

PEELing Argumentation: Scaffolding Elementary Argumentative Writing with Decomposed LLM Feedback

Stella Xin Yin^{1*,†}, Geyu Lin^{2,†}, Zhengyuan Liu³, Brian Y. Lim² and Nancy F. Chen³

¹Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore

²National University of Singapore, 21 Lower Kent Ridge Rd, 119077, Singapore

³Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way, 138634, Singapore

Abstract

Argumentative writing is an essential skill that fosters logical reasoning and analytical thinking in early education. Traditionally, mastering this skill requires one-on-one teacher feedback, which is often hindered by large class sizes. Although educational technology has emerged to support argumentative writing practice, existing systems predominantly target secondary and higher education, overlooking the unique cognitive and literacy needs of elementary school students who require more granular, decomposed scaffolding and engaging interfaces. To address these gaps, we propose an LLM writing system integrated with the PEEL (Point, Evidence, Explanation, and Link) framework. Grounded in Cognitive Load Theory, the system decomposes argumentation into structural components and leverages LLMs to deliver immediate, age-appropriate feedback tailored to each PEEL element. We conducted a four-week experiment with 130 fifth-grade students, comparing the PEEL-guided system against an unstructured LLM feedback condition and a traditional teacher-led control group. Results demonstrated that the PEEL-guided group achieved significant gains in overall writing quality. Although both experiment groups saw increased self-efficacy, only the PEEL-guided group showed significant improvements in student engagement and satisfaction. Furthermore, students using the PEEL-guided system demonstrated more revision behavior and a 26.3% increase in word count compared to the unstructured AI group. These findings highlight that pedagogical framing and decomposed scaffolding are essential for effective writing education.

Keywords

Argumentative Writing, Generative AI, Co-writing, Human-AI collaboration, Writing Assistant

1. Introduction

Argumentation skill refers to the ability to construct, analyze, and evaluate claims supported by evidence and logic, which is essential for developing critical thinking and complex reasoning [1]. Writing practice is one approach to cultivating this skill, where students externalize and refine their reasoning through composition [2]. In traditional writing classes, effective argumentation pedagogy has relied on intensive, one-on-one teacher feedback to provide the necessary scaffolding for student growth [3]. Particularly, students often submit a completed draft and receive feedback only after the assignment is graded, leaving little to no opportunity for iterative refinement in the writing process [4]. This “one-shot” submission model also places a burden on teachers in large classes and limits the frequency and quality of personalized guidance students receive.

Despite the emergence of educational technology to alleviate this burden [5, 6], current writing system on argumentation has focused predominantly on secondary and higher education, leaving the needs of elementary school students (ages 7 to 11) largely unexplored [7, 8]. Unlike adult learners, children in this age group possess developing cognitive and literacy skills that require more granular, decomposed scaffolding to process instructional feedback effectively [9]. Furthermore, many existing systems rely on plain-text interactions that fail to engage younger learners or support their specific information-processing needs [10]. In addition, without pedagogical framing, students may be overwhelmed by unstructured AI prompts, leading to “passive outsourcing” rather than active learning.

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea

*Corresponding author.

[†]These authors contributed equally.

✉ xin013@e.ntu.edu.sg (S. X. Yin)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To address these challenges, we propose an LLM-powered writing system integrated with the PEEL (Point, Evidence, Explanation, and Link) framework [11, 12]. Grounded in cognitive load theory [13], our system scaffolds the writing process by 1) decomposing argumentation into four structural components to reduce extraneous cognitive load, 2) using color-coded analysis to make abstract rhetorical structures tangible for young writers, and 3) leveraging LLMs to provide immediate, age-appropriate feedback tailored to each specific PEEL element. We conducted an experimental study with 130 elementary school students to evaluate the effectiveness of this approach. Participants were randomly assigned to three conditions: PEEL-guided LLM feedback, unstructured LLM feedback, or traditional teacher-led control group. By measuring outcomes in writing performance, engagement, satisfaction, and self-efficacy, this study aims to explore how pedagogically aligned, decomposed scaffolding can bridge the gap in elementary-level argumentative writing education. This study aims to answer three questions: **RQ1:** To what extent does the PEEL-guided writing system improve students’ argumentative writing skills? **RQ2:** How does the PEEL-guided writing system influence students’ writing processes? **RQ3:** How does the PEEL-guided writing system enhance students’ engagement, self-efficacy, and learning satisfaction in argumentative writing practice?

