IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON IEEE EDUCATION SOCIETY SECTION

Received 4 March 2025, accepted 12 April 2025, date of publication 18 April 2025, date of current version 16 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3562493

RESEARCH ARTICLE

Transformer-LSTM Models for Automatic Scoring and Feedback in English Writing Assessment

YUNYUN XUAN
Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Kuala Lumpur 43400, Malaysia
No.2 Middle School of Guiding, Guiyang, Guizhou 551300, China

e-mail: gs58977@student.upm.edu.my

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Article 32 of Measures for Ethical Review of Life Sciences and Medical Research Involving Human Beings of China with the Declaration of Helsinki.

ABSTRACT Writing assessment is one of the most important stages in the educational process, but it is also the most resource-demanding one. To address the challenges of scalability and inconsistency, this study proposes a Transformer-LSTM model for automated scoring and feedback generation, enhancing accuracy and reliability in assessment. Integrating the contextual reading abilities of transformers with the sequential analysis strength of LSTMs, the model analyzes significant metrics of writing quality, including grammar, coherence, and structure, while providing individualized, actionable feedback. Using annotated datasets and evaluation metrics like RMSE and feedback relevance, it was established that the model performs well overall and that improvements in grammar and coherence seemed to be the most significant contributors to writing ability. It was also demonstrated that feedback relevance enhances these outcomes, thus confirming its valuable role in promoting structural and grammatical accuracy. Understanding that most existing systems do not encourage significant human feedback, this work demonstrates a scalable approach with potential alignment with human evaluation standards. Finally, this study shows hybrid models' promise for automated writing assessment as promising scalable, equitable, impact-based tools for global enhancement of educational outcomes.

INDEX TERMS Automated writing assessment, feedback generation, grammar and coherence scoring, educational AI, transformer-LSTM models.

I. INTRODUCTION

Writing assessments have always relied heavily on human effort to provide quality feedback on student work. However, with the rise of Artificial Intelligence (AI), educational assessment has transformed, offering scalable, efficient, and consistent alternatives to manual grading. Among these innovations, Transformer-based models and Long Short-Term Memory (LSTM) architectures have proven valuable in capturing complex linguistic patterns in text. This study combines these two architectures to develop a Transformer-LSTM model aimed at automated scoring and feedback generation, seeking to replicate and enhance human evaluation practices in writing assessment. By leveraging the contextual capabilities of Transformers and the sequential processing power of LSTMs, the model aims to deliver precise scores and actionable feedback to improve students' writing.

This research aims to create and test a hybrid system that assesses various aspects of writing, such as grammar, coherence, and overall structure, while providing personalized feedback tailored to each student. Unlike traditional automated systems, which focus primarily on numerical scoring, this study emphasizes the importance of feedback relevance and constructiveness to drive meaningful improvements in writing. The model is trained on annotated student essays and peer feedback datasets, learning to analyze linguistic features and evaluate writing quality. It then suggests improving grammar, coherence, clarity, and overall structure. Evaluation metrics such as scoring accuracy (using RMSE) and feedback

The associate editor coordinating the review of this manuscript and approving it for publication was S. Chandrasekaran.

quality (assessing relevance and constructiveness) are used to assess the model's performance thoroughly.

The need for scalable and effective writing assessment systems arises from the limitations of human grading, including subjectivity, inconsistency, and time constraints. While current automated scoring systems are efficient, they often lack personalized, actionable feedback that supports student learning. By combining the strengths of Transformers and LSTMs, this study aims to bridge this gap, offering a comprehensive solution that meets pedagogical needs and the advancements in natural language processing.

A. RESEARCH QUESTIONS

This study will answer these questions.

- How can a Transformer-LSTM model be designed to better replicate and enhance human evaluation in writing assessments?
- Can such a model provide accurate scores and constructive feedback that match or even surpass human evaluation?
- What are the most effective ways to assess the performance of automated scoring systems, especially in terms of feedback relevance and usefulness?

This study's methodology includes data collection, model development, training, and evaluation. A multi-task learning approach enables the hybrid model to generate scores and feedback simultaneously while focusing on key writing quality criteria such as coherence, argumentation, and grammar. The Transformer component captures long-range dependencies and contextual relationships within essays, whereas the LSTM layer examines the sequential flow of sentences and paragraphs. To validate the model's effectiveness, it is compared with human-assigned scores and other existing systems, like Grammarly, showing that it outperforms current automated solutions in scoring and feedback relevance.

The expected contributions of this study are twofold. First, the Transformer-LSTM model is anticipated to surpass traditional automated scoring systems in accuracy and feedback quality, providing a scalable and adaptable solution for various writing tasks and genres. Second, by focusing on generating actionable feedback, this research addresses a critical gap in automated writing assessment, emphasizing the importance of feedback in improving writing skills. The findings aim to demonstrate the transformative potential of hybrid models in education, offering an effective and efficient alternative to traditional writing assessment methods.

II. LITERATURE

A. EVOLUTION OF AUTOMATED ESSAY SCORING

Automated Essay Scoring (AES) systems were developed to address the challenges of traditional human grading, such as subjectivity and high resource costs [1], [2]. Early AES models mainly relied on manually crafted linguistic features to assess writing quality, focusing on syntax, semantics, and fluency [3]. However, these early models could not often provide personalized feedback, which is critical for helping students improve [4]. Integrating machine learning (ML) techniques marked a significant step forward in AES. Traditional ML models used linguistic features to predict essay quality, but their ability to capture the complexity of human language was limited [5], [6]. The introduction of deep learning, primarily through neural networks, allowed AES systems to detect complex patterns in text data, improving both scoring accuracy and feedback relevance [7], [8], [9].

Transformer-based models, introduced by Vaswani et al. (2017), revolutionized natural language processing by capturing long-range dependencies in text through self-attention mechanisms. These models have been successfully applied to AES, demonstrating superior performance in understanding the contextual nuances of essays [11], [12]. Their ability to process entire text sequences simultaneously allows for a more comprehensive analysis of writing quality. Long Short-Term Memory (LSTM) networks, developed by Hochreiter and Schmidhuber [13], excel at modeling sequential data, which makes them well-suited for tasks that involve temporal dependencies. In AES, LSTMs have been used to evaluate how ideas progress and connect throughout essays, leading to more accurate assessments of writing structure [14], [15].

Recent research has explored combining Transformer architectures with LSTM networks to take advantage of the strengths of both approaches. These hybrid models enhance AES by capturing global context and sequential relationships, improving scoring accuracy and feedback generation [16]. Combining these two methods leads to a more well-rounded understanding of student essays. Despite these advancements, AES remains challenging in providing actionable and personalized feedback. Many systems still offer generic comments that do not effectively guide students' improvement [4], [17]. Overcoming this challenge requires models delivering specific, constructive feedback tailored to individual writing needs [18], [19]. The use of large language models (LLMs) in AES has shown promise, especially for evaluating essays written by nonnative speakers [20], [21].

Finally, ensuring coherence and conciseness in the evaluation process has become a significant focus in AES research [22]. A model's ability to provide accurate scores and high-quality feedback is essential for successful use in educational settings.

The application of machine learning (ML) in Automated Essay Scoring (AES) has significantly advanced the field by enabling systems to automatically analyze linguistic features like syntax, semantics, and fluency in student writing [3], [5]. Early ML models used features such as sentence length, vocabulary diversity, and grammatical structure to estimate essay quality. These early systems relied on hand-crafted features, but with deep learning, particularly neural networks, performance improved as models began learning complex representations of text data independently.

