adjusted rubric score, subject to the maximum score allocated to the question.

The hybrid mechanism combines semantic robustness through holistic comparison with interpretability through explicit concept-level assessment. By integrating these components, the framework reduces dependence on surface-level lexical matching while maintaining a direct link between the assigned score and the knowledge components represented in the reference answer. The resulting concept-level information is also used to identify sufficiently covered and missed concepts, which forms the basis for the personalized feedback generated by the system.

3.9. Personalised Feedback Generation

The system uses the rubric evaluation results to generate interpretable and personalized feedback. The feedback identifies concepts that are sufficiently covered in the student response and highlights important concepts that are not adequately addressed. The missed concepts are then presented as recommended areas for improvement, providing targeted guidance based on the specific content of the response. This approach links the automated score to concept-level evidence, allowing the feedback to provide more actionable information than a numerical score alone.

4. Results and Discussion

4.1. Human Rater Analysis

The human-rater scores provide an important reference for examining the consistency of subjective grading before evaluating the automated models. Differences in central tendency and score dispersion can be observed among the three human raters. As shown in Figure 2, HR1 and HR3 show similar central tendencies, with median scores around 7 and relatively higher means, indicating a tendency toward more generous grading. In contrast, HR2 has a lower median of approximately 5 and a lower mean, suggesting a comparatively stricter scoring pattern. The score distributions also show noticeable differences in dispersion. The interquartile ranges indicate variability in the scores assigned to the same set of responses, while HR2 exhibits a relatively more compact distribution despite its lower scoring tendency. Several high-value outliers are also observed, particularly for HR2, indicating that some responses received substantially higher scores despite its generally stricter scoring pattern. Overall, the overlap among the distributions indicates that the raters generally evaluated the responses within a similar range, but with different levels of severity.

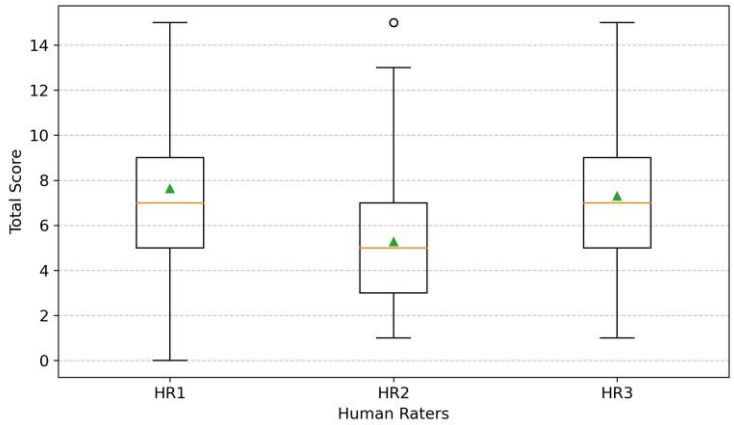

Figure 2. Distribution of scores across human raters.

The Friedman test results in Table 3 provide further evidence of these differences. Statistically significant differences were observed among the three human raters for both the overall scores and each question ($p < 0.001$). The overall Kendall's W was 0.4517, indicating a moderate level of concordance among the raters while also reflecting systematic differences in their scoring patterns. At the question level, Kendall's W values were 0.2654 for Q1, 0.2359 for Q2, and 0.1996 for Q3. The lower values at the question level indicate that the degree of agreement varied across the assessment items.

Table 3. Friedman Test and Effect Size for Human Raters

Analysis Level	Chi-square (χ^2)	p-value	Kendall's W
Overall (Total Scores)	265.612	<0.001	0.4517
Q1	156.049	<0.001	0.2654
Q2	138.712	<0.001	0.2359
Q3	117.388	<0.001	0.1996

To identify the specific pairwise differences, a post-hoc Wilcoxon signed-rank test with Bonferroni correction was conducted. As reported in Table 4, all pairwise comparisons were statistically significant. This indicates that the observed differences were not limited to a particular pair of raters but were present across all three comparisons. The results therefore suggest that the raters had distinct scoring tendencies, consistent with the differences observed in their score distributions.

Table 4. Post-hoc Pairwise Comparisons Between Human Raters

Rater Pair	p-value	Significance
HR1 vs. HR2	<0.001	Significant
HR1 vs. HR3	0.00183	Significant
HR2 vs. HR3	<0.001	Significant

Taken together, the human-rater analysis shows that subjective grading in the dataset contains measurable variation in both score level and scoring consistency. HR2 generally assigned lower scores than HR1 and HR3, while the statistically significant pairwise comparisons confirm that these differences were systematic. This finding is relevant when interpreting the subsequent AI results because the human consensus score represents an aggregated reference derived from judgments that are not perfectly identical. Thus, agreement between an AI model and the consensus should be interpreted in the context of the underlying variability in human assessment.

4.2. Sentence-BERT Model Comparison

The five Sentence-BERT models show broadly similar score distributions, although differences in central tendency and score dispersion can be observed. As shown in Fig. 3, the

median scores are generally clustered around 9–11. MPNet, RoBERTa, and DistilBERT exhibit slightly higher median scores, indicating a tendency toward more generous scoring, whereas all-MiniLM-L12-v2 and ParaMiniLM show marginally lower medians. Lower-end outliers are particularly evident for MPNet, RoBERTa, and DistilBERT, while the concentration of scores near the upper range indicates that the models frequently assign high scores to responses with strong semantic correspondence to the reference answers.