This study makes several contributions to the field of educational technology and writing instruction. First, we provide empirical evidence that structured, framework-aligned AI scaffolding is more effective than unstructured LLM feedback for improving argumentative writing skills. By analyzing student iterative revision data, we demonstrate that pedagogical framing prevents “passive outsourcing” and instead fosters active revision behavior and revision depth. Beyond performance gains, this research illustrates how decomposed AI feedback can bridge the gap in one-on-one instruction for large classrooms while maintaining high levels of student engagement and satisfaction. Ultimately, this work offers a scalable model for designing age-appropriate AI tutors that respect the unique cognitive and developmental needs of young learners, suggesting that the effectiveness of Generative AI in education depends less on the underlying model and more on the pedagogical constraints integrated into its design.

2. Related Work

Children at ages 7-11 demonstrate emerging abilities in abstract reasoning [14, 15, 16], and early argumentation training has been shown to significantly improve critical thinking, metacognitive awareness, and analytical skills across disciplines [17, 18, 4]. Traditional teacher-centered approaches produced positive effects but often limited engagement and hindered iterative improvement through delayed feedback [19, 20, 21, 22]. Educational technology has partially addressed these gaps through online argumentation platforms and automated writing evaluation systems [23, 24, 25, 26, 27]. More recently, LLM-based writing systems, such as prompt-based [28, 29], domain-specific [30, 31], RAG [32, 33], and multi-agent architectures [34, 35, 36], have demonstrated effectiveness in supporting argumentative writing. However, existing research has mainly focused on secondary and higher education [7, 10], leaving elementary level students under-explored. Unlike adults, children aged 7-11 rely heavily on tangible examples [14, 37] and require substantial scaffolding to bridge concrete experiences and abstract reasoning [38, 6, 39], suggesting that systems designed for adults may not address their developmental needs. The full version of Related Work is provided in Appendix A.

3. System Design and Implementation

Our system contains four interconnected modules designed to deliver adaptive argumentation instruction for elementary students (Figure 1). The Expert Model encodes domain knowledge by operationalizing the PEEL framework as the structural foundation for decomposing students’ writing tasks. The Student Model maintains learner profiles that track progress across PEEL elements, enabling personalized adaptation to each child’s developing cognitive abilities. The Tutoring Model Generate writing topics and leverages LLM capabilities to deliver real-time, iterative feedback through scaffolding strategies, simulating one-on-one teacher guidance at scale. The User Interface engages learners through age-appropriate design and structured interactions that support their information-processing needs.

PEEL Writing Guide

Student ID: user001
Date: 18/06/2025

Topic
Choose the topic following teacher's instruction.

Reminder: You can click "Analyze Paragraph" whenever you finish writing to get the feedback and improve your writing!

Shoplifters can change for the better.

2. Student Input

Paragraph

Shoplifters can change for the better as after they steal, when they get caught, people can help them.
For example, when they get caught, people can help them seek financial help.
When shoplifters get caught, they are also more likely to change their mindset and stop stealing. For example, the shoplifters can also go to psychologists to help change their mindset of stealing.
As a result, I agree that shoplifters can change for the better.

Analyze Paragraph

3. PEEL Structure

PEEL Analysis

I agree that Shoplifters can change for the better.
Shoplifters can change for the better as after they steal, when they get caught, people can help them.
For example, when they get caught, people can help them seek financial help.
When shoplifters get caught, they are also more likely to change their mindset and stop stealing.
For example, the shoplifters can also go to psychologists to help change their mindset of stealing.
As a result, I agree that shoplifters can change for the better.

4. Decomposed Feedback

★★★★★

> Point: Well done! The point is clear.

> Evidence: The evidence is convincing.

> Explanation: The explanation is clear but could be expanded. For example, in your first example, think about "How can financial support change someone's mind about stealing?"

> Link: The link is a good summary. Make sure it connects back to the main point clearly.

Figure 1: The Interface of PEEL-guided Argumentative Writing System

PEEL Writing Guide

Student ID: user002
Date: 18/06/2025

Topic
Choose the topic following teacher's instruction.

Reminder: You can click "Analyze Paragraph" whenever you finish writing to get the feedback and improve your writing!

Shoplifters can change for the better.

PEEL Analysis

Point: I agree.
Evidence: Probably the shoplifter have no money to support his family.
Explanation: If the shoplifter's family had support from the government the shoplifter would not have to shoplift.
Evidence: They will learn their lesson in prison.
Explanation: They will know that their family is in more trouble and they will not do it again.
Link: As a result, shoplifters can change for the better.

★★★★★

Feedback

Please tell the reader exactly what topic you agree with at the start. You have good evidence, but you could add more detail to make them more convincing. In your examples, explain how the government helps families or what exactly people learn in prison. Finally, try to make your last sentence a short summary of your main point.