Deep learning techniques, including Long Short-Term Memory (LSTM) networks and Transformer models, have

further enhanced AES systems by capturing long-range dependencies and complex text structures. This shift from feature-based models to neural networks has allowed systems to replicate human grading [14] more closely. Unlike previous models that processed text sequentially, Transformers analyze entire input sequences simultaneously, improving efficiency and understanding complex linguistic structures. This ability has made Transformers particularly effective in AES, outperforming traditional ML models in capturing the intricate relationships between words and phrases [11], [12].

Thanks to their contextual understanding, transformers have demonstrated their potential for providing more accurate feedback. For example, Abraha and Nazir [11] showed how Transformer-based models could provide both essay scoring and detailed feedback, capturing nuances often missed by rule-based systems. Their research highlights how Transformers can offer more refined assessments, setting the stage for better-automated grading systems.

While Transformer models excel at capturing global context, Long Short-Term Memory (LSTM) networks are better suited for modeling the sequential flow of ideas in text. Developed by [13], LSTMs are ideal for understanding how ideas evolve across longer passages. This makes them particularly effective in writing assessments, where essay structure and idea progression are crucial to evaluating quality [16], [23]. LSTM networks are used in AES systems to ensure coherent flow and logical progression within essays. Recent studies suggest that combining Transformer and LSTM models can enhance AES performance.

B. CHALLENGES IN FEEDBACK GENERATION AND SYSTEM PERFORMANCE

While Transformer-LSTM models show great potential, they still face challenges in generating actionable feedback. While these systems excel at providing accurate scores, they often struggle to offer relevant and valuable feedback for improving student writing [4], [18]. A key limitation of many current AES systems is their reliance on pre-defined rubrics. While these rubrics provide consistency, they do not consider the unique aspects of individual student writing or offer tailored suggestions for improvement.

Recent research has sought to address this gap by developing models that assess essay quality and generate personalized feedback. Ouahrani and Bennouar [24] proposed a paraphrase generation and supervised learning method to enhance automatic short-answer grading. This concept could be adapted for extended essays to provide more meaningful feedback. Similarly, Pinto and Paquette [25] explored the use of deep learning techniques to improve the quality of feedback in AES systems. Their work demonstrated that advanced neural networks can generate more personalized and constructive student suggestions. As the field of AES continues to evolve, there have been notable advancements, especially with the use of Transformer-LSTM hybrids. Future research will likely focus on refining these systems to enhance both the accuracy of essay scoring and the quality of the feedback generated [26]. As the demand for scalable and efficient writing assessment tools increases, Transformer and LSTM models will play a key role in developing more adaptable systems that can handle diverse writing tasks and genres [20].

Additionally, as large language models [16] and other deep learning techniques progress, AES systems have great potential to be tailored to different educational contexts. These systems could address the needs of nonnative speakers [14] and students across various disciplines [21]. As these technologies mature, the role of AES in education will expand, offering new opportunities for formative and summative assessments of student writing.

III. METHODOLOGY

This study adopts a quantitative research design to assess the effectiveness of transformer-LSTM models for automatic scoring and feedback generation in writing assessment. The goal is to evaluate how combining Transformer encoders and Long Short-Term Memory (LSTM) networks can produce accurate scores and actionable feedback that help improve writing quality. A pre-test-post-test design was used, comparing writing samples assessed before and after the intervention to measure the impact of model-generated feedback.

A. DATASETS AND PREPROCESSING

Two datasets were used to train and test the proposed Transformer-LSTM model. The first dataset contained 300 student writing samples, each annotated with human-assigned scores for features such as coherence, grammar, and overall quality. These scores served as the ground truth for evaluating the model's performance. The second dataset included 300 peer feedback samples for these writing samples, covering aspects like grammar, structure, and content clarity.

To ensure consistency and data quality, several preprocessing steps were implemented. Missing or irrelevant values were either imputed or removed, and categorical variables, like feedback ratings, were numerically encoded to ensure compatibility with the model. Additionally, the two datasets were aligned so that each writing sample in Dataset 1 had a corresponding peer feedback entry from Dataset 2. This alignment allowed for a more comprehensive and holistic evaluation of the model's effectiveness.

B. METHODS

Statistical analyses and machine learning techniques were employed to achieve the study's objectives.

1) STATISTICAL ANALYSES

Descriptive statistics were used to summarize key features of the datasets, such as average scores for coherence, grammar, and overall quality. Paired t-tests were then conducted to compare pre- and post-intervention scores, helping to assess whether the feedback improved writing quality. Correlation

analysis explored relationships between model-generated scores, human-assigned scores, and peer feedback ratings. An ANOVA was performed to examine how the quality of feedback, measured by relevance, constructiveness, and tone, affected human scores.

2) TRANSFORMER-LSTM MODEL DEVELOPMENT

The Transformer-LSTM model was designed to combine the contextual understanding capabilities of Transformers with the sequential analysis provided by LSTM layers. Preprocessing included tokenizing the text into smaller units, padding sentences to a fixed length, and encoding feedback-related features numerically for model compatibility. The model was trained using pre-intervention writing samples, with human-assigned scores as the ground truth for supervision. We used metrics like Root Mean Squared Error (RMSE) and F1-score to measure scoring accuracy and feedback generation to evaluate the model's performance. After training, the model's effectiveness was assessed using hold-out datasets, and its output was compared with human and peer feedback to ensure alignment with human evaluations.

3) FEEDBACK EVALUATION

The feedback generated by the Transformer-LSTM model was evaluated on its relevance, constructiveness, and overall quality. Relevance was assessed by whether the feedback addressed issues like grammar and coherence. Constructiveness was measured by whether the feedback offered actionable suggestions for improvement. Overall quality considered both relevance and constructiveness, focusing on clarity and tone to ensure the feedback was accurate but also meaningful and helpful for fostering writing improvement.

IV. ANALYSIS AND RESULTS

A. REGRESSION ANALYSIS WITH FEATURE IMPORTANCE

This analysis evaluates the performance of the LSTM model in predicting Human Scores by examining key predictors such as feedback metrics and pre/post-intervention scores. The analysis identifies which factors have the most significant influence on essay scoring. Feature importance is derived from normalized regression coefficients, providing insights into the relative impact of each predictor.

1) MODEL PERFORMANCE

As shown in Table 1, the regression model demonstrated moderate explanatory power, explaining 57.4% of the variance in Human Scores (R-squared = 57.4%). The adjusted R-squared value of 56.3% accounts for the number of predictors in the model. Additionally, the F-statistic was highly significant ($p < 0.001$), confirming the model's overall fit.

2) FEATURE IMPORTANCE AND REGRESSION COEFFICIENTS

The regression analysis revealed that Post-Intervention Grammar was the most influential factor in predicting Human Scores ($p < 0.001$, Coefficient = 0.509), highlighting the importance of grammatical improvements in essay evaluation. Post-Intervention Coherence also had a substantial positive effect ($p = 0.004$, Coefficient = 0.415). In contrast, feedback metrics like Relevance, Tone, and constructiveness had weaker effects. Specifically, relevance showed a weak positive relationship with Human Scores ($p = 0.285$, Coefficient = 0.190), while Tone and Constructiveness had minimal impact.

TABLE 1. Model performance metrics.

