Scientific Journal of Informatics
Vol. 13, No. 1, February 2026

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Automated Essay-Answer Grading using Transformer and GenAI-Based Error Analysis on Answer Sentence

Mohammad Mirza Qusyairi[1*], Yusep Rosmansyah[2]
[1,2]Master Program of Electrical Engineering, Institut Teknologi Bandung, Indonesia

Abstract.
Purpose: This study aims to address the limitations of manual essay assessment, which is time-consuming and prone to subjectivity, by developing an automated essay grading system that is faster, more efficient, and more objective. In addition to producing scores, the system is intended to support learning by providing meaningful feedback on students' writing errors.
Methods: The research applies a transformer-based approach to automated essay scoring in order to capture deeper semantic relationships between student answers and reference answers. Unlike conventional word embedding methods, transformers are used to model contextual meaning at the word, phrase, and sentence levels. The system is further enhanced with an AI-based error analysis module that identifies categories of student errors, including content omissions, conceptual misunderstandings, and structural inaccuracies, enabling the generation of corrective feedback alongside numerical scores.
Result: The experimental results show that the proposed model achieves strong agreement with human scoring, indicated by a Pearson correlation coefficient (r) of 0.8432 and a Mean Absolute Error (MAE) of 0.6013. These results demonstrate an improvement in prediction accuracy compared to word embedding-based approaches. Furthermore, the system successfully generates relevant and actionable error feedback that supports students in understanding their mistakes and improving their essay quality.
Novelty: The novelty of this research lies in the integration of transformer-based automated essay scoring with an AI-driven error analysis module. Rather than focusing solely on score prediction accuracy, the proposed approach combines quantitative scoring with qualitative error feedback, enabling the system to function not only as an assessment tool but also as a learning support mechanism that helps students understand and improve their writing.

Keywords: Automatic essay grading, Transformer, Error analysis, Accuracy improvement
Received December 2025 / **Revised** January 2026 / **Accepted** February 2026

INTRODUCTION

The development of educational technology in the last decade is marked by the emergence of the Smart Learning Environment (SLE), which focuses on personalizing learning thru the integration of smart technology [1]. One of the key components of SLE is the learning module, which contains assessment results and plays an important role in ensuring fair assessment and the effective use of time and resources. In this context, assessment becomes an essential element for measuring students' understanding of the learning material [2].

Assessment in education is generally conducted thru objective tests and essay tests. Objective tests excel in consistency and ease of scoring, while essay tests are more suitable for measuring higher-order thinking skills such as synthesis and evaluation [3], [4], [5]. However, essay assessment is complex, time-consuming, and prone to subjectivity, which can lead to bias and unfair grading [6], [7], [8]. This condition encourages the use of e-assessment as a solution to improve efficiency and objectivity.

Computer-Based Testing (CBT) is a widely used form of e-assessment, but it is generally limited to multiple-choice questions due to ease of scoring. Essay assessment automatically demands text analysis skills because student responses can vary syntactically but be equivalent in meaning [9]]. In the literature, this study is divided into Automated Essay Scoring (AES) and Automated Short-Answer Scoring (ASAS) [10], [11], with ASAS focusing on assessing the relevance of students' answer content meaning to reference answers [12], [13].

Automated essay scoring essentially measures text similarity thru lexical and semantic approaches [14]. Lexical similarity assesses the similarity of characters or word sequences, while semantic similarity evaluates the closeness of meaning between texts [15], [16]. The development of NLP and AI allows for

*Corresponding author.
Email addresses: 23223312@mahasiswa.itb.ac.id (Qusyairi)
DOI: 10.15294/sji.v13i1.39859

the use of more advanced techniques, ranging from traditional methods like TF-IDF and cosine similarity to transformer-based deep learning models [17], [18], [19]. Models like SBERT have proven capable of representing sentence meaning contextually and improving the accuracy of automated assessment.

Previous research in Indonesia has shown the evolution of automated essay scoring approaches, ranging from string and keyword similarity without semantics [20], syntactic and transfer learning approaches [21], hybrid STS methods [22], to the utilization of BERT and SBERT embeddings with data imbalance handling [23], [24], [25]. Although the accuracy of the scores has improved, most studies still focus on how closely the machine's scores align with human assessments and have not yet emphasized the diagnostic aspects of learner errors.