Point	I (strongly agree/ agree/ disagree/ strongly disagree) that...
Evidence	Shoplifting is a victimless crime. First Supporting Point: The shoplift workers might get framed by the shoplifters. Elaboration / Explanation / Example: The manager might think that the worker stole something because there are no CCTV cameras for proof and accuse the worker who is a victim. Second Supporting Point: The shop will lose money. Elaboration / Explanation / Example: The shop will not be able to get the money back of the stolen item back if they shoplifter isn't caught.
Link	As a result... / In a nutshell... Thus, shoplifting is not a victimless crime.

A) Experiment Group 2: LLM-based Writing System with unstructured Feedback

B) Control Group: Paper-based Writing Practice with Teacher's Feedback

Figure 2: (left)) The Interface of Writing System_base and (right) Paper-based Writing Practice Sheet

3.1. PEEL Framework and Learning Objective

The PEEL framework (Point, Evidence, Explanation, and Link) provides a structure by decomposing argumentative paragraphs into four components [11]. This framework has gained widespread adoption in elementary level writing instruction as an accessible heuristic that guides students through the logical progression of constructing a persuasive argument [40]. Therefore, we integrate this framework as a foundation to instruct argumentative writing. The **Point** requires students to articulate a clear claim or topic sentence that establishes the central argument. The **Evidence** component prompts learners to support their claims with convincing reasons. The **Explanation** element is the most cognitively demanding for young writers, requiring students to provide concrete facts, anecdotes or examples to explain why they agree or disagree with statements. Finally, the **Link** asks students to connect their argument back to the main idea. Based on this PEEL framework, we defined the learning objective as "Students are able to write a clear argument that answers give topic, explains their thinking, uses relevant examples to support their ideas, and connects their explanation back to the main point." This object is embedded within the Expert Model to guide feedback generation and within the Student Model to track learner progress.

3.2. Writing Practice Implementation and Deployment

In conventional classrooms, teachers utilize structured worksheets that guide students through the PEEL framework using a tabular graphic organizer [21]. As shown in Figure 2-B, the task presents students with a debatable statement (e.g., “Shoplifting is a victimless crime”) and asks them to: (1) state their position by selecting an agreement level, (2) provide supporting points with reasons, (3) elaborate with details, anecdotes, or examples, and (4) link their argument back to their stated position. We translated this paper-based structure into an interactive digital interface that preserves the pedagogical affordances of organized layout and enables real-time, adaptive feedback, as shown in Figure 1. Additionally, we developed a baseline version of the system that provides automated feedback in a single, unstructured plain-text block without the PEEL framework (Figure 2-A). This base system allows us to examine the effectiveness of our proposed design. The Student Model monitors student inputs in real time and decomposes the student’s input, mapping their writing onto the four structural components.

3.3. Feedback Generation

We employed OpenAI’s GPT-4o API to generate feedback because of its outstanding performance in educational content generation and reasoning tasks [41]. The base prompt contains pedagogical guidelines aligned with curriculum standards for elementary-level English language instruction. Core constraints included: (a) using an encouraging, age-appropriate tone suitable for learners aged 9 to 11, (b) limiting feedback length to 3-5 sentences to avoid cognitive overload, (c) using simple vocabulary and sentence structures accessible to developing readers, and (d) employing scaffolding strategy rather than direct correction. Component-specific prompts were designed to evaluate students’ writing based on the PEEL rubric (Appendix Table 2), which is provided by elementary school teachers. All prompts underwent iterative refinement through pilot studies and teacher consultation to verify age appropriateness and curricular relevance.

To establish the consistency and reliability of automated scoring (we use five stars to represent the writing score), we compared system-generated scores and feedback with teacher assessments of 28 students’ past PEEL writing samples. Correlation analysis revealed a moderate-to-strong agreement between human and system scoring ($r = .67$, $p < .01$). The agreement level of four components is as follows: Point ($r = .64$), Evidence ($r = .67$), Explanation ($r = .74$), and Link ($r = .62$). These findings suggest acceptable criterion validity for formative feedback deployment.

