

Analysis of Direct Scoring and Similarity-Based Scoring Approaches in Automatic Short Answer Scoring (ASAS)

Bayu Wicaksono^{1*}, Rasim², Yaya Wihardi³

^{1,2,3} Universitas Pendidikan Indonesia, Indonesia

¹bwbayuuu@upi.edu, ²rasim@upi.edu, ³yayawihardi@upi.edu



*Corresponding Author

Article History:

Submitted: 27-06-2025

Accepted: 30-06-2025

Published: 11-07-2025

Keywords:

Automatic Short Answer Scoring;
Cross-Prompt; Direct Scoring;
Outlier; Similarity-Based
Scoring; Specific-Prompt.

Brilliance: Research of

Artificial Intelligence is licensed
under a Creative Commons
Attribution-Non Commercial 4.0
International (CC BY-NC 4.0).

ABSTRACT

In the era of digital education, the need for automated scoring systems for short text answers has been steadily increasing. Automatic Short Answer Scoring (ASAS) aims to automate this assessment process with efficient and consistent approaches. Two commonly used approaches in ASAS are direct scoring and similarity-based scoring. Although these two approaches have been widely used, previous research has mostly focused on metrics like RMSE and Pearson Correlation to assess model performance. This study aims to provide a more in-depth analysis by comparing both approaches in two evaluation scenarios, specific-prompt and cross-prompt, by evaluating the accuracy and stability of the models. The dataset used in this study is the Rahutomo dataset. The results of the analysis show that direct scoring outperforms similarity-based scoring in terms of lower RMSE, higher Pearson Correlation, and fewer outliers. In the specific-prompt scenario, an RMSE of 0.0817 and a Pearson Correlation of 0.9504 were obtained, while in the cross-prompt scenario, the RMSE was 0.0917 and the Pearson Correlation was 0.9286. This study provides a more comprehensive insight into model performance by not only relying on evaluation metrics but also examining the distribution of residuals and outliers, which offers a more complete picture of model stability. Based on these findings, direct scoring is recommended for implementation in ASAS systems and for future research that can extend the analysis to other datasets or languages.

INTRODUCTION

In the era of digital education, the need for automated assessment systems has been increasing, particularly for efficiently, objectively, and consistently scoring short text answers on a large scale. Automatic Short Answer Scoring (ASAS) aims to automate this assessment process reliably. The commonly used approach is to compare student answers with pre-defined reference answers. However, the diversity of language structures and vocabulary in student responses makes the comparison process complex (Chamidah et al., 2023).

To address this issue, various Natural Language Processing (NLP) techniques have been used to extract semantic information from the text, which is then used in machine learning models. Recent research has adopted fine-tuning techniques on pre-trained models based on transformer architectures, such as BERT, which can generate meaningful vector representations despite variations in answers, by understanding the surrounding context (Devlin et al., 2018).

In fine-tuning-based model training, two commonly used approaches are direct scoring and similarity-based scoring. The direct scoring approach utilizes raw vector representations of student answers and reference answers to predict scores, assuming that these vectors contain sufficient semantic information (Kaya & Cicekli, 2024; Salim et al., 2022). On the other hand, the similarity-based scoring approach uses semantic similarity features, such as cosine similarity between the vector representations of student answers and reference answers, to predict scores, assuming that the level of semantic similarity is a key indicator in the assessment for ASAS models (Ahmed et al., 2022; Chamidah et al., 2023).

In addition to the differences in approaches, the choice of dataset splitting scheme also plays an important role in ASAS research. Two commonly used dataset splitting scenarios are specific-prompt and cross-prompt. The specific-prompt scenario evaluates the model's ability to score student answers that have not been seen before, while the cross-prompt scenario evaluates the model's ability to score answers from questions that have not been seen before. Both of these scenarios are important for measuring model performance in various real-world conditions.

Although various studies have discussed both approaches, the majority have focused on only one dataset splitting scenario. Some previous studies, such as (Ahmed et al., 2022; Salim et al., 2022), have focused more on evaluating the specific-prompt scenario. Recent research by (Ferreira Mello et al., 2025), have started to apply both scenarios. However, no similar studies have compared both BERT-based approaches in these two different scenarios within the same experimental environment, with a consistent dataset and data split.

This study is important because it provides systematic insights for selecting the more effective approach in both



scenarios and offers a holistic view of model performance. Previous studies have often focused on evaluating models based on metrics such as RMSE and Pearson Correlation, which are useful for measuring overall accuracy. However, these metrics do not assess model stability, particularly in terms of extreme predictions (outliers). In ASAS systems, the model's ability to generate highly extreme score prediction is crucial, as it indicates an unreliable model.

Therefore, this study conducts a more in-depth analysis of model performance. In addition to accuracy, this research also evaluates model stability through residual and outlier analysis, which has been less discussed in previous studies. Thus, this study provides a more comprehensive understanding of the model's ability to handle variations and extreme conditions in the data.

To evaluate model performance, this study uses three analysis methods, evaluation metrics to analyze overall model accuracy, residual analysis to identify outliers and assess model stability, and the distribution of outliers per question and the average absolute residual per question to evaluate the effectiveness of the model at the question level. The results of this study can serve as a solid foundation to maximize the best approach for each scenario in future research, as well as provide practical guidance for developers looking to integrate ASAS models into systems.

