**Reviewer #1 Comments**

This manuscript presents a timely and technically competent study of an AI-driven intelligent feedback system designed to improve students’ self-assessment accuracy (SAA) in higher education writing. The integration of transformer-based models with handcrafted linguistic features reflects a thoughtful hybrid approach, and the authors make a commendable effort to bridge technical execution with educational relevance. That said, several aspects of the manuscript would benefit from further elaboration or clarification to enhance transparency and reader confidence.

1. Terminology such as “sustainable” and “inclusive” is used repeatedly, but its operational meaning within the scope of AI feedback systems remains vague. A more concrete explanation, such as whether it refers to educational reach, technological scalability, or cost-effectiveness, would help improve clarity.

|  |
| --- |
| Response:  We thank the reviewer for this valuable observation. We have added explicit definitions of “sustainable” and “inclusive” in the Introduction, clarifying that “sustainable” refers to the system’s ability to deliver consistent, high-quality feedback over time with minimal additional resources, and “inclusive” refers to its ability to provide adaptive, personalized feedback for diverse learner needs. |

1. Although the study reports a non-significant main effect, the compensatory effect among students with low initial SAA is a notable finding. It would be helpful if the authors could further elaborate on the practical implications of this differential effect, particularly in relation to instructional targeting or educational equity.

|  |
| --- |
| Response:  We appreciate the reviewer’s insight. We have expanded the Discussion to explain that early identification of learners with low initial SAA enables targeted allocation of feedback resources, thereby enhancing metacognitive monitoring and addressing disparities in self-regulated learning skills. |

1. The experimental design uses a neutral prompt for the control group instead of human feedback, which is reasonable given the focus. Nonetheless, this makes it difficult to position the LLM system in relation to existing human-centered formative feedback practices. A brief discussion of this tradeoff would be appropriate.

|  |
| --- |
| Response:  We thank the reviewer for raising this point. We acknowledge that using a neutral prompt limits direct comparison with human feedback. This was an intentional choice to isolate the effects of LLMF and avoid variability from instructor style. We have added this trade-off to the Conclusion, noting the limitation and recommending that future studies compare the system with multiple forms of human-provided feedback. |

1. The dataset section is well explained, but the rationale for using 64 participants, and whether this size offers sufficient power for SEM and regression, is not justified. A brief justification or citation would strengthen methodological credibility.

|  |
| --- |
| Response:  We appreciate the reviewer’s comment. The sample size of 64 was determined by the maximum course enrollment and equal group allocation. Although modest, it aligns with widely cited recommendations for exploratory SEM and regression that suggest approximately 10–15 cases per estimated parameter (Kline, 2015), and is consistent with empirical findings from Sideridis et al. (2014), who reported adequate model fit with samples of 50–70 participants. This justification has been added to the 4.1 Study Sample section. |

1. Table 1 provides representative feedback examples, but these mainly cover conventional writing issues. Including examples of ambiguous or borderline feedback cases would allow readers to better assess the model’s limitations and pragmatic depth.

|  |
| --- |
| Response:  We thank the reviewer for this suggestion. To address it, we have expanded Table 1 to include an additional category, Ambiguous / Borderline Cases, which presents examples of feedback for less clear or context-dependent issues, such as redundant expressions, unnatural collocations, and unclear pronoun references. This addition highlights the system’s limitations in handling nuanced language and illustrates the pragmatic depth of the feedback process. We have also noted in the table that the original student texts were in Chinese and examples are translated for clarity. |

1. The limitations section does not address how the system might behave in more complex writing genres or with non-native writers. This is important given the claim of scalability and broad educational applicability.

|  |
| --- |
| Response:  We agree with the reviewer’s point. A statement has been added to the limitations section noting that the current evaluation was conducted on short, general-purpose essays written by native speakers. The system’s behavior with more complex genres or non-native writers remains unexplored. We now suggest future research to examine genre-specific adaptations and multilingual capability enhancements to support broader educational applicability |

1. The manuscript mentions the use of RoBERTa and BERT embeddings, but does not compare performance between models or explain why both were used in tandem. Including a short rationale would help clarify design decisions.

|  |
| --- |
| Response:  We thank the reviewer for raising this point. A short rationale has been added to the methodology section explaining that BERT and RoBERTa embeddings were used in combination to exploit their complementary strengths in contextual and fine-grained linguistic representation. As the study aimed to evaluate the pedagogical impact of the integrated system rather than conduct model benchmarking, no direct performance comparison between the models was performed. |

1. The path analysis (SEM) results are clearly presented, but the indirect effects, if any, are not discussed. It may be worth noting whether mediated effects through self-assessment accuracy influence learning performance more strongly than direct effects.

|  |
| --- |
| Response:  We appreciate this insightful suggestion. As noted in the revised discussion, the current study did not test mediation effects due to its focus on primary path relationships. A statement has been added to acknowledge this limitation and to suggest that future work examine whether SAA mediates the relationship between feedback and learning performance, potentially yielding stronger effects than direct pathways. |

1. Figures 5 and 6, which illustrate the structural equation model, are conceptually informative. However, the visual design could be improved. Specifically, the labels inside the circular nodes are cramped, with some text spilling outside the boundaries. Increasing the size of the circular elements slightly relative to the font size would enhance both readability and professionalism of presentation.

|  |
| --- |
| Response:  Thank you for the helpful suggestion. We have revised Figures 5 and 6 to improve visual clarity by increasing the size of the circular elements and adjusting the font-to-shape ratio. The updated figures now enhance readability and overall presentation quality. |

1. While the use of the abbreviation “LLMF” for LLM-generated feedback is appropriate, it appears inconsistently throughout the manuscript. After the term is first introduced and defined, subsequent references should consistently adopt the abbreviation. In some places, the full term is used again without need, while in others the abbreviation appears without clear prior definition. Maintaining consistency in abbreviation usage will improve clarity and reduce potential confusion for readers. Please review the manuscript for other similar cases of inconsistent or unexplained abbreviations. Ensuring that all abbreviations are clearly defined upon first use and used consistently thereafter would enhance the manuscript’s readability and professionalism.

|  |
| --- |
| Response:  Thank you for the valuable suggestion. We have thoroughly reviewed the manuscript to ensure that all abbreviations, including “LLMF,” are clearly defined upon first use and used consistently throughout. In addition to LLMF, other terms that warranted abbreviation have also been standardized to improve clarity, readability, and overall professionalism of the manuscript. |