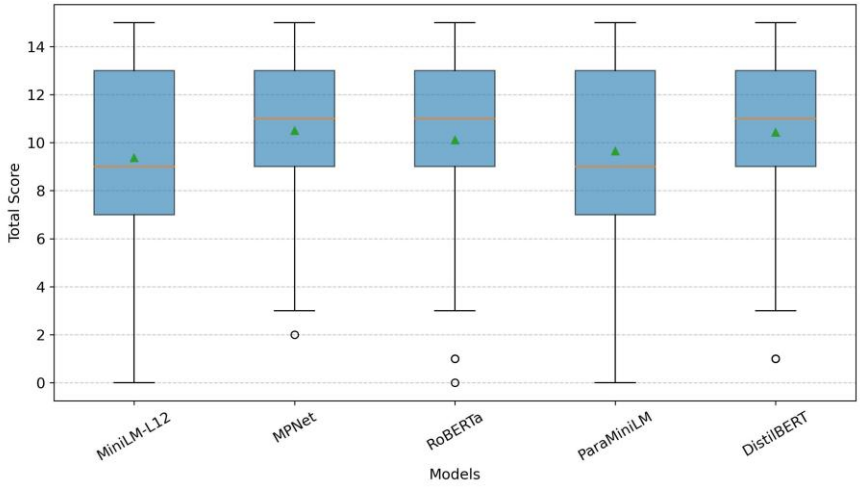

Figure 3. Distribution of total scores across Sentence-BERT models.

The Friedman test results in Table 5 indicate statistically significant differences in the scores produced by the five models ($p < 0.001$). However, the relatively small Kendall's W for the overall scores (0.121) suggests that these differences are limited in magnitude. At the question level, the smallest and largest concordance values were observed for Q2 (W=0.0563) and Q3 (W= 0.1768), respectively, indicating that the extent of variation among model outputs depends on the assessment question.

Table 5. Friedman Test and Effect Size for Sentence-BERT Models

Analysis Level	Chi-square (χ^2)	p-value	Kendall's W
Overall (Total Scores)	141.999	<0.001	0.121
Q1	175.176	<0.001	0.149
Q2	66.228	<0.001	0.0563
Q3	207.945	<0.001	0.1768

The combination of statistical significance and relatively small effect sizes is important for interpreting these results. Although the Friedman test indicates that the models do not produce identical score distributions, the Kendall's W values indicate that the magnitude of the differences is limited. Therefore, the statistical significance should not be interpreted as evidence that one architecture is categorically superior to the others. Instead, the results indicate that the evaluated models exhibit some differences in scoring behavior while maintaining broadly similar overall assessment patterns. This supports the use of multiple complementary criteria for model selection rather than relying on statistical significance alone.

The observed differences may be associated with variations in transformer architecture, model capacity, embedding dimensionality, and sentence-embedding training objectives. Although all five models generate semantic representations, their different training characteristics may influence how relationships between student responses and reference answers are represented. Greater model capacity does not necessarily result in lower grading error or stronger agreement with human scores for a particular assessment task. The relatively small effect sizes therefore suggest that these model characteristics contribute to scoring variation, but do not independently determine assessment performance.

The post-hoc Wilcoxon signed-rank comparisons provide further detail on the differences identified by the Friedman test. As shown in Table 6, significant differences were

observed for most model pairs, while the comparisons between all-MiniLM-L12-v2 and ParaMiniLM and between MPNet and DistilBERT were not statistically significant. This indicates that some models produced similar scoring behavior despite differences in their underlying architectures and training characteristics.

Table 6. Post-hoc Wilcoxon Pairwise Comparisons with Bonferroni Correction

Model Pair	p-value	Significance
all-MiniLM-L12-v2 vs. MPNet	<0.001	Significant
all-MiniLM-L12-v2 vs. RoBERTa	<0.001	Significant
all-MiniLM-L12-v2 vs. ParaMiniLM	0.00697	Not Significant
all-MiniLM-L12-v2 vs. DistilBERT	<0.001	Significant
MPNet vs. RoBERTa	0.00151	Significant
MPNet vs. ParaMiniLM	<0.001	Significant
MPNet vs. DistilBERT	0.61176	Not Significant
RoBERTa vs. ParaMiniLM	0.00082	Significant
RoBERTa vs. DistilBERT	0.00492	Significant
ParaMiniLM vs. DistilBERT	<0.001	Significant

At the question level, the post-hoc results in Table 7 show that the degree of differentiation among models varies across the three questions. Q1 and Q3 generally show more significant pairwise differences, whereas Q2 contains several non-significant comparisons. This pattern is consistent with the smaller effect size observed for Q2 in Table 5. The results suggest that differences among models may partly depend on the characteristics of the assessment question rather than being uniform across all responses.

Table 7. Per-Question Post-hoc Pairwise Comparisons Using the Wilcoxon Signed-Rank Test with Bonferroni Correction

Model Pair	Q1	Q2	Q3
all-MiniLM-L12-v2 vs. MPNet	Significant	Not Significant	Significant
all-MiniLM-L12-v2 vs. RoBERTa	Significant	Not Significant	Significant
all-MiniLM-L12-v2 vs. ParaMiniLM	Significant	Significant	Significant
all-MiniLM-L12-v2 vs. DistilBERT	Significant	Not Significant	Significant
MPNet vs. RoBERTa	Significant	Not Significant	Not Significant
MPNet vs. ParaMiniLM	Significant	Significant	Significant
MPNet vs. DistilBERT	Not Significant	Not Significant	Significant
RoBERTa vs. ParaMiniLM	Significant	Significant	Significant
RoBERTa vs. DistilBERT	Significant	Significant	Significant
ParaMiniLM vs. DistilBERT	Significant	Significant	Significant

Overall, the results in Tables 5–7 indicate that the five Sentence-BERT models follow broadly comparable grading patterns but differ in the magnitude and distribution of their assigned scores. The statistical differences are more apparent at the pairwise and question levels than in the overall effect size. Consequently, subsequent model selection considers not only score distributions and statistical differences, but also agreement with human ratings, grading error, reliability, and computational efficiency. This provides a more balanced basis for identifying a model for the subsequent hybrid scoring framework.

