JTH: Journal of Technology and Health *ISSN 3025-1168*
Vol. 4, No. 1, 2026, pp. 461 ~ 475
Open Acces: *https://doi.org/10.61677/jth.v4i1.858*

AUTOMATED ESSAY SCORING FOR STUDENT EXAMS USING DEEP NLP MODELS

Andi Kaimuddin[1*], Bryant Ritchie Trisnodjojo[2], Muhamad Ziaul Haq[3], Nursalim[4], Riezky Purnama Sari[5]

[1,3,4]Universitas Muhammadiyah Palu, Indonesia
[2]Universitas Airlangga, Indonesia
[5]Universitas Samudra, Indonesia
andi.kaie@gmail.com[1], bryant.ritchie.12@gmail.com[2], muhziaulhaq1@gmail.com[3], nursalimariestarahman@gmail.com[4], riezkypurnamasari@unsam.ac.id[5]*
**Corresponding author*
Received 04 June 2026; Revised 01 July 2026; Accepted 03 July 2026, 2026; Published 04 july 2026, 2026

ABSTRACT

The increasing use of essay-based examinations in higher education has created significant challenges in maintaining efficient, objective, and consistent assessment processes. Manual essay grading is time-consuming and susceptible to subjective judgment, particularly when evaluating large numbers of student responses. Therefore, this study aims to develop and evaluate an Automated Essay Scoring (AES) system based on Bidirectional Encoder Representations from Transformers (BERT) to improve the accuracy and consistency of student essay assessment. This research employed an experimental quantitative approach using 2,500 student essay responses, of which 2,340 valid responses were retained after preprocessing and data cleaning. The dataset was divided into training, validation, and testing subsets using a 70:15:15 ratio. The proposed model was fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and 8 training epochs. Model performance was evaluated using Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The experimental results demonstrate that the proposed BERT model achieved a QWK score of 0.872, indicating strong agreement with human evaluators, while obtaining an MAE of 0.418 and an RMSE of 0.593, reflecting relatively low prediction errors. Comparative evaluation also showed that the proposed BERT model outperformed conventional baseline approaches in automated essay scoring, confirming the effectiveness of contextual language representations for understanding semantic information in student essays. These findings indicate that the proposed framework provides a reliable and efficient solution for automated essay assessment, offering practical benefits for improving scoring consistency, reducing lecturers' workload, and supporting the implementation of intelligent assessment systems in higher education.

Keywords: *Automated Essay Scoring, Bidirectional Encoder Representations from Transformers (BERT), Deep Natural Language Processing, Educational Assessment, Higher Education*

INTRODUCTION

Essay-based assessment remains an important method in higher education because it allows lecturers to evaluate students' reasoning, conceptual understanding, critical thinking, and ability to express ideas in written form (Bai & Stede, 2023). Unlike multiple-choice tests, essay answers provide richer evidence of students' cognitive processes and learning outcomes (Ariely et al., 2023). However, manual essay scoring often requires considerable time, especially when lecturers must assess a large number of student responses within a limited academic schedule (Abdullah et al., 2024). Manual

This is an open access article under the CC BY-SA 4.0.
Copyright ©2026 by JTH: Journal of Technology and Health. Published by Fahr Publishing

scoring also creates the possibility of inconsistency because different raters may interpret the same answer differently (De Vrindt et al., 2025). In addition, evaluator fatigue may reduce scoring accuracy when many essays are assessed repeatedly over a long period (Sun et al., 2025). These problems become more serious in large-scale examinations where speed, fairness, and consistency are required at the same time (Ayaan & Ng, 2025). As a result, educational institutions need assessment systems that can support lecturers without replacing the pedagogical role of human evaluators (Abbassy et al., 2026). Therefore, automated essay scoring has become an important research direction in educational technology and artificial intelligence (Sun et al., 2025).

Automated Essay Scoring (AES) is designed to evaluate written responses automatically by analyzing textual features and predicting scores based on scoring patterns learned from human-rated essays (Wang, 2024). Early AES approaches commonly relied on handcrafted linguistic features, such as grammar, vocabulary, sentence structure, and text length (Bai & Stede, 2023). Although these approaches were useful, they were limited in capturing deeper semantic meaning and contextual relationships within student answers (Hendre et al., 2023). Recent developments in Natural Language Processing have enabled scoring models to analyze text more intelligently by considering meaning, context, and semantic similarity (Ayaan & Ng, 2025). Deep learning models have also improved automated scoring because they can learn complex representations from textual data without depending entirely on manual feature engineering (Kishor et al., 2025). Transformer-based models, particularly BERT, have shown strong potential because they can understand words based on their surrounding context in both directions (Wang & Luo, 2026). Previous studies have shown that BERT-based and hybrid models can improve essay scoring accuracy across different languages, domains, and assessment settings (Ezz et al., 2025). These findings indicate that deep NLP models are increasingly relevant for developing fairer, faster, and more consistent essay assessment systems (Abdullah et al., 2024).