This study aims to analyze and compare two approaches across two evaluation scenarios, in order to gain empirical insights into the effectiveness of each approach in predicting student answer scores. By employing a more in-depth analysis approach, this research can contribute to the development of more flexible, stable, and effective ASAS systems under various conditions.

LITERATURE REVIEW

Previous studies provide various perspectives on the use of direct scoring and similarity-based scoring in ASAS systems. Salim et al. (2022) used the direct scoring approach in the specific-prompt scenario by comparing several variants of the IndoBERT model on the Rahutomo dataset. Based on the experimental results on the Rahutomo dataset, the best model was indobert-lite-base-p2, which achieved a Pearson correlation of 0.9483 and an RMSE of 0.4277 (on a 0-5 score scale). Another study by Kaya & Cicekli (2024) used the direct scoring approach in the specific-prompt scenario with a hybrid architecture that combines BERT, CNN, and Bi-LSTM. However, the experimental results on the Mohler dataset showed that this hybrid model was unable to outperform the BERT Regressor architecture. The hybrid model produced a Pearson correlation of 0.747 and an RMSE of 0.852, while the BERT Regressor achieved a Pearson correlation of 0.777 and an RMSE of 0.732 (on a 0-5 score scale).

Another study by Ahmed et al. (2022) used the similarity-based scoring approach in the specific-prompt scenario by comparing several sentence-BERT models on the Mohler dataset. The best model based on the experiments was all-distilroberta-v1, which achieved a Pearson correlation of 0.82 and an RMSE of 0.69 (on a 0-5 score scale). The study by Haidir & Purwarianti (2020) also used the similarity-based scoring approach in the cross-prompt scenario by comparing word embeddings such as Word2Vec and contextual embeddings from several BERT models. The experimental results showed that the combination of contextual embeddings using the bert-large variant with element-wise multiplication provided the best performance, with an MAE of 0.547 and an RMSE of 0.922, compared to Word2Vec with concatenation, which resulted in an MAE of 0.84 and an RMSE of 1.234 (on a 0-5 score scale).

A recent study by Ferreira Mello et al. (2025) adopted the GPT-4 prompt engineering method in two scenarios, specific-prompt and cross-prompt. The experimental results showed that the few-shot method produced the best performance. In the specific-prompt scenario, the model achieved an MAE of 1.02 and an RMSE of 1.48, while in the cross-prompt scenario, the model obtained an MAE of 0.69 and an RMSE of 1.01. Although the results were intriguing, this study primarily focused on evaluation metrics and did not delve into model stability, particularly regarding the emergence of extreme predictions (outliers).

The study by Chamidah et al. (2023) used the similarity-based scoring approach in the specific-prompt scenario, comparing several sentence-BERT models. Two distance functions were tested, Manhattan LSTM and cosine similarity. Based on the experiments, the combination of the distiluse-base-multilingual-cased-v2 model with Manhattan LSTM showed the best performance, achieving a Pearson correlation of 0.92 and an RMSE of 11.26 (on a 0-100 score scale) on the Rahutomo dataset. This study not only considered evaluation metrics but also conducted a qualitative analysis by examining the test data and its score predictions.

Although these various studies contribute significantly to the development of ASAS, the majority have focused solely on evaluating metrics such as RMSE and Pearson Correlation, without providing in-depth analysis of model stability. Additionally, most studies only tested models in a single dataset scenario, either specific-prompt or cross-prompt, but did not test both in a uniform experimental setup. This gap is what this study aims to address, aiming to provide a thorough analysis of model accuracy and stability across both dataset scenarios to ensure its reliability in various conditions.

METHOD

Research design refers to the stages, plan, or procedures involved in conducting research. The research design used is based on the method proposed by Santoso et al. (2020), which can be seen in Fig. 1. This research design consists of data collection, data preprocessing, model development, model evaluation, and analysis and conclusion.



Detailed explanations for each stage are provided as follows.

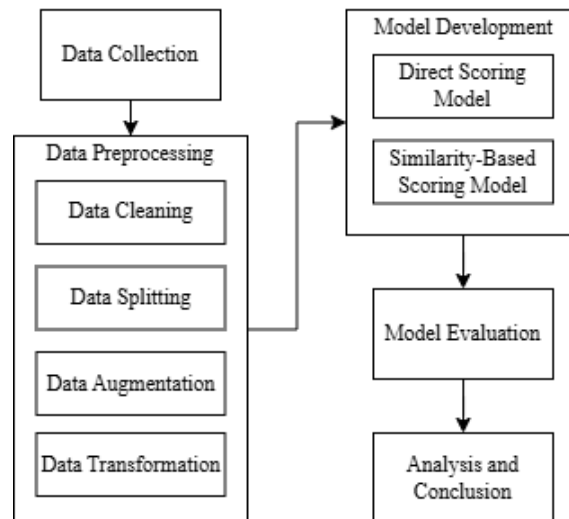


Figure 1. Research Design

Data Collection

This dataset is a collection of essay grading data in Indonesian from Rahutomo et al. (2018). The dataset consists of 40 questions divided into four categories, politics, lifestyle, sports, and technology, with 10 questions in each category. Each question is accompanied by a reference answer and answers from approximately 50 students. The total available data is 2,162, with scores from three evaluators using a 0–100 scale, where 100 represents the most relevant answer and 0 represents a completely wrong answer. This dataset meets the data characteristics required for the study, namely, it includes columns for reference answers, student answers, and numerical score labels.