4.3. Inter-Rater and Inter-Model Agreement

To further examine grading consistency, agreement was evaluated separately among human raters and among the Sentence-BERT models. Table 8 reports the quadratic weighted Cohen's Kappa values for the three human-rater pairs across the three questions. The values range from 0.496 to 0.788, indicating moderate to substantial agreement, with the highest agreement observed between HR1 and HR3. The variation across rater pairs confirms that human grading involves measurable differences in scoring severity and interpretation.

Therefore, the human consensus score provides a practical reference for evaluating the automated models while still reflecting the variability inherent in subjective assessment.

Table 8. Inter-Rater Agreement Among Human Raters Based on Quadratic Weighted Cohen's Kappa

Question	HR1–HR2	HR1–HR3	HR2–HR3
Q1	0.527	0.788	0.559
Q2	0.509	0.660	0.607
Q3	0.496	0.743	0.543

Table 9 shows the agreement among the five evaluated models. Pearson correlations range from 0.750 to 0.869, while Spearman correlations range from 0.712 to 0.858. The quadratic weighted Kappa values are also consistently high, ranging from 0.748 to 0.875. These results indicate that the models generally produced similar relative grading patterns, despite the differences observed in their agreement and error measures against the human consensus. Thus, inter-model consistency alone is not sufficient to establish correspondence with human assessment.

Table 9. Inter-Model Agreement

Model Pair	Pearson	Spearman	Kappa
all-MiniLM-L12-v2 vs. MPNet	0.869	0.858	0.816
all-MiniLM-L12-v2 vs. RoBERTa	0.777	0.756	0.748
all-MiniLM-L12-v2 vs. ParaMiniLM	0.862	0.852	0.875
all-MiniLM-L12-v2 vs. DistilBERT	0.862	0.849	0.835
MPNet vs. RoBERTa	0.750	0.712	0.765
MPNet vs. ParaMiniLM	0.836	0.830	0.804
MPNet vs. DistilBERT	0.854	0.840	0.864
RoBERTa vs. ParaMiniLM	0.764	0.745	0.751
RoBERTa vs. DistilBERT	0.827	0.804	0.834
ParaMiniLM vs. DistilBERT	0.859	0.853	0.838

The comparison with human consensus in Table 10 provides a different perspective. The models do not exhibit the same relative performance across all evaluation criteria. DistilBERT achieves the highest Pearson (0.692) and Spearman (0.703) correlations, whereas all-MiniLM-L12-v2 records the lowest MAE (3.171) and RMSE (3.917). All-MiniLM-L12-v2 also obtains the highest Kappa (0.429) among the evaluated models in this comparison. These results illustrate that stronger correlation with human scores does not necessarily correspond to smaller numerical grading errors or higher agreement. Consequently, model selection requires consideration of complementary measures rather than reliance on correlation alone.

Table 10. AI–Human Performance Comparison

Model	MAE	RMSE	Pearson	Spearman	Kappa
all-MiniLM-L12-v2	3.171	3.917	0.591	0.592	0.429
MPNet	3.892	4.623	0.609	0.595	0.337
RoBERTa	3.577	4.109	0.681	0.685	0.406
ParaMiniLM	3.380	4.156	0.584	0.574	0.402
DistilBERT	3.822	4.473	0.692	0.703	0.403

The intraclass correlation coefficients in Table 11 range from 0.467 to 0.527. Under the interpretation adopted in this study, these values indicate limited to moderate reliability across the evaluated models. all-MiniLM-L12-v2 records the highest ICC (0.527), whereas MPNet records the lowest (0.467). The ICC results complement the error and correlation measures by assessing the consistency of the automated scores relative to the human reference. Taken together, Tables 8–11 show that high inter-model agreement does not necessarily translate

into equivalent agreement with human assessment, highlighting the importance of evaluating automated grading from multiple perspectives.

Table 11. Intraclass Correlation Coefficient (ICC) Analysis

Model	ICC
all-MiniLM-L12-v2	0.527
MPNet	0.467
RoBERTa	0.512
ParaMiniLM	0.510
DistilBERT	0.502

4.4 Computational Efficiency

Figure 4 compares the processing time of the five Sentence-BERT models. The results indicate clear differences in computational cost, with all-MiniLM-L12-v2 requiring the shortest processing time, followed by DistilBERT and ParaMiniLM. RoBERTa requires substantially more processing time than the other models, while MPNet falls between RoBERTa and the lighter models.

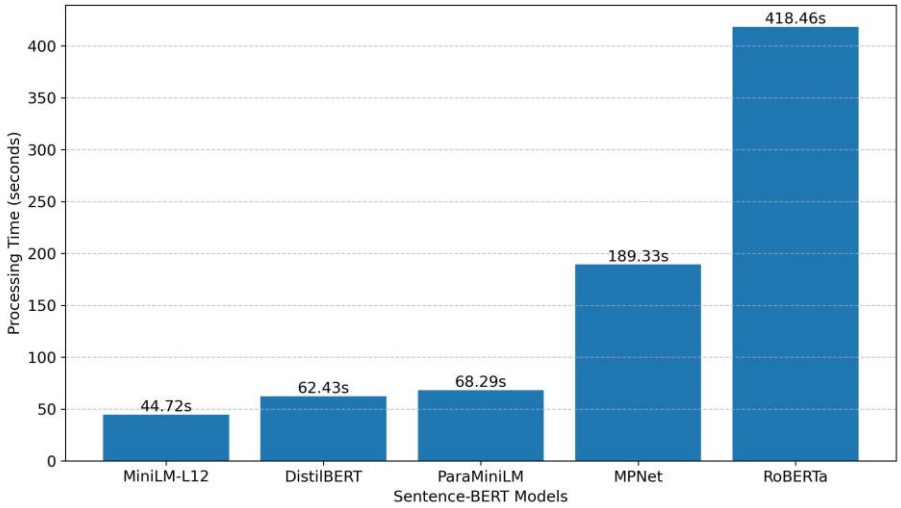

Figure 4. Processing Time Comparison Across Sentence-BERT Models.

