**Reviewer #3 Comments**

The study investigates the effects of LLM-generated feedback on students’ self-assessment accuracy and writing improvement, offering an innovative and scalable solution to feedback delivery in higher education. The integration of deep NLP techniques with educational theory is a significant strength of the paper. However, a few clarifications and expansions would enhance its clarity and practical contribution.

1. The study finds no main effect of LLM feedback but highlights significant moderation by initial self-assessment accuracy. This is insightful, but more interpretation is needed regarding why high-SAA students show little gain. This pattern may be due to ceiling effects or a lack of sufficient task challenge.

|  |
| --- |
| Response: The authors appreciate the reviewer’s insightful comment. In response, the Discussion section has been revised to provide a more detailed interpretation of the limited improvement observed among students with high ISAA. One possible explanation is that these students may have already possessed relatively accurate self-monitoring skills, resulting in limited room for measurable gains. Another possibility is that the task complexity or the depth of feedback provided was not sufficiently challenging to provoke further metacognitive engagement in this subgroup. The revised paragraph also cites Winstone and Carless (2019), suggesting that learners with high ISAA may require more nuanced, domain-specific, or reflective prompts in order to benefit meaningfully from feedback. These additions aim to clarify the boundary conditions of LLMF effectiveness. The authors thank the reviewer for raising this important point. |

1. The SEM model structure is explained, but the decision to use SEM rather than multilevel modeling, especially considering the within-subjects essay-revision design, is not justified. Providing a brief rationale would help readers who are less familiar with PLS-SEM in educational research contexts.

|  |
| --- |
| Response: The authors thank the reviewer for raising this important methodological point. In the revised manuscript, Section 4.4 has been expanded to provide a clear rationale for selecting PLS-SEM over multilevel modeling (MLM). Although the experimental design involved a within-subjects essay revision procedure, the primary focus of the analysis was not on modeling nested or longitudinal change within individuals. Rather, the objective was to examine how pre-existing learner characteristics (e.g., ISAA and IP) moderated the effects of LLMF on post-intervention performance. For this purpose, PLS-SEM was considered more appropriate, as it is well suited for exploratory models involving latent constructs, interaction terms, and small sample sizes, while MLM is primarily intended for hierarchical or longitudinal data structures. |

1. The literature review is extensive, yet the discussion could more explicitly position this system in contrast to rule-based or teacher-scripted feedback systems. What are the comparative tradeoffs in accuracy, adaptability, and student trust?

|  |
| --- |
| Response: We thank the reviewer for this valuable suggestion. In response, we have revised both the literature review and the Discussion to explicitly contrast the proposed LLMF system with traditional rule-based or teacher-scripted approaches. Specifically, in Section 2.3, we now highlight that rule-based and scripted systems, while ensuring consistency, are often criticized for their rigidity and limited adaptability (Zawacki-Richter et al., 2019). This addition helps situate our system within the broader evolution of automated feedback technologies. Furthermore, in the Discussion section, we note that the proposed LLMF system offers greater adaptability and scalability with comparable accuracy, while sustaining student trust remains an essential design consideration. These revisions aim to clarify the comparative tradeoffs and position the contribution of the present system more explicitly. |

1. While the feedback engine is technically described, the tone control or pedagogical alignment of generated feedback is not evaluated. As tone is crucial in educational messaging, a brief discussion or example would be useful.

|  |
| --- |
| Response: The authors thank the reviewer for raising this important point. We agree that tone and pedagogical alignment are crucial in educational messaging, as emphasized in prior work on effective feedback. While the present study focused primarily on the effects of LLMF on SAA and writing improvement, the design did not systematically evaluate tone or pedagogical fit. To address this limitation, the Limitations section has been revised to explicitly acknowledge that the tone and pedagogical alignment of generated feedback were not assessed and that future research should explore how these dimensions influence student uptake and trust. This addition ensures that the manuscript transparently recognizes the boundary of the current design and outlines a concrete avenue for further investigation. |

1. There is no discussion of system robustness in real-world conditions—e.g., how does it respond to irrelevant input, unusually short essays, or adversarial prompts? A mention of system safeguards or edge case handling would be reassuring.

|  |
| --- |
| Response: The authors thank the reviewer for raising the important point regarding system robustness in real-world conditions. In response, the section 3.2 has been revised to clarify the error-handling and safeguard mechanisms integrated into the system. Specifically, beyond standard preprocessing operations such as tokenization, normalization, and input validation, the revised description now notes that the system incorporates minimum-length requirements, rejection of irrelevant input, and basic filtering rules for adversarial prompts. These design considerations are intended to enhance the reliability and robustness of the system when confronted with atypical or low-quality input. |

1. Participants rated their writing using a single-item Likert scale. This is a practical choice, but it may lack nuance in capturing students’ evaluative reasoning. The authors might consider this as a limitation or suggest future use of multi-item scales.

|  |
| --- |
| Response: We appreciate the reviewer’s thoughtful observation. To address this, Section 4.3 has been revised to clarify the rationale for using a single-item measure. Acknowledging the potential limitations of this approach, we now explain that it was deliberately chosen to minimize participant fatigue and preserve cognitive focus during repeated assessment cycles. Additionally, we highlight that this design choice was appropriate for capturing overall evaluative judgment in our experimental context. The revised paragraph also acknowledges that future studies may benefit from adopting multi-item scales to better capture nuanced evaluative reasoning. |

1. The distinction between students’ perceived usefulness of feedback and their actual gains in self-assessment accuracy is not fully disentangled. Including a brief note on this divergence, possibly with reference to expectancy-value theory, would add conceptual depth to the discussion.

|  |
| --- |
| Response: Thank you for raising this important point. In response, we have revised Section 4.3 to acknowledge that while students’ perceived usefulness of feedback was not directly measured in the current study, expectancy-value theory (Eccles & Wigfield, 2002) suggests that learners may still perceive feedback as helpful even in the absence of measurable improvements in self-assessment accuracy. This addition offers conceptual clarification and anticipates potential divergences between subjective perception and actual performance metrics, as suggested by the reviewer. |

1. Finally, the conclusion makes a strong case for personalization, but the current system does not seem to adapt feedback tone or strategy based on learner profile. A brief roadmap toward more adaptive dialogue capabilities would be a valuable addition.

|  |
| --- |
| Response: We appreciate the reviewer’s thoughtful suggestion. To address this point, we have revised the concluding paragraph to explicitly mention the need for adaptive feedback mechanisms that tailor tone and strategy to individual learner profiles. While the current study did not implement such dynamic personalization, we recognize its importance for maximizing learner engagement and trust. The revised sentence underscores this direction by situating it within broader concerns of scalability, interoperability, and pedagogical robustness in real-world educational settings. Thank you for encouraging us to clarify this future pathway. |

Overall, the manuscript is theoretically rich, methodologically sound, and highly relevant. The clarifications above will help to communicate its contributions more effectively to both technical and pedagogical audiences.