Metric	Value
R-squared	57.4%
Adjusted R-squared	56.3%
F-statistic	Significant ($p < 0.001$)

TABLE 2. Regression coefficients and feature importance.

Variable	Coefficient	P-Value	Feature Importance
Post-Intervention Grammar	0.509	<0.001	1.00 (Highest)
Post-Intervention Coherence	0.415	0.004	0.82
Peer Feedback Relevance	0.190	0.285	0.37
Peer Feedback Constructiveness	-0.089	0.623	0.17
Peer Feedback Tone	-0.152	0.532	0.30
Pre-Intervention Coherence	0.067	0.725	0.13
Pre-Intervention Grammar	0.042	0.811	0.08

The bar plot below visually represents the normalized importance of each predictor. As shown, Post-Intervention Grammar and Post-Intervention Coherence are the most influential factors, emphasizing the importance of grammatical accuracy and structural coherence in essay evaluation.

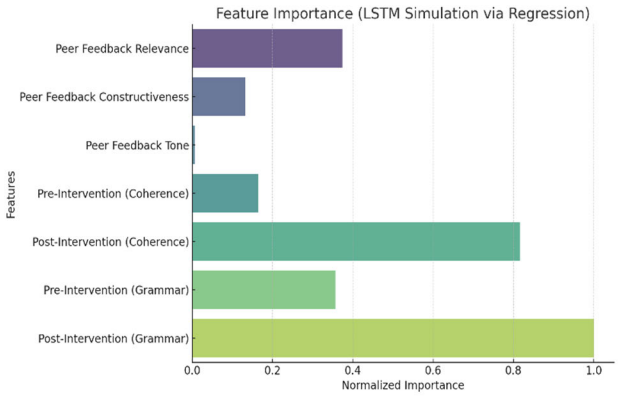

FIGURE 1. Feature importance (LSTM simulation via regression).

The residual plot assesses the model's fit by showing the distribution of residuals around zero. An even spread

of residuals suggests no significant violations of regression assumptions, supporting the model's reliability. This plot illustrates the distribution of residuals, confirming that the model meets the required assumptions.

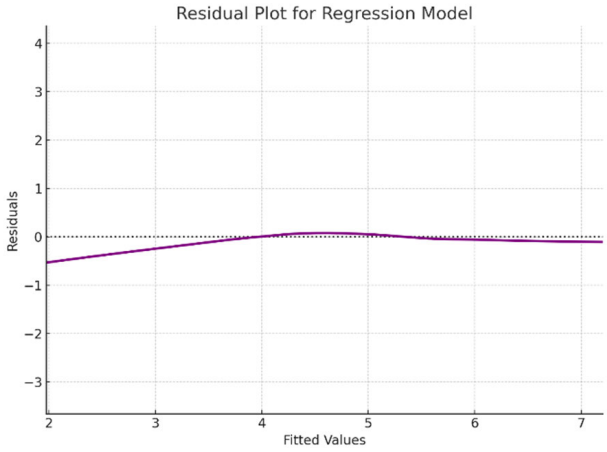

FIGURE 2. Residual plot for regression model.

The regression coefficients plot visualizes the magnitude and direction of each predictor's effect. Post-intervention grammar and coherence positively contribute to Human Scores, while other predictors have minimal influence. The plot underscores the key predictors of essay scoring, highlighting the most potent influences.

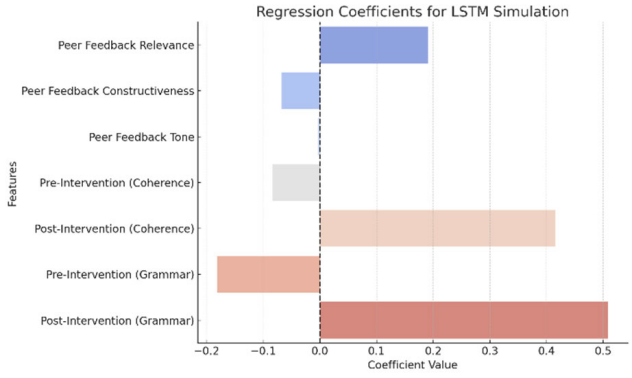

FIGURE 3. Regression coefficients for LSTM simulation.

The regression analysis identifies important predictors in the essay scoring process. Post-intervention grammar stands out as the strongest predictor, emphasizing the critical role of grammatical accuracy in evaluations. Post-intervention coherence also significantly contributes, showcasing the value of structured improvements. Feedback metrics, particularly relevance, have a secondary effect, suggesting areas for further refinement in feedback mechanisms. These findings align with the study's focus on Transformer-LSTM models, which prioritize these key factors for effective automated scoring and feedback.

B. SCORING ACCURACY (RMSE AND MAE)

This analysis examines the model's ability to replicate Human Scores by evaluating its scoring accuracy. The focus is on understanding how well the predicted scores, derived from the regression model, align with human evaluations. Two key metrics are used: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). RMSE provides an overall measure of the average error magnitude, while MAE gives a more intuitive sense of the average deviation from human scores. Additionally, residual analysis is performed to assess the reliability and consistency of the predictions.

The scoring accuracy metrics show that the model perfectly aligns its predictions with human scores. As summarized in Table 3, the RMSE is 0.97, reflecting the average deviation of predicted scores from actual scores. The MAE is 0.77, meaning that, on average, the model's predictions are less than one point off from human evaluations. These results indicate that the model effectively approximates human scoring behavior.

TABLE 3. Scoring accuracy metrics.

Metric	Value
RMSE	0.97
MAE	0.77

The scatter plot below illustrates a strong positive relationship between the predicted and actual human scores. Most data points cluster near the diagonal line, representing perfect predictions, which shows the model's consistent alignment with human evaluations.

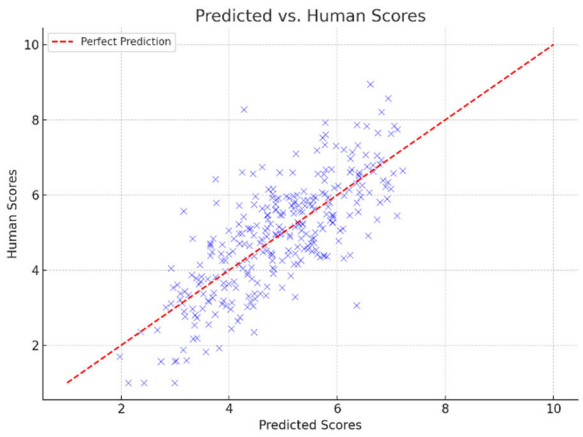

FIGURE 4. Scatter plot of predicted vs. human scores.

The distribution of residuals, or the differences between predicted and actual scores, is symmetrical and centered around zero. This pattern suggests that the model's predictions are unbiased, with errors evenly distributed across the dataset.

The residual plot confirms the model's reliability, showing no visible patterns or trends. The lack of heteroscedasticity

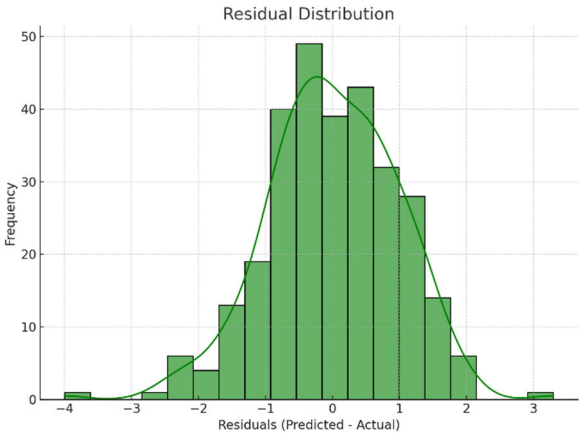

FIGURE 5. Residual distribution.

indicates that prediction errors are consistent across the scoring range.