The computational results provide an additional dimension for interpreting the model-level performance reported in Tables 10 and 11. Although DistilBERT achieves the highest Pearson and Spearman correlations with the human consensus, all-MiniLM-L12-v2 combines lower MAE and RMSE with the highest ICC and the shortest processing time. The reported processing time also shows that all-MiniLM-L12-v2 is approximately nine times faster than RoBERTa, while RoBERTa requires approximately twice the processing time of MPNet. These differences are relevant when considering deployment scenarios in which grading latency and computational resources are important.

These results indicate that all-MiniLM-L12-v2 provides a favorable balance among grading error, reliability, and computational efficiency under the conditions of this study. The higher correlation values obtained by DistilBERT do not eliminate the importance of error magnitude and processing cost. Similarly, the higher computational cost of RoBERTa may limit its suitability for real-time or resource-constrained deployment settings. Accordingly, all-MiniLM-L12-v2 was selected for the subsequent hybrid scoring and web-based prototype implementation based on the combined evidence from grading error, agreement, reliability, and computational efficiency rather than on a single evaluation criterion.

4.5 Response-Length Bias Analysis

Figure 5 presents the relationship between response length and human-assigned scores. A positive association is observed, although the increasing dispersion at higher word counts

indicates that the relationship is not strictly linear. Longer responses may contain more relevant concepts and supporting details, which can contribute to higher scores; however, response length alone does not guarantee stronger performance. The presence of relatively short responses receiving high scores further indicates that concise answers can perform well when they adequately address the expected concepts.

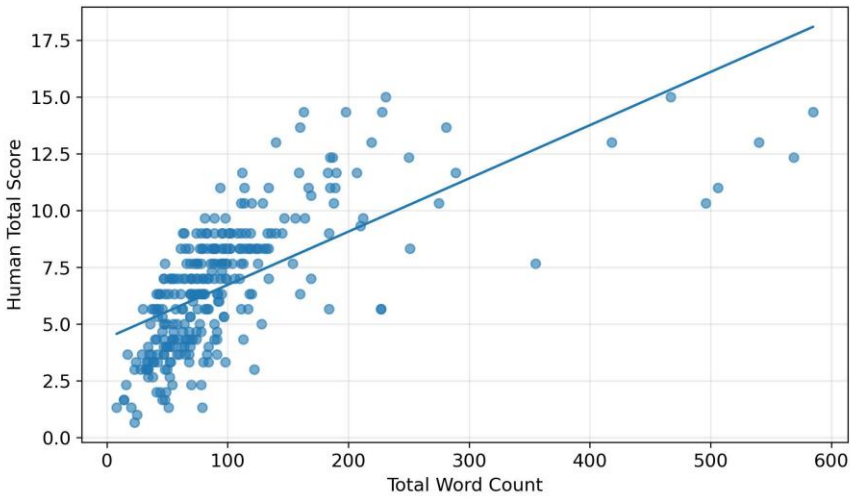

Figure 5. Response length (word count) versus human score.

Figure 6 shows the corresponding relationship between response length and the scores produced by all-MiniLM-L12-v2. The association remains positive but is weaker than that observed for human scores. The wider dispersion of model scores among responses with similar word counts suggests that responses of comparable length can receive different scores, indicating that factors beyond length also contribute to the model output. Nevertheless, the upward trend indicates that response length continues to have some association with automated scores. Thus, the results suggest reduced response-length sensitivity relative to human scoring, rather than complete length invariance or evidence of overall algorithmic fairness.

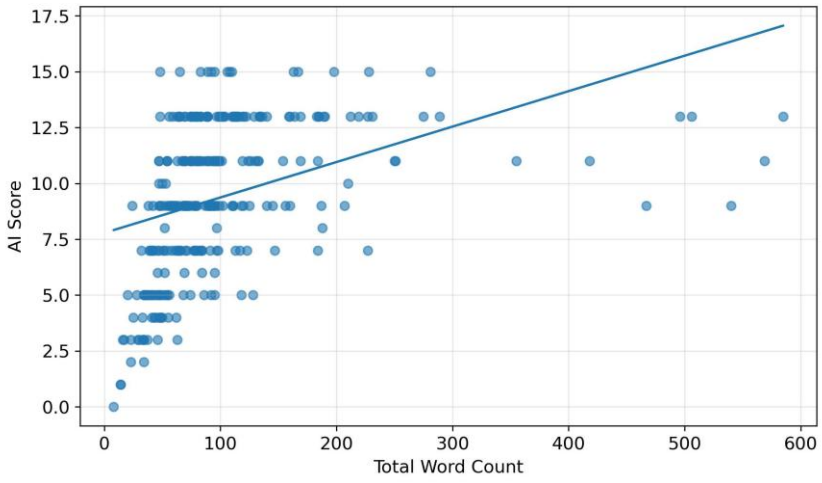

Figure 6. Response length (word count) versus all-MiniLM-L12-v2 score.

Response length was examined because it represents a documented construct-irrelevant feature that can influence grading outcomes [20], [31]. The analysis is intentionally limited to response-length sensitivity and should not be interpreted as a comprehensive fairness evaluation. Other potential sources of bias, including demographic, linguistic, and disciplinary factors, were outside the scope of the present study and require evaluation using more diverse datasets.