Beside accuracy, the main challenge of essay assessment is students' low understanding of where they went wrong when feedback is only in the form of a score [26], [27]. The integration of AI-based error analysis has proven capable of reducing error repetition and improving the effectiveness of revisions [28], [29], [30] Therefore, this study proposes the development of a transformer-based automated essay scoring system integrated with learner error analysis to support more comprehensive, reflective, and SLE-aligned formative assessment.

METHODS

A. Dataset

This study utilizes the same dataset as used in the studies by [20], [24]. The use of identical data aims to make it easier and more objective to compare the accuracy between the solution developed in this study and the approaches from previous research. Sample data from the dataset are presented in Tables 1 and 2.

The data comes from the results of daily tests in the Basic Programming subject for students in Class X of the Computer and Network Engineering (TKJ) department at SMKN 8 Semarang. There are seven essay questions with reference answers, each completed by 32 students, resulting in a total of 224 responses. Each answer was manually graded by two teachers on a scale of 0-4, where a score of 0 indicates an incorrect answer and a score of 4 represents a correct answer. For the purpose of evaluating the correlation between the automated scoring system and manual scoring, the average scores from both raters were used. Table 1 shows the reference questions and answers.

Table 1. Sample question and reference answer from dataset

No	Question	Answer
1	Apa yang kalian ketahui tentang algoritma? *(What do you know about algorithm?)*	Algoritma adalah urutan langkah-langkah logis penyelesaian masalah yang disusun secara sistematis dan logis (Algorithm is a sequence of logical steps for solving problems that are arranged systematically and logically)
2	Tuliskan algoritma untuk menyelesaikan masalah perhitungan luas persegi panjang *(Write the algorithm to find the area of a rectangle)*	Start, Masukkan panjang dan lebarnya, Hitung Luas panjang kali lebar, Hasil kali panjang dan lebar = luas, Finish (Start, enter length and width, Calculate area length times width, Product length and width = area, Finish)
3	Apa yang kalian ketahui tentang flowchart? *(What do you know about flowchart?)*	Bagan (chart) yang menunjukkan alir (flow) di dalam program atau prosedur sistem secara logika (A chart that shows the flow in a program or system procedure logically)
4	Ada berbagai macam simbol-simbol pada flowchart. Berfungsi untuk apa simbol decision pada flowchart? *(There are various kinds of symbols on a flowchart. What is the function of the decision symbol on a flowchart?)*	Pemilihan proses berdasarkan kondisi yang ada. (Process selection based on existing conditions.)
5	Berfungsi untuk apa simbol terminator pada flowchart? *(What is the function of the terminator symbol on a flowchart)*	Simbol terminator digunakan untuk permulaan (start) atau akhir (finish) dari suatu kegiatan. (The terminator symbol is used for the start or end of an activity.)
6	Apa yang dimaksud dengan inisialisasi variabel? *(What is variable initialization?)*	Inisialisasi variabel adalah mengisi nilai untuk pertama kalinya ke dalam variabel. (Variable initialization is filling a value for the first time into a variable.)
7	Apa fungsi operator pada pemrograman? *(What are the functions of operators in programming?)*	Operator berfungsi untuk memanipulasi nilai dari suatu variabel (Operators function to manipulate the value of a variable)

Two evaluators assessed each response individually on an integer scale ranging from zero (entirely incorrect) to four (flawless solution). Table 2 presents the scores allocated by the two instructors to the responses of four students about Question 3. Tables 1 and 2 illustrate that the dataset comprises diverse questions, reference answers, learner responses, and the scores assigned by both instructors for each learner response. The correlation value and mean absolute error (MAE) were calculated by comparing the instructors' scores and their average, which represent manual grading, with the automatic answer grading.

Table 2. Sample student answer and evaluators' manual grading for question 3 from dataset

No.	Student Answer	Score 1*	Score 2*	The Avg. Score
3.1	bagan yang menunjukkan program / prosedur secara logika *(A diagram showing the program/procedure logically)*	4	3,5	3,75
3.2	bagan yang menunjukkan flow dalam suatu proses didalam sistem prosedur secara logika *(A diagram showing the logical flow within a process in a system of procedures.)*	4	4	4
3.3	bagian yang menunjukan flow dalam suatu proses di dalam suatu prosedur secara logika *(The part that logically shows the flow within a process in a procedure.)*	4	4	4
3.4	Flowchart adalah bagan yang menunjukkan alir didalam program/prosedur sistem secara logika. *(A flowchart is a diagram that logically shows the flow within a system program/procedure.)*	4	4	4