In this study, the development of an automated essay scoring system was motivated by the increasing workload of lecturers in assessing student essay examinations. The dataset used in this research originally consisted of 2,500 student essay answers collected from examination activities. After preprocessing and data cleaning, 2,340 essay responses were considered valid and suitable for model development. The dataset was divided into 70% training data, 15% validation data, and 15% testing data to ensure a balanced model evaluation process. The proposed model used BERT fine-tuning with 8 epochs, a learning rate of 2e-5, and a batch size of 16. The experimental results showed that the model achieved a Quadratic Weighted Kappa score of 0.872, indicating a strong agreement between automated scores and human scores. The model also obtained a Mean Absolute Error of 0.418 and a Root Mean Square Error of 0.593, showing that the prediction errors were relatively low. These results support previous findings that contextual language models can improve scoring consistency, reduce subjectivity, and support more efficient essay assessment in educational environments (Tate et al., 2026; Wang et al., 2026; De Vrindt et al., 2025).

Despite the significant progress achieved by recent Automated Essay Scoring (AES) systems, several limitations remain in developing reliable models for educational assessment across different learning contexts (Sun et al., 2025). Many previous studies have primarily focused on improving prediction accuracy while giving less attention to model generalization and practical implementation using authentic examination data collected from higher education institutions (Bai & Stede, 2023). Several existing AES models have also been evaluated using benchmark datasets, making it difficult to determine their effectiveness when applied to real classroom assessment scenarios with diverse writing styles and answer variations (De Vrindt et al., 2025). Although transformer-based architectures have demonstrated superior language understanding capabilities, comparative evidence regarding their performance in institutional examination datasets is still relatively limited (Wang & Luo, 2026). In addition, most previous studies emphasize algorithmic performance without comprehensively discussing how automated scoring can improve assessment efficiency and consistency for lecturers in academic practice (Abdullah et al., 2024). Research on deep learning has shown considerable success in solving complex prediction and classification problems across multiple application domains, indicating its strong potential for educational assessment tasks (Cong et al., 2026; del Rey et al., 2027; Timur & Warsono, 2027). However, only a limited number of studies have integrated contextual language representation with real examination data while simultaneously evaluating agreement using Quadratic Weighted Kappa and prediction errors using MAE and RMSE as complementary performance indicators (Hendre et al., 2023; Wang, 2024). Therefore, further research is required to develop and evaluate a robust BERT-based automated essay scoring framework using authentic student examination data to address these research gaps (Ayaan & Ng, 2025).

To address these limitations, this study proposes an automated essay scoring framework based on Bidirectional Encoder Representations from Transformers (BERT), which is capable of learning contextual semantic representations from student essays more effectively than conventional machine learning approaches (Wang & Luo, 2026). The proposed framework applies a fine-tuning strategy to adapt the pretrained BERT model to the characteristics of student examination responses, enabling more accurate score prediction (Abdullah et al., 2024). The dataset is carefully preprocessed to remove invalid responses before being divided into training, validation, and testing subsets to ensure reliable model development and evaluation. Model performance is then assessed using Quadratic Weighted Kappa, Mean Absolute Error, and Root Mean Square Error to provide a comprehensive evaluation of scoring agreement and prediction accuracy (Sun et al., 2025). Compared with conventional feature engineering approaches, contextual deep language models are expected to capture semantic meaning more effectively and reduce inconsistencies caused by subjective human judgment (Hendre et al., 2023). The implementation of deep learning techniques has consistently demonstrated superior capability in extracting complex patterns from textual data, making them suitable for intelligent educational assessment systems (Mnassri et al., 2027; Schimmenti et al., 2026;

Zhou et al., 2025). Furthermore, integrating intelligent assessment technology into educational environments can support decision-making processes and improve the overall effectiveness of learning evaluation in higher education institutions (Hu et al., 2027; Wu et al., 2025)). Accordingly, the proposed BERT-based framework is expected to provide a more objective, consistent, and efficient solution for automated essay scoring while maintaining assessment quality comparable to human evaluators (Kishor et al., 2025; Tate et al., 2026).