The response-length findings should also be interpreted within the study's evaluation setting. The dataset was obtained from a single introductory university course at one

institution and comprised three open-ended questions, which may limit generalizability across disciplines, educational levels, question types, and assessment contexts. In addition, the models were evaluated without task-specific fine-tuning, and their performance may differ following domain adaptation or supervised training. The human consensus score was used as the reference standard despite the inter-rater variability identified in Section 4.1. Finally, the comparative results are specific to the evaluated models, reference answers, scoring thresholds, and experimental conditions. The selection of all-MiniLM-L12-v2 should therefore be interpreted as a study-specific multi-criteria outcome rather than a claim of universal superiority.

4.6. Error Analysis

The error analysis in Table 12 reveals a clear asymmetry in model–human disagreement, with over-scoring occurring substantially more often than severe under-scoring. For all-MiniLM-L12-v2, Q2 produced the largest mean absolute error (MAE = 1.601), followed by Q3 (MAE = 1.417), whereas Q1 showed considerably lower discrepancy (MAE = 0.822). Severe over-scoring was observed in 31 Q2 responses (10.54%) and 22 Q3 responses (7.48%), where the model assigned the maximum score of 5 despite human consensus scores of ≤2. No severe under-scoring cases were identified across the three questions.

The recurring high-discrepancy cases further indicate that the observed errors were not limited to a single model architecture. Cases 152 and 269 were among the five largest discrepancies for all five evaluated models, while cases 251, 198, and 216 recurred across three or four models. This recurrence suggests that certain response–reference relationships pose a common challenge for embedding-based grading. Inspection of these cases indicated that responses could achieve high semantic similarity by capturing the central concept of the reference answer while omitting some of its expected content.

Table 12. Error Analysis and Representative Model–Human Scoring Discrepancies for all-MiniLM-L12-v2

Analysis	Question/ Case	Result	Interpretation
Mean error	Q1	0.253	Lowest mean over-scoring
Mean error	Q2	1.354	Greater over-scoring
Mean error	Q3	1.032	Greater over-scoring
Mean absolute error	Q1	0.822	Lowest discrepancy
Mean absolute error	Q2	1.601	Highest discrepancy
Mean absolute error	Q3	1.417	High discrepancy
Severe over-scoring	Q1	0 (0.00%)	Not observed
Severe over-scoring	Q2	31 (10.54%)	Observed
Severe over-scoring	Q3	22 (7.48%)	Observed
Severe under-scoring	Q1	0 (0.00%)	Not observed
Severe under-scoring	Q2	0 (0.00%)	Not observed
Severe under-scoring	Q3	0 (0.00%)	Not observed
Recurring high-discrepancy case	152	Human = 3.00; AI = 13; error = +10.00; 5/5 models	Strong recurring over-scoring
Recurring high-discrepancy case	269	Human = 3.67; AI = 13; error = +9.33; 5/5 models	Strong recurring over-scoring
Recurring high-discrepancy case	251	Human = 2.33; AI = 11; error = +8.67; 4/5 models	Strong recurring over-scoring
Recurring high-discrepancy case	198	Human = 1.33; AI = 11; error = +9.67; 3/5 models	Strong recurring over-scoring
Recurring high-discrepancy case	216	Human = 3.00; AI = 13; error = +10.00; 3/5 models	Strong recurring over-scoring

Note: A positive error indicates over-scoring relative to the human consensus. Severe over-scoring was defined as an automated score of 5 with a human consensus score ≤2, whereas severe under-scoring was defined as an automated score ≤2 with a human consensus score ≥4. Recurring cases were identified among the five largest total-score absolute discrepancies for each model.

These patterns highlight a limitation of using semantic similarity as the sole grading criterion for open-ended responses. Sentence-BERT can capture broad semantic correspondence, but a high similarity score does not necessarily indicate that all required concepts have been adequately addressed. Consequently, the fixed similarity-to-score thresholds may produce over-scoring when a response conveys the general meaning of the reference answer but lacks sufficient conceptual coverage. This finding provides a direct rationale for integrating rubric-based concept coverage with semantic similarity in the subsequent hybrid scoring mechanism. Table 12 summarizes the magnitude and direction of the observed errors, together with representative cases that exhibited large and recurring model–human discrepancies. The results support the use of concept-level rubric information as a complementary signal and reinforce the role of instructor review for ambiguous or borderline responses.

4.7. Weight Sensitivity and Component Ablation

The weight sensitivity analysis in Table 13 examines how the relative contributions of the semantic and rubric components affect automated grading performance. Across the eight tested configurations, the 25% semantic–75% rubric configuration (25S–75R) produced the lowest MAE (1.8499) and RMSE (2.3980), indicating the lowest observed scoring error. The 40S–60R configuration achieved the highest weighted Kappa (0.4305), but its MAE (1.9280) and RMSE (2.4267) were slightly higher than those of the 25S–75R configuration. Pearson and Spearman correlations remained nearly unchanged across the continuous-score configurations, suggesting that the weighting primarily affected error magnitude and categorical agreement rather than the rank ordering of responses. Based on the lowest observed MAE and RMSE, the 25S–75R configuration was selected for subsequent implementation.

Table 13. Weight Sensitivity Analysis of Semantic–Rubric Weighting.