*Score 1 from evaluator one and Score 2 from evaluator two

B. Proposed Solution

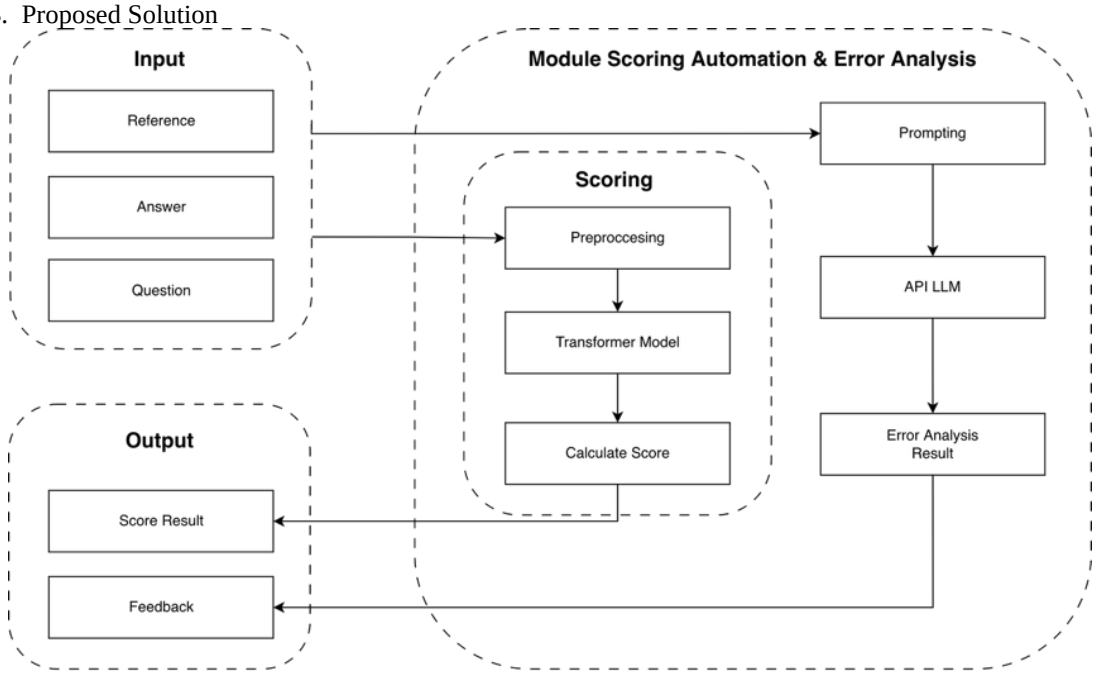

Figure 1. Design of an automated essay answer grading system and error analysis

This section describes the proposed solution, which is an automated essay scoring system design using transformers equipped with learner error analysis based on a Large Language Model (LLM), as shown in Figure 1. The system consists of three main components: text preprocessing, semantic mapping using the best transformer model, and error analysis using an LLM. Preprocessing is done lightly to maintain semantic context, only including excessive space cleaning, punctuation normalization, and irrelevant character removal without stemming or stopword removal. Once the text is ready, the system uses a single best transformer model to generate sentence embeddings that represent the meaning of student responses and reference answers. This representation is used to calculate semantic similarity with cosine similarity, resulting in consistent and objective automatic scores.

The next step is to analyze learner errors based on an LLM by inputting the question text, reference answer, and student answer. Considering the context of the question, LLMs can identify forms of errors such as conceptual misconceptions, incomplete answers, deviations from the question's focus, the use of irrelevant information, and imprecise terminology. Beside detecting errors, LLMs also generate explanatory summaries and brief feedback that help teachers understand where mistakes are without having to manually read the entire answer. Thus, the proposed system not only provides automatic assessment scores but also offers more meaningful error information for the learning process.

C. Experimental Design

- Preprocessing

This stage serves to prepare the text so that it can be processed optimally by the language model without changing the semantic structure of the sentence. The process is very light and only includes cleaning excess spaces, normalizing punctuation, removing irrelevant characters, and standardizing writing formats. Techniques that could potentially remove meaning, such as stemming or stopword removal, are not used so that word order, syntactic relations, and sentence context remain intact.