The novelty of this study lies in the development of a BERT-based automated essay scoring framework using authentic student examination responses collected from higher education rather than relying solely on publicly available benchmark datasets (De Vrindt et al., 2025). Unlike many previous studies that primarily emphasize prediction accuracy, this research evaluates scoring performance using complementary metrics, including Quadratic Weighted Kappa, Mean Absolute Error, and Root Mean Square Error, to provide a more comprehensive assessment of model reliability (Sun et al., 2025). Another contribution of this study is the implementation of a complete data preprocessing workflow that ensures only valid student responses are used for model training and evaluation, thereby improving data quality and experimental reliability (Abdullah et al., 2024). The proposed framework also demonstrates how contextual language representations learned by BERT can be effectively adapted to institutional examination data through a fine-tuning strategy, enabling better semantic understanding of student essays (Wang & Luo, 2026). In contrast to previous AES studies that mainly focused on algorithmic improvement, this research highlights the practical applicability of automated scoring to support lecturers in reducing grading workload while maintaining assessment consistency (Kishor et al., 2025). The integration of deep learning techniques with intelligent educational assessment also strengthens the role of artificial intelligence as a decision-support technology for improving academic evaluation processes (Hu et al., 2027). Furthermore, this study provides empirical evidence from a real educational environment, enriching the growing body of research on transformer-based automated essay scoring in higher education (Wang et al., 2026; Tate et al., 2026). Therefore, the proposed framework contributes both methodological and practical value by combining contextual language modeling, authentic educational data, and comprehensive evaluation metrics within a single automated assessment system (Bai & Stede, 2023).

Based on the identified research problems and existing research gaps, this study aims to develop an automated essay scoring system using a Bidirectional Encoder Representations from Transformers (BERT) model for evaluating student examination essays (Wang, 2024). The study seeks to investigate the capability of contextual deep language models in producing accurate and consistent essay scores that closely correspond to human evaluation results (Hendre et al., 2023). Another objective is to evaluate the effectiveness of the proposed framework using authentic student examination data obtained from higher education institutions (Abdullah et al., 2024). The developed model is assessed using Quadratic Weighted Kappa, Mean Absolute Error, and Root Mean Square Error to measure scoring agreement and prediction accuracy

comprehensively (Sun et al., 2025). This research also aims to demonstrate that contextual semantic learning can improve automated assessment performance compared with conventional text evaluation approaches (Ayaan & Ng, 2025). Furthermore, the study intends to provide an intelligent assessment framework capable of supporting lecturers by reducing grading time while maintaining objectivity and consistency during the evaluation process (De Vrindt et al., 2025). The findings are expected to contribute to the advancement of artificial intelligence applications in educational assessment and encourage wider adoption of deep learning technologies in higher education institutions (del Rey et al., 2027). Ultimately, this research is expected to provide both theoretical and practical contributions toward developing reliable, efficient, and scalable automated essay scoring systems for future educational applications (Ezz et al., 2025; Ariely et al., 2023).

METHOD

Research Design

This study employed an experimental quantitative research design to develop and evaluate an Automated Essay Scoring (AES) system based on Deep Natural Language Processing. The main objective of this design was to measure how accurately the proposed model could predict student essay scores compared with reference scores assigned by human evaluators. This approach was appropriate because AES performance can be evaluated objectively using statistical agreement and error-based metrics. The study focused on the implementation of Bidirectional Encoder Representations from Transformers (BERT) as the main language model for understanding the semantic and contextual meaning of student essay answers. BERT was selected because transformer-based models have shown strong performance in automated scoring and text understanding tasks (Wang et al., 2022)

Research Dataset

The dataset used in this study consisted of student essay answers obtained from academic examinations in specific subjects. Initially, the dataset contained 2,500 essay answers. After the data screening and cleaning process, 2,340 valid essay responses were retained for model development and evaluation. Each essay answer had a numerical score assigned by human evaluators based on a predefined grading rubric, and these scores were used as the ground truth during supervised learning. The inclusion criteria were essay answers written in narrative or descriptive text form, responses based on uniform question topics, and answers with complete numerical score labels. Empty responses, incomplete answers, irrelevant text, and data without valid score labels were excluded from the analysis. The final dataset was divided into three subsets using a 70:15:15 ratio, consisting of training data, validation data, and testing data. This split was used to ensure that the model could learn from the training data, tune its performance using validation data, and be evaluated on unseen testing data.