Data Cleaning

In this study's dataset, a label column was created by calculating the average score from the three evaluators and normalizing the scores to a 0-1 scale. In the student answers and reference answer columns, duplicate data was removed, inconsistent scores were addressed, punctuation was eliminated, and all text was converted to lowercase. After the cleaning process, the dataset contained 2,008 entries.

Data Splitting

At this stage, the data was split into training, validation, and testing sets with an 80:10:10 ratio for two scenarios, specific-prompt and cross-prompt. The specific-prompt scenario involved splitting the data based on student answers for each question, while the cross-prompt scenario involved splitting the data based on the questions. In the cross-prompt scenario, there are 1,618 training data points from 32 questions, 182 validation data points from 4 questions, and 208 testing data points from 4 questions. In the specific-prompt scenario, there are 1,588 training data points, 200 validation data points, and 220 testing data points.

Data Augmentation

The distribution of scores after the data split was inconsistent, with most of the data falling within the low to medium score range. Therefore, data augmentation was applied at this stage. The study applied a prompt engineering-based data augmentation method by replacing some words in the student answers with semantically relevant synonyms (Ferreira Mello et al., 2025). This process was performed at the word level using the Gemma 3 model (4B parameter variant). After augmentation, in the specific-prompt scenario, the number of training data points before augmentation was 1,588, and after augmentation, it became 3,892. In the cross-prompt scenario, the number of training data points before augmentation was 1,618, and after augmentation, it became 3,810.

Data Transformation

In this study, data transformation was applied to the student answers and reference answers columns by converting them into a format compatible with the model's requirements, followed by tokenization and encoding. In the direct scoring approach, the reference answer and student answer were merged into a single text input in the format (Kaya & Cicekli, 2024). "[CLS] reference answer [SEP] student answer [SEP]." This format was used because this approach only accepts one text input, so merging the answers allows the model to consider both parts of the text simultaneously. Tokenization was performed using BertTokenizer with a max_length of 512 and padding. Then, encoding was performed to generate three input feature components, input_ids, attention_mask, and token_type_ids.



In the similarity-based scoring approach, the reference answer and student answer were not merged into one text but were processed separately. Each text was packaged in the input format “[CLS] answer [SEP].” This format was used because the architecture of the model used accepts two separate texts as inputs. Tokenization was done using AutoTokenizer with a max_length of 512 and padding. Encoding was then performed to generate two input feature components, input_ids and attention_mask. Token_type_ids were not used because this approach does not require information regarding the distinction between segments in the input text.

Model Development

This study uses two approaches, direct scoring and similarity-based scoring. Detailed explanations of each approach are provided below.

Direct Scoring Model Approach

The direct scoring approach uses the IndoBERT model, specifically the indobert-lite-base-p2 variant. IndoBERT is a model trained using the BERT architecture on 220 million Indonesian words obtained from various sources such as Wikipedia, news, and the Indonesian web corpus. This model undergoes the same training stages as BERT, specifically in Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks (Koto et al., 2020).

In this approach, the model receives a single input formed by combining the student answer and reference answer, which is processed by the tokenizer into three features, input_ids, attention_mask, and token_type_ids. These three features serve as input to the IndoBERT encoder. The embedding from the pooling layer produced by the encoder is used as a semantic representation of the entire input and passed to the dense (regressor) layer to produce the predicted score. This approach allows for end-to-end training, where the feature extraction by the encoder and prediction by the regressor occur simultaneously in one architecture. The architecture of this approach is shown in Fig. 2.