(a)

```
0      langkah langkah yang logis untuk menyelesaikan...
1      algoritma adalah langkah langkah yang disusun ...
2      algoritma adalah langkah langkah yang disusun ...
3      Algoritma adalah urutan langkah-langkah logis ...
4      Urutan langkah langkah logis penyelesaian masa...
                            ...
219    simbol-simbol khusus yang digunakan untuk meng...
220    fungsi operator = mengelola segala bentuk pemr...
221    digunakan untuk untuk memanipulasi data sepert...
222                 untuk menyatakan hitungan pada angka
223    untuk mengoperasikan suatu nilai dalam bentuk ...
```

(b)

```
0      langkah langkah yang logis untuk menyelesaikan...
1      algoritma adalah langkah langkah yang disusun ...
2      algoritma adalah langkah langkah yang disusun ...
3      algoritma adalah urutan langkahlangkah logis p...
4      urutan langkah langkah logis penyelesaian masa...
                            ...
219    simbolsimbol khusus yang digunakan untuk mengo...
220    fungsi operator mengelola segala bentuk pemrog...
221    digunakan untuk untuk memanipulasi data sepert...
222                 untuk menyatakan hitungan pada angka
223    untuk mengoperasikan suatu nilai dalam bentuk ...
```

Figure 2. Preprocessing text (a) before (b) after

- Transformer

This component is responsible for generating meaning representations from student answer texts and reference answers. The best model transformer is selected based on experimental results and used to form sentence embeddings that capture relationships between words, syntactic structure, and semantic context. This vector representation is then used to calculate semantic similarity using cosine similarity as the basis for automatic scoring. With this approach, the system no longer explicitly relies on lexical matching, but instead leverages semantic similarity to identify whether a student's answer contains core ideas comparable to the reference answers prepared by the teacher.

In this study, the transformer model used is SBERT *paraphrase-multilingual-MiniLM-L12-v2*, which is a lightweight and efficient variant with a 12-layer encoder capable of handling multilingual processing, including Indonesian. This model adopts a Siamese architecture, allowing it to convert sentences into vector embedding spaces that can be easily compared using distance functions. Beside having relatively small parameter sizes, which makes inference fast, this model is also specifically trained on paraphrase tasks,

making it excellent at capturing semantic relationships between phrases. This makes it suitable for the context of evaluating essay answers, as the system can recognize variations in students' expressions without relying on literal word similarity.

Beside determining the model, the system also normalizes the scores so that the learning targets are within a consistent range, i.e., 0-1. This normalization is necessary because the final scores given by teachers are generally within different ranges (e.g., 0-4 or 0-100), so they must be standardized before being used as training labels. In its implementation, normalization is done by dividing the original value (n_rata), which is the average teacher score, by the maximum score (max_score) to obtain a proportional value. With this normalization, the model can learn the semantic relationships between answers more stably because the target values are always on a uniform scale and are easily predictable by the regression loss function.

- Error Analysis

This component utilizes the natural language understanding capabilities of LLMs to analyze the quality of answers based on the context of the question and the reference answer. LLMs take all three elements as input and identify error types such as misconceptions, incompleteness, the use of irrelevant information, or deviations from the question's focus. Beside detecting errors, LLMs also generate brief explanations and corrective feedback, making the evaluation process more informative. This error analysis component leverages the natural language understanding capabilities of the *GPT-4o* model to evaluate the quality of student answers contextually. Evaluation is not only based on word matching, but also considers meaning, semantic relationships, and the relevance of content to the essay question. To achieve accurate assessment, the system employs three prompt engineering techniques: Role-Based Prompting, Hidden Chain-of-Thought (CoT), and Context Injection. In Role-Based Prompting, the model is instructed to act as a teacher, resulting in an educational analytical style that focuses on misconceptions and provides direct improvement guidance to students using direct language. This role assignment allows the model not only to identify errors but also to help students understand what needs to be corrected.

Hidden CoT is used to guide the model to perform internal reasoning concisely without displaying it in the final output. This mechanism allows *GPT-4o* to first identify the core concepts in the reference answer, compare them to the student's answer, determine the main errors, and select the most relevant improvement suggestions. This hidden reasoning keeps the output concise, focused, and safe, without explicitly revealing the internal thot process. Next, Context Injection ensures the model receives three essential elements as input: the essay question, the key answer, and the student's answer. This context injection serves as the foundation for evaluation, enabling the model to distinguish main ideas, identify incompleteness, detect irrelevant information, and assess whether the response remains within the scope of the question.