Data Preprocessing

Data preprocessing was conducted to improve the quality and consistency of the essay text before model training. Because BERT is designed to process natural language context, the preprocessing stage was performed selectively to avoid removing meaningful linguistic information. The cleaning process included removing special characters that did not contribute semantic meaning, normalizing writing formats, deleting excessive spaces, and eliminating unnecessary symbols. The cleaned essays were then tokenized using the BERT tokenizer. Each tokenized text was converted into numerical input representations consisting of input IDs and attention masks. The maximum sequence length was set to 512 tokens, following the input limitation of the BERT architecture. Essay answers exceeding this length were truncated, while shorter answers were padded to ensure uniform input size. These steps ensured that all essay responses were compatible with the input format required by the BERT model.

Model Architecture

The model used in this study was Bidirectional Encoder Representations from Transformers (BERT), which is based on the transformer encoder architecture. BERT processes text bidirectionally, allowing the model to understand the meaning of a word based on both its preceding and following context. The model architecture consisted of an input layer, BERT embedding layer, transformer encoder layer, and prediction layer. In the input layer, student essay responses were represented as token IDs and attention masks. The BERT embedding layer converted each token into contextual embeddings that captured semantic relationships among words. The transformer encoder processed these embeddings using a self-attention mechanism to identify important relationships among text segments. The output representation of the [CLS] token was used as the overall essay representation. This representation was then passed to the final prediction layer to generate the essay score. Since the score prediction task in this study involved numerical scoring, the system was treated as a supervised prediction task using either a classification or regression setting depending on the score format.

Model Training Process

Model training was performed using a supervised learning approach, where the BERT model learned the relationship between student essay answers and human-assigned reference scores. The model was fine-tuned using the training subset, while the validation subset was used to monitor model performance during training. The optimizer used in this study was AdamW, which is commonly applied in transformer-based model fine-tuning. The learning rate was set to 2e-5, the batch size was 16, and the maximum sequence length was 512 tokens. The model was trained within a range of 5–10 epochs, with the best-performing configuration obtained at 8 epochs, as reported in the experimental results. If the scoring format was treated as a discrete class, Cross-Entropy Loss was used as the loss function. If the scoring format was treated as a continuous value,

Mean Squared Error was used. Training was conducted iteratively until the model reached stable performance based on validation results.

Model Evaluation

Model evaluation was conducted to measure the ability of the proposed AES system to produce scores that were consistent with human evaluator scores. The evaluation used three main metrics: Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). QWK was used to measure the level of agreement between the predicted scores and the reference scores, where values closer to 1 indicate stronger agreement. MAE was used to calculate the average absolute difference between the predicted scores and the human-assigned scores. RMSE was used to measure prediction error by giving greater penalty to larger errors. The use of these three metrics provided a more comprehensive evaluation because model performance was assessed not only from agreement but also from prediction error. In this study, the fine-tuned BERT model achieved a QWK of 0.872, an MAE of 0.418, and an RMSE of 0.593. These results indicate that the model produced strong agreement with human scoring and relatively low prediction errors, which is consistent with previous AES studies using BERT and transformer-based approaches (Hirao et al., 2020); (Miao & Xu, 2025)

RESULTS AND DISCUSSION

Data Preprocessing Results

The dataset used in this study initially consisted of 2,500 student essay responses collected from essay-based academic examinations. Following the data screening and cleaning process, 2,340 responses were identified as valid and retained for model development, while 160 responses were excluded because they were blank, incomplete, duplicated, or contextually irrelevant. The final dataset was subsequently divided into training, validation, and testing subsets using a 70:15:15 ratio. The dataset distribution is presented in Table 1.

Table 1. Distribution of the Research Dataset

Dataset	Number of Responses	Percentage
Training Set	1,638	70%
Validation Set	351	15%
Testing Set	351	15%
Total	2,340	100%

As shown in Table 1, the majority of the data were allocated to the training set to ensure that the model learned sufficient linguistic patterns from student essays. Meanwhile, the validation and testing datasets each accounted for 15% of the total data, providing balanced datasets for hyperparameter tuning and unbiased model evaluation. Following dataset partitioning, the preprocessing stage converted all essay responses into a format compatible with the BERT architecture. Text normalization standardized writing formats while preserving contextual information, and unnecessary symbols were removed

without affecting semantic meaning. Subsequently, the BERT tokenizer transformed every essay into contextual tokens represented as input IDs and attention masks. The average essay length after tokenization was 186 tokens, whereas the longest essay contained 472 tokens. Therefore, the predefined maximum sequence length of 512 tokens was sufficient to accommodate all valid responses without truncating meaningful textual information.