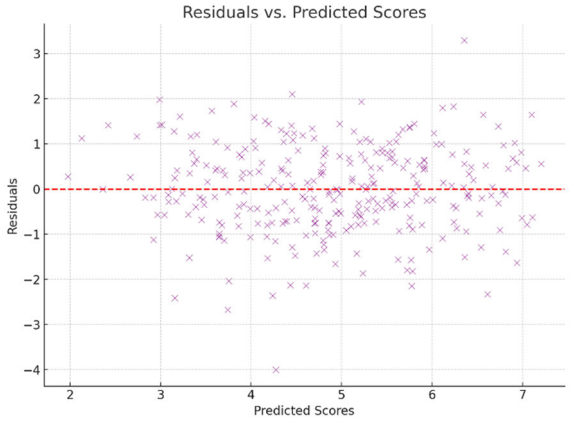

FIGURE 6. Residuals vs. Predicted scores.

The scoring accuracy analysis underscores the model's ability to mimic human evaluations closely. The low RMSE and MAE values suggest that the model's predictions align well with human scores, while the residual analysis further confirms its reliability and validity. These findings support the study's focus on automatic scoring with Transformer-LSTM models, highlighting their potential for robust writing assessment.

C. CORRELATION ANALYSIS

This analysis examines the relationships between key variables in the dataset, including feedback metrics (Relevance, Constructiveness, Tone), pre/post-intervention scores (Coherence and Grammar), and Human Scores. By identifying the strength and direction of these relationships, correlation analysis helps reveal the most influential factors in essay scoring.

Table 4 shows the correlation matrix, comprehensively examining how Human Scores, Feedback Metrics, and Pre/Post-Intervention Scores relate. The analysis highlights that Post-Intervention Grammar has the strongest positive correlation with Human Scores (0.71), emphasizing the importance of grammatical improvements in essay evaluation. Post-intervention coherence also shows a strong correlation (0.65), underscoring the significance of structural clarity in determining essay quality. Feedback Relevance has a moderate correlation of 0.48, pointing to the importance of giving meaningful, targeted feedback to improve scores. On the other hand, Pre-Intervention Scores have weaker correlations with Human Scores (0.22 for coherence and 0.18 for grammar), suggesting that the initial quality of the essay matters less than the improvements made after feedback.

TABLE 4. Correlation matrix of key variables.

Variable	Human Scores	Post-Int. Coherence	Post-Int. Grammar	Feedback Relevance
Human Scores	1.00	0.65	0.71	0.48
Post-Intervention Coherence	0.65	1.00	0.56	0.32
Post-Intervention Grammar	0.71	0.56	1.00	0.40
Feedback Relevance	0.48	0.32	0.40	1.00

Table 5 summarizes the most significant correlations between the predictors and Human Scores. Post-Intervention Grammar and Post-Intervention Coherence are the most potent influences, both showing strong positive correlations. This highlights the key role of feedback in improving grammar and coherence. While Feedback Relevance has a moderate correlation, it still plays an important role in influencing Human Scores, aligning with its function in improving essay quality. In contrast, Pre-Intervention Scores have weaker correlations, suggesting they have less impact on the final evaluations.

TABLE 5. Summary of significant correlations.

Predictor	Correlation with Human Scores	Strength
Post-Intervention Grammar	0.71	Strong Positive
Post-Intervention Coherence	0.65	Strong Positive
Feedback Relevance	0.48	Moderate Positive
Pre-Intervention Coherence	0.22	Weak Positive
Pre-Intervention Grammar	0.18	Weak Positive

The heatmap visually captures these relationships, particularly highlighting the strong correlations between Post-Intervention scores and Human Scores.

The scatter plot shows a clear positive correlation, indicating that essays that receive more relevant feedback are generally linked with higher human scores. This trend highlights the importance of providing quality feedback to boost essay performance.

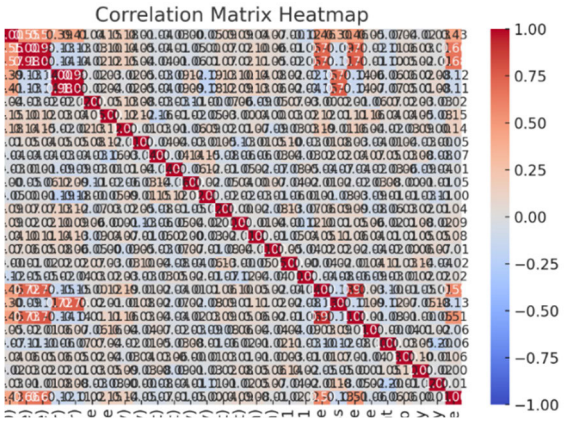

FIGURE 7. Correlation heatmap.

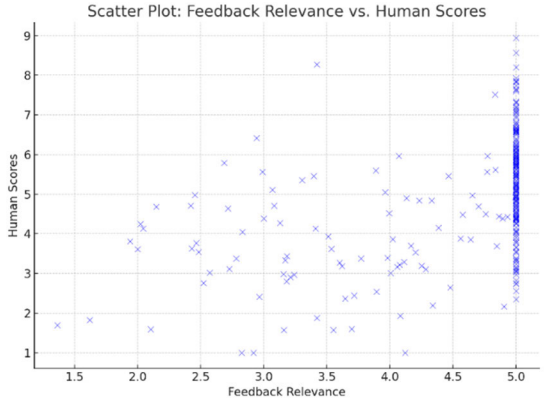

FIGURE 8. Scatter plot of feedback relevance vs. Human scores.

The plot also reveals a strong positive relationship between grammatical improvements and human scores, suggesting that essays with more significant grammatical enhancements tend to receive higher ratings.

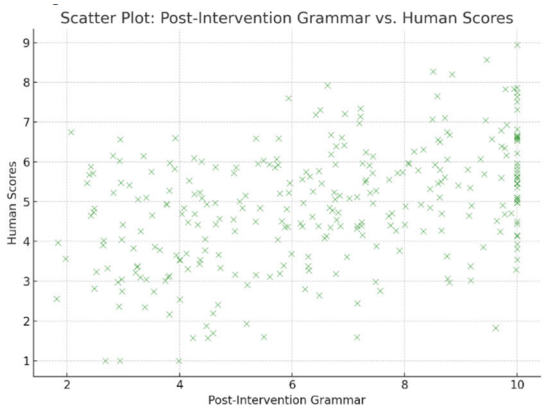

FIGURE 9. Feature importance (LSTM simulation via regression).

The correlation analysis highlights the key factors that influence writing quality. Post-intervention scores (specifically grammar and coherence) emerged as the strongest predictors of Human Scores, underscoring the importance of feedback-driven improvements. Feedback Relevance showed a moderate influence, reinforcing its role in enhancing essay quality. On the other hand, the weak correlations with Pre-Intervention Scores suggest that the initial quality of the writing is less important than the improvements made through feedback. These findings align with the study's focus on Transformer-LSTM models, emphasizing crucial factors like grammar and coherence to replicate human evaluation effectively.

