**AI-Driven Intelligent Feedback System for Enhancing Self-Assessment Accuracy in Higher Education Writing**

**Abstract**

With the rapid advancement of generative artificial intelligence, large language models (LLMs) have become increasingly integrated into education, particularly for automated formative feedback and writing assessment. This study introduces and evaluates an AI-driven intelligent feedback system designed to enhance sustainable and inclusive higher education, leveraging transformer-based models (BERT and RoBERTa) to provide scalable, adaptive, and personalized writing support. The system aims to improve students’ self-assessment accuracy (SAA), a critical factor for self-regulated learning, while addressing the challenge of delivering high-quality feedback efficiently in under-resourced contexts. A quasi-experimental design was employed to examine the effects of LLM-generated feedback (LLMF) on students’ SAA and to investigate how these effects vary by initial ability. Results indicated no significant group-level difference in posttest SAA between the experimental and control groups. However, interaction analysis revealed a significant effect between feedback type and Initial Self-Assessment Accuracy (ISAA), indicating that students with lower baseline accuracy benefited more from LLMF, whereas those with higher baseline SAA showed no significant change. These findings highlight the potential of AI-driven intelligent feedback systems as cost-effective and sustainable solutions to reduce cognitive calibration gaps and foster metacognitive development. By embedding adaptive and personalized mechanisms, such systems advance educational equity, promote scalable personalized learning, and contribute to the broader agenda of intelligent and sustainable education.

**Keywords:** AI-driven intelligent feedback system, self-assessment accuracy, higher education, natural language processing, transformer-based models, sustainable education, expert systems.

1. Introduction

As artificial intelligence (AI) plays an increasingly critical role in education, AI-driven intelligent learning systems are emerging as essential tools for enhancing educational quality and equity. These systems not only reshape instructional methods but also offer innovative solutions for achieving sustainable education and scalable personalized support (Luckin, 2025). In rapidly evolving and autonomy-driven learning environments, students must exhibit a high level of self-regulated learning (SRL), including effective time and resource management, individualized goal setting, and continuous monitoring and evaluation of their strategies and progress (Chang et al., 2023). Within this framework, self-assessment becomes an indispensable component of SRL. The ability of learners to accurately judge their own progress and performance directly influences how they adjust their learning strategies and allocate resources. When such judgments are misaligned with actual performance, learning decisions may be misguided, ultimately compromising academic performance. To address this challenge, this study not only examines the pedagogical effects of AI feedback but also proposes a scalable AI-driven intelligent feedback system that integrates transformer-based models (BERT and RoBERTa) for automated essay evaluation, personalized writing support, and sustainable educational implementation. Here, sustainable refers to the system’s ability to deliver consistent, high-quality feedback over time with minimal additional human and financial resources, enabling large-scale deployment in under-resourced settings. Inclusive denotes its capacity to provide adaptive, personalized feedback that meets diverse learner needs, regardless of performance level, language background, or access to instructional support. Embedding these principles in the design promotes educational equity while ensuring scalability and cost-effectiveness.

Self-assessment accuracy (SAA) refers to the degree of alignment between students’ subjective evaluations and their actual academic performance (Wang et al., 2025). As a core element of SRL, SAA significantly affects whether students can realistically recalibrate their goals and strategies based on their learning status. However, research has shown that many students tend to overestimate their performance, leading to biased judgments that impede effective regulation (Panadero et al., 2016). This has led to increasing calls for concrete and systematic interventions to enhance students’ SAA (Luo and Zhou, 2024). Among various strategies, targeted feedback has been identified as one of the most effective methods. It helps learners gain a clearer understanding of task requirements and performance standards, thereby improving SAA, particularly in higher education settings (Braumann et al., 2024). Students with relatively low performance tend to demonstrate weaker SAA and are more dependent on feedback from instructors or AI systems. By leveraging a layered system architecture that combines handcrafted linguistic features with contextualized transformer embeddings, the proposed feedback system not only delivers accurate and adaptive feedback but also reduces instructor workload, contributing to the scalability and sustainability of educational practices.

Through timely and concrete feedback, students receive clearer learning guidance, correct misconceptions, and modify strategies accordingly, which in turn fosters both motivation and metacognitive development (Liu et al., 2025). For educators, however, providing high-quality feedback is a cognitively and time-intensive task, especially for complex activities such as writing (Meyer et al., 2024). With the rapid development of large language models (LLMs), AI has shown great potential in augmenting feedback processes. LLMs can deliver immediate and personalized suggestions at lower marginal cost, expanding the availability and sustainability of feedback mechanisms in under-resourced educational settings. For instance, Meyer et al. (2024) demonstrated that LLM-generated feedback (LLMF) positively influences learning performance, including writing performance, motivation, and emotional engagement.

Nonetheless, evidence regarding the effectiveness of LLMF in supporting SAA remains inconclusive. Although LLMs can generate fast and consistent responses, concerns remain about their accuracy, contextual appropriateness, and cognitive scaffolding compared to human teachers (Meyer et al., 2024). LLM feedback often lacks situational sensitivity and may fall short in guiding strategic thinking or identifying nuanced errors, which can limit its impact on SAA. Lew et al. (2010) emphasized that providing students with explicit performance standards or structured feedback is especially important for those with initially poor performance, as such guidance can help reconstruct their understanding and strategic awareness.

This study aims to investigate whether LLMF can be as effective as teacher-provided feedback in improving students’ SAA. Specifically, it examines the capacity of intelligent AI systems to identify learners’ Initial Self-Assessment Accuracy (ISAA) levels and performance gaps, and to provide differentiated and compensatory feedback accordingly. Through a randomized controlled trial, we explore changes in students’ SAA after receiving LLM feedback, and the relationship between these changes and their academic performance. Additionally, we examine how students’ initial performance (IP) and ISAA serve as moderators in this process. The study aims to develop an adaptive and differentiated AI feedback support model, with practical and sustainable potential for large-scale implementation in educational environments. By developing a robust system architecture that combines linguistic feature extraction with transformer-based embeddings, this research demonstrates how intelligent feedback systems can enhance educational equity, self-regulated learning, and scalable automated assessment. The study addresses the following research questions:

1. Can LLM-generated feedback improve students’ self-assessment accuracy?
2. Compared to higher-performing students, can lower-performing students achieve greater improvements in self-assessment accuracy through LLM-generated feedback?
3. Compared to students with higher initial self-assessment accuracy, can those with lower initial self-assessment accuracy benefit more from LLM-generated feedback in enhancing their self-assessment accuracy?

2. Related Works

**2.1. Empirical Foundations of Self-Assessment Accuracy and Feedback in Scalable Learning Contexts**

The concept of SAA, also referred to as calibration accuracy (Hacker and Bol, 2019) or metacognitive monitoring accuracy (de Bruin and van Merriënboer, 2017), denotes the degree of alignment between students’ self-evaluations and their actual academic performance. Self-assessment encompasses a variety of techniques and formats that support learners in monitoring their learning processes and evaluating their progress, thereby facilitating learning adjustments and improving performance (Yan and Brown, 2017). When students can accurately assess their learning status, they are more likely to set realistic goals, monitor progress effectively, and make informed adjustments to their learning strategies (Rickey et al., 2025). As a key component of self-regulated learning (SRL), SAA plays a central role in cognitive processes. The act of self-assessing has been shown to enhance learners’ reflective abilities and awareness of self-monitoring, serving as a critical developmental process for SRL skills (Andrade, 2019). For example, Thiede et al. (2010) demonstrated in the context of reading comprehension that students with higher SAA were better able to identify material that required further review, resulting in superior learning performance (LP). These findings underscore that SAA is not only a metacognitive indicator but also a prerequisite for effective learning. Students with high SAA are better equipped to recognize weaknesses, revise their strategies accordingly, and improve academic performance. Ernst et al. (2025) further noted that high SAA fosters more realistic understanding of one’s learning processes, reducing the risk of overconfidence or strategic misjudgments, and enhancing both motivation and learning efficiency.

Despite its importance, inaccurate self-assessment has been widely documented in educational research (León et al., 2023; Panadero et al., 2016), emphasizing the need for effective interventions to support SAA. According to cue utilization theory, students rely on various cues related to their performance when making self-judgments. However, the quality of these cues varies, which may compromise the accuracy of their evaluations (Kakaria et al., 2024). Koriat (1997) differentiated between diagnostic cues, which reliably predict performance, and nondiagnostic cues, which do not. For instance, students may judge their comprehension of a text based on reading speed, a nondiagnostic cue, resulting in flawed self-assessments. In contrast, when students receive textual feedback from teachers, considered a diagnostic cue, they are better positioned to make accurate evaluations.