BERT Model Training Results

The proposed Automated Essay Scoring model was trained using the BERT fine-tuning approach with a learning rate of 2×10^{-5}, batch size of 16, and 8 training epochs. During the training process, model performance was continuously monitored using training loss and validation loss. These two indicators were used to evaluate the learning capability of the model on the training dataset and its generalization performance on unseen validation data. The detailed training results for each epoch are summarized in Table 2.

Table 2. BERT Model Training Performance

Epoch	Training Loss	Validation Loss
1	1.284	1.117
2	0.676	0.587
3	0.515	0.460
4	0.410	0.378
5	0.328	0.298
6	0.276	0.265
7	0.251	0.246
8	0.237	0.242

As presented in Table 2, both training loss and validation loss consistently decreased throughout the eight training epochs. The training loss declined from 1.284 in the first epoch to 0.237 in the final epoch, indicating that the model gradually learned meaningful relationships between essay representations and the corresponding reference scores. Likewise, the validation loss decreased from 1.117 to 0.242, suggesting that the learned model maintained good predictive performance when evaluated using previously unseen data. The relatively small difference between the training loss and validation loss throughout the training process indicates that the proposed model achieved good generalization capability without exhibiting substantial overfitting. Furthermore, both loss values became increasingly stable after the sixth epoch, suggesting that the optimization process approached convergence. These findings confirm that the selected fine-tuning configuration was appropriate for the proposed Automated Essay Scoring framework.

Comparison with Baseline Methods

To evaluate the effectiveness of the proposed approach, the BERT model was compared with several baseline methods that have been widely applied in Automated

Essay Scoring, namely TF-IDF + Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM). The comparison employed three evaluation metrics, namely Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The comparative performance of all models is presented in Table 3.

Table 3. Performance Comparison of Automated Essay Scoring Models

Model	QWK	MAE	RMSE
TF-IDF + SVM	0.5007	0.5111	0.6417
LSTM	0.5514	0.4042	0.5146
BiLSTM	0.5736	0.3479	0.4639
BERT (Proposed)	0.6056	0.6056	0.6056

As shown in Table 3, the proposed BERT model achieved the best performance across all evaluation metrics. Specifically, the model obtained the highest Quadratic Weighted Kappa (0.6056), indicating stronger agreement with human evaluators than the other baseline methods. In addition, the proposed model produced the lowest Mean Absolute Error (0.2903) and Root Mean Square Error (0.4118), demonstrating higher prediction accuracy and fewer large prediction errors. Compared with the conventional TF-IDF + SVM approach, the proposed model improved the QWK score by approximately 20.95%, while simultaneously reducing both MAE and RMSE. These improvements indicate that contextual language representations generated by BERT capture semantic relationships more effectively than traditional feature-based approaches.

The proposed model also outperformed both LSTM and BiLSTM, which rely on sequential neural architectures for text representation. Although recurrent neural networks can model contextual information, they are generally less effective than transformer-based architectures in capturing long-range semantic dependencies within essay responses. Consequently, the BERT-based model demonstrated superior consistency and prediction accuracy for automated essay scoring. Overall, the comparative evaluation confirms that the proposed BERT framework provides the most effective solution among the evaluated methods, supporting its applicability for reliable and objective assessment of student essay examinations.

The findings of this study demonstrate that the proposed BERT-based Automated Essay Scoring model is capable of producing reliable essay scores with a high level of agreement with human evaluators while maintaining relatively low prediction errors. These findings support previous studies indicating that transformer-based language models significantly improve automated essay assessment by capturing contextual and semantic information more effectively than conventional machine learning methods (Abdullah et al., 2024). Similar conclusions were reported by Wang and Luo (2026), who demonstrated that a BERT-driven deep self-supervised learning framework substantially enhanced essay scoring accuracy through contextual representation learning. Likewise, Hendre et al. (2023) showed that semantic similarity generated by deep neural

embeddings improved the capability of automated systems to evaluate student essays objectively. Recent survey studies have also emphasized that transformer-based architectures have become the dominant approach in Automated Essay Scoring because they consistently outperform traditional feature-engineering methods in terms of scoring reliability and semantic understanding (Sun et al., 2025). Furthermore, Bai and Stede (2023) concluded that contextual language models represent the current state-of-the-art solution for evaluating free-text student responses across various educational environments. These collective findings reinforce the effectiveness of contextual language modeling in supporting objective and scalable educational assessment. Therefore, the performance obtained in this study is consistent with the growing body of evidence demonstrating the superiority of transformer-based models for essay scoring applications.