3.4. Interface and System Architecture

Figure 1 illustrates the main writing interface, which comprises four parts. First, **Topic Selection Panel** at the top allows students to select from teacher-assigned argumentative topics via a dropdown menu. Second, The central workspace is divided into two side-by-side panels. The **Paragraph Panel** (left) provides a free-text input area where students compose their argumentative paragraphs. The **PEEL Analysis Panel** (right) displays a color-coded decomposition of the student’s writing, with each sentence automatically classified and highlighted according to its PEEL component: Point (red), Evidence (blue), Explanation (green), and Link (yellow). This visual segmentation makes the abstract structure of argumentation tangible for young learners, enabling them to immediately see how their writing maps onto the target PEEL framework. Once the student clicks “Analyze Paragraph”, an Overall Score is displayed using a five-star rating scale, providing an indicator of writing quality appropriate for the target age group. The **Component Feedback Section** presents expandable cards for each PEEL element, color-coded to match the analysis panel. Each card contains actionable feedback, ranges from positive reinforcement (e.g., Well done! The point is clear.) to scaffolded guidance prompting deeper thinking (e.g., The explanation is clear but could be expanded. For example, think about “How can financial support change someone’s mind about stealing?”).

To build this system, a ReactJS frontend communicates via Axios with a Python backend serving as the central orchestrator. After receiving student submissions, the backend assembles prompts that integrate student text with PEEL grading rubrics and contextual instructions before routing them to the

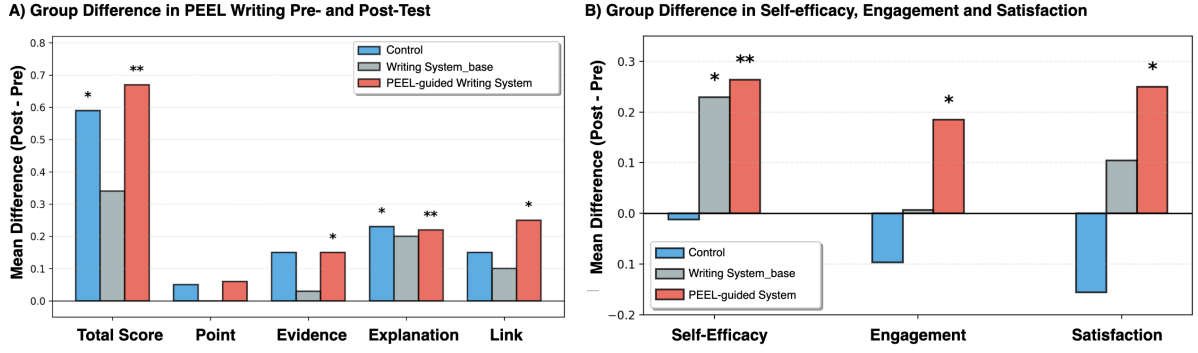


Figure 3: Comparison of A) pre- and post-test and B) self-efficacy, engagement and satisfaction across three group conditions.

GPT-4o. API calls are configured with a low temperature parameter to ensure analytical consistency and specify JSON-formatted output for streamlined parsing. The generated evaluations are serialized and stored in a MongoDB database, whose schema flexibility accommodates semi-structured data including original text, engineered prompts, component-level feedback, scores, and session metadata.

4. Experimental Methodology

We recruited 130 fifth-grade students (66 female, 64 male; Mean age = 10.79 years), randomly assigned to three conditions: a PEEL-guided LLM system ($N = 52$), an unstructured LLM system (Writing System_base) ($N = 39$), and a traditional teacher-led control group ($N = 39$), where teachers provided written feedback according to the PEEL framework to ensure comparability. Writing practice was integrated into students' regular English lessons, with each session requiring a 25-minute argumentative paragraph composition in response to teacher-approved topics.

We conducted a four-week experiment with pre- and post-tests measuring argumentative writing skills, alongside validated surveys assessing self-efficacy [42], engagement [43] across behavioral ($\alpha = .82$), emotional ($\alpha = .90$), and cognitive ($\alpha = .79$) dimensions, and satisfaction [44] ($\alpha = .90$) [45]. Baseline equivalence across groups was confirmed via one-way ANOVA on pre-test writing scores ($F(2, 127) = 0.74, p = .481$) and all affective measures (all $p > .05$). Writing outcomes were analyzed using paired-samples t -tests and Cohen's d , while system logs tracked word count growth, text similarity between consecutive revisions, and feedback adoption rates. Open-ended survey responses were examined through thematic analysis to capture students' qualitative perceptions of the learning experience. The full version of Experimental Methodology is provided in Appendix C.