To enhance SAA, students must be provided with effective diagnostic cues (Winstone et al., 2017). This highlights the critical role of feedback in delivering such cues to learners (Butler and Winne, 1995; Panadero et al., 2016). In a recent meta-analysis, Gutierrez de Blume (2022) found that feedback has a moderate positive effect on improving students’ SAA. Nonetheless, empirical studies exploring the interaction between students’ performance levels and feedback effectiveness remain relatively scarce (Maier and Klotz, 2025). Prior research has often examined SAA as a dependent outcome of feedback but has rarely considered SAA as a potential moderator that shapes how feedback operates. When students have low ISAA, it reflects underlying difficulties in calibration and monitoring. In such cases, structured feedback can serve as a calibration benchmark, helping students identify learning gaps and revise their strategies (Ernst et al., 2025; Nederhand et al., 2019). This aligns with findings on the pivotal role of diagnostic feedback in supporting self-monitoring and strategic regulation, particularly for high-risk learners (Wille et al., 2025). Investigating how students’ IP and ISAA jointly moderate the impact of feedback provides deeper insight into the mechanisms of feedback and offers practical guidance for designing personalized interventions in higher education settings.

**2.2. Feedback Literacy and Self-Assessment as Synergistic Components in Intelligent Systems**

Feedback literacy is a critical factor that determines whether students can engage in meaningful self-assessment and benefit from it. Although this study does not aim to redefine feedback literacy comprehensively, we examine it as a closely related capability that complements self-assessment. Feedback literacy encompasses not only the interpretation of evaluative information but also emotional regulation when responding to feedback, and an interactive understanding that transforms feedback into learning resources (Molloy et al., 2020). Carless and Boud (2018) define feedback literacy as the combination of understandings, skills, and dispositions that enable learners to interpret feedback effectively and use it to enhance their academic or professional practices. Subsequent research further identifies three core dimensions of embedding feedback literacy in curriculum design: proactively seeking relevant information, processing feedback effectively, and taking action based on received feedback (Malecka et al., 2020). Nicol (2021) introduces the concept of internal feedback, emphasizing that students should develop self-generated cognitive evaluation mechanisms throughout the feedback process to deepen their comprehension and application of feedback.

In intelligent learning environments, the development of feedback literacy no longer relies solely on teachers or peers. Rather, it involves the integration of real-time feedback, strategic prompts, and behavioral analytics provided by AI systems, allowing students to repeatedly engage in the processes of receiving, interpreting, and applying feedback. For instance, large language models or intelligent learning platforms can offer immediate comparisons to exemplars, guiding questions, or highlight discourse structures and argumentative gaps, fostering students' self-monitoring during writing or revision. Such feedback interfaces function not merely as knowledge transmission tools but as strategic environments for cultivating feedback literacy. Although feedback literacy and self-assessment are conceptually distinct (Kang et al., 2025), both serve essential roles in supporting self-regulated and lifelong learning (Boud, 1999; Winstone and Carless, 2019). To understand and implement them effectively in practice, an integrative perspective is necessary—one that examines the dynamic interplay between these two constructs. The process of self-assessment inherently provides multiple opportunities for developing feedback literacy. Students with higher levels of feedback literacy are more likely to engage in meaningful self-assessment. Like other literacies, feedback literacy develops progressively over time. With AI systems offering structured and responsive feedback scaffolds, students can independently practice interpreting and applying feedback even without continuous teacher involvement. This promotes enhanced self-monitoring and evaluative sensitivity.

However, how self-assessment contributes to the development of feedback literacy remains underexplored. The role of feedback literacy in facilitating self-assessment can be understood on two levels. First, self-assessment does not involve the self alone. The presence of others, including peers, teachers, or AI systems, is equally important in shaping how students seek and interpret feedback during the process (Yan and Brown, 2017). When "others" are replaced by intelligent agents such as AI-based feedback engines, students’ behavioral patterns in seeking and interpreting feedback may shift. Boud (1999) also noted that self-assessment requires students to actively seek feedback from their learning environment, including teachers, peers, and family members. Students with strong feedback literacy are more inclined to seek feedback proactively and are more aware of how factors such as expertise, credibility, and interpersonal dynamics influence the quality and reception of feedback. This awareness increases the likelihood that students will access feedback that genuinely supports self-assessment (Malecka et al., 2022).

Second, self-assessment generates internal feedback. Students compare their work against reference standards, resulting in self-generated feedback that supports various aspects of the evaluative process, such as setting criteria, identifying strengths and weaknesses, and adjusting learning strategies. Students with strong feedback literacy are more capable of generating high-quality internal feedback that is oriented toward learning, thereby making more effective use of self-assessment results to guide improvement (Yan, 2020). When AI systems are capable of dynamically generating evaluation criteria, offering exemplar-based comparisons, providing revision suggestions, and responding to students’ reflective inputs, internal feedback is no longer solely student-generated. Instead, it evolves into a co-constructed cognitive process between human and machine.

Feedback seeking refers to the learner's proactive acquisition of information related to their work or academic performance. It is regarded as a core behavioral component of feedback literacy, as it effectively bridges learners’ internal cognitive processes with external informational resources. Feedback seeking can be categorized into two strategies: inquiry and monitoring. Inquiry involves directly asking others for feedback on one’s progress or on perceived issues. Monitoring refers to gathering information from the environment, such as comparing one’s own performance with that of others, reviewing examples, examining assessment criteria, or consulting other resources (Ashford and Cummings, 1983; Joughin et al., 2021; Leenknecht et al., 2019). In intelligent learning systems, these strategies can be embedded as feedback triggers, self-selected learning pathways, or prompt-based guidance mechanisms that promote students’ agency and feedback awareness. External feedback alone is insufficient to promote learning gains. Only when students can process and apply such information to generate internal feedback does it contribute to meaningful learning improvement. Thus, internal feedback constitutes another key behavioral component of feedback literacy. Students construct meaning from feedback through self-insight and apply it to adjust future actions. When AI systems can simulate human-like feedback strategies and incorporate student inputs and reactions into the feedback generation process, the overall feedback experience becomes more personalized and interactive, further enhancing students’ capacity for learning transformation.

**2.3. The Transition from Human Feedback to AI-Based Intelligent Feedback Systems**

Formative assessment is designed to continuously adjust instructional content to meet students' needs (Filsecker and Kerres, 2012), although some scholars remain cautious regarding its effectiveness (Bennett, 2010). Within this framework, formative feedback plays a central role by linking assessment with learning improvement. It enables students to recognize performance gaps, revise learning strategies, and enhance their overall performance (Shute, 2008). According to Hattie and Timperley (2007), effective feedback should include three core components: feed up, feed back, and feed forward. These elements operate across different levels, including task, process, self-regulation, and personal dimensions.

Beyond content quality, several linguistic features—such as tone, clarity, prompt relevance, and message length—also influence the effectiveness of feedback. The tone of feedback should be encouraging while avoiding overly positive language (Kluger and DeNisi, 1996). A supportive tone fosters a constructive learning environment and helps maintain critical insight essential for growth (Brookhart, 2017). Overly vague messages can hinder understanding, making it crucial for feedback to be specific, direct, and easy to comprehend (Ossenberg et al., 2019). Regarding message length, while Kulhavy et al. (1985) suggest brevity and clarity, Van der Kleij et al. (2015) found that detailed feedback can lead to greater learning gains than oversimplified messages. These pragmatic characteristics determine the readability, uptake, and educational value of feedback and should be considered in the language generation design of AI-based feedback systems.

**Traditional automated feedback systems, such as rule-based or teacher-scripted approaches, ensure consistency and transparency but are often criticized for their rigidity and limited adaptability (Zawacki-Richter et al., 2019).** By contrast, AI-based intelligent feedback systems employ machine learning or large language models to generate, adapt, and guide student learning processes dynamically. These systems are characterized by real-time responsiveness, personalization, interactivity, and scalability. They are capable of simulating human-like feedback strategies and adjusting recommendations based on students' behavioral data through adaptive orchestration. For example, Shute (2007) noted that automated feedback is particularly beneficial for lower-performing students, while Zawacki-Richter et al. (2019) emphasized the potential of AI systems to reduce human bias and improve consistency in feedback delivery.

Recent studies have demonstrated the diverse applications and potential of large language models in educational settings. Nguyen et al. (2023) showed that large language models can accurately identify misconceptions in mathematical reasoning and offer appropriate guidance. Seßler et al. (2023) used GPT to provide writing feedback and observed improvements in students' composition quality. Gabbay and Cohen (2024) found that while large language models can detect coding errors, they still face limitations in generating high-quality feedback. Estévez-Ayres et al. (2024) reported that these models struggle with exercises involving concurrency errors, highlighting the difficulty of understanding complex programming concepts. Koutcheme et al. (2024) observed that large language models tend to offer excessively positive feedback in introductory programming courses, potentially overlooking critical issues. These findings suggest that although AI feedback systems hold great promise, they face ongoing challenges related to task specificity, pragmatic depth, and contextual integration.