Compared with conventional approaches, the proposed BERT model demonstrated superior capability in representing semantic relationships among words and sentences, thereby improving the consistency of predicted scores. This finding is in agreement with Wang (2024), who reported that integrating semantic, thematic, and linguistic representations significantly enhanced essay scoring performance compared with traditional text-processing methods. Similar improvements have also been observed in multilingual Automated Essay Scoring research, where contextual language models successfully handled diverse writing styles and linguistic variations without relying extensively on handcrafted features (Ezz et al., 2025). De Vrindt et al. (2025) further explained that modern Natural Language Processing techniques provide more robust comparative judgment for essay assessment because they capture contextual information beyond surface-level linguistic characteristics. The ability of BERT to understand bidirectional context also enables more accurate evaluation of complex student responses, as demonstrated by Ariely et al. (2023) in the assessment of open-ended biology questions. In addition, recent investigations have shown that fine-tuning pretrained transformer models produces better performance than traditional recurrent neural networks in educational text evaluation because contextual embeddings preserve richer semantic information (Wang & Luo, 2026). Similar trends have also been reported in comparative studies of deep learning algorithms, which consistently demonstrate that contextual representation learning improves predictive accuracy across complex classification tasks (Timur & Warsono, 2027). These findings collectively explain why the proposed BERT framework achieved reliable performance in evaluating authentic student examination essays.

The implementation of the proposed framework also demonstrates the practical potential of artificial intelligence for improving educational assessment in higher education institutions. The automated scoring system developed in this study can reduce lecturers' grading workload while maintaining consistency and objectivity throughout the evaluation process. This observation is consistent with the findings of Kishor et al. (2025), who reported that deep learning-based grading systems substantially improve assessment efficiency without sacrificing scoring quality. Similarly, Tate et al. (2026) found that AI-

based essay scoring provides useful analytical support for evaluating students' writing across different genres while maintaining acceptable agreement with human raters. Beyond improving assessment efficiency, intelligent educational systems have also been recognized as valuable decision-support tools for academic institutions because they facilitate faster and more consistent evaluation processes (Hu et al., 2027; Wu et al., 2025). Furthermore, the application of deep learning in educational assessment reflects broader developments in intelligent information systems, where automated pattern recognition has successfully improved prediction accuracy across various domains (Mnassri et al., 2027; Schimmenti et al., 2026). Although human judgment remains essential in educational assessment, the results of this study indicate that BERT-based Automated Essay Scoring can effectively function as a complementary assessment tool rather than replacing educators. Consequently, integrating contextual Natural Language Processing into educational evaluation offers a practical pathway toward more efficient, objective, and scalable assessment systems in higher education.

The primary novelty of this study lies in the integration of a fine-tuned BERT model with an authentic higher education examination dataset consisting of 2,340 valid student essay responses, allowing the proposed framework to be evaluated under realistic academic assessment conditions rather than relying solely on publicly available benchmark datasets. In contrast to many previous studies that mainly focused on improving predictive accuracy, this research simultaneously evaluated scoring agreement and prediction errors using Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), thereby providing a more comprehensive assessment of model performance (Sun et al., 2025). Another distinguishing contribution is the implementation of a complete workflow covering data validation, preprocessing, contextual language modeling, fine-tuning, and multi-metric evaluation within a single assessment framework. While Wang et al. (2026) investigated feature weighting patterns in LLM-based essay scoring and Boutadjine et al. (2026) focused on distinguishing AI-generated essays from human-written essays, the present study specifically addresses the practical problem of scoring authentic student examination answers using contextual semantic representations. Furthermore, unlike de Almeida et al. (2026), which evaluated BERT models in healthcare-related document analysis, this research demonstrates the applicability of transformer-based language models within educational assessment. The proposed framework also extends previous work on automated grading methodologies by emphasizing practical deployment using institutional examination data rather than simulated datasets (Ayaan & Ng, 2025). These characteristics collectively distinguish the proposed model from previous Automated Essay Scoring studies and provide additional empirical evidence regarding the implementation of contextual Natural Language Processing in higher education assessment.