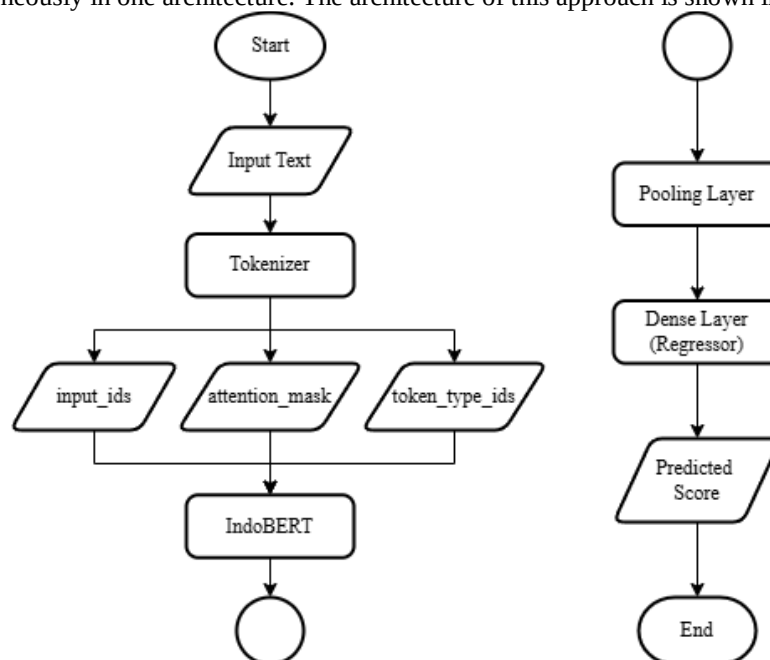


Figure 2. Architecture of the Direct Scoring Approach with IndoBERT

In this study, various hyperparameters were used, such as a batch size of 16, a learning rate of 2e-5, and a maximum of 100 epochs. The maximum number of epochs was set to ensure that the model reached convergence. However, to avoid overfitting due to the large number of epochs, early stopping was also applied to halt training if the validation loss did not show further reduction after 5 epochs. The optimizer used was AdamW with a weight decay of 0.01. In addition to hyperparameters, in the specific-prompt scenario, the pooling strategy used was CLS with a classifier dropout of 0.3. In the cross-prompt scenario, Attention Pooling was used with a classifier dropout of 0.1.

Similarity-Based Scoring Model Approach

The similarity-based scoring approach uses the Sentence-BERT model with the variant sentence-transformers/distiluse-base-multilingual-cased-v2. Sentence-BERT is a modification of the BERT architecture that employs siamese and triplet networks to generate sentence-level embeddings with semantic meaning. The siamese network allows the parallel processing of two input texts with shared weights (Reimers & Gurevych, 2019).

The input consists of two separate texts, the reference answer and the student answer. Both inputs are processed in parallel using the same encoder (shared weights), producing two fixed representation vectors through a pooling layer.



The embeddings of both texts are then compared using cosine similarity, which represents the semantic similarity between the student answer and the reference answer. This similarity value is used as a feature in a linear regression model to predict the final score.

The training process consists of two stages, fine-tuning the Sentence-BERT model on the manual scores and training the linear regression model using the cosine similarity features. Fig. 3 displays the architecture of the Sentence-BERT encoder. The hyperparameters and optimizer used in this approach are the same as those in the direct scoring approach. In addition to the hyperparameters, in the specific-prompt scenario, the pooling strategy used is Mean Pooling. In the cross-prompt scenario, CLS is used.

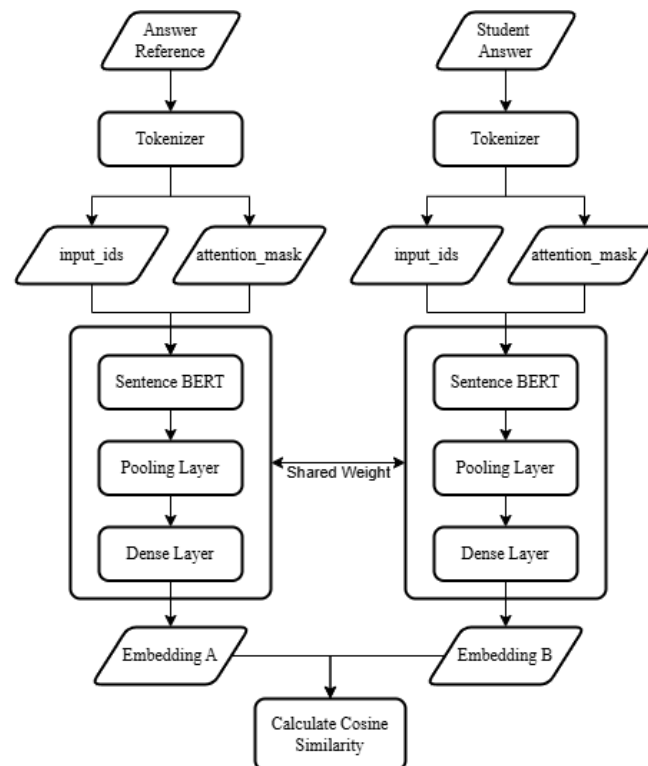


Figure 3. Architecture of the Sentence-BERT Encoder in the Similarity-Based Scoring Approach

Evaluation Scenarios

To evaluate the model's generalization ability, two evaluation scenarios were used, specific-prompt and cross-prompt. In the specific-prompt scenario, the data is split based on the student answers for each question. This scenario evaluates the model's ability to assess new answers for questions that have appeared during training. In contrast, in the cross-prompt scenario, the data is split based on the questions. This scenario evaluates the model's ability to predict scores for answers to questions that have not been seen before.

Model Evaluation

In this study, model evaluation was conducted through analysis of evaluation metrics and residuals to detect outliers.

1. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a metric that measures the difference between the values predicted by the model and the actual values. In RMSE, the closer the score is to 0, the better the model is at predicting the actual value (Mardini G et al., 2024).

2. Pearson Correlation Coefficient

Pearson's Correlation Coefficient (Pearson's r) is a metric used to measure the linear relationship between two variables, such as predicted values and actual values. Pearson's r values range from $[-1, 1]$, where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation (Wijanto & Yong, 2024).

3. Residual

Residuals are used to identify extreme prediction errors or outliers. A residual represents the difference between the predicted value and the actual value. In this study, residuals are calculated as absolute differences, so the results are always positive. Data is considered an outlier if it has an absolute residual value that exceeds a certain

threshold (Verma, 2025).