5. Results

5.1. RQ1: Impact on Argumentative Writing

Table 3 (in Appendix D) presents descriptive results of student writing performance and survey on self-efficacy, engagement, and satisfaction for the Control, System_base, and PEEL-guided LLM conditions at both pre- and post-test intervals. Figure 3-A presents within-group paired t -test results. The PEEL-guided group achieved the most overall gains ($M_{diff} = 0.67, t(51) = 4.10, p < .01, d = 0.57$), while the Control group also showed significant growth ($M_{diff} = 0.59, t(38) = 2.55, p = .02, d = 0.41$). Notably, the Writing System_base group failed to reach significance ($M_{diff} = 0.34, t(38) = 1.20, p = .24, d = 0.19$), indicating that unstructured feedback may not be sufficient for skill acquisition. Subscore analysis revealed no significant Point improvements across any group (all $p > .47, d < 0.11$), suggesting that elementary students already possess the ability to clearly state their main ideas. Only the PEEL-guided group showed significant improvement in Evidence ($M_{diff} = 0.15, t(51) = 2.68, p < .01, d = 0.37$) and Link ($M_{diff} = 0.25, t(51) = 2.44, p = .02, d = 0.34$). For Explanation, both the Control and PEEL-guided

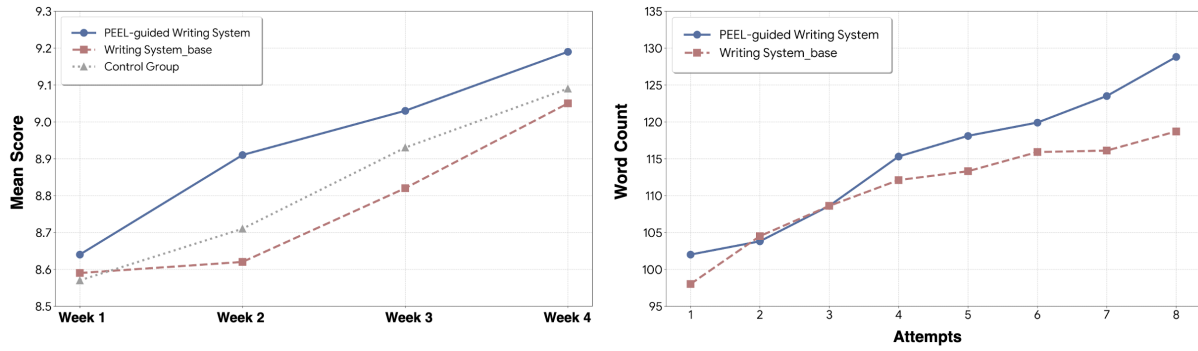


Figure 4: (left) Mean scores of writing practice across the four-week experiment for the PEEL-guided, Writing System_base, and Control groups. (right) Average word count growth per revision attempt for students in the PEEL-guided and Writing System_base conditions.

Writing groups showed significant improvements ($p = .01, d = 0.48$ and $p < .01, d = 0.32$, respectively). The pattern suggests that students require intensive, structured guidance to master evidence integration and logical linking. In addition to pre- and post-test scores, all three groups showed steady improvement over four weeks, with the PEEL-guided group demonstrating greater gains (Figure 4 left). Students in the PEEL-guided group also averaged more revision attempts per topic (5.58 vs. 3.92), indicating that structured scaffolding not only improves learning outcomes but also increases iterative engagement.

5.2. RQ2: Impact on Students' Writing Process

First, we compared word count growth per revision attempt across up to eight attempts. As shown in Figure 4 (right), both groups started at similar baseline lengths (PEEL-guided: 102 words; Writing System_base: 98 words), but diverged over successive attempts. The PEEL-guided group demonstrated continuous growth to 128.8 words (26.3% increase), whereas the Writing System_base group plateaued after Attempt 4, reaching only 118.7 words by Attempt 8 (21.1% increase). This suggests that decomposed feedback helped students identify specific component to edit, while unstructured feedback left students uncertain about what to improve.

Second, we calculated text similarity between consecutive attempts using cosine similarity to assess revision depth (Figure 5 (left)). Both groups began with similar revision intensity (PEEL-guided: 70.2%; Writing System_base: 72.7%). However, the PEEL-guided group maintained consistent revision depth of 16-17% per iteration, whereas the Writing System_base group's revisions regressed toward minimal changes, reaching 91-92% similarity by Attempts 7-8. These findings demonstrate that without structured guidance, revision behavior naturally declines, whereas the PEEL framework sustains active engagement.

Finally, we analyzed feedback distributions between two systems (Table 1). The PEEL-guided system generated 11,862 feedback instances compared to 8,626 in the Writing System_base. Evidence-related feedback dominated both systems ($\approx 40\%$), followed by Explanation feedback (36.6%). Common feedback types targeted adding examples, clarifying evidence, and strengthening links to the main idea. Notably, both systems autonomously generated social-emotional support without explicit prompting.