Despite these advancements, the implementation of AI-based intelligent feedback systems continues to face several limitations and risks. While many systems can generate grammatically correct feedback, they often lack pedagogical intent, strategic focus, and contextual alignment, which can undermine feedback uptake (Zhai et al., 2020). The feedback generation logic generally fails to adjust for students' prior knowledge, learning styles, or emotional responses, thus lacking true personalization. Moreover, the communicability and reciprocity of the feedback process remain underdeveloped. Most AI systems currently deliver one-way feedback, limiting their ability to facilitate dynamic teacher-student interactions. For instance, Wu et al. (2023) applied a pretrained BERT model with zero-shot prompting to evaluate student writing, demonstrating the feasibility of using large language models for scoring tasks. Guo et al. (2024) developed a multi-agent system based on large language models to provide automated feedback in science education. While innovative, this system was not compared with teacher-generated feedback and lacked empirical validation in real classroom environments, thereby limiting its generalizability. Latif and Zhai (2024) also noted that GPT-based models outperform traditional BERT systems in automatic scoring tasks but carry risks of overgeneralization and misleading feedback.

In summary, AI-based intelligent feedback systems demonstrate notable strengths in providing real-time, scalable, and personalized learning support. However, their pragmatic depth, pedagogical alignment, and systematic design still require further development. Most existing studies focus on model accuracy and system performance, with limited attention to how feedback strategies can be integrated with learning process data or how content should be adapted for diverse learners. The mechanisms for human-AI collaboration and field-level implementation also remain immature, suggesting that AI feedback cannot yet fully replace human expertise or contextual sensitivity. As the literature indicates, realizing the full potential of AI systems for learning support will require advancements in multimodal feedback design, dynamic adaptation, and educational integration. These gaps form the basis of the current study’s research focus.

3. System Architecture

**3.1 Dataset and Preprocessing**

The dataset used in this study consists of 7,158 essays written by undergraduate students enrolled in Mandarin courses at the university’s General Education Center between January 2024 and July 2025. These texts include argumentative and narrative essays, providing a representative sample of students’ natural Chinese writing performance in an authentic educational context.

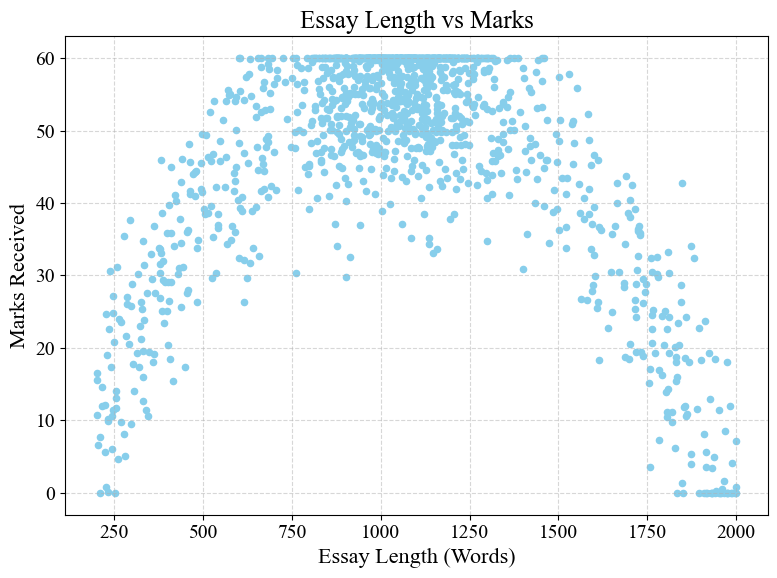
To reflect variations in task types and rhetorical goals, the dataset was organized into eight clusters based on essay topics and writing purposes. Each cluster represents a distinct discourse genre, emphasizing dimensions such as semantic organization, argumentative structure, or narrative technique. This categorization facilitates the language model’s ability to identify and respond to diverse rhetorical demands when generating feedback. Figure 1 presents the distribution of (a) essay lengths and (b) word counts in the training set, highlighting the importance of equipping the model with sufficient contextual processing capabilities to generate targeted and coherent feedback for extended texts.

|  |  |
| --- | --- |
|  | |
| (a) Distribution of Essay Lengths | (b) Distribution of Essay Words |

**Figure 1.** Distributions of (a) Essay Lengths and (b) Word Counts, Showing Input Variability That Necessitates Contextual Processing.

To examine the relationship between linguistic features and writing scores, lexical diversity was calculated using the Type-Token Ratio (TTR), as shown in Equation (1). TTR is a standard metric for measuring vocabulary variation within a text (Richards, 1987). It is defined as the ratio of the number of unique words (word types) to the total number of words (word tokens), providing a proxy for lexical richness.

The distribution of TTR values is illustrated in Figure 2. The mean TTR for the entire dataset is 0.34, with a standard deviation of 0.06. These results suggest that most students demonstrate a moderate level of lexical diversity. Although the relationship between TTR and essay scores is not strictly linear, higher-scoring essays generally exhibit greater lexical variation. This indicates that lexical richness may be associated with the persuasiveness and depth of content development in student writing.

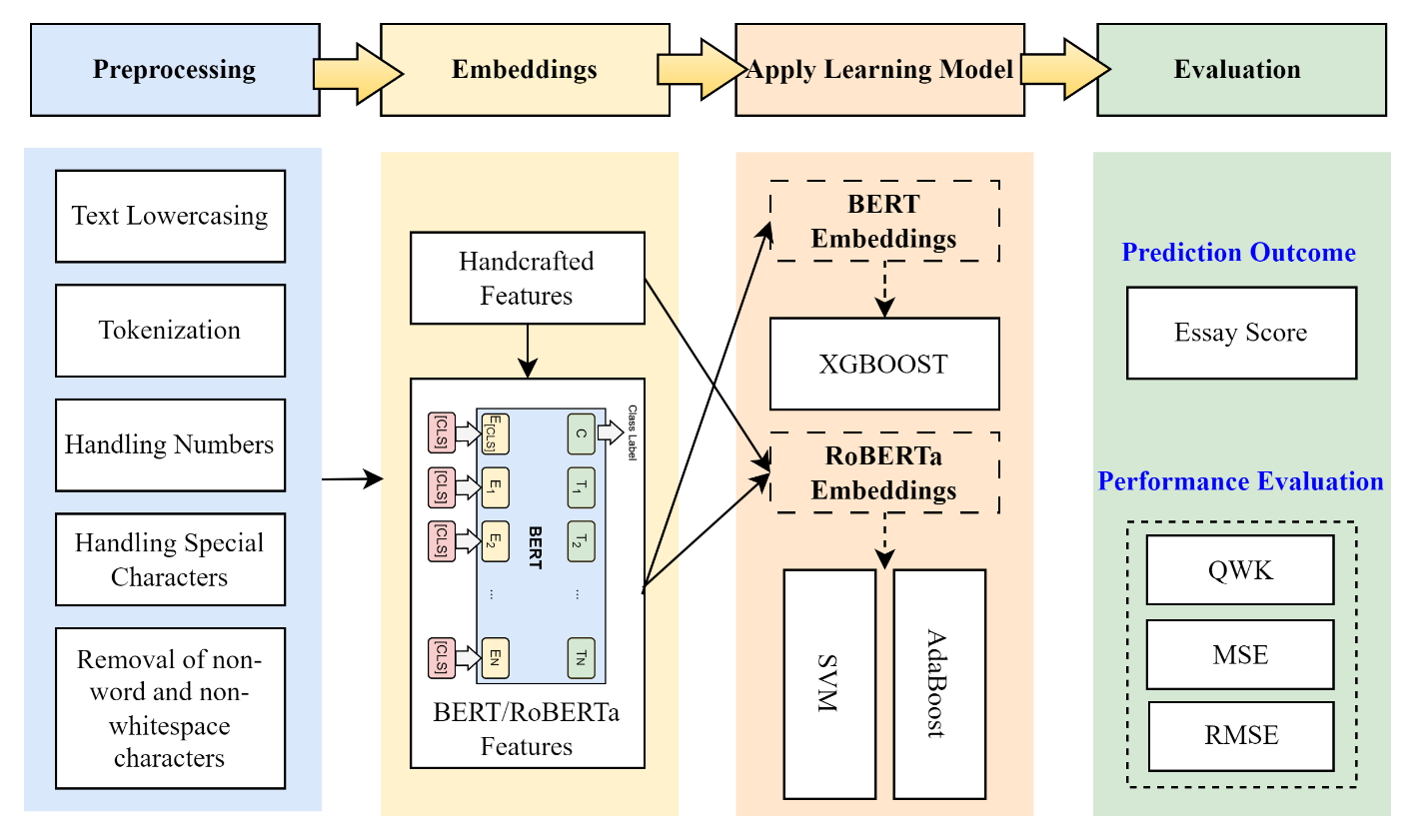


**Figure 2.** Average Essay Length vs. Marks.

**3.2 System Overview**

The overall architecture of the writing assessment system developed in this study is illustrated in Figure 3. During the preprocessing stage, several operations are applied to ensure that the text data are cleaned, standardized, and properly prepared for feature extraction and model training. These operations include text lowercasing, tokenization, number normalization, handling of special characters, and the removal of non-word and non-whitespace symbols.