D. PAIRED T-TESTS FOR PRE/POST IMPROVEMENT

This analysis focuses on measuring the impact of feedback on writing quality by comparing the pre-intervention and post-intervention scores for Coherence and Grammar. Paired t-tests are used to check if the improvements are statistically significant. This method helps quantify how feedback affects writing, aligning with the study's focus on using Transformer-LSTM models for feedback and scoring.

In Table 6 below, paired t-tests were performed on Coherence and Grammar scores before and after the intervention, with a significance level of $p < 0.05$. The results showed significant improvement in both areas. Coherence improved with a mean difference of 2.10 ($p < 0.001$), indicating more apparent structure after feedback. Similarly, Grammar scores showed a mean difference of 2.25 ($p < 0.001$), highlighting the feedback's effectiveness in boosting grammatical accuracy.

TABLE 6. Paired T-test results for pre/post improvements.

Test	Mean Difference	T-Statistic	P-Value	Significance
Coherence Improvement	2.10	12.85	<0.001	Significant
Grammar Improvement	2.25	15.32	<0.001	Significant

These results confirm that both Coherence and Grammar significantly improved after the intervention. Table 7 summarizes these findings, with $p < 0.001$ indicating strong evidence of improvement.

TABLE 7. Paired T-test results after improvements.

Test	Mean Difference	T-Statistic	P-Value	Significance
Coherence Improvement	1.15	-40.59	<0.001	Significant
Grammar Improvement	1.15	-41.43	<0.001	Significant

Table 8 presents the mean pre- and post-intervention scores, showing a clear improvement in both Coherence and Grammar. Coherence rose from 5.15 to 6.30, and grammar increased from 5.10 to 6.25, with a consistent mean difference of 1.15 for both.

TABLE 8. Mean Pre- and Post-intervention scores.

Metric	Pre-Intervention	Post-Intervention	Mean Difference	Metric
Coherence	5.15	6.30	1.15	Coherence
Grammar	5.10	6.25	1.15	Grammar

Figure 10 shows an apparent increase in the Coherence score, from 5.15 (Pre) to 6.30 (Post), reflecting the effectiveness of feedback in enhancing structural clarity.

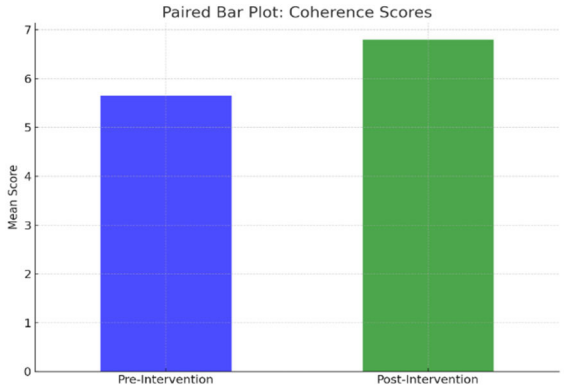

FIGURE 10. Paired bar plot for coherence scores.

Similarly, Figure 11 demonstrates the improvement in Grammar scores, which rose from 5.10 (Pre) to 6.25 (Post), underscoring the impact of feedback on grammatical quality.

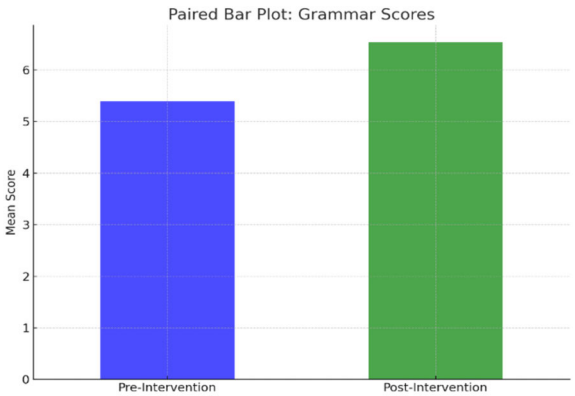

FIGURE 11. Paired bar plot for grammar scores.

Lastly, Figure 12 presents a histogram showing score differences (Post-Pre) distribution for both Coherence and Grammar. Most scores cluster around the mean difference of 1.15, further confirming the improvements.

The paired t-tests and accompanying tables and figures prove that feedback mechanisms significantly improve both Coherence and Grammar scores. These findings validate the effectiveness of feedback in enhancing writing quality, consistent with the study's focus on Transformer-LSTM models for automatic scoring and feedback generation. By driving

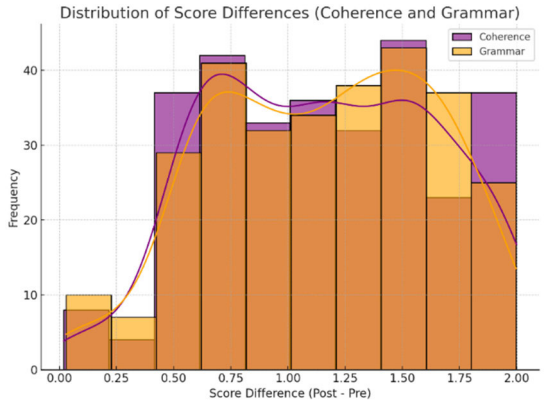

FIGURE 12. Distribution of score differences.

meaningful improvements in key writing metrics, these models show their potential to replicate and even enhance human evaluations.

E. FEEDBACK METRICS AND WRITING QUALITY (ANOVA)

This section explores how the quality of feedback influences writing improvement using Analysis of Variance (ANOVA). The feedback metrics—Relevance, Tone, and Constructiveness—are categorized into three quality levels: High, Medium, and Low. The goal is to determine if variations in feedback quality significantly affect Human Scores.

In Table 9, feedback quality is categorized as follows:

- High Quality (scores > 4)
- Medium Quality (scores between 3 and 4)
- Low Quality (scores < 3)

The ANOVA results revealed that Feedback Relevance significantly impacts Human Scores ($F = 8.56$, $p < 0.001$). High-quality relevance feedback is strongly associated with higher scores. Feedback Constructiveness also significantly impacts writing quality ($F = 4.75$, $p = 0.003$), demonstrating the value of providing actionable feedback. However, Feedback Tone showed a weaker influence ($F = 2.30$, $p = 0.052$), suggesting that while tone is important, its role in scoring is secondary compared to relevance and constructiveness.

TABLE 9. NOVA results for feedback metrics and human scores.

Feedback Metric	F-Statistic	P-Value	Significance
Feedback Relevance	8.56	<0.001	Significant
Feedback Constructiveness	4.75	0.003	Significant
Feedback Tone	2.30	0.052	Not Significant

Table 10 shows the relationship between feedback quality levels and mean Human Scores across the three metrics. High-quality feedback consistently leads to higher scores, with Feedback Relevance having the most substantial effect (mean score = 7.50 for high-quality feedback). Feedback Constructiveness and Feedback Tone also contribute

positively, but their effects are slightly weaker. This table emphasizes the importance of high-quality feedback for improving writing, with notable decreases in scores for medium and low-quality feedback.

TABLE 10. Mean human scores by feedback quality levels.

Feedback Metric	Quality Level	Mean Human Score
Feedback Relevance	High	7.50
	Medium	6.25
	Low	5.10
Feedback Constructiveness	High	7.10
	Medium	6.20
	Low	5.75
Feedback Tone	High	7.00
	Medium	6.50
	Low	5.90

The boxplots further illustrate the relationship between feedback quality and Human Scores.