By combining these three approaches, *GPT-4o* is able to identify various forms of errors such as conceptual misconceptions, incomplete answers, deviations in focus, or the use of irrelevant information [31]. Beside detecting errors, *GPT-4o* also generates concise explanations and corrective feedback that are immediately actionable for students, such as highlighting missing core points or suggesting revision steps. Thus, the evaluation process becomes more formative and informative because students not only receive a grade but also gain an understanding of where they went wrong and how to improve.

- Calculating Score and Testing Process

Before the system automatically generates a score, the first step is to calculate the semantic similarity between the student's answer and the reference answer. This process begins by building a numerical representation of the text in the form of document vectors. Each word in the answer is converted into a word vector, and then the average value of all those word vectors is calculated to form a single document vector [23], as explained in Equation (1). In this way, both the student's answer (A) and the reference answer (R) have vector representations (V_d) that can be mathematically compared.

$$V_d = \frac{\sum_{i=1}^{n} v_{w_i}}{n} \quad (1)$$

After the vector representation is obtained, the level of semantic similarity between the students' answers and the reference answers is calculated using cosine similarity as shown in Equation (2) [23]. This measure

yields values in the range of 0 to 1, where larger values indicate a smaller distance between vectors and a higher degree of semantic similarity between responses. This is the value that forms the basis of the automatic score mapping process.

$$sim(A,R) = \frac{V_{d_A} \cdot V_{d_R}}{\|V_{d_A}\| \cdot \|V_{d_R}\|} = \frac{\sum_{i=1}^{n} v_{d_{Ai}} \cdot v_{d_{Ri}}}{\sqrt{\sum_{i=1}^{n} v_{d_{Ai}}^2} \cdot \sqrt{\sum_{i=1}^{n} v_{d_{Ri}}^2}} \quad (2)$$

where:

V_{d_A} = document vector for the student's answer.

V_{d_R} = document vector for the reference answer.

To obtain a score that meets teacher standards, the cosine similarity value from Equation (2) is linearly projected into the maximum score range of the assessment rubric. Since cosine similarity is already within the interval of 0-1, the mapping is simply done by multiplying it by the maximum score, resulting in the system output score(A) being directly within the domain of manual assessment values, as shown in Equation (3). This approach ensures that an increase in semantic similarity value according to Equation (2) is directly proportional to an increase in predictive score, and maintains consistency in score interpretation against teacher standards.

$$score(\mathrm{A}) = \ sim(\mathrm{A}, \mathrm{R}) \quad (3)$$

To simplify the testing process, the data consisting of questions, reference answers, and student responses is first entered thru the GUI interface and then stored in the database. Subsequently, this data is sent as input to the automated essay scoring model via API. The model then calculates the final score for each student response in a two-decimal-place format within the range of 0 - 4. An example of the interaction between the system and student answers can be seen in Table 3, which shows how the API provides scores as well as error feedback.

Table 3. Examples of request and response from API

Request	{
	"question": "Apa yang kalian ketahui tentang algoritma?",
	"reference_answer": "Algoritma adalah urutan langkah-langkah logis penyelesaian masalah yang disusun secara sistematis dan logis",
	"answer": "algoritma adalah metode efektif sebagai rangkaian terbatas cara untuk menyelesaikann suatu masalah secara berurutan"
	}

Response	{
	"timestamp": "2025-12-22 01:08:04",
	"answer": "algoritma adalah metode efektif sebagai rangkaian terbatas cara untuk menyelesaikann suatu masalah secara berurutan",
	"jawaban_kunci": "Algoritma adalah urutan langkah-langkah logis penyelesaian masalah yang disusun secara sistematis dan logis",
	"score": 2.71,
	"error_analysis ": "Kamu belum menjelaskan bahwa algoritma adalah urutan langkah-langkah logis yang disusun secara sistematis. Hati-hati, konsep \"metode efektif\" yang kamu gunakan kurang tepat karena algoritma lebih berfokus pada urutan langkah daripada efektivitas. Untuk memperbaiki jawabanmu, coba tambahkan penjelasan tentang bagaimana algoritma disusun secara logis dan sistematis."
	}

The scores obtained are stored back into the database to facilitate the evaluation and monitoring of results. In addition to storing values, the system can also provide feedback in the form of error analysis, so students not only know their final score but also understand which parts of their answers were incorrect, had a different meaning than the reference answers, or did not meet the main concept elements. This information helps students reflect and improve their answers for the next opportunity.