Proper preprocessing is essential for ensuring model reliability. Inadequate processing may introduce bias, reduce feature quality, and compromise prediction accuracy. For instance, incorrect tokenization, improper handling of special characters, or missing punctuation can distort semantic meaning and diminish the effectiveness of subsequent embeddings. To address these issues, the system incorporates multiple error-handling mechanisms. These include input validation to allow only valid lexical items to enter subsequent stages, placeholder substitution or imputation methods to handle missing or erroneous data, and the normalization or removal of edge cases such as numerical values or special characters. In addition, basic safeguards such as minimum-length requirements, rejection of irrelevant input, and simple filtering rules for adversarial prompts were implemented to enhance robustness in real-world use.



**Figure 3.** Layered Architecture of the Proposed Essay Scoring Model.

To compute semantic similarity, the system employs a pretrained BERT model to generate embeddings for both student essays and reference texts. Cosine similarity is then used to quantify the semantic closeness between vectors, yielding a score ranging from 0 to 1. In terms of textual coherence, the system detects the presence and distribution of transition words such as "however" and "therefore" to evaluate logical progression across sentences. Lexical richness is further assessed using the metric described in Equation (2), which reflects the author's vocabulary diversity and cognitive maturity.

Once all features are computed, they are appended column-wise to the original dataset. This process is repeated for each essay until the entire dataset is fully annotated. Subsequently, the RoBERTa model is employed to extract contextualized embeddings that capture fine-grained semantic information within the essays. RoBERTa demonstrates strong capabilities in encoding linguistic structures and semantic associations. By integrating handcrafted features with RoBERTa embeddings, the system achieves a more comprehensive understanding of both the structural and semantic quality of student writing, thereby improving overall assessment accuracy.

**3.3 Transformer-Based Models**

3.3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is capable of capturing intricate contextual relationships and subtle semantic nuances within a text, which significantly enhances the accuracy of essay evaluation. Its bidirectional architecture enables the model to simultaneously process preceding and succeeding contexts, thereby facilitating the interpretation of sentence meaning, identification of linguistic cues, and maintenance of textual coherence. This allows for a more precise representation of the essay’s content.

Trained on extensive corpora, BERT demonstrates a strong ability to comprehend diverse linguistic conventions and writing styles. The contextualized embeddings generated by BERT encapsulate the essay content holistically, thereby improving the model’s capacity to assess writing quality and provide in-depth analysis. Prior to model input, handcrafted features are extracted and concatenated with BERT-based embeddings to form the final feature vector. This process is described in Equations (3), (4), and (5):

Here, ​ denotes the handcrafted feature vector of the -th essay, ​ represents the embedding produced by the BERT model, is the concatenated feature vector, and ϕ is the feature extraction function. During model training, the prediction of essay scores is formulated as shown in Equation (6):

Where is the predicted score for the -th essay based on the combined feature vector ​, and denotes the model parameters to be optimized.

3.3.2 RoBERTa

The Robustly Optimized BERT Approach (RoBERTa) is an improved variant of BERT designed to address its limitations. RoBERTa is trained on larger datasets and with longer input sequences, making it particularly well-suited for tasks such as Automated Essay Scoring (AES), which require the integration of both local and global contextual information. Equations (7), (8), and (9) describe the process of integrating RoBERTa-based feature vectors.

In this study, RoBERTa is implemented as a standalone model. It receives both RoBERTa-generated embeddings and handcrafted features as input, including syntactic error counts, semantic similarity scores, and lexical richness measures. The combined input allows the model to capture comprehensive linguistic attributes.

*3.3.3 Rationale for Dual-Model Embeddings*

To leverage the complementary strengths of different contextual language models, both BERT and RoBERTa embeddings were incorporated into the system. BERT provides robust general-purpose contextual representations, while RoBERTa, trained on a larger corpus with dynamic masking, captures more nuanced linguistic and syntactic features. By combining both models, the system benefits from a richer and more comprehensive representation of textual input.

The embeddings from both models were concatenated with handcrafted linguistic features to form a unified input vector for the regression model. No explicit weighting scheme was used. Instead, the model was trained to automatically learn the optimal contribution of each feature type. Since the main objective of this study was to examine the pedagogical impact of the feedback system, a direct performance comparison between BERT and RoBERTa was not conducted.

**3.4 Evaluation Metrics**

Evaluation metrics play a crucial role in assessing the performance of writing evaluation systems. These metrics provide a quantitative basis for measuring a model’s precision, consistency, and reliability, thereby determining whether its assessments of content and quality align with human expectations. In this study, the primary performance indicator is the Quadratic Weighted Kappa (QWK), which evaluates the level of agreement between predicted scores and human-assigned scores. The QWK is calculated as shown in Equation (10). In the equation, represents the observed score matrix (i.e., the confusion matrix between actual and predicted scores), denotes the expected score matrix assuming random distribution, is the weight matrix, and indicates the number of possible score levels.

Compared to traditional accuracy measures that only assess the correctness of classification, QWK accounts for the ordinal nature of the score categories and incorporates the degree of closeness between predicted and actual values. This provides a more discriminative and informative basis for model evaluation. To further examine the model's performance in predicting continuous variables, this study also includes Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as supplementary metrics. These are defined in Equations (11) and (12), where denotes the actual score, represents the predicted score, and is the total number of essays:

Together, QWK, MSE, and RMSE provide a comprehensive evaluation of the model’s performance in terms of both categorical agreement and numerical precision.

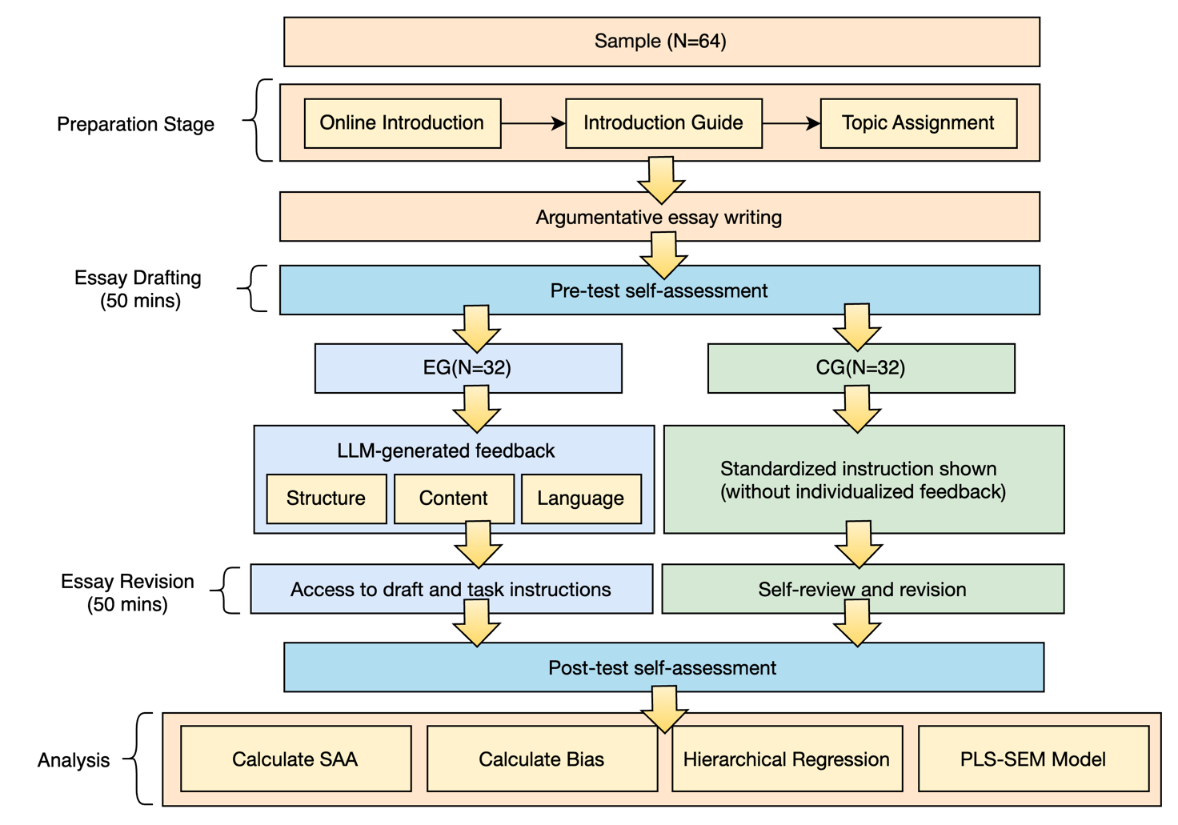
4. Method

**4.1 Study Sample**

The participants in this study consisted of 64 undergraduate students (*N* = 64) enrolled in the College of Engineering at a university in southern Taiwan. Data collection was conducted between April and July 2024. The experimental course was part of the university’s general education program, focusing on the development of humanities literacy. To ensure baseline equivalence in academic background and foundational abilities, students were assigned to either the experimental group (EG, *N* = 32) or the control group (CG, *N* = 32) using a stratified random assignment procedure. Among the participants, 56.75% were male with a mean age of 18.2 years, while 44.25% were female with a mean age of 18.7 years. This sample size was determined by the maximum course enrollment and equal group allocation. Although modest, it aligns with standard 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 sample sizes of 50–70 participants.