Two thresholds are used, 0.1 and 0.2. The value of 0.1 was selected because it is close to the third quartile (Q3) of the residual distribution, representing the upper limit of frequent prediction errors. The value of 0.2 was chosen because it is close to the upper limit of the IQR method, which represents a stricter threshold for detecting truly extreme prediction errors (Amin et al., 2022).

Models that generate fewer outliers are considered more stable because they tend to provide predicted scores closer to the actual scores. Additionally, a fixed threshold across all models is used to ensure that the evaluation process between models is fair and balanced.

RESULT

The evaluation was conducted separately for the two scenarios, specific-prompt and cross-prompt, by comparing two model approaches. Each evaluation includes analysis of accuracy metrics, residual/outlier analysis, and the distribution of outliers per question.

Direct Scoring vs Similarity-Based Scoring in the Specific-Prompt Scenario

Model performance evaluation was conducted by comparing two main metrics, Root Mean Squared Error (RMSE) and Pearson Correlation. Table 1 shows the comparison of evaluation metrics and the number of outliers from both approaches on the specific-prompt test data.

Table 1. Comparison of Evaluation Metrics and Number of Outliers for Each Approach (Specific-Prompt)

Approach	RMSE ↓	Pearson ↑	Threshold 0.2	Threshold 0.1
Direct Scoring	0.0817	0.9504	4	42
Similarity-based Scoring	0.0996	0.9285	16	50

To evaluate extreme errors in more detail, an analysis of the outlier distribution was also conducted in the specific-prompt scenario. Fig. 4(a) shows the distribution of the number of outliers per question in the specific-prompt scenario. Fig. 4(b) displays the average absolute residual per question, which is used to reinforce the results. The average absolute residual provides an overview of how far the model's predictions deviate from the actual scores.

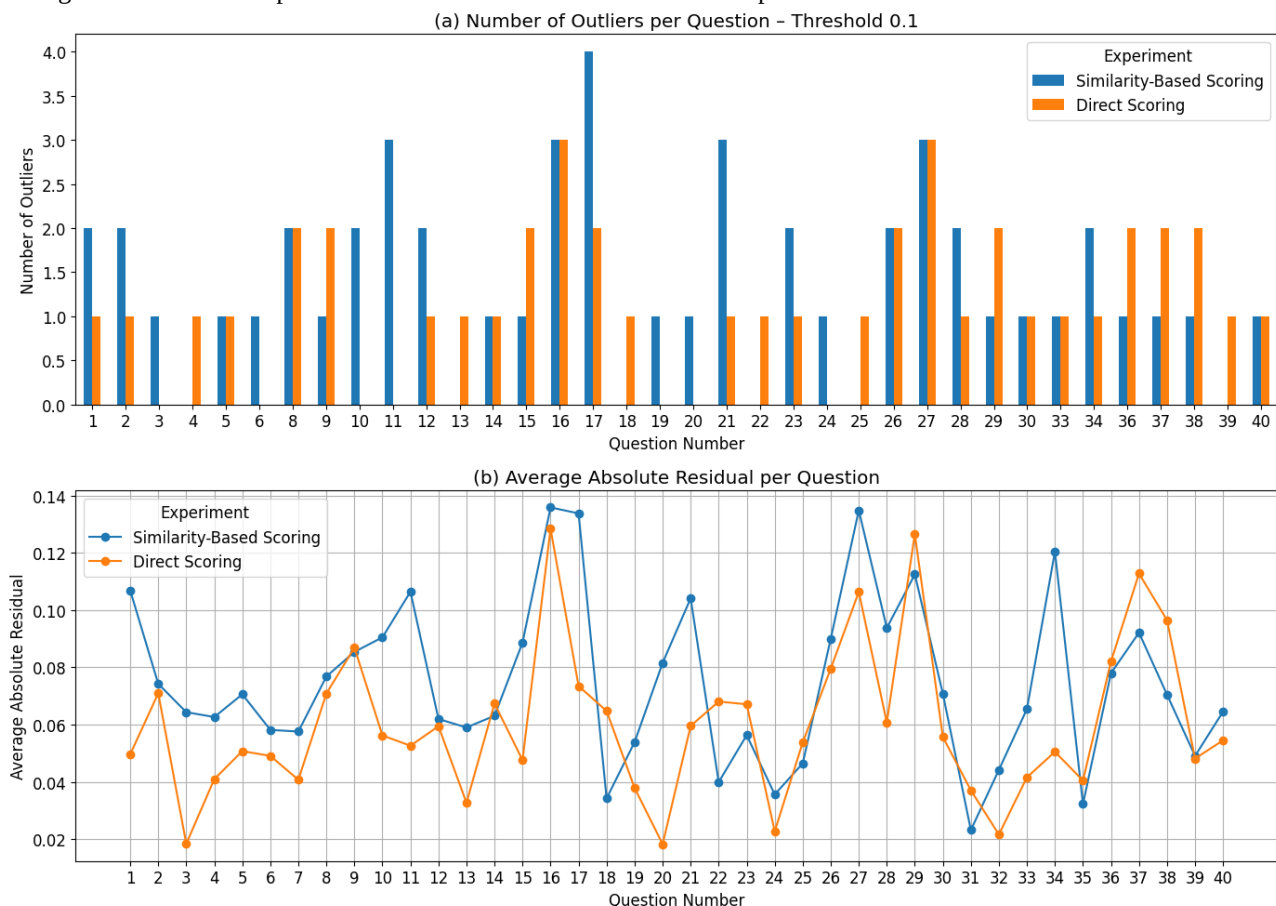


Figure 4. Comparison of the number of outliers per question at the 0.1 threshold (a) and the average absolute residual per question (b) for all models in the specific-prompt scenario.