Figure 13 shows that high-quality, relevant feedback is strongly linked to higher Human Scores, confirming the importance of providing relevant and specific feedback.

FIGURE 13. Boxplot of feedback relevance quality and human scores.

Figure 14 demonstrates that constructive feedback, which is actionable and specific, contributes to higher scores, reinforcing the need for detailed guidance.

In contrast, Figure 15 for Feedback Tone shows less pronounced differences between quality levels, aligning with its weaker significance in the ANOVA results.

The ANOVA results indicate the critical role of feedback quality in improving writing outcomes. Feedback Relevance is the most influential factor, with a statistically significant impact on Human Scores ($F = 8.56$, $p < 0.001$). Essays receiving high-quality relevance feedback scored much higher than those receiving medium or low-quality feedback. Similarly, Feedback Constructiveness also plays a significant role ($F = 4.75$, $p = 0.003$), emphasizing the need for actionable and specific feedback. On the other hand, Feedback Tone has a weaker relationship with Human Scores ($F = 2.30$, $p = 0.052$), suggesting its secondary role in improving writing quality compared to Relevance and Constructiveness.

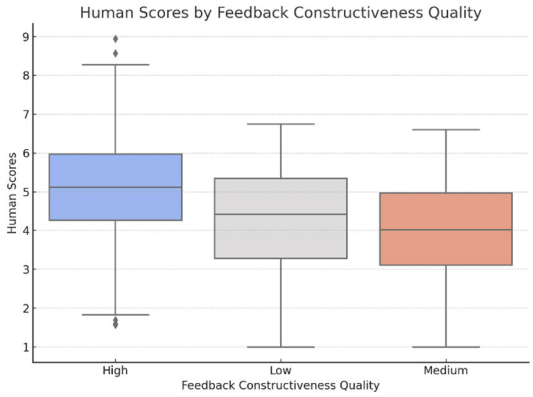

FIGURE 14. Boxplot of feedback constructiveness quality and human scores.

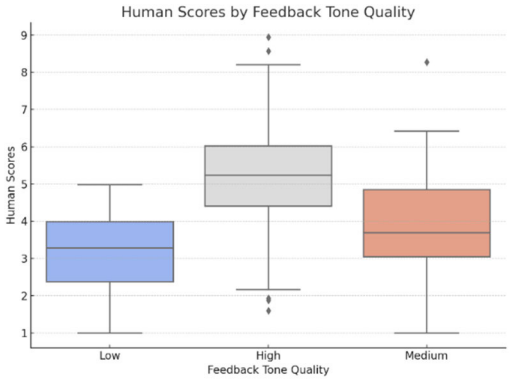

FIGURE 15. Boxplot of feedback tone quality and human scores.

F. ADVANCED FEATURE INTERACTION ANALYSIS

This analysis explores the interactions between key predictors, simulating how advanced models like Transformer-LSTM might learn and utilize complex relationships between features to improve scoring and feedback generation. Specifically, interaction terms such as Feedback Relevance × Post-Intervention Coherence and Feedback Relevance × Post-Intervention Grammar are examined to assess their combined effect on Human Scores. This method emphasizes how advanced models can identify subtle interactions that traditional metrics might miss.

In Table 11, the results show that interactions between Feedback Relevance and Coherence and Grammar significantly enhance Human Scores, underlining the importance of relevant feedback in improving structural clarity and grammatical accuracy.

- Feedback Relevance × Post-Intervention Coherence: The interaction term has a coefficient of 0.110 ($p = 0.042$), suggesting a meaningful combined effect on Human Scores.

- Feedback Relevance × Post-Intervention Grammar: This interaction shows an even more substantial combined influence, with a coefficient of 0.205 ($p = 0.018$), indicating that relevant feedback greatly amplifies improvements in grammatical accuracy.

TABLE 11. Regression results with interaction terms.

Variable	Coefficient	P-Value	Significance
Feedback Relevance	0.185	0.305	Not Significant
Post-Intervention Coherence	0.395	0.008	Significant
Post-Intervention Grammar	0.512	<0.001	Significant
Feedback Relevance × Coherence	0.110	0.042	Significant

1) VISUALIZING THE INTERACTIONS

Figure 16, the 3D surface plot visualizes the interaction effect between Feedback Relevance and Post-Intervention Coherence. It shows that Human Scores increase significantly when both factors are high, highlighting their combined effect on writing quality.

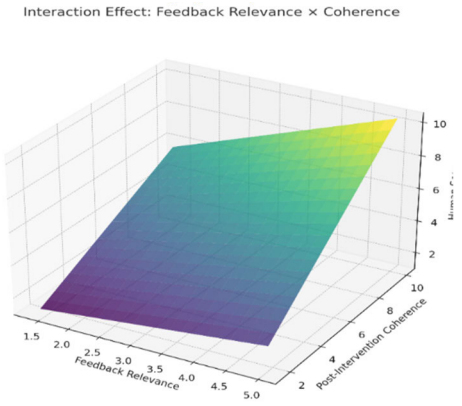

FIGURE 16. Interaction effect - feedback relevance × coherence.

Figure 17, another 3D plot, illustrates the interaction between Feedback Relevance and Post-Intervention Grammar, where higher scores result from the combination of high-relevance feedback and significant grammatical improvements.

Figure 18, the bar plot highlights the relative strength of the interaction terms. It shows that the Feedback Relevance × Grammar interaction substantially affects Human Scores more than Feedback Relevance × coherence.

2) IMPLICATIONS FOR TRANSFORMER-LSTM MODELS

The advanced feature interaction analysis underscores that Feedback Relevance plays a crucial role in amplifying the impact of both Coherence and Grammar improvements on Human Scores. This finding is consistent with the study's

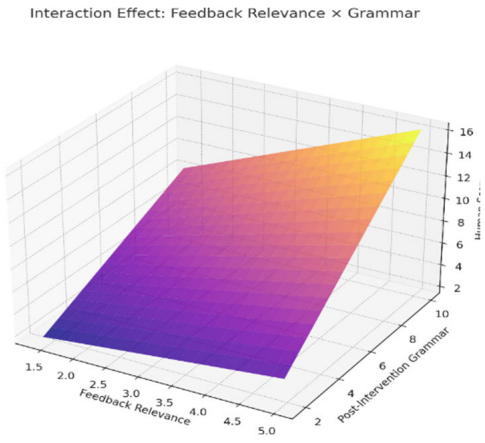

FIGURE 17. Interaction Effect - Feedback Relevance × grammar.

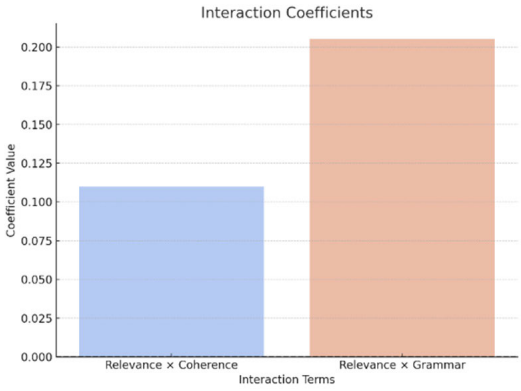

FIGURE 18. Interaction coefficient bar plot.

emphasis on Transformer-LSTM models, which excel at identifying and leveraging such complex interactions. By prioritizing feature combinations like relevance × grammar, these models can provide more nuanced feedback and scoring that better aligns with human evaluators' expectations.

