**Reviewer #2 Comments**

This paper contributes to the growing literature on AI-powered formative feedback by developing a transformer-based intelligent feedback system aimed at improving self-assessment accuracy. The theoretical alignment with feedback literacy and self-regulated learning is strong, and the system architecture is well-articulated. The empirical study is well executed, although certain aspects would benefit from elaboration or clarification to better contextualize the findings and system performance.

1. While the authors claim the system is “privacy-preserving” by avoiding camera-based input, this deserves more nuanced discussion. CSI signals, linguistic analysis, and writing style can still contain identifying traits. A clearer articulation of privacy boundaries would be welcome.

|  |
| --- |
| Response: Thank you for this important observation. While we did not explicitly claim that the system is privacy-preserving, we recognize that the use of textual data, including writing style and linguistic patterns, can carry identifiable traits. To prevent potential misunderstanding, we have revised the manuscript to clarify the data modalities used and to acknowledge the broader considerations of privacy in language-based educational systems. We also added a note indicating that future extensions may consider privacy-preserving mechanisms such as anonymization and secure data handling practices. |

1. The feedback pipeline includes both regression-based essay scoring and classification of revision quality or types of errors, but the two tasks are not clearly separated. A diagram or explicit description of how these modules interact would improve the methodological clarity.

|  |
| --- |
| Response: Thank you for your observation. We would like to clarify that the present study focuses solely on the task of automated essay scoring, formulated as a regression problem. While we recognize that revision classification is an important component in broader feedback systems, our pipeline does not currently include a separate module for revision type classification. The model architecture in Figure 3 has been updated slightly to clarify this scope, and additional textual clarification has been added to prevent possible misinterpretation. |

1. The authors report using both handcrafted features and contextual embeddings, but it is unclear how these are weighted or fused in the final model. Some additional explanation or ablation results would be useful for readers interested in system performance.

|  |
| --- |
| Response: We thank the reviewer for pointing out this issue. In our design, handcrafted features and contextual embeddings were concatenated into a single feature vector, which was then fed into the regression models (e.g., XGBoost, SVM, AdaBoost). No additional weighting scheme was applied, as the regression learners inherently optimized the contribution of each feature dimension. We have revised the Methodology section to clarify this fusion strategy. |

1. The authors discuss feedback literacy and internal feedback mechanisms, but student engagement with the LLM feedback is not measured directly. Including data on students’ perception, uptake, or revision behaviors (even qualitatively) would enrich the analysis.

|  |
| --- |
| Response: We thank the reviewer for this valuable suggestion. While the present study focused on the cognitive effects of LLMF as reflected in SAA, we agree that future work should further examine learners' engagement behaviors, including perception, uptake, and revision actions. This direction has been noted in the revised conclusion. |

1. The short duration of the intervention (single 50-minute revision) may not allow sufficient time for reflection or behavioral change. The authors acknowledge this, but it may be useful to briefly compare with longer-term feedback studies.

|  |
| --- |
| Response: We thank the reviewer for highlighting the limitation related to the short intervention duration. This concern has been explicitly addressed in the revised conclusion, where we acknowledge the limited opportunity for reflection and note that future research should consider longer-term interventions to better capture behavioral change and feedback uptake. |

1. While the system architecture is rigorous, practical implementation issues such as processing time, scalability, and integration with LMS platforms are not discussed. These aspects are important considerations for actual deployment in educational settings.

|  |
| --- |
| Response: We thank the reviewer for highlighting important considerations related to real-world implementation, such as processing time, system scalability, and integration with LMS. While the current study focuses on pedagogical outcomes and does not empirically evaluate engineering performance or deployment logistics, we agree that these aspects are crucial for broader adoption. We have acknowledged this need for future exploration in the revised conclusion section, particularly in discussing the system’s scalability and applicability across diverse educational environments. |

1. The sample consists of students from engineering backgrounds writing in Chinese, but this restricts generalizability. The authors might consider commenting on how language-specific linguistic features were handled or how transferable the system is to other domains/languages.

|  |
| --- |
| Response: We appreciate the reviewer’s observation regarding the limitations of using a sample drawn from engineering students writing in Chinese. While this context aligns with the system’s current design and testing scope, we acknowledge that language-specific linguistic features may influence both model performance and feedback interpretation. Although this study did not examine domain or language transferability directly, we have explicitly added this as a direction for future research in the revised conclusion section. Specifically, we note the importance of validating system effectiveness across genre-specific writing and multilingual contexts to ensure broader applicability. |

1. The strong correlation between SAA and learning performance is a valuable insight, but the causal relationship is not explored. Some speculative discussion on the directionality (e.g., whether improved SAA leads to better outcomes or vice versa) would be intellectually stimulating.

|  |
| --- |
| Response: We appreciate the reviewer’s thoughtful comment regarding the correlation between SAA and learning performance. While our current study focused on identifying predictive relationships, we agree that exploring the directionality between SAA and learning outcomes would enrich the theoretical understanding of AI-assisted self-regulated learning. As noted in the revised conclusion, future research should incorporate mediation analyses to examine whether SAA functions as an intermediary mechanism through which AI-generated feedback influences broader learning outcomes, or conversely, whether more proficient learners are better at self-monitoring. Such analyses would provide a deeper understanding of the causal pathways underlying feedback effectiveness. |

In sum, the manuscript is well-developed and theoretically sound. The suggested revisions aim to enhance generalizability, technical transparency, and practical relevance.