Configuration	Semantic Weight (%)	Rubric Weight (%)	MAE	RMSE	Pearson	Spearman	Weighted Kappa
0S–100R	0	100	2.2346	2.8882	0.6383	0.6388	0.2816
25S–75R	25	75	1.8499	2.3980	0.6383	0.6388	0.4091
40S–60R	40	60	1.9280	2.4267	0.6383	0.6388	0.4305
50S–50R	50	50	2.1107	2.5859	0.6382	0.6388	0.4189
55S–45R	55	45	2.2331	2.7013	0.6382	0.6388	0.4074
60S–40R	60	40	2.3675	2.8370	0.6382	0.6388	0.3768
75S–25R	75	25	2.8719	3.3392	0.6382	0.6388	0.3334
100S–0R	100	0	3.8912	4.3750	0.6382	0.6388	0.2505

Note: S = semantic similarity component; R = adjusted rubric component. Scores are continuous total scores obtained by summing Q1–Q3. The 25S–75R configuration was selected based on the lowest observed continuous-score MAE and RMSE. The highest weighted Kappa was obtained with the 40S–60R configuration.

The component ablation results in Table 14 further clarify the contribution of the two scoring components. Semantic-only scoring produced the largest errors, with an MAE of 3.8912 and an RMSE of 4.3750. Incorporating the adjusted rubric component alone reduced these values to 2.2346 and 2.8882, respectively. The hybrid configuration using the selected 25S–75R weighting reduced the errors further, yielding an MAE of 1.8499 and an RMSE of 2.3980, while also producing the highest weighted Kappa among the three ablation conditions (0.4091). Relative to semantic-only scoring, the hybrid approach reduced MAE by 52.46% and RMSE by 45.19%. The near-identical Pearson and Spearman correlations across the three conditions indicate that the main improvement was associated with reduced scoring error and stronger agreement rather than a substantial change in response ranking.

At the question level, Table 15 shows that the selected 25S–75R configuration maintained relatively similar error levels across Q1–Q3. MAE ranged from 0.7494 to 0.7888, while RMSE ranged from 0.9624 to 1.0096. Q2 produced the highest MAE and RMSE for the selected configuration, consistent with the greater scoring variability observed for Q2 in the preceding analyses. The 40S–60R configuration produced slightly lower error for Q1 and a higher weighted Kappa overall, whereas 25S–75R maintained the lowest overall MAE and RMSE. These question-level results indicate that the observed performance of the selected

configuration was distributed across the three questions rather than being driven exclusively by a single item.

Table 14. Component Ablation Analysis of the Automated Scoring Framework

Scoring Condition	MAE	RMSE	Pearson	Spearman	Weighted Kappa
Semantic-only	3.8912	4.3750	0.6382	0.6388	0.2505
Adjusted-rubric-only	2.2346	2.8882	0.6383	0.6388	0.2816
Hybrid (25% Semantic + 75% Rubric)	**1.8499**	**2.3980**	**0.6383**	**0.6388**	**0.4091**

Note: The adjusted-rubric component includes the rubric-score adjustment used in the proposed framework. Evaluation is based on the summed human-consensus score across Q1–Q3.

Table 15. Question-Level Performance of Selected Weighting Configurations

Configuration	Question	MAE	RMSE	Pearson	Spearman	Weighted Kappa
25S–75R	Q1	0.7494	0.9624	0.6692	0.7033	0.2974
25S–75R	Q2	0.7888	1.0096	0.5225	0.5178	0.3552
25S–75R	Q3	0.7766	0.9680	0.5314	0.5241	0.3759
25S–75R	Total	1.8499	2.3980	0.6383	0.6388	0.4091
40S–60R	Q1	0.7317	0.9322	0.6692	0.7033	0.4996
40S–60R	Q2	0.8460	1.0634	0.5224	0.5178	0.3223
40S–60R	Q3	0.7704	0.9724	0.5314	0.5241	0.3485
40S–60R	Total	1.9280	2.4267	0.6383	0.6388	0.4305
50S–50R	Q1	0.7465	0.9471	0.6692	0.7033	0.4739
50S–50R	Q2	0.9276	1.1376	0.5223	0.5178	0.2800
50S–50R	Q3	0.8153	1.0181	0.5314	0.5241	0.3486
50S–50R	Total	2.1107	2.5859	0.6382	0.6388	0.4189

Overall, the sensitivity and ablation analyses provide complementary evidence for the role of the hybrid scoring design. The sensitivity analysis identifies the weighting configuration associated with the lowest observed error, while the ablation analysis shows that combining semantic similarity with rubric-based concept coverage reduces grading error relative to either component used independently. The results therefore support the use of a hybrid scoring mechanism in the subsequent prototype evaluation, with 25S–75R adopted as the operating configuration under the conditions of this study.

4.8 Web-Based Prototype

The web-based prototype shown in Figure 7 demonstrates the practical implementation of the selected Sentence-BERT model and the proposed automated assessment framework. The system provides automated scores and interpretable feedback while allowing instructors to review the generated assessment criteria. The prototype is intended to demonstrate the operational feasibility of the framework rather than serve as a separate experimental evaluation.

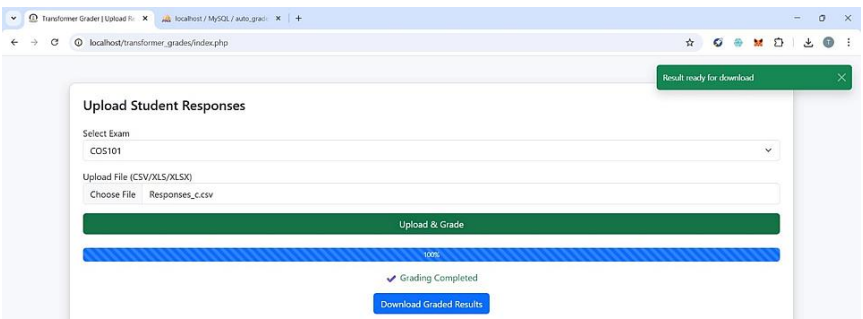

Figure 7. Web-based prototype demonstrating the implementation of the proposed automated assessment framework.