5.3. RQ3: Impact on Engagement, Self-efficacy, and Satisfaction

As illustrated in Figure 3-B, self-efficacy significantly increased in both experimental conditions, with the PEEL-guided group showing the largest gains ($M_{diff} = 0.26, t(69) = 3.30, p = .002, d = 0.39$) followed by the Writing System_base group ($M_{diff} = 0.23, t(59) = 2.17, p = .03, d = 0.28$), while the Control group showed no significant change ($M_{diff} = -0.01, t(39) = -0.13, p = .90, d = -0.02$). Notably, increased self-efficacy only translated into higher engagement ($M_{diff} = 0.19, t(69) = 2.19, p = .03, d = 0.26$) and satisfaction ($M_{diff} = 0.25, t(69) = 2.22, p = .03, d = 0.27$) for the PEEL-guided group, suggesting that structured scaffolding fosters a more positive learning experience.

Table 1

Comparison of Component Distributions between PEEL-guided and Base Writing Systems

Theme	PEEL-guided Writing System		Writing System_base	
	Count	Distribution	Count	Distribution
Explanation: Give example	2,735	23.0%	2,053	23.8%
Explanation: Be specific / detail	1,613	13.6%	958	11.1%
Evidence: Clarify	2,395	20.2%	1,712	19.8%
Evidence: Explain why	2,381	20.1%	1,886	21.9%
Link: Connect back	1,613	13.6%	1,129	13.1%
Point: Provide main idea	673	5.7%	513	5.9%
Social emotional support	452	3.8%	375	4.3%

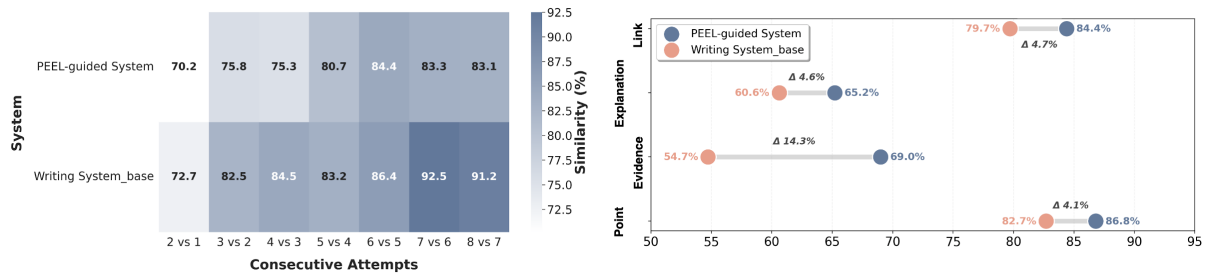


Figure 5: (left) Text similarity between consecutive revision attempts in two groups. Higher percentages indicate the text changed very little, while lower percentages mean more revisions were made. (right) Feedback adoption rate across four PEEL components between the PEEL-guided System and the Writing System_base. The percentage indicates how much proportion of feedback was successfully incorporated into the revised text.

Thematic analysis of open-ended responses revealed three key patterns. First, nearly half of respondents (47.3%) described the tool as helpful, appreciating its concrete, actionable feedback—“*The AI was nice because it told me what to add or what was wrong*”—which was closely linked to improved PEEL comprehension (31.0%). Second, one-third of students (32.6%) described the experience as fun or engaging, with students noting they “*liked playing with the robot to write better*” and found it “*super fun to write my work and let the robot look at it*,” suggesting the tool successfully balanced pedagogical effectiveness with affective appeal. However, a small portion (4.7%) reported inconsistencies in scoring, with students expressing frustration—“*I was sad because I had 4 stars but the robot made me get less*”—highlighting a tension between instructional feedback and students’ performance expectations.

6. Discussion

6.1. Pedagogical Alignment for Effective AI-Assisted Writing

Our results suggest that adult-oriented AI feedback designs do not straightforwardly transfer to elementary students, as confirmed by the Writing System_base condition. Cognitive load theory [13] helps explain why the PEEL framework succeeded: unstructured feedback overwhelmed students’ developing executive functions by requiring them to simultaneously interpret holistic feedback, identify weaknesses, and plan revisions. The PEEL structure reduced this extraneous load by constraining the task into four concrete components, channeling cognitive resources toward targeted skill building.

While both the PEEL-guided and control groups achieved significant writing gains, only the PEEL-guided system fostered engagement and satisfaction. The control condition relied on a passive submit-and-wait cycle that, while effective for skill development, did little to foster intrinsic motivation. In contrast, the color-coded PEEL analysis and five-star rating activated what self-determination theory [46] describes as competence satisfaction—students could see exactly which components they had mastered and act on feedback in real time, instead of waiting days for teacher comments.