RESULT AND DISCUSSION

The evaluation conducted in this study applied parametric statistical tests using the Pearson correlation coefficient (r), as the type of data measurement scale used was at the ratio level. The data analyzed consists of the assessment scores generated by the automated essay answer grading system and the manual assessment scores provided by teachers, both of which fall within the range of 0-4 and have met the assumption of normal distribution in numerical form. The equation for calculating the correlation coefficient (r) is as follows in Equation (4):

$$r_{xy} = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{\sqrt{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \quad (4)$$

Additionally, the testing process in this study also follows the same evaluation mechanism as previous research [20], [24], [25] namely by calculating the Mean Absolute Error (MAE) value. One of the main objectives of applying this error measure is to facilitate the comparison process between the research results developed in this study and the accuracy achievement results from previous studies. The Mean Absolute Error (MAE) calculated as follows in Equation (5):

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} \quad (5)$$

Based on the results in Table 4, it can be understood that compared to previous studies conducted by [20], [24], [25], the performance of automated essay grading in this study shows better results. The indication of this increase is evident from the increase in the Pearson correlation coefficient (r) value by 0.1475 and the decrease in the MAE value by 0.1244 compared to the research findings [25]. This finding indicates that utilizing the transformer model can improve the accuracy of automated essay scoring systems, particularly in generating more precise semantic representations.

In addition, the correlation value (r) of 0.8432 indicates that the relationship between manual assessment scores and automated assessment scores is in the fairly strong category. The MAE value of 0.6013 indicates a lower average absolute error compared to previous studies. The smaller or closer the value is to zero (0), the better the MAE value is considered because it indicates a smaller average error rate.

Table 4. Result of the evaluation

Comparison	Correlation (r)	Mean Absolute Error (MAE)
[20]	0,6542	0,9499
[24]	0,6829	0,8414
[25]	0,6957	0,7257
This Study	**0,8432**	**0,6013**

This finding indicates that utilizing the transformer model can improve the accuracy of automated essay scoring systems, particularly in generating more precise semantic representations and enabling better synthesis of contextual meaning within students' answers.

The details of the assessment results between the teachers' average score and each student's automatic score are shown in graphical form in Figure 2. The orange curve on the graph represents the average scores given by Teacher 1 and Teacher 2, while the blue curve represents the automated assessment results (scores from the robot). Based on the visual pattern in the graph, it can be observed that most of the scores generated by the automated scoring system are still below the average manual scores given by teachers. However, there are some cases where the automated scores are actually higher than the manual scores.

This phenomenon occurs because some student answers contain keywords identical to the reference answers, leading the system to rate them highly. Low scores from teacher assessments are likely influenced by differences in the intended meaning and context of students' answers compared to the concepts expected in the reference answers. In this condition, automated scoring systems tend to focus more on the level of semantic similarity between words without thoroughly assessing contextual relevance.

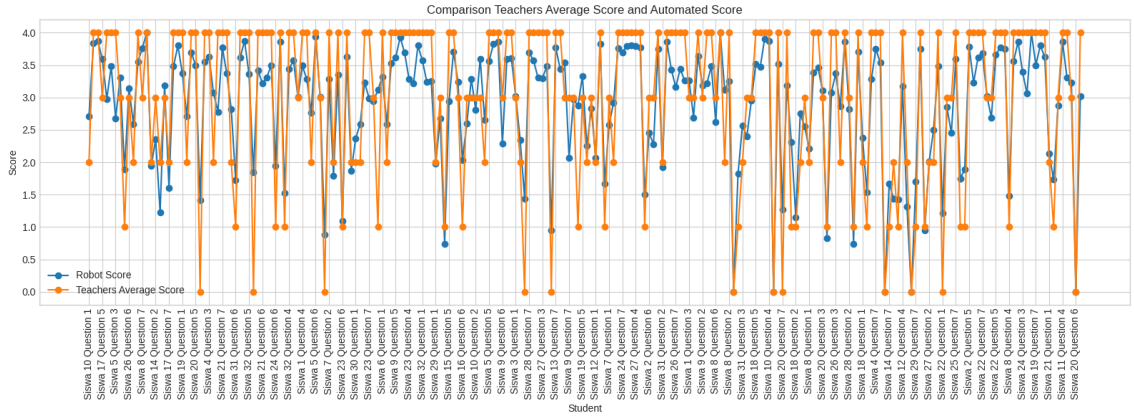

Figure 3. Comparison graph of average teacher scores and automated scores for each student