**4.2 Study Design and Procedure**

An overview of the experimental procedure is provided in Figure 4. The entire study was administered via computer using an online survey format. Before the experiment began, students were given a brief instructional overview. Following the instructions, both groups were asked to write an argumentative essay in Traditional Chinese. The writing session was limited to 50 minutes. Participants were randomly assigned one of two essay prompts, both requiring them to express personal opinions and provide supporting arguments and examples. The first prompt stated: “Do you agree with the following statement? A person’s success is determined by choice rather than talent. Provide specific reasons and examples to support your view.” The second prompt was: “Do you agree with the following statement? AI will eventually replace most human jobs. Provide specific reasons and examples to support your view.”



**Figure 4.** Experimental Procedure.

After completing the initial draft, all participants performed a self-assessment to evaluate their writing performance. Students in the experimental group then received feedback generated by the LLM system developed for this study and were instructed to revise their essays accordingly. The revision prompt was: “Please revise your essay based on the feedback provided by the system. Aim to improve it as much as possible and take sufficient time to complete the revision.” Representative examples of the feedback are shown in Table 1. Students in the control group also revised their essays but did not receive any feedback from the system. Instead, they were given a standardized instruction that asked them to reread their essays and revise to the best of their ability. This neutral prompt served to maintain procedural consistency across both groups while ensuring that only the experimental group received targeted feedback.

**Table 1.** Feedback Examples Generated by the LLM System.

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Revision Prompt** | **Specific Revision Suggestions** |
| Structure | The essay lacks clear openings and conclusions, making the overall structure difficult to follow. Readers may find it hard to grasp the flow of content. It is recommended to add transitional phrases between paragraphs to improve coherence. | Add introductions and concluding paragraphs. Arrange related ideas within the same paragraph and use transition or causal words such as "however," "therefore," or "on the other hand" to guide readers through the content. |
| Content | The arguments in the essay lack sufficient support and specific examples. The stance is not clearly stated, and the overall expression lacks clarity and consistency. | Add concrete examples to reinforce argumentation. Clearly express the stance to avoid vague or abstract statements, helping readers better understand the author's viewpoint. |
| Language | The essay contains several spelling and grammatical errors. Vocabulary is too simple, and some sentences are overly long. It is suggested to break down long sentences and enhance structure to improve readability. | Carefully check and correct spelling and grammatical errors. Try to use more precise and rich vocabulary. Break down long sentences using appropriate linking words such as "however," "in addition," or "for example." |
| Ambiguous / Borderline Cases | Sometimes it might rain, so it’s better to bring an umbrella. / He has a deep affection for his phone. / After watching the movie, Xiao Zhang told Xiao Li he should read more books. | For “sometimes it might,” suggest revising to either “sometimes it will” or “it might” to avoid redundancy. For “deep affection” with “phone,” suggest alternatives like “close relationship” or “strong interest” for more natural collocation. For the pronoun “he” in the last example, prompt the writer to clarify the reference, e.g., “Xiao Zhang told Xiao Li that Xiao Li should read more books” or “…that he himself should read more books.” |

Note*.* Original student compositions were written in Chinese; examples are translated for clarity.

During the revision process, students in both groups were allowed to access the original task prompt and their initial draft at any time. The maximum revision time was also set at 50 minutes for both groups. The only difference between the two conditions was that the experimental group received structured feedback and revision guidance, while the control group was provided only with a general instruction. After completing their revisions, all participants completed a post-feedback self-assessment to evaluate their updated writing performance. At the end of the experiment, demographic information such as age, gender, and final grades in the university’s Chinese language course was collected. The entire procedure was completed over three class sessions, totaling 150 minutes.

**4.3 Measures**

To assess self-perceived writing quality, students responded to a single-item measure after each essay task: “Please evaluate the essay you just completed. How would you rate its overall quality?” Ratings were given on a five-point Likert scale (1 = *very poor*, 5 = *very good*). To calculate SAA, the absolute difference between each student’s self-assigned score and the LLM-generated score was computed. While SAA served as the primary outcome variable, subjective perceptions of feedback usefulness were not directly measured. In line with expectancy-value theory (Eccles & Wigfield, 2002), it is possible that students perceived the feedback as helpful even if no measurable improvement occurred in SAA. A single-item scale was employed to minimize participant fatigue and preserve cognitive resources during the repeated assessment cycles. This approach was deemed sufficient for capturing overall evaluative judgment within the context of the current experimental design. The calculation followed the formula proposed by Schraw (2009), as shown in Equation (13), where N represents the number of student participants, denotes the self-assessed score of student *i*, and denotes the score assigned by the LLM system for the same student.

In addition to SAA, this study also examined self-assessment bias, defined as the simple difference between a student's self-assessed score and the score generated by the LLM system. A positive value indicates overestimation, while a negative value suggests underestimation. The bias was calculated using Schraw’s (2009) formula in Equation (14). Including bias in the analysis helps to determine whether changes in SAA reflect a reduction in either overestimation or underestimation. In this formula, represents the discrepancy between the self-assessed and system-generated scores. However, bias should not be interpreted as a direct indicator of absolute accuracy, as it may lead to misleading conclusions if considered in isolation.

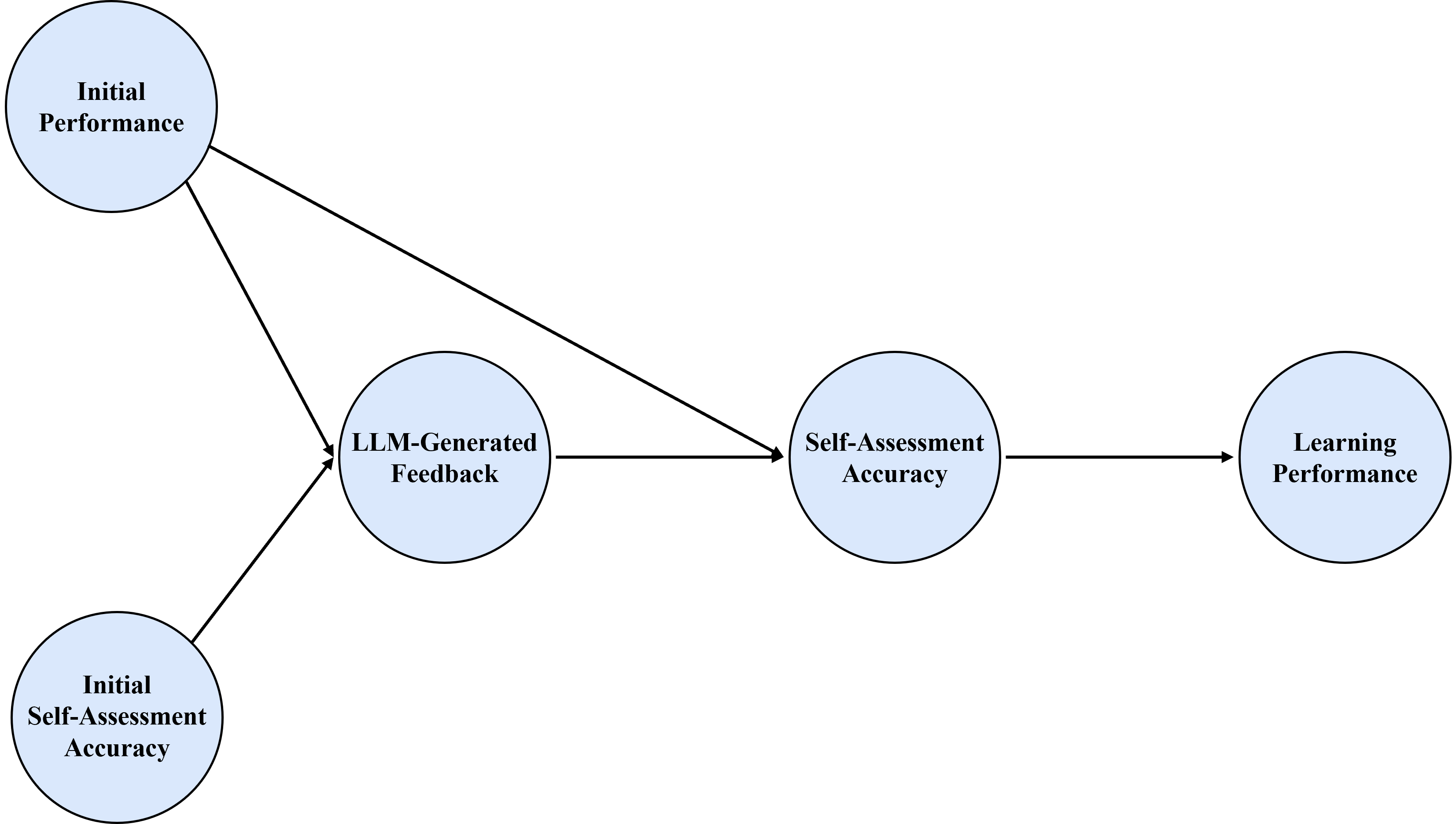
**4.4 Analytic Approach**

This study employed structural equation modeling (SEM) using SmartPLS 4 to analyze the proposed model. Partial least squares SEM is well-suited for complex models involving mediating and moderating variables. It accommodates both formative and reflective latent constructs and is appropriate for prediction-oriented research, especially with limited sample sizes.