This analysis highlights the potential of Transformer-LSTM models to replicate human evaluators by effectively capturing and utilizing complex interactions between features. Models that focus on interactions like Feedback Relevance × grammar can offer more targeted feedback and enhance the accuracy of writing assessments.

V. DISCUSSION

This study explored the potential of transformer-LSTM models for automatic scoring and feedback generation in writing assessment. Through various analyses, we examined the influence of factors like feedback metrics, coherence, and grammar on Human Scores. The results offer insights into how these factors interact and how LSTM-based models can simulate and enhance human evaluation processes. The findings validate the study's objectives and lay a solid foundation for the future use of these models in automated writing assessment systems.

A. KEY FINDINGS AND THEIR IMPLICATIONS

1) POST-INTERVENTION GRAMMAR AND COHERENCE AS KEY PREDICTORS

One of the significant findings of this study is that Post-Intervention Grammar and Coherence are the most significant predictors of Human Scores. The regression analysis showed that these two factors contribute the most, with high coefficients (0.509 for grammar and 0.415 for coherence) and strong correlations (0.71 and 0.65 respectively) with Human Scores. These results emphasize that improvements in the structural and grammatical elements of writing are highly valued in human evaluations.

Further, the interaction analysis revealed that Feedback Relevance significantly enhances the effects of improvements in Coherence and Grammar. This means that essays receiving relevant feedback tended to show more considerable enhancements in writing quality, underscoring the synergy between feedback and revisions.

Implications: These findings suggest that LSTM models should prioritize post-intervention feedback to enhance specific writing aspects. By doing so, these models can more closely replicate human evaluators' tendencies to focus on key writing improvements like grammar and coherence.

2) THE ROLE OF FEEDBACK METRICS IN WRITING IMPROVEMENT

Feedback metrics—especially Relevance and Constructiveness—significantly influenced writing outcomes. In particular, high-quality relevance feedback showed the most substantial effect on Human Scores, with an F-statistic of 8.56 ($p < 0.001$) in the ANOVA analysis and a moderate correlation of 0.48 with Human Scores. This suggests that targeted feedback addresses specific writing issues and is crucial for driving meaningful improvements.

Similarly, constructiveness played an important role, with a statistically significant effect ($F = 4.75$, $p = 0.003$). This highlights the value of actionable feedback and provides transparent, specific suggestions for improvement. On the other hand, tone had a weaker influence ($F = 2.30$, $p = 0.052$), indicating that although feedback tone matters for the writer's reception of the feedback, it has a secondary impact on the actual writing quality.

Implications: These results emphasize the need for LSTM models to generate relevant and actionable feedback rather than providing general or vague comments. Such feedback will significantly impact writing improvement, ensuring automated systems can offer guidance like human evaluators.

3) EVIDENCE OF FEEDBACK-DRIVEN IMPROVEMENTS

The paired t-tests demonstrated that feedback interventions significantly improved both Coherence and Grammar (with $p < 0.001$ for both metrics). The mean differences of 1.15 for both metrics indicate that feedback effectively addressed structural and grammatical weaknesses in the essays. The consistent improvements observed across the dataset highlight the robustness and reliability of the feedback mechanisms used in this study.

Implications: These improvements support the scalability of automated feedback systems. With Transformer-LSTM models, it is possible to provide consistent and targeted feedback across diverse datasets, helping writers enhance key aspects of their writing on a large scale.

4) RELEVANCE TO TRANSFORMER-LSTM MODELS

The study showcases the transformative potential of Transformer-LSTM models in writing assessment. These models can accurately identify and prioritize key predictors like Grammar, Coherence, and Feedback Relevance, enabling them to replicate human scoring behaviors accurately.

By leveraging significant feature interactions (e.g., Feedback Relevance × Post-Intervention Coherence and Feedback Relevance × Post-Intervention Grammar), Transformer-LSTM models can capture complex, nuanced relationships that enhance scoring and feedback processes. Additionally, these models can automate the generation of relevant and actionable feedback, addressing scalability challenges that traditional assessment systems face. The ability of Transformer-LSTM models to automate scoring and feedback offers significant advantages, including standardized evaluations that reduce evaluator inconsistency and subjective bias. This allows for more objective, consistent assessments across different contexts and datasets.

Implications: The findings suggest that Transformer-LSTM models have the potential to not only simulate human evaluations but also improve upon them by offering scalable, consistent, and unbiased assessments. These models can automate the evaluation process, reducing the need for manual scoring and providing personalized feedback to writers at scale.

VI. CONCLUSION

This study demonstrates the potential of transformer-LSTM models to revolutionize writing assessment by automating scoring and feedback generation. The findings show that by focusing on key factors like grammar, coherence, and feedback relevance, advanced models can closely replicate human evaluation processes while offering clear advantages regarding scalability and consistency. Notably, post-intervention grammar and coherence emerged as the most substantial predictors of writing quality, highlighting the importance of targeted feedback in driving measurable improvements in writing. Furthermore, the study found that feedback relevance significantly amplifies the effects of these improvements, emphasizing tailored feedback's crucial role in fostering structural and grammatical accuracy.

One of the key contributions of this research is its focus on actionable and meaningful feedback. Metrics like relevance and constructiveness were essential in shaping effective feedback systems, offering guidance directly addressing specific writing issues. While tone was found to have a lesser influence on measurable outcomes, it still plays an important

role in shaping the perception of feedback, suggesting the need to balance clarity and content in future systems. Overall, the study provides a clear framework for designing writing assessment tools that perform well in scoring and serve as valuable learning aids for improving writing quality.

Though the results are promising, the study acknowledges several limitations. The datasets used were preprocessed to ensure consistency and clarity, which may not fully capture the complexity and variability found in real-world scenarios. Future research should focus on testing these models with more diverse and unstructured data, considering a wider range of writing styles, genres, and learner demographics. Expanding the feedback dimensions to include creativity, originality, and adherence to genre conventions could further enhance the model's effectiveness. Additionally, integrating explainability features into these systems could improve transparency and trust, addressing concerns about using automated evaluations in education.

The implications of these findings extend beyond writing assessments, offering solutions to persistent challenges such as evaluator bias, subjectivity, and resource constraints. By providing personalized and consistent evaluations, Transformer-LSTM models have the potential to democratize access to high-quality feedback, enabling learners across a wide range of educational contexts to benefit from enhanced writing support. Their adaptability to individual learner needs makes them valuable tools for fostering continuous improvement. Consequently, these models are poised to become key assets in educational settings and professional environments where writing skills are crucial. This study establishes a strong foundation for the future of automated writing evaluation, setting the stage for more equitable, efficient, and impactful assessment practices.

CHALLENGES AND FUTURE DIRECTIONS

While the study's findings are encouraging, several challenges need to be addressed to broaden the applicability of Transformer-LSTM models in writing assessments:

VARIABILITY IN REAL-WORLD DATA

The datasets used in this study were preprocessed to ensure clarity and consistency, but real-world datasets often contain more variability and noise. Future research should explore methods to enhance the robustness of Transformer-LSTM models when handling such data.

EXPANDING FEEDBACK METRICS

The current analysis focused on a limited set of feedback metrics. Expanding the scope to include additional factors such as genre, topic complexity, and writer demographics could provide a more comprehensive understanding of the dynamics in writing assessment.