Direct Scoring vs Similarity-Based Scoring in the Cross-Prompt Scenario

Model performance evaluation in the cross-prompt scenario was also conducted by comparing two main metrics, Root Mean Squared Error (RMSE) and Pearson Correlation. Table 2 shows the comparison of evaluation metrics and the number of outliers from both approaches on the cross-prompt test data.

Table 2. Comparison of Evaluation Metrics and Number of Outliers for Each Approach (Cross-Prompt)

Approach	RMSE ↓	Pearson ↑	Threshold 0.2	Threshold 0.1
Direct Scoring	0.0917	0.9286	8	50
Similarity-based Scoring	0.1238	0.8766	19	88

To evaluate extreme errors in more detail, an analysis of the outlier distribution was also conducted in the cross-prompt scenario. Fig. 5(a) and 5(b) show the distribution of the number of outliers per question in the cross-prompt scenario. Fig. 5(c) displays the average absolute residual per question, which is used to reinforce the results.

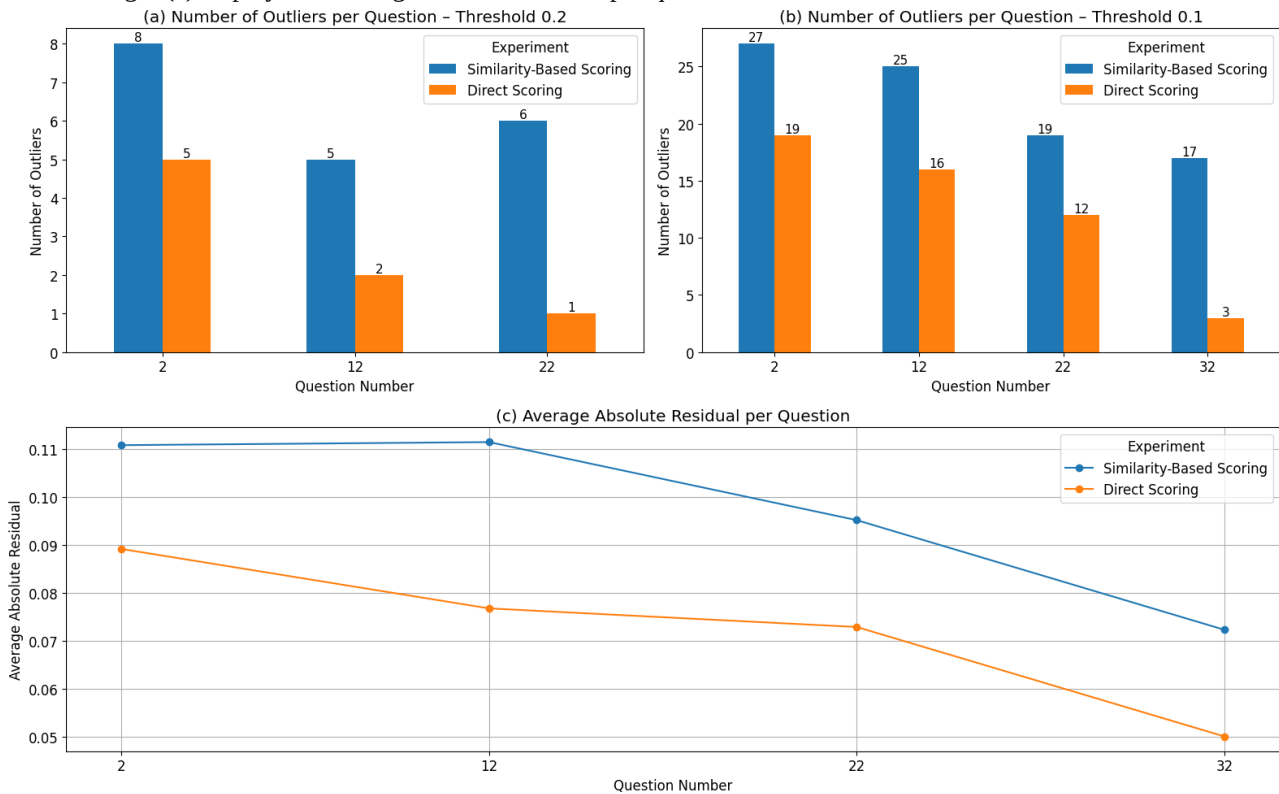


Figure 5. Comparison of the number of outliers per question at the thresholds of 0.2 (a) and 0.1 (b), as well as the average absolute residual per question (c), for all models in the cross-prompt scenario.

DISCUSSION

This chapter presents a detailed discussion of the experimental results above, divided into two subsections, comparing the two approaches in the specific-prompt and cross-prompt scenarios. Each subsection is divided into three parts, analysis of evaluation metrics, outliers at both threshold levels, and the distribution of outliers per question.