Although the experimental procedure involved a within-subjects essay revision design, the primary analytical focus was not on modeling time-based or nested change at the individual level, but rather on examining how pre-existing learner characteristics moderated the effects of LLMF on post-intervention performance. Therefore, multilevel modeling (MLM) was not adopted, as it is primarily designed for analyzing hierarchical or longitudinal data structures. Instead, PLS-SEM provided a more appropriate framework for estimating interaction effects and latent variable relationships in the current design. Figure 5 presents the SEM model. To test the first research question, a direct path from LLMF to SAA was specified, and its significance was evaluated through bootstrapping.

To investigate how IP and ISAA influence the perceived usefulness of LLMF, two interaction terms were incorporated into the model. The first interaction examines the direct path from IP to LLMF, corresponding to Research Question 2. The second interaction tests the direct path from ISAA to LLMF, corresponding to Research Question 3. To reduce multicollinearity, all continuous variables were standardized before constructing the interaction terms.

In addition to the SEM analysis, hierarchical regression analyses were conducted using SPSS to provide supplementary validation. A three-step model was employed to assess changes in explanatory power. Model 1 included only the group variable to examine the main effect. Model 2 added IP and ISAA as pretest control variables. Model 3 introduced the two interaction terms to assess moderation effects.

****

**Figure 5.** Proposed Structural Equation Model.

5. Results

**5.1 Descriptive Statistics**

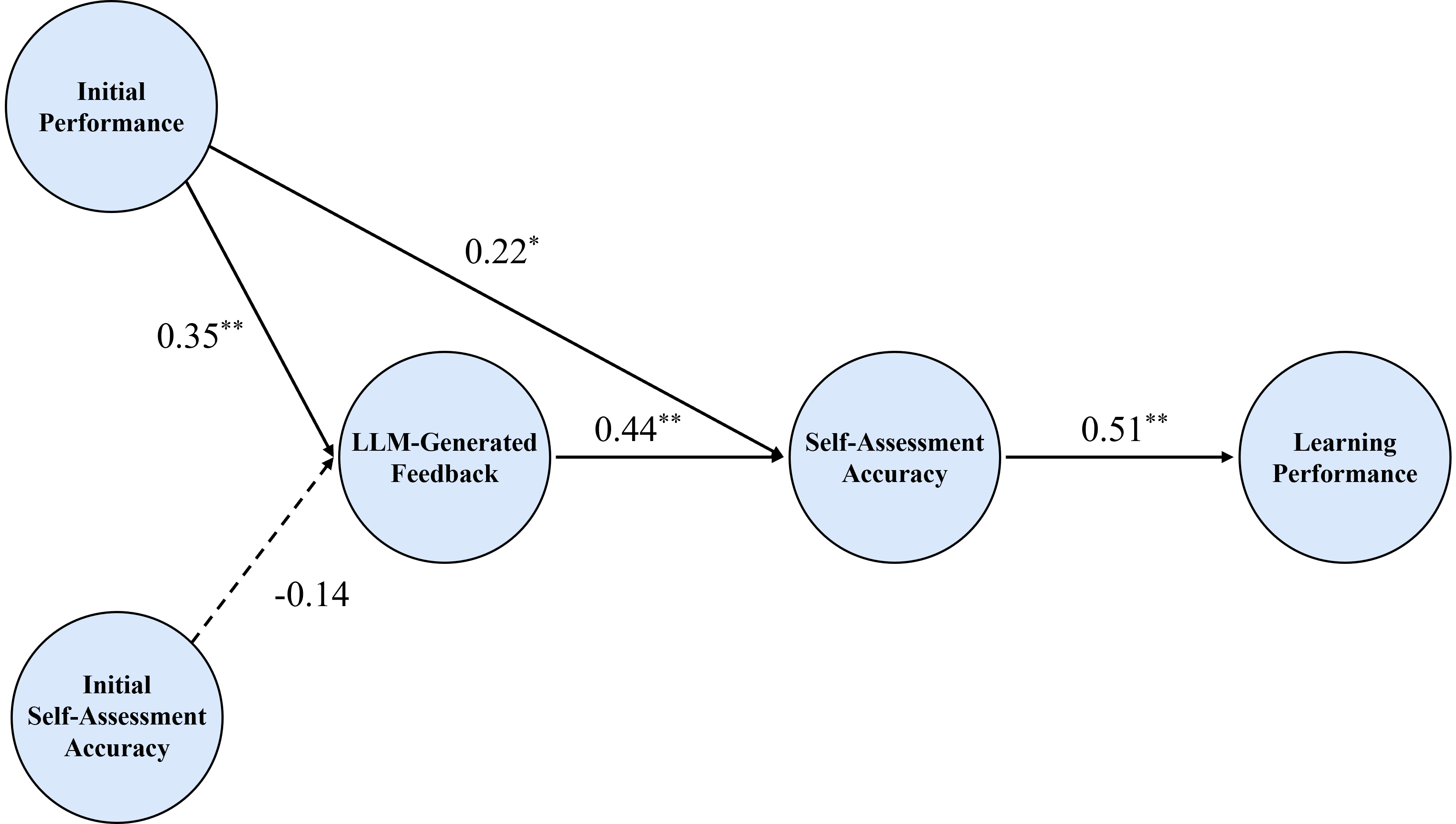
The means, standard deviations, and partial correlations of each construct in the structural equation model are presented in Table 2 and Figure 6. The average IP score was 2.43 (*SD* = 0.85), while the average ISAA was 1.82 (*SD* = 0.73), indicating a notable gap between performance and perceived accuracy. After interacting with the writing assessment system, students’ scores on LLMF reached an average of 3.01 (*SD* = 0.64), and SAA also improved compared to ISAA (*M* = 2.10, *SD* = 0.81). The final LP score averaged 2.89 (*SD* = 0.76).

Regarding partial correlations, IP showed a positive partial correlation with both LLMF (*r* = 0.35, *p* < .01\*\*) and LP (*r =* 0.40, *p* < .01\*\*), indicating that students with stronger initial abilities were more likely to receive extensive feedback and achieve better learning performance. ISAA was positively correlated with SAA (*r* = 0.39, *p* < .01\*\*), suggesting that students with higher initial assessment accuracy were more likely to retain accurate self-evaluation skills after learning. LLMF exhibited moderate positive partial correlations with both SAA (*r* = 0.44, *p* < .01\*\*) and LP (*r* = 0.36, *p* < .01\*\*), implying that feedback generated by the LLM system may provide meaningful support in improving SAA and learning performance. The highest partial correlation was observed between SAA and LP (*r* = 0.51, *p* < .01\*\*), revealing a stable and positive relationship between accurate self-assessment and enhanced LP.

**Table 2.** Means, Standard Deviations, and Partial Correlations.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | *Mean* | *SD* | IP | ISAA | LLMF | SAA | LP |
| Initial Performance (IP) | 2.43 | 0.85 |  |  |  |  |  |
| Initial Self-Assessment Accuracy (ISAA) | 1.82 | 0.73 | -0.28\*\* |  |  |  |  |
| LLM-Generated Feedback (LLMF) | 3.01 | 0.64 | 0.35\*\* | -0.14 |  |  |  |
| Self-Assessment Accuracy (SAA) | 2.10 | 0.81 | 0.22\* | 0.39\*\* | 0.44\*\* |  |  |
| Learning Performance (LP) | 2.89 | 0.76 | 0.40\*\* | -0.10 | 0.36\*\* | 0.51\*\* |  |

Note. \**p* < .05; \*\**p* < .01; \*\*\**p* < .001



*Note.* Solid lines represent significant paths (\**p* < .05; \*\**p* < .01; \*\*\**p* < .001); dashed line represents a nonsignificant path.

**Figure 6.** Path Coefficients of the Structural Model.

**5.2 Evaluation of Hypotheses**

To test the research hypotheses, a three-step hierarchical regression analysis was conducted. The results are presented in Table 3 and Table 4. In Model 1, only the group variable was included to examine the direct effect of LLMF on SAA, addressing RQ1. The result showed that the main effect of LLMF was not statistically significant (*β* = 0.10, *p* = .098), indicating that RQ1 was not supported. In Model 2, IP and ISAA were added as pretest covariates. The analysis revealed that both IP (*β* = 0.17, *p* = .011) and ISAA (*β* = 0.43, *p* < .001\*\*\*) were significant positive predictors of SAA, and the model's explanatory power improved significantly (*ΔR²* = 0.12). Model 3 introduced two interaction terms, LLMF \* IP and LLMF \* ISAA, to examine whether feedback effects varied based on students' IP measures. The interaction between LLMF and IP was not significant (*β* = 0.07, *p* = .243), providing no support for RQ2. However, the interaction between LLMF and ISAA was significant (*β* = 0.22, *p* < .01\*\*), supporting RQ3. The final model explained 35 percent of the variance in SAA (*R²* = .35), showing a notable improvement over Model 2 (*ΔR²* = 0.12).