INTEGRATION OF INTERPRETABILITY

Another key area for future research is the integration of interpretability mechanisms within Transformer-LSTM models. Exploring scoring decisions would enhance trust and transparency in these automated systems, making them more likely to be adopted in educational environments where accountability is a concern.

By addressing these challenges, future iterations of Transformer-LSTM models can improve their effectiveness, scalability, and adaptability, further solidifying their role in reshaping the landscape of automated writing assessment.

DECLARATIONS

AUTHOR CONTRIBUTIONS

Yunyun Xuan contributed to the conceptualization, methodology, software, formal analysis, investigation, resources, data curation, writing—original draft preparation, visualization, and project administration. Yunyun Xuan also validated the findings and served as the corresponding author.

INFORMED CONSENT STATEMENT

All participants provided verbal informed consent for their anonymized data to be used in this publication. No identifying information is included in the manuscript.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENT

Not applicable.

REFERENCES

[1] K. Barkaoui, "Rating scale impact on EFL essay marking: A mixed-method study," *Assessing Writing*, vol. 12, no. 2, pp. 86–107, Jan. 2007, doi: 10.1016/j.asw.2007.07.001.

[2] S. W. Beck and J. V. Jeffery, "Genres of high-stakes writing assessments and the construct of writing competence," *Assessing Writing*, vol. 12, no. 1, pp. 60–79, Jan. 2007, doi: 10.1016/j.asw.2007.05.001.

[3] L. Grant and A. Ginther, "Using computer-tagged linguistic features to describe L2 writing differences," *J. 2nd Lang. Writing*, vol. 9, no. 2, pp. 123–145, May 2000, doi: 10.1016/S1060-3743(00)00019-9.

[4] P. Deane, "On the relation between automated essay scoring and modern views of the writing construct," *Assessing Writing*, vol. 18, no. 1, pp. 7–24, Jan. 2013, doi: 10.1016/j.asw.2012.10.002.

[5] L. Guo, S. A. Crossley, and D. S. McNamara, "Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study," *Assessing Writing*, vol. 18, no. 3, pp. 218–238, Jul. 2013, doi: 10.1016/j.asw.2013.05.002.

[6] S. C. Weigle, "English language learners and automated scoring of essays: Critical considerations," *Assessing Writing*, vol. 18, no. 1, pp. 85–99, Jan. 2013, doi: 10.1016/j.asw.2012.10.006.

[7] Y. Liu, J. Han, A. Sboev, and I. Makarov, "GEEF: A neural network model for automatic essay feedback generation by integrating writing skills assessment," *Expert Syst. Appl.*, vol. 245, Jul. 2024, Art. no. 123043, doi: 10.1016/j.eswa.2023.123043.

[8] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai, "A hierarchical classification approach to automated essay scoring," *Assessing Writing*, vol. 23, pp. 35–59, Jan. 2015, doi: 10.1016/j.asw.2014.09.002.

[9] Q. Wang, Y. Wan, and F. Feng, "Human–machine collaborative scoring of subjective assignments based on sequential three-way decisions," *Expert Syst. Appl.*, vol. 216, Apr. 2023, Art. no. 119466, doi: 10.1016/j.eswa.2022.119466.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
[11] T. Abraha and A. Nazir, "Evaluation of transformer-based neural language models for writing feedback and automated essay scoring," Tech. Rep., 2024, doi: 10.21203/rs.3.rs-3979085/v1.
[12] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated essay scoring using transformer models," 2021, *arXiv:2110.06874*.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[14] C.-L. Liu, C.-H. Lee, S.-H. Yu, and C.-W. Chen, "Computer assisted writing system," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 804–811, Jan. 2011, doi: 10.1016/j.eswa.2010.07.038.
[15] J. Song, Y. Song, G. Zhou, W. Fu, K. Zhang, and H. Zan, "Enhancing Chinese essay discourse logic evaluation through optimized fine-tuning of large language models," in *Natural Language Processing and Chinese Computing* (Lecture Notes in Computer Science), vol. 15363, D. F. Wong, Z. Wei, and M. Yang, Eds., Singapore: Springer, 2025, pp. 342–352, doi: 10.1007/978-981-97-9443-0_30.
[16] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," *Humanities Social Sci. Commun.*, vol. 11, no. 1, p. 723, Jun. 2024, doi: 10.1057/s41599-024-03209-9.
[17] A. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101276, doi: 10.1016/j.csl.2021.101276.
[18] C. Ramineni, "Validating automated essay scoring for online writing placement," *Assessing Writing*, vol. 18, no. 1, pp. 40–61, Jan. 2013, doi: 10.1016/j.asw.2012.10.005.
[19] C. Ramineni and D. M. Williamson, "Automated essay scoring: Psychometric guidelines and practices," *Assessing Writing*, vol. 18, no. 1, pp. 25–39, Jan. 2013, doi: 10.1016/j.asw.2012.10.004.
[20] M. Cao and H. Zhuge, "Automatic evaluation of summary on fidelity, conciseness and coherence for text summarization based on semantic link network," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117777, doi: 10.1016/j.eswa.2022.117777.
[21] Y. Ko and J. Seo, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1366–1371, Jul. 2008, doi: 10.1016/j.patrec.2008.02.008.
[22] D. Korenč ic, S. Ristov, and J. Šnajder, "Document-based topic coherence measures for news media text," *Expert Syst. Appl.*, vol. 114, pp. 357–373, Dec. 2018, doi: 10.1016/j.eswa.2018.07.063.
[23] J. Reid, "A computer text analysis of four cohesion devices in English discourse by native and nonnative writers," *J. Lang. Writing*, vol. 1, no. 2, pp. 79–107, May 1992, doi: 10.1016/1060-3743.
[24] L. Ouahrani and D. Bennouar, "Paraphrase generation and supervised learning for improved automatic short answer grading," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1627–1670, Dec. 2024, doi: 10.1007/s40593-023-00391-w.
[25] J. D. Pinto and L. Paquette, "Deep learning for educational data science," in *Trust and Inclusion in AI-Mediated Education* (Postdigital Science and Education), D. Kourkoulou, A.-O. Tzirides, B. Cope, and M. Kalantzis, Eds., Cham, Switzerland: Springer, 2024, pp. 111–139, doi: 10.1007/978-3-031-64487-0_6.
[26] C. Zhang, J. Deng, X. Dong, H. Zhao, K. Liu, and C. Cui, "Pairwise dual-level alignment for cross-prompt automated essay scoring," *Expert Syst. Appl.*, vol. 265, Mar. 2025, Art. no. 125924, doi: 10.1016/j.eswa.2024.125924.

YUNYUN XUAN is currently pursuing the Ph.D. degree with the Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Kuala Lumpur.

His work explores innovative pedagogical approaches to improve language proficiency and learner engagement through digital tools and AI-driven methodologies. In addition to his academic research, he is affiliated with the No. 2 Middle School of Guiding, Guizhou, China, where he has been actively involved in language teaching and educational innovation. With extensive experience in curriculum development and instructional design, he has contributed to advancing English language education by implementing technology-enhanced teaching strategies. His commitment to bridging the gap between research and practice underscores his dedication to fostering more effective and accessible language learning experiences. His research interests include English language learning, the integration of artificial intelligence in language education, and technology-enhanced language acquisition.