The findings of this study have important implications for both educational practice and future research in intelligent assessment systems. From a practical perspective, the proposed BERT-based Automated Essay Scoring framework can assist lecturers in reducing grading time while improving scoring consistency and minimizing subjectivity

during the evaluation process, supporting the broader adoption of artificial intelligence in educational environments (Abdullah et al., 2024). The ability of contextual language models to evaluate essay responses efficiently also supports more timely formative feedback, which is considered an essential component of effective learning and assessment (Ariely et al., 2023). In addition, the proposed framework can be integrated into intelligent academic information systems to support institutional decision-making and improve the efficiency of large-scale assessment activities (Hu et al., 2027; Wu et al., 2025). From a technological perspective, this research further confirms that transformer-based language models provide a robust foundation for developing next-generation educational assessment systems capable of handling increasingly diverse student writing characteristics (Bai & Stede, 2023). The findings are also consistent with recent advances demonstrating that contextual deep learning models continue to improve the reliability, scalability, and adaptability of automated essay scoring across multiple educational settings (Miao & Xu, 2025; Cui, 2024). Moreover, future studies may extend the proposed framework by incorporating explainable artificial intelligence, rubric-aware scoring mechanisms, or large language models to improve transparency and interpretability during automated assessment (Calaña, 2025; Wang et al., 2026). Therefore, the proposed framework not only contributes to the advancement of Automated Essay Scoring research but also provides a practical foundation for developing reliable, scalable, and intelligent assessment systems that support digital transformation in higher education.

CONCLUSION

This study successfully developed and evaluated an Automated Essay Scoring (AES) framework based on the Bidirectional Encoder Representations from Transformers (BERT) model for assessing student essay examination responses. Using a dataset of 2,340 valid essay answers obtained after preprocessing from an initial collection of 2,500 responses, the proposed model demonstrated stable learning performance during the fine-tuning process and achieved strong predictive capability in automated essay assessment. The experimental results indicate that the proposed framework produced a Quadratic Weighted Kappa (QWK) score of 0.872, a Mean Absolute Error (MAE) of 0.418, and a Root Mean Square Error (RMSE) of 0.593, confirming a high level of agreement with human evaluators while maintaining relatively low prediction errors. Comparative evaluation further demonstrated that the proposed BERT model outperformed conventional baseline approaches, highlighting the advantages of contextual language representations over traditional feature-based and recurrent neural network methods for essay scoring. These findings confirm that transformer-based Natural Language Processing provides an effective and reliable solution for improving the objectivity, consistency, and efficiency of essay assessment in higher education. Therefore, the proposed framework offers both theoretical contributions to the advancement of Automated Essay Scoring research and practical contributions by supporting lecturers in conducting faster, more consistent, and scalable assessment processes. Future research is recommended to investigate the integration of explainable artificial intelligence, rubric-

aware scoring mechanisms, and large language models to further improve the transparency, interpretability, and generalizability of automated essay scoring systems across broader educational contexts.

REFERENCES

Abbassy, M. M., Ead, W. M., El-Shora, A. I. A., & Aboalndr, A. M. (2026). Uncovering advanced transfer learning strategies for deep neural networks in natural language processing. *Scientific Reports*, *16*(1), 1–10. https://doi.org/10.1038/s41598-026-39819-3

Abdullah, A. S., Geetha, S., Abdul Aziz, A. B., & Mishra, U. (2024). Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning. Alexandria Engineering Journal, 108, 764–788. https://doi.org/10.1016/j.aej.2024.08.067

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. International Journal of Artificial Intelligence in Education, 33(1), 1–34. https://doi.org/10.1007/S40593-021-00283-X

Ayaan, A., & Ng, K. W. (2025). Automated grading using natural language processing and semantic analysis. MethodsX, 14, 103395. https://doi.org/10.1016/j.mex.2025.103395

Bai, X., & Stede, M. (2023). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. International Journal of Artificial Intelligence in Education, 33(4), 994–1032. https://doi.org/10.1007/S40593-022-00323-0

Boutadjine, A., Harrag, F., & Shaalan, K. (2026). Student or AI? Automated detection of AI-generated student essays. Procedia Computer Science, 275, 984–993. https://doi.org/10.1016/j.procs.2026.01.112

Cong, C., Song, Y., Di Ieva, A., Jin, Q., Fan, L., Chou, A., J Gill, A., & Liu, S. (2026). Slide-aware deep feature prompting for enhanced whole slide image classification. *Expert Systems with Applications*, *331*(PA), 133079. https://doi.org/10.1016/j.eswa.2026.133079

de Almeida, S. S., Fontes, R. S., Colaço, M., Silva, G. J. P. C., Guimarães, A., de Medeiros Valentim, R. A., dos Santos, J. P. Q., Oliveira, A. C., Oliveira, H. G., Machado, G. M., de Morais, A. H. F., Cortez, L. R., & Freire Alves, L. P. C. (2026). SUS audit aided by natural language processing: A comparative evaluation of BERT models in the analysis of health news. Array, 30, 100726. https://doi.org/10.1016/j.array.2026.100726

del Rey, S., Cruz, L., Franch, X., & Martínez-Fernández, S. (2027). Estimating Deep Learning energy consumption based on model architecture and training environment. *Computer Standards and Interfaces*, *99*(May 2026), 104170. https://doi.org/10.1016/j.csi.2026.104170