CONCLUSION

This research successfully developed an automated essay scoring system that utilizes a transformer-based semantic equivalence approach thru sentence representation modeling. The evaluation results show that the scores generated by the system have a strong and consistent correlation with manual assessments by teachers. This indicates that the model is capable of accurately mimicking human assessment patterns, and produces relatively small differences in scores, bringing the system's performance closer to manual assessments. In addition to generating assessment scores, the system also integrates a learner error analysis module capable of identifying various forms of deviation in answers from the reference key. This analysis includes irregularities in sentence structure, incomplete concepts, and inaccuracies in meaning that appear in student responses. This mechanism allows the system to provide a fairer assessment based on semantic consistency and the flow of answers. This finding reinforces that context-based semantic representations are more effective in capturing the meaning of answers compared to static embedding approaches that rely solely on surface text similarity. As a suggestion, future research could develop a more detailed and adaptive error analysis so that the feedback provided by the system becomes increasingly personalized and supports the continuous learning process.

REFERENCES

[1] Y. Rosmansyah, B. L. Putro, A. Putri, N. B. Utomo, and Suhardi, "A simple model of smart learning environment," *Interactive Learning Environments*, vol. 31, no. 9, pp. 5831–5852, Dec. 2023, doi: 10.1080/10494820.2021.2020295.

[2] F. G. K. Yılmaz, A. B. Ustun, and R. Yılmaz, "Investigation of pre-service teachers' opinions on advantages and disadvantages of online formative assessment: an example of online multiple-choice exam," *Journal of Teacher Education and Lifelong Learning*, vol. 2, no. 1, pp. 1–8, 2020.

[3] I. Levy-Feldman, "The Role of Assessment in Improving Education and Promoting Educational Equity," *Educ. Sci. (Basel).*, vol. 15, no. 2, p. 224, Feb. 2025, doi: 10.3390/educsci15020224.

[4] T. Mahmud, Md. F. Rabbi, M. T. Hossain, A. A. Talukder, and Md. K. Hasan, "Portfolio assessment for developing higher order thinking skills in Bangladeshi undergraduate EFL writing classes," *Language Testing in Asia*, vol. 15, no. 1, p. 61, Oct. 2025, doi: 10.1186/s40468-025-00404-6.

[5] E. S. Lestari and B. Widi Pratolo, "Analysis of Higher Order Thinking Skills (HOTs) in English Final Examination Questions at Junior High School," *ETERNAL (English Teaching Journal)*, vol. 16, no. 1, pp. 24–36, Feb. 2025, doi: 10.26877/eternal.v16i1.892.

[6] D. Deepshikha, "A systematic review on the future of educational assessment: AI-driven grading and personalised feedback in higher education," *Artificial Intelligence in Education*, vol. 2, no. 2, pp. 75–115, Dec. 2026, doi: 10.1108/AIIE-03-2025-0036.

[7] E. Susilawati, E. Supriani Siregar, S. Ekalestari, and E. S. Harahap, "Fitrah: Journal of Islamic Education DEVELOPMENT OF A DIGITAL ASSESSMENT MODEL FOR AUTOMATED

SHORT ESSAY SCORING IN ISLAMIC EDUCATION STUDY PROGRAM," *Fitrah: Journal of Islamic Education*, vol. 6, no. 1, pp. 208–223, doi: 10.53802/fitrah.v6i1.1195.

[8] M. Ridwan, A. Najib, H. Ruzakki, R. Rofiqi, and M. Muslimin, "Bias in Peer Assessment: Challenges, Solutions, and Best Practices for Fair Student Evaluation," *AL-ISHLAH: Jurnal Pendidikan*, vol. 17, no. 2, Jun. 2025, doi: 10.35445/alishlah.v17i2.7086.

[9] Y. Al Moaiad, M. Alobed, M. Alsakhnini, and W. Al-Haithami, "A Review of Automated Essay Scoring (AES)," *International Journal on Contemporary Computer Research (IJCCR)*, vol. 1, no. 1, pp. 68–86, 2025.

[10] S. Kumar, S. Chakrabarti, and S. Roy, "Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 2046–2052. doi: 10.24963/ijcai.2017/284.

[11] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.

[12] C. Brew and C. Leacock, "Automated short answer scoring: Principles and prospects," in *Handbook of automated essay evaluation*, Routledge, 2013, pp. 136–152.

[13] A. Sahu and P. K. Bhowmick, "Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance," *IEEE Transactions on Learning Technologies*, vol. 13, no. 1, pp. 77–90, Jan. 2020, doi: 10.1109/TLT.2019.2897997.