**Table 3.** Hierarchical Regression Analysis Predicting Students’ Self-Assessment Accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Predictor | *B* | *SE* | *β* | *t* | *p* |
| LLMF (EG = 1) | 0.10 | 0.06 | 0.10 | 1.67 | .098 |
| Initial Performance (IP) | 0.18 | 0.07 | 0.17 | 2.57 | <.05\* |
| Initial Self-Assessment Accuracy (ISAA) | 0.41 | 0.05 | 0.43 | 8.20 | <.001\*\*\* |
| LLMF \* IP | 0.07 | 0.06 | 0.07 | 1.17 | .243 |
| LLMF \* ISAA | 0.22 | 0.06 | 0.22 | 3.67 | <.01\*\* |

Note. \**p* < .05; \*\**p* < .01; \*\*\**p* < .001

**Table 4.** Hierarchical Regression Model Summary Predicting Students’ Self-Assessment Accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | *R²* | *ΔR²* | *F(df)* | *ΔF* | *p* |
| Model 1 | 0.17 |  | 10.28(1, 63) |  |  |
| Model 2 | 0.29 | 0.12 | 18.94(1, 63) | 15.3 | <.001\*\*\* |
| Model 3 | 0.35 | 0.06 | 19.76(1, 63) | 8.20 | <.001\*\*\* |

Note. Model 1 included only the group variable. Model 2 added IP and ISAA as predictors. Model 3 further included two interaction terms: LLMF \* IP and LLMF \* ISAA.

To further illustrate the interaction effects, Figure 7 presents plots of group differences by pretest SAA and bias. Subfigure (a) shows that among students with higher ISAA, those in the experimental group exhibited greater improvement in posttest accuracy than those in the control group, indicating a positive interaction. Subfigure (b) illustrates the trend differences in bias change between the two groups.

|  |  |
| --- | --- |
| (a) Interaction between Pre SAA and Group | (b) Interaction between Pre Bias and Group |

*Note.* Lower SAA values indicate higher self-assessment accuracy. Negative bias indicates underestimation, while positive bias indicates overestimation.

**Figure 7.** Interaction Effects Between Group and Pre-intervention Variables on Post-assessment Performance.

6. Discussion

Although students in the experimental group showed an overall improvement in post-intervention SAA, the difference compared to the control group was not statistically significant. This finding suggests that LLMF does not exert uniform positive effects across all learners. It aligns with Hattie and Timperley's (2007) assertion that the effectiveness of feedback depends heavily on students' ability to interpret and apply it. This also echoes Winstone et al. (2017), who emphasized the prerequisites for effective feedback uptake. Even when feedback is structurally aligned with quality standards, it may fail to prompt internalized self-regulatory processes if learners lack sufficient feedback literacy (Carless & Boud, 2018).

The study further examined whether IP moderated the effects of LLMF on SAA. Results showed no significant interaction between IP and LLMF. This indicates that students' IP did not influence how much they benefited from the feedback. Learners with lower performance but strong reflection and calibration skills were still able to self-correct effectively, whereas high-performing students lacking reflective habits might not respond adequately to feedback (Lew et al., 2010).

A significant interaction was found between ISAA and LLMF, indicating that students with initially lower SAA demonstrated notable improvements after receiving feedback from the LLM system. This finding aligns with earlier work by Butler and Winne (1995) and Koriat (1997), who emphasized that externally provided calibration cues help learners identify cognitive biases and adjust their strategies accordingly. Students with poor ISAA often lack internal benchmarks and evaluation strategies, making high-quality external feedback particularly crucial for effective self-regulation (Ernst et al., 2025). In contrast, students who already possess accurate self-monitoring skills may exhibit limited improvement due to having less room to grow. It is also possible that the feedback or task complexity was not sufficiently demanding to trigger deeper recalibration in these students. As Winstone and Carless (2019) noted, such learners may require more nuanced, domain-specific, or reflective prompts to engage meaningfully with feedback. Future feedback systems should therefore incorporate adaptive scaffolds that adjust both in depth and specificity according to individual learner profiles.

Building on this finding, a key practical implication is that identifying learners with lower ISAA early in the learning process allows educators and intelligent systems to allocate feedback resources more strategically. Providing these students with detailed guidance, exemplars, and revision prompts can enhance their metacognitive monitoring and close calibration gaps more effectively. Such targeted support addresses disparities in self-regulated learning skills and aligns with Winstone et al. (2017) in advocating feedback processes that are responsive to learner profiles. This approach positions AI-based feedback systems not merely as generic performance enhancers but as equity-oriented tools capable of adapting to diverse student needs. In line with prior reviews, while rule-based or teacher-scripted feedback is often rigid or resource-intensive, LLM-based systems offer greater adaptability and scalability with comparable accuracy, though sustaining student trust remains essential (Zawacki-Richter et al., 2019).

Overall, this study offers nuanced empirical evidence regarding the role of LLMF in supporting self-assessment. The findings suggest that such feedback should not be treated as universally effective, but rather as a targeted support tool that is especially beneficial for students prone to overestimating their performance. The study expands the current understanding of LLMF by incorporating individual differences into the analysis of self-assessment performance (Meyer et al., 2024; Panadero et al., 2016). Although the feedback is automatically generated, it can still provide timely and personalized support to learners with weaker self-assessment skills. Therefore, implementing LLM feedback mechanisms should avoid a one-size-fits-all approach. Instead, differentiated and adaptive feedback strategies should be developed to maximize the educational potential of LLMs.

7. Conclusion

This study empirically examined the design and effects of an AI-driven intelligent feedback system powered by LLMs on SAA, with particular attention to how learner characteristics shape the effectiveness of such feedback. Although the overall treatment effect was not statistically significant, the results revealed a compensatory benefit for students with lower ISAA. These findings highlight the importance of considering individual differences when deploying AI-based feedback systems in educational settings.

From a theoretical perspective, this study contributes to the literature on formative feedback, feedback literacy, and self-assessment processes. It demonstrates that LLMF, when tailored to learner needs, can support cognitive calibration and enhance metacognitive engagement. These findings reinforce the potential of LLMs as adaptive tools for fostering reflective learning. From a practical perspective, the results suggest that intelligent feedback should not be applied uniformly. Effective feedback design must be responsive to students’ baseline competencies and self-monitoring abilities. Systems should incorporate diagnostic and reasoning mechanisms capable of dynamically adjusting the level, tone, and content of feedback according to individual learner profiles and performance patterns. Such personalization has the potential to support scalable, equitable, and sustainable learning performance, aligning with the broader goals of intelligent education.

Several limitations should be acknowledged to inform future research. First, while the feedback design was grounded in established theories, it did not incorporate knowledge of results, which may have reduced its value for students with already high SAA. Second, the study’s short intervention period, coupled with its analytical focus, may have limited the opportunity to capture deeper reflective change or explore mediation effects such as the role of SAA in shaping learning performance. Third, although the use of a neutral prompt in the control group enhanced internal validity, it constrained comparison with traditional instructor feedback and limited exploration of tone or pedagogical alignment—both essential for effective educational messaging. Finally, while the system avoids biometric data, its reliance on linguistic input raises potential privacy concerns, and its robustness across more complex genres or among non-native writers remains to be validated.

To address these limitations, future research should incorporate longer-term interventions, comparative studies with human-delivered feedback, and the integration of privacy-preserving mechanisms. Larger and more diverse samples will also be needed to enable mediation analyses that clarify the role of SAA in shaping learning performance. Moreover, extending the system to support genre-specific writing and multilingual learners, while simultaneously addressing deployment issues such as computational efficiency, real-time feedback delivery, and interoperability with learning management systems, and adaptive feedback mechanisms that personalize tone and strategy based on learner profiles, will be essential to ensure the robustness, scalability, and practical adoption of AI-driven feedback systems in authentic educational environments.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data that support the findings of this study are unavailable due to privacy and ethical restrictions. The nature of this research involves sensitive data which, if shared publicly, could compromise the privacy of individuals or groups studied. For further information on the data and its limitations, interested researchers may contact the corresponding authors.