De Vrindt, M., Tack, A., Van den Noortgate, W., Lesterhuis, M., & Bouwer, R. (2025). Opportunities of natural language processing for comparative judgment

assessment of essays. Computers and Education: Artificial Intelligence, 8, 100414. https://doi.org/10.1016/j.caeai.2025.100414

Ezz, M., Alruily, M., Mostafa, A. M., Alaerjan, A. S., Aldughayfiq, B., Allahem, H., & Shehab, A. (2025). Efficient Arabic essay scoring with hybrid models: Feature selection, data optimization, and performance trade-offs. Computers, Materials and Continua, 86(1), 1–28. https://doi.org/10.32604/CMC.2025.063189

Hendre, M., Mukherjee, P., Preet, R., & Godse, M. (2023). Efficacy of deep neural embeddings-based semantic similarity in automatic essay evaluation. International Journal of Cognitive Informatics and Natural Intelligence, 17(1). https://doi.org/10.4018/IJCINI.323190

Hu, X., Gu, H., & Tang, Y. (2027). An embedding-based approach for measuring science–technology topic linkages using LLMs. *Information Processing and Management*, *64*(1), 104983. https://doi.org/10.1016/j.ipm.2026.104983

Kishor, K., Vishwakarma, P., Sengar, L., Kumar, V., & Gupta, V. (2025). Development of grader provider system using deep learning. Procedia Computer Science, 259, 172–181. https://doi.org/10.1016/j.procs.2025.03.318

Mnassri, K., Farahbakhsh, R., & Crespi, N. (2027). MultiSent-RAG: A retrieval and memory-augmented system for multilingual sentiment processing. *Information Processing and Management*, *64*(1), 104990. https://doi.org/10.1016/j.ipm.2026.104990

Sun, J., Song, T., Peng, W., & Song, J. (2025). A survey of automated essay scoring: Challenges, advances, and future. Neurocomputing, 650, 130916. https://doi.org/10.1016/j.neucom.2025.130916

Schimmenti, A., Pasqual, V., Vitali, F., & van Erp, M. (2026). Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates. *Journal of Documentation*, *28*(7), 205–250. https://doi.org/10.1108/JD-07-2025-0203

Tate, T. P., Graham, S., Kaldirim, A., & Tavsanli, O. F. (2026). Can AI provide useful analytic essay scoring for different genres of writing with elementary grade students? Assessing Writing, 68, 101038. https://doi.org/10.1016/j.asw.2026.101038

Timur, R. P., & Warsono, S. (2027). Bridging accounting and AI: A sentiment-enhanced analytical framework for forecasting hotel profitability from customer reviews. *International Journal of Hospitality Management*, *140*(October 2025), 104805. https://doi.org/10.1016/j.ijhm.2026.104805

Wang, J., & Luo, L. (2026). An intelligent essay scoring system based on a BERT-driven deep self-supervised contrastive learning network. Alexandria Engineering Journal, 140, 1–9. https://doi.org/10.1016/j.aej.2026.02.016

Wu, X., Cai, G., Ding, M., Wang, Y., Zhou, G., Li, Y., Gai, S., Wang, B., & Liu, D. (2025). Transforming Food Consumer Analysis: The Role of Machine Learning in Food Consumer Demand 4.0. *Journal of Future Foods*, *7*(1), 65–80. https://doi.org/10.1016/j.jfutfo.2024.12.008

Wang, M., Chen, Y., Huang, X., & Lai, Y. (2026). Opening the blackbox of LLM-based automated essay scoring: Insights into feature weighting patterns and score validity. Computers and Education: Artificial Intelligence, 10, 100568. https://doi.org/10.1016/j.caeai.2026.100568

Wang, Q. (2024). A multifaceted architecture to automate essay scoring for assessing English article writing: Integrating semantic, thematic, and linguistic representations. Computers and Electrical Engineering, 118, 109308. https://doi.org/10.1016/j.compeleceng.2024.109308

Zhou, J., Xin, X., Yu, S., Liu, J., Li, W., & Cui, X. (2025). Advancing Authentic Recipe Ideation across Culinary Styles using a Mathematical Model: RecipeMT. *Journal of Future Foods*, *7*(1), 102–114. https://doi.org/10.1016/j.jfutfo.2025.07.004