[14] H. A. Abdeljaber, "Automatic Arabic Short Answers Scoring Using Longest Common Subsequence and Arabic WordNet," *IEEE Access*, vol. 9, pp. 76433–76445, 2021, doi: 10.1109/ACCESS.2021.3082408.

[15] D. D. Prasetya, A. Prasetya Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, p. 63, Mar. 2018, doi: 10.26555/ijain.v4i1.152.

[16] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, Apr. 2013, doi: 10.5120/11638-7118.

[17] D. Ifenthaler, "Automated Essay Scoring Systems," in *Handbook of Open, Distance and Digital Education*, Singapore: Springer Nature Singapore, 2023, pp. 1057–1071. doi: 10.1007/978-981-19-2080-6_59.

[18] M. Hakiki and C. Fatichah, "Klasifikasi Kemampuan Mahasiswa Berdasarkan Automatic Essay Scoring terhadap Jawaban Essay Ujian Kompetensi dengan Metode Machine Learning," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 6, no. 3, pp. 1532–1546, Sep. 2025, doi: 10.63447/jimik.v6i3.1325.

[19] Y. Nie, "Automated essay scoring with SBERT embeddings and LSTM-Attention networks," *PeerJ Comput. Sci.*, vol. 11, p. e2634, Feb. 2025, doi: 10.7717/peerj-cs.2634.

[20] U. Hasanah, A. E. Permanasari, S. S. Kusumawardani, and F. S. Pribadi, "A scoring rubric for automatic short answer grading system," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 2, p. 763, Apr. 2019, doi: 10.12928/telkomnika.v17i2.11785.

[21] A. Wiratmo and C. Fatichah, "Indonesian Short Essay Scoring Using Transfer Learning Dependency Tree LSTM," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, pp. 278–285, Apr. 2020, doi: 10.22266/ijies2020.0430.27.

[22] U. Hasanah and B. P. Hartato, "Assessing Short Answers in Indonesian Using Semantic Text Similarity Method and Dynamic Corpus," in *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*, IEEE, Oct. 2020, pp. 312–316. doi: 10.1109/ICITEE49829.2020.9271696.

[23] R. A. Rajagede, "Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 11–18, Feb. 2021, doi: 10.22219/kinetik.v6i1.1196.

[24] F. F. Lubis *et al.*, "Automated Short-Answer Grading using Semantic Similarity based on Word Embedding," *International Journal of Technology*, vol. 12, no. 3, p. 571, Jul. 2021, doi: 10.14716/ijtech.v12i3.4651.

[25] T. Efendi, "PENILAIAN JAWABAN ESAI OTOMATIS MENGGUNAKAN KESAMAAN SEMANTIK DAN ANALISIS SINTAKSIS UNTUK MENDETEKSI URUTAN KATA PADA KALIMAT JAWABAN," Institute Teknologi Bandung. Accessed: Mar. 17, 2026. [Online]. Available: https://digilib.itb.ac.id/gdl/view/71402/23221047

[26] D. L. Butler and P. H. Winne, "Feedback and Self-Regulated Learning: A Theoretical Synthesis," *Rev. Educ. Res.*, vol. 65, no. 3, pp. 245–281, Sep. 1995, doi: 10.3102/00346543065003245.

[27] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: a model and seven principles of good feedback practice," *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218, Apr. 2006, doi: 10.1080/03075070600572090.

[28] M. Macías Borrego, "Towards a Digital Assessment: Artificial Intelligence Assisted Error Analysis in ESL," *Integrated Journal for Research in Arts and Humanities*, vol. 3, no. 4, pp. 76–84, Jul. 2023, doi: 10.55544/ijrah.3.4.10.

[29] M. Macías-Borrego, "ErrorAnalysisandArtificialIntelligence:PreliminaryexplorationofEnglishasaForeignLanguagewritenproductions," *I+D Revista de Investigaciones*, vol. 19, no. 1, pp. 45–55, Jun. 2024, doi: 10.33304/revinv.v19n1-2024004.

[30] M. A. Wijayanti, M. N. A. Asnur, and A. F. Syaputra, "AI as a Learning Partner: Error Analysis, Text Revision, and Student Perceptions in German Writing," *Interference: Journal of Language, Literature, and Linguistics*, vol. 6, no. 2, p. 166, Oct. 2025, doi: 10.26858/interference.v6i2.76536.

[31] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing," Aug. 2021.