**Conflicts of Interest:** The authors declare no conflict of interest.

References

Andrade, H. L. (2019, August). A critical review of research on student self-assessment. In *Frontiers in education* (Vol. 4, p. 87). Frontiers Media SA. <https://doi.org/10.3389/feduc.2019.00087>

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance*, *32*(3), 370-398. <https://doi.org/10.1016/0030-5073(83)90156-3>

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2-3), 70-91. <https://doi.org/10.1080/15366367.2010.508686>

Boud, D. (1999). Avoiding the traps: Seeking good practice in the use of self assessment and reflection in professional courses. *Social work education*, *18*(2), 121-132. <https://doi.org/10.1080/02615479911220131>

Braumann, S., van de Pol, J., Kok, E., Pijeira-Díaz, H. J., van Wermeskerken, M., de Bruin, A. B., & van Gog, T. (2024). The role of feedback on students’ diagramming: Effects on monitoring accuracy and text comprehension. *Contemporary Educational Psychology*, *76*, 102251. <https://doi.org/10.1016/j.cedpsych.2023.102251>

Brookhart, S. M. (2017). *How to give effective feedback to your students*. Ascd.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245-281. <https://doi.org/10.3102/00346543065003245>

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>

Chang, D. H., Lin, M. P. C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, *15*(17), 12921. <https://doi.org/10.3390/su151712921>

de Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, *51*, 1-9. <https://doi.org/10.1016/j.learninstruc.2017.06.001>

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, *53*(1), 109-132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Ernst, H. M., Prinz-Weiß, A., Wittwer, J., & Voss, T. (2025). Discrepancy between performance and feedback affects mathematics student teachers’ self-efficacy but not their self-assessment accuracy. *Frontiers in Psychology*, *15*, 1391093. <https://doi.org/10.3389/fpsyg.2024.1391093>

Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., Alario-Hoyos, C., & Delgado Kloos, C. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International journal of artificial intelligence in education*, 1-17. <https://doi.org/10.1007/s40593-024-00406-0>

Filsecker, M., & Kerres, M. (2012). Repositioning Formative Assessment from an Educational Assessment Perspective: A Response to Dunn & Mulvenon (2009). *Practical Assessment, Research & Evaluation*, *17*(16), n16.

Gabbay, H., & Cohen, A. (2024, July). Combining LLM-generated and test-based feedback in a MOOC for programming. In *Proceedings of the eleventh ACM conference on learning@ scale* (pp. 177-187). <https://doi.org/10.1145/3657604.3662040>

Guo, S., Latif, E., Zhou, Y., Huang, X., & Zhai, X. (2024). Using generative AI and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*. <https://doi.org/10.48550/arXiv.2411.07407>

Gutierrez de Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology*, *114*(4), 681.

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. <https://doi.org/10.1017/9781108235631.026>

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112. <https://doi.org/10.3102/003465430298487>

Joughin, G., Boud, D., Dawson, P., & Tai, J. (2021). What can higher education learn from feedback seeking behaviour in organisations? Implications for feedback literacy. *Assessment & Evaluation in Higher Education*, *46*(1), 80-91. <https://doi.org/10.1080/02602938.2020.1733491>

Kakaria, S., Simonetti, A., & Bigne, E. (2024). Interaction between extrinsic and intrinsic online review cues: perspectives from cue utilization theory. *Electronic Commerce Research*, *24*(4), 2469-2497. <https://doi.org/10.1007/s10660-022-09665-2>

Kang, C., Huang, J., Liu, Y., & Yin, H. (2025). Development and validation of a generic self-assessment scale for K-12 teachers as feedback givers: Insights from item response theory and factor analysis. *Humanities and Social Sciences Communications*, *12*(1), 1-10. <https://doi.org/10.1057/s41599-025-04927-4>

Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). New York, NY: Guilford Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General*, *126*(4), 349. <https://doi.org/10.1037/0096-3445.126.4.349>

Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J., & Denny, P. (2024). Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (pp. 52-58). <https://doi.org/10.1145/3649217.3653612>

Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. *Contemporary Educational Psychology*, *10*(1), 28-37. <https://doi.org/10.1016/0361-476X(85)90003-7>

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. <https://doi.org/10.1016/j.caeai.2024.100210>

Leenknecht, M., Hompus, P., & van der Schaaf, M. (2019). Feedback seeking behaviour in higher education: the association with students’ goal orientation and deep learning approach. *Assessment & Evaluation in Higher Education*, *44*(7), 1069-1078. <https://doi.org/10.1080/02602938.2019.1571161>

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review*, *35*(4), 106. <https://doi.org/10.1007/s10648-023-09819-0>

Lew, M. D., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self‐assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, *35*(2), 135-156. <https://doi.org/10.1080/02602930802687737>

Liu, C. C., Hwang, G. J., Yu, P., Tu, Y. F., & Wang, Y. (2025). Effects of an automated corrective feedback-based peer assessment approach on students’ learning achievement, motivation, and self-regulated learning conceptions in foreign language pronunciation. *Educational technology research and development*, 1-22. <https://doi.org/10.1007/s11423-025-10484-z>

Luckin, R. (2025). Nurturing human intelligence in the age of AI: rethinking education for the future. *Development and Learning in Organizations: An International Journal*, *39*(1), 1-4. <https://doi.org/10.1108/DLO-04-2024-0108>

Luo, R. Z., & Zhou, Y. L. (2024). The effectiveness of self‐regulated learning strategies in higher education blended learning: A five years systematic review. *Journal of Computer Assisted Learning*, *40*(6), 3005-3029. <https://doi.org/10.1111/jcal.13052>

Maier, U., & Klotz, C. (2025). Students ignore their mistakes: Elaborated error feedback processing in a digital learning system. *Contemporary Educational Psychology*, 102395. <https://doi.org/10.1016/j.cedpsych.2025.102395>

Malecka, B., Boud, D., & Carless, D. (2022). Eliciting, processing and enacting feedback: mechanisms for embedding student feedback literacy within the curriculum. *Teaching in Higher Education*, *27*(7), 908-922. <https://doi.org/10.1080/13562517.2020.1754784>

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students’ text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. <https://doi.org/10.1016/j.caeai.2023.100199>

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, *45*(4), 527-540. <https://doi.org/10.1080/02602938.2019.1667955>

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, *33*(6), 1068-1079. <https://doi.org/10.1002/acp.3548>

Nguyen, H. A., Stec, H., Hou, X., Di, S., & McLaren, B. M. (2023, August). Evaluating chatgpt’s decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning* (pp. 278-293). Cham: Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-42682-7_19>

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in higher education*, *46*(5), 756-778. <https://doi.org/10.1080/02602938.2020.1823314>

Ossenberg, C., Henderson, A., & Mitchell, M. (2019). What attributes guide best practice for effective feedback? A scoping review. *Advances in Health Sciences Education*, *24*(2), 383-401. <https://doi.org/10.1007/s10459-018-9854-x>

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational psychology review*, *28*, 803-830. <https://doi.org/10.1007/s10648-015-9350-2>

Richards, B. (1987). Type/token ratios: What do they really tell us?. *Journal of child language*, *14*(2), 201-209. <https://doi.org/10.1017/S0305000900012885>

Rickey, N., Panadero, E., & DeLuca, C. (2025). How do students self-assess? examining the metacognitive processes of student self-assessment. *Metacognition and Learning*, *20*(1), 1-29. <https://doi.org/10.1007/s11409-025-09430-4>

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning*, *4*, 33-45. <https://doi.org/10.1007/s11409-008-9031-3>

Seßler, K., Xiang, T., Bogenrieder, L., & Kasneci, E. (2023, August). Peer: Empowering writing with large language models. In *European conference on technology enhanced learning* (pp. 755-761). Cham: Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-42682-7_73>

Shute, V. J. (2007). Focus on formative feedback. *ETS Research Report Series*, *2007*(1), i-47. <https://doi.org/10.1002/j.2333-8504.2007.tb02053.x>

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189. <https://doi.org/10.3102/0034654307313795>

Sideridis, G., Simos, P., Papanicolaou, A., & Fletcher, J. (2014). Using structural equation modeling to assess functional connectivity in the brain: Power and sample size considerations. *Educational and psychological measurement*, *74*(5), 733-758. [https://doi.org/10.1177/0013164414525](https://doi.org/10.1177/0013164414525397)

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331-362. <https://doi.org/10.1080/01638530902959927>

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis. *Review of educational research*, *85*(4), 475-511. <https://doi.org/10.3102/0034654314564881>

Wang, W. S., Lin, C. J., Lee, H. Y., Huang, Y. M., & Wu, T. T. (2025). Enhancing self-regulated learning and higher-order thinking skills in virtual reality: the impact of ChatGPT-integrated feedback aids. *Education and Information Technologies*, 1-27. <https://doi.org/10.1007/s10639-025-13557-x>

Wille, E., Opheim, H. M. S., Kisa, S., & Hjerpaasen, K. J. (2025). Building Resilience and Competence in Bachelor Nursing Students: A Narrative Review of Clinical Education Strategies. <https://doi.org/10.20944/preprints202506.2171.v1>

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational psychologist*, *52*(1), 17-37. <https://doi.org/10.1080/00461520.2016.1207538>

Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge. <https://doi.org/10.4324/9781351115940>

Wu, X., He, X., Liu, T., Liu, N., & Zhai, X. (2023, June). Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In *International conference on artificial intelligence in education* (pp. 401-413). Cham: Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-36272-9_33>

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education*, *45*(2), 224-238. <https://doi.org/10.1080/02602938.2019.1629390>

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247-1262. <https://doi.org/10.1080/02602938.2016.1260091>

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247-1262. <https://doi.org/10.1080/02602938.2016.1260091>

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 1-27. <https://doi.org/10.1186/s41239-019-0171-0>

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111-151. <https://doi.org/10.1080/03057267.2020.1735757>