6.2. Structured Feedback Shapes Productive Writing Processes

Our findings reveal that decomposed structure not only improves outcomes but also shapes the writing process. Students in the PEEL-guided group made more effective revisions with increased word count and higher feedback adoption. However, Evidence and Explanation remained the most challenging components, as reasoning and metacognitive monitoring in grades 4–5 are still developing [18, 47, 48]. To better support these components, future systems should incorporate graduated scaffolding strategies. For example, curated example banks or retrieval-augmented prompts for Evidence, and interactive sentence frames (e.g., “This example shows that...because...”) or worked examples for Explanation. Additionally, adaptive fading mechanisms that gradually reduce scaffolding as students demonstrate proficiency would prevent over-reliance while ensuring sufficient guidance during early skill acquisition.

6.3. Reliability of Feedback

A notable concern is that 4.7% of students reported inconsistencies where incorporating AI suggestions lowered their scores. It is a known challenge in automated writing evaluation [23, 25, 31] that carries particular weight for elementary learners whose trust in instructional tools is still forming. Our system included safeguards such as age-appropriate language and harmful content prevention, yet automated scoring misalignments can still create confusion. Future systems should incorporate transparency mechanisms, such as brief explanations alongside score changes (e.g., “Your score decreased because the new sentence moved away from the main point”) and visual progress indicators that track component-level growth over time.

Our system demonstrated acceptable criterion validity ($r = .67$ with teacher scoring), though the narrow 1–3 point rubric scale and short paragraph lengths inherent to elementary writing limit scoring granularity and inflate disagreement. This raises the question of whether numerical scoring is the most appropriate evaluation method for this age group. Progress-oriented evaluation methods that track component-level growth trajectories across multiple submissions may better capture learning gains than single-submission scores.

7. Conclusion and Limitations

This study investigated whether an LLM-powered writing system integrated with the PEEL framework could effectively scaffold elementary students’ argumentative writing. Through a four-week experiment with 130 fifth-grade students, we compared a PEEL-guided system against unstructured LLM feedback and traditional teacher-led instruction. Our results demonstrate that pedagogical framing and decomposed scaffolding drive learning effectiveness. The PEEL-guided group achieved the largest gains in overall writing quality and positively enhanced students’ affective outcomes and revision behaviors. These findings contribute to AI for education field by providing empirical evidence that LLM-based writing systems for elementary learners require developmentally appropriate structural scaffolding, and that process-level metrics—revision depth, feedback adoption, and iteration persistence—are valuable indicators of learning quality beyond pre-post score comparisons.

We acknowledge several limitations in this study. First, the unequal group sizes (52 students in the PEEL-guided condition but only 39 in each of the other two groups) may have afforded greater statistical power to the PEEL-guided condition. Second, the four-week duration may limit causal claims and provide limited evidence regarding long-term retention. It remains unclear whether improvements in writing quality and self-efficacy would persist once system support is withdrawn. Future work should explore adaptive scaffolding that gradually fades PEEL structure as students demonstrate mastery, examine whether gains transfer to independent writing without system support, and replicate the approach across diverse linguistic and curricular contexts. Additionally, investigating how to better support the adoption of high-difficulty feedback types, such as through worked examples or interactive prompts for Evidence and Explanation, would address the gap between feedback frequency and student actionability identified in this study.

Declaration on Generative AI

The authors employed Claude and Grammarly for editing and proofreading for grammar and clarity. The authors evaluated the output by carefully reviewing and editing all AI-generated text to ensure accuracy, originality, and alignment with the research findings. The authors assume all responsibility for the content of this submission.