Direct Scoring vs Similarity-Based Scoring in the Specific-Prompt Scenario

Based on Table 1, direct scoring outperforms similarity-based scoring. This is marked by a 17.97% reduction in RMSE and a 2.36% increase in Pearson Correlation. The reduction in RMSE indicates that direct scoring predicts the scores more accurately, while the increase in Pearson Correlation shows a stronger relationship between the predicted and actual scores.

In terms of stability in Table 1, direct scoring also has a lower proportion of outliers at both tested thresholds. The direct scoring approach has 1.82% outliers at the 0.2 threshold and 19.09% outliers at the 0.1 threshold. In contrast, similarity-based scoring shows 7.27% outliers at the 0.2 threshold and 22.73% outliers at the 0.1 threshold. These results suggest that direct scoring is more stable in generating scores, while similarity-based scoring is more prone to fluctuations.

Based on the distribution of the number of outliers per question in Figure 4 and the average absolute residuals in Fig. 5, it can be seen that both approaches perform effectively on certain questions. Questions 31, 32, and 35 did not generate outliers at the 0.1 threshold for both approaches, indicating that both approaches were effective on those



questions. This result is reinforced by the average absolute residual values for these three questions, which were less than 0.04 for both approaches. However, questions 16 and 27 show similar levels of difficulty for both approaches, as they produced the same number of outliers (3) and had average absolute residuals greater than 0.1 for both approaches.

Some questions, such as numbers 3, 20, and 21, show an advantage for the direct scoring approach, with minimal outliers (between 0 and 1), while similarity-based scoring shows 2 to 3 outliers. The difference in average absolute residuals between the two approaches on these three questions is also quite large, greater than 0.02. However, questions 18, 22, and 37 appear to be more effective for similarity-based scoring, as no outliers were found on these questions, even though direct scoring produced only one outlier. The difference in average absolute residuals between the two approaches on these three questions is also significant, more than 0.04. Similarity-based scoring excels on questions with factual, standardized answers that do not have much variation, such as the text of the proclamation (question number 22) or abbreviations (question number 37). On the other hand, direct scoring is more effective on free-form questions, as it is better able to capture variations in context and sentence structure through raw vector representations.

Based on accuracy and model stability, the direct scoring approach proves superior compared to similarity-based scoring. The key advantage of direct scoring lies in the use of raw vector representations produced by the model for the regression layer, which is more effective in predicting scores than the semantic similarity features between two vector representations. These raw vector representations provide richer and more relevant features, enabling the model to better capture context and sentence structure. In contrast, similarity-based scoring is limited to the similarity value between representations, which is less capable of capturing broader context, making it more prone to errors in predictions.

Overall, the results indicate that the direct scoring approach significantly outperforms similarity-based scoring in terms of model stability and prediction accuracy. However, both approaches have their strengths on certain questions. Table 3 compares the results of both approaches with other studies using the Rahutomo dataset in the specific-prompt scenario.

Table 3. Comparison of Both Approaches with Models from Other Studies

Approach	RMSE ↓	Pearson ↑
Direct Scoring	0.0817	0.9504
Similarity-based Scoring	0.0996	0.9285
(Salim et al., 2022) (on a 0-1 score scale)	0.08554	0.9483
(Chamidah et al., 2023) (on a 0-1 score scale)	0.1065	0.92

Although the direct scoring approach shows superior performance, there are some limitations to this study. The model was trained and tested on the Rahutomo dataset, which is in Indonesian, and the specific-prompt scenario, so generalization to other datasets or languages still needs to be further validated. Additionally, for some questions or answers, the performance of direct scoring can be lower compared to similarity-based scoring.

Direct Scoring vs Similarity-Based Scoring in the Cross-Prompt Scenario

In the cross-prompt scenario, the direct scoring approach again shows better performance compared to similarity-based scoring. Based on Table 2, the direct scoring approach experienced a 25.93% reduction in RMSE and a 5.93% increase in Pearson Correlation, indicating that direct scoring is effective in predicting scores even in different question contexts. This result aligns with the findings in the specific-prompt scenario, where direct scoring excels in capturing the relationship between the answers and scores. Table 2 shows that direct scoring consistently delivers accurate results on both evaluation metrics used.

This advantage is also visible in the number of outliers. Based on Table 2, direct scoring is better at minimizing prediction errors compared to similarity-based scoring, which is more prone to outliers. In terms of the proportion of outliers, direct scoring has 3.85% outliers at the 0.2 threshold and 24.04% outliers at the 0.1 threshold, while similarity-based scoring has 9.13% outliers at the 0.2 threshold and 42.31% outliers at the 0.1 threshold. This suggests that direct scoring is more stable in providing accurate scores, even in different question contexts.

Figure 5. shows that direct scoring generally generates fewer outliers across all questions. For example, in question number 2, which recorded the highest number of outliers in both approaches, direct scoring still maintained its more stable performance, producing 19 outliers at the 0.1 threshold, while similarity-based scoring produced 27 outliers. Meanwhile, in question number 32, which seemed easier to predict, direct scoring only generated 3 outliers at the 0.1 threshold, while similarity-based scoring produced 17 outliers, demonstrating that direct scoring is more stable and accurate in generating correct predictions. Based on the average absolute residual in Fig. 5, the difference between the two approaches is around 0.02, reinforcing the superiority of direct scoring.