The findings indicate that Sentence-BERT-based grading can function as an assistive assessment technology rather than as a replacement for human judgment, particularly in high-stakes assessment contexts. The model comparison showed that selection should consider grading error, reliability, correlation, agreement, and computational efficiency jointly. Under the conditions evaluated in this study, all-MiniLM-L12-v2 provided a favorable balance across these criteria, while the 25S–75R hybrid configuration reduced scoring error by combining holistic semantic similarity with rubric-based concept coverage. The prototype translates these components into an operational workflow that retains instructor review for cases requiring additional judgment.

Several limitations should be considered when interpreting these findings. First, model and scoring-configuration selection and comparative evaluation were performed on the same study dataset; therefore, the reported results should not be interpreted as an independent external validation estimate. Validation using independent datasets from other institutions and assessment contexts is necessary to assess generalizability. Second, the cosine-similarity thresholds used for automated score mapping were specified a priori as rubric-informed operational cutoffs but were not empirically optimized or subjected to a dedicated threshold-sensitivity analysis. Third, the response-length analysis examined length-related associations but did not constitute a comprehensive fairness assessment across demographic, linguistic, or other protected characteristics. Future studies should therefore incorporate independent external validation, threshold calibration, and broader fairness evaluation across diverse educational settings and learner populations.

5. Conclusions

This study developed and evaluated a Sentence-BERT-based framework for automated assessment of open-ended short-answer responses. The results demonstrated measurable variability among human raters and statistically detectable differences among the evaluated models, while also showing broadly consistent grading patterns across models. Among the evaluated alternatives, all-MiniLM-L12-v2 provided a favorable balance of grading error, reliability, and computational efficiency under the conditions of this study.

The response-length analysis showed a positive association between response length and automated scores, but the association was weaker than that observed for human scores. Thus, the findings suggest reduced response-length sensitivity rather than complete length invariance or comprehensive algorithmic fairness. The weight sensitivity analysis identified the 25% semantic–75% rubric configuration as having the lowest observed continuous-score MAE and RMSE, while the ablation results demonstrated the complementary contribution of holistic semantic similarity and rubric-based concept coverage.

The selected model and scoring configuration were subsequently implemented in a web-based prototype that provides automated scoring and interpretable feedback while retaining instructor review. Taken together, the findings support positioning the framework as an assistive grading technology rather than a replacement for human judgment, particularly in high-stakes assessment. Future research should evaluate the framework on independent datasets and across broader educational contexts, investigate threshold calibration and response-length normalization strategies, and examine instructor usability and additional sources of potential bias.

Author Contributions: Conceptualization: T.O.O., F.H.O. and Y.A.I.; Methodology: T.O.O., A.T.E. and Y.A.I.; Software: T.O.O.; Validation: T.O.O. and Y.A.I.; Formal analysis: T.O.O. and A.T.E.; Investigation: T.O.O.; Resources: T.O.O. and Y.A.I.; Data curation: T.O.O., F.H.O. and A.T.E.; Writing—original draft preparation: T.O.O., F.H.O., A.T.E. and Y.A.I.; Writing—review and editing: T.O.O., F.H.O., A.T.E. and Y.A.I.; Visualization: T.O.O. and Y.A.I.; Supervision: T.O.O.; Project administration: F.H.O.; Funding acquisition: T.O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tertiary Education Trust Fund (TETFUND) 2025 intervention, grant number 2025/FX302-IBR.

Data Availability Statement: The dataset used in this study is available at https://doi.org/10.6084/m9.figshare.32109718.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] L. Rodrigues, C. Xavier, N. Costa, D. Gasevic, and R. F. Mello, "Is GPT-4 fair? An empirical analysis in automatic short answer grading," *Comput. Educ. Artif. Intell.*, vol. 8, no. June, p. 100428, Jun. 2025, doi: 10.1016/j.caeai.2025.100428.

[2] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions," *Interact. Learn. Environ.*, vol. 30, no. 1, pp. 177–190, Jan. 2022, doi: 10.1080/10494820.2019.1648300.

[3] F. Morley, E. Walland, and C. V. Rodeiro, "Auto-marking short answer questions in science: The foundational years of transformer-based models from BERT to GPT-4," *Int. J. Artif. Intell. Educ.*, vol. 36, no. 1–2, p. 100005, Mar. 2026, doi: 10.1016/j.ijaied.2026.100005.

[4] S. A. Mahmood and M. A. Abdulsamad, "Automatic assessment of short answer questions: Review," *Edelweiss Appl. Sci. Technol.*, vol. 8, no. 6, pp. 9158–9176, Dec. 2024, doi: 10.55214/25768484.v8i6.3956.

[5] F. Rodrigues and P. Oliveira, "A system for formative assessment and monitoring of students' progress," *Comput. Educ.*, vol. 76, pp. 30–41, Jul. 2014, doi: 10.1016/j.compedu.2014.03.001.

[6] S. Tobler, "Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments," *MethodsX*, vol. 12, p. 102531, Jun. 2024, doi: 10.1016/j.mex.2023.102531.

[7] A. Kashi, S. Shastri, A. R. Deshpande, J. Doreswamy, and G. Srinivasa, "A Score Recommendation System Towards Automating Assessment In Professional Courses," in *2016 IEEE Eighth International Conference on Technology for Education (T4E)*, Dec. 2016, pp. 140–143. doi: 10.1109/T4E.2016.036.

[8] R. Weegar and P. Idestam-Almquist, "Reducing Workload in Short Answer Grading Using Machine Learning," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 247–273, Jun. 2024, doi: 10.1007/s40593-022-00322-1.