These results show that the model's ability to generalize to different question contexts depends on the features used for score prediction. The direct scoring approach, which directly maps the vector representations of answers to scores, proves to be more adaptive in capturing semantic complexity and structural variations across questions. This allows the model to remain accurate even when faced with unseen contexts. In contrast, the similarity-based scoring approach heavily relies on the similarity between the student's answer and the reference answer, making its



performance significantly worse when the structure or language style changes. This difference indicates that raw vector representations are overall superior in scenarios that demand high generalization ability.

However, this study has some limitations. First, it used the Rahutomo dataset in Indonesian, and no other studies have used the same dataset in the cross-prompt scenario for comparison. Therefore, further validation is needed to ensure the generalization of the model to other datasets or languages in the cross-prompt scenario.

CONCLUSION

Based on the evaluation metrics and residual analysis, the direct scoring approach consistently shows superior performance compared to the similarity-based scoring approach, both in the specific-prompt and cross-prompt scenarios. This superiority is marked by low RMSE values, high Pearson correlation, and fewer outliers at various error thresholds. These results indicate that the use of raw vector representations provides richer semantic information, enabling the model to predict scores more accurately. Therefore, the direct scoring approach is recommended for use in ASAS systems and for the future development of ASAS models. This study not only proves the effectiveness of the direct scoring approach in ASAS but also opens opportunities for building more adaptive and reliable ASAS systems in the future.

However, the similarity-based scoring approach still holds potential for certain questions. Future research is recommended to explore both approaches on datasets in other languages and analyze the relationship between question characteristics and approach effectiveness. If certain approaches prove to be more optimal for specific question characteristics, the development of an adaptive system that can recognize question characteristics and select the appropriate approach would be a promising direction for future research.

REFERENCES

- Ahmed, A., Joorabchi, A., & Hayes, M. J. (2022). On the Application of Sentence Transformers to Automatic Short Answer Grading in Blended Assessment. *2022 33rd Irish Signals and Systems Conference, ISSC 2022, June 2022*, 1–6. <https://doi.org/10.1109/ISSC55427.2022.9826194>
- Amin, M., Afzal, S., Akram, M. N., Muse, A. H., Tolba, A. H., & Abushal, T. A. (2022). Outlier detection in gamma regression using Pearson residuals: Simulation and an application. *AIMS Mathematics*, 7(8), 15331–15347. <https://doi.org/10.3934/math.2022840>
- Chamidah, N., Yulianti, E., & Budi, I. (2023). Evaluating the Impact of Sentence Tokenization on Indonesian Automated Essay Scoring Using Pretrained Sentence Embeddings. *Revue d'Intelligence Artificielle*, 37(5), 1101–1108. <https://doi.org/10.18280/ria.370502>
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019, Mlm*, 4171–4186. <https://aclanthology.org/N19-1423.pdf>
- Ferreira Mello, R., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., & Gasevic, D. (2025). Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models? *15th International Conference on Learning Analytics and Knowledge, LAK 2025*, 93–103. <https://doi.org/10.1145/3706468.3706481>
- Haidir, M. H., & Purwarianti, A. (2020). Short Answer Grading Using Contextual Word Embedding and Linear Regression. *Jurnal Linguistik Komputasional*, 3(2), 54–61. <https://inacil.id/journal/index.php/jlk/article/view/38>
- Kaya, M., & Cicekli, I. (2024). A Hybrid Approach for Automated Short Answer Grading. *IEEE Access*, 12(May), 96332–96341. <https://doi.org/10.1109/ACCESS.2024.3420890>
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757–770. <https://doi.org/10.18653/v1/2020.coling-main.66>
- Mardini G, I. D., Quintero M, C. G., Vilorio N, C. A., Percybrooks B, W. S., Robles N, H. S., & Villalba R, K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies*, 29(4), 4565–4590. <https://doi.org/10.1007/s10639-023-11890-7>
- Rahutomo, F., Ari Roshinta, T., Rohadi, E., Siradjuddin, I., Ariyanto, R., Setiawan, A., & Adhisuwignjo, S. (2018). Open Problems in Indonesian Automatic Essay Scoring System. *International Journal of Engineering & Technology*, 7(4.44), 156. <https://doi.org/10.14419/ijet.v7i4.44.26974>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. <https://doi.org/10.18653/v1/d19-1410>
- Salim, H. R., De, C., Pratamaputra, N. D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. *Bulletin of Electrical Engineering and Informatics*, 11(3), 1586–1603. <https://doi.org/10.11591/eei.v11i3.3531>



-
- Santoso, R. R., Megasari, R., & Hambali, Y. A. (2020). Implementasi Metode Machinelearning. *Jurnal Aplikasi Dan Teori Ilmu Komputer*, 3(2), 85–97. <https://ejournal.upi.edu/index.php/JATIKOM>
- Verma, V. (2025). *A Comprehensive Framework for Residual Analysis in Regression and Machine Learning*. January. <https://doi.org/10.52783/jisem.v10i31s.4958>
- Wijanto, M. C., & Yong, H. S. (2024). Combining Balancing Dataset and SentenceTransformers to Improve Short Answer Grading Performance. *Applied Sciences (Switzerland)*, 14(11). <https://doi.org/10.3390/app14114532>