[9] J. Pecuchova, Ľ. Benko, and M. Drlik, "Automated Grading of Open-Ended Questions in Higher Education Using GenAI Models," *Int. J. Artif. Intell. Educ.*, vol. 35, no. 6, pp. 3813–3846, Dec. 2025, doi: 10.1007/s40593-025-00517-2.

[10] E. del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation," *Knowl. Inf. Syst.*, vol. 65, no. 10, pp. 4295–4334, Oct. 2023, doi: 10.1007/s10115-023-01892-9.

[11] T. O. Omotehinwa, R. A. Ezekiel, M. Onoja, E. C. Omeye, and Y. A. Ibrahim, "Comparative Evaluation of Transformer-Based Models for Automated Short-Answer Grading," *NIPES J. Sci. Technol. Res.*, vol. 7, no. 1, pp. 843–849, Oct. 2025, doi: 10.37933/nipes/7.4.2025.SI97.

[12] S. Bonthu, S. R. Sree, and M. H. M. Krishna Prasad, "Framework for automation of short answer grading based on domain-specific pre-training," *Eng. Appl. Artif. Intell.*, vol. 137, p. 109163, Nov. 2024, doi: 10.1016/j.engappai.2024.109163.

[13] J. Garg, J. Papreja, K. Apurva, and G. Jain, "Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Jun. 2022, pp. 1–6. doi: 10.1109/CONIT55038.2022.9847754.

[14] A. Amalia, M. S. Lydia, M. A. Muchtar, F. Y. Manik, Sinu, and D. Gunawan, "Mitigating Bias and Assessment Inconsistencies with BERT-Based Automated Short Answer Grading for the Indonesian Language," *IAENG Int. J. Comput. Sci.*, vol. 52, no. 3, pp. 533–545, 2025.

[15] S. Konade, Y. Hirgude, A. Kulkarni, A. Jadhav, and L. Sonar, "Implementation of an Automated Answer Evaluation System," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Feb. 2024, pp. 457–462. doi: 10.1109/IC2PCT60090.2024.10486693.

[16] O. Bolgova, P. Ganguly, M. F. Ikram, and V. Mavrych, "Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders," *Med. Educ. Online*, vol. 30, no. 1, Dec. 2025, doi: 10.1080/10872981.2025.2550751.

[17] G. Kuling *et al.*, "Assessment of Short-Answer Questions by ChatGPT in a Medical School Course," *NEJM AI*, vol. 3, no. 2, Jan. 2026, doi: 10.1056/AIcs2500239.

[18] M. Kaya and I. Cicekli, "A Hybrid Approach for Automated Short Answer Grading," *IEEE Access*, vol. 12, no. June, pp. 96332–96341, 2024, doi: 10.1109/ACCESS.2024.3420890.

[19] J. Osaka, A. Maeda, H. Oka, Y. Mori, T. Ishioka, and H. Suyari, "Reliable and efficient automated short-answer scoring for a large dataset using active learning and deep learning," *Interact. Learn. Environ.*, vol. 33, no. 6, pp. 3776–3787, Jul. 2025, doi: 10.1080/10494820.2025.2452005.

[20] F. E. A. Hassanein, R. R. Hussein, Y. Ahmed, J. El-Guindy, D. E. Ahmed, and A. Abou-Bakr, "Calibration of AI large language models with human subject matter experts for grading of clinical short-answer responses in dental education," *BMC Oral Health*, vol. 26, no. 1, p. 286, Feb. 2026, doi: 10.1186/s12903-026-07665-4.

[21] C. Anghel, E. Pecheanu, A. A. Anghel, M. V. Craciun, and A. Cocu, "ExamQ-Gen: Instructor-in-the-Loop Generation of Self-Contained Exam Questions from Course Materials and Decision-Support Grading," *Computers*, vol. 15, no. 3, p. 177, Mar. 2026, doi: 10.3390/computers15030177.

[22] H. M. T. W. Seneviratne and S. S. Manathunga, "Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings," *BMC Med. Educ.*, vol. 25, no. 1, p. 1146, Aug. 2025, doi: 10.1186/s12909-025-07718-2.

[23] Y. Oh Lee, B. Bang, J. Lee, and S. Oh, "Personalized Auto-Grading and Feedback System for Constructive Geometry Tasks Using Large Language Models on an Online Math Platform," *IEEE Access*, vol. 14, no. December 2025, pp. 17788–17802, 2026, doi: 10.1109/ACCESS.2026.3657726.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.
[27] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.
[28] D. Ortiz Martes, E. Gunderson, C. Neuman, and N. N. Kachouie, "Transformer Models for Paraphrase Detection: A Comprehensive Semantic Similarity Study," *Computers*, vol. 14, no. 9, p. 385, Sep. 2025, doi: 10.3390/computers14090385.
[29] H. Gusdevi, A. Setyanto, K. Kusrini, and E. Utami, "Cosine Similarity-Based Evidences Selection for Fact Verification Using SBERT on the FEVER Dataset," *CogITo Smart J.*, vol. 11, no. 1, pp. 52–66, Jun. 2025, doi: 10.31154/cogito.v11i1.917.52-66.
[30] A. Kumar and H. Kumar, "scaLAR SemEval-2024 Task 1: Semantic Textual Relatednes for English," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 2024, pp. 902–906. doi: 10.18653/v1/2024.semeval-1.129.
[31] M. Acar Güvendir, A. F. Kılıç, E. Güvendir, and T. Kaçak, "Bridging minds and machines: a comparative study of AI and human rater agreement and reliability in educational assessment," *Educ. Inf. Technol.*, vol. 31, no. 11, pp. 3781–3804, Jul. 2026, doi: 10.1007/s10639-026-13949-7.