References

- [1] D. Kuhn, W. Udell, The development of argument skills, *Child development* 74 (2003) 1245–1260.
- [2] M. Billig, *Arguing and thinking: A rhetorical approach to social psychology*, Cambridge University Press, 1996.
- [3] R. P. Ferretti, S. Graham, *Argumentative writing: Theory, assessment, and instruction*, *Reading and Writing* 32 (2019) 1345–1357.
- [4] G. E. Newell, R. Beach, J. Smith, J. VanDerHeide, Teaching and learning argumentative reading and writing: A review of research, *Reading research quarterly* 46 (2011) 273–304.
- [5] Z. Xu, M. Banerjee, G. Ramirez, G. Zhu, K. Wijekumar, The effectiveness of educational technology applications on adult english language learners' writing quality: a meta-analysis, *Computer Assisted Language Learning* 32 (2019) 132–162.
- [6] Z. Liu, S. X. Yin, C. Lee, N. F. Chen, Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions, in: *Proceedings of 2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 1258–1265.
- [7] C. Strobl, E. Ailhaud, K. Benetos, A. Devitt, O. Kruse, A. Proske, C. Rapp, Digital support for academic writing: A review of technologies and pedagogies, *Computers & education* 131 (2019) 33–48.
- [8] Z. Liu, G. Lin, H. L. Tan, H. Zhang, Y. Lu, X. Gao, S. X. Yin, H. Sun, H. H. Goh, L. H. Wong, N. F. Chen, Singakids: A multilingual multimodal dialogic tutor for language learning, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 1244–1253. URL: <https://aclanthology.org/2025.acl-industry.86/>.
- [9] C. S. Brown, Language and literacy development in the early years: Foundational skills that support emergent readers., *Language and Literacy Spectrum* 24 (2014) 35–49.
- [10] C. Williams, S. Beam, Technology and writing: Review of research, *Computers & education* 128 (2019) 227–242.
- [11] S. Gibbons, “death by peel?” the teaching of writing in the secondary english classroom in england, *English in Education* 53 (2019) 36–45.
- [12] S. Humphrey, T. Sharpe, T. Cullen, Peeling the peel: Integrating language and literacy in the middle years, *Literacy Learning: The Middle Years* 23 (2015) 53–62.
- [13] J. Sweller, Cognitive load theory, in: *Psychology of learning and motivation*, volume 55, Elsevier, 2011, pp. 37–76.
- [14] J. Piaget, M. Cook, et al., *The origins of intelligence in children*, volume 8, International universities press New York, 1952.
- [15] B. Inhelder, J. Piaget, *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*, Routledge, 2013.
- [16] Z. Liu, S. X. Yin, D. H. L. Goh, N. F. Chen, Cogent: A curriculum-oriented framework for generating grade-appropriate educational content, in: *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, 2025, pp. 129–143. URL: <https://aclanthology.org/2025.bea-1.10/>.
- [17] D. Kuhn, A. Crowell, Dialogic argumentation as a vehicle for developing young adolescents' thinking, *Psychological science* 22 (2011) 545–552.
- [18] D. Kuhn, W. Moore, Argumentation as core curriculum, *Learning: Research and practice* 1 (2015) 66–78.

- [19] L. Flower, J. R. Hayes, A cognitive process theory of writing, *College Composition & Communication* 32 (1981) 365–387.
- [20] S. Graham, D. Perin, What we know, what we still need to know: Teaching adolescents to write, *Scientific studies of reading* 11 (2007) 313–335.
- [21] S. Graham, D. Perin, A meta-analysis of writing instruction for adolescent students., *Journal of educational psychology* 99 (2007) 445.
- [22] J. Hattie, H. Timperley, The power of feedback, *Review of educational research* 77 (2007) 81–112.
- [23] Z. V. Zhang, Engaging with automated writing evaluation (awe) feedback on l2 writing: Student perceptions and revisions, *Assessing Writing* 43 (2020) 100439.
- [24] Q.-K. Fu, C.-J. Lin, G.-J. Hwang, L. Zhang, Impacts of a mind mapping-based contextual gaming approach on efl students' writing performance, learning perceptions and generative uses in an english course, *Computers & Education* 137 (2019) 59–77.
- [25] J. Wilson, R. D. Roscoe, Automated writing evaluation and feedback: Multiple metrics of efficacy, *Journal of Educational Computing Research* 58 (2020) 87–125.
- [26] K. Guo, J. Wang, S. K. W. Chu, Using chatbots to scaffold efl students' argumentative writing, *Assessing Writing* 54 (2022) 100666.
- [27] Y. Su, Y. Lin, C. Lai, Collaborating with chatgpt in argumentative writing classrooms, *Assessing Writing* 57 (2023) 100752.
- [28] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, in: *The eleventh international conference on learning representations*, 2022.
- [29] W. Song, K. Zhang, R. Fu, L. Liu, T. Liu, M. Cheng, Multi-stage pre-training for automated Chinese essay scoring, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 6723–6733. URL: <https://aclanthology.org/2020.emnlp-main.546/>.
- [30] A. Mizumoto, M. Eguchi, Exploring the potential of using an ai language model for automated essay scoring, *Research Methods in Applied Linguistics* 2 (2023) 100050.
- [31] Q. Wang, J. M. Gayed, Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting, *Computer Assisted Language Learning* (2024) 1–29.
- [32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in neural information processing systems* 33 (2020) 9459–9474.
- [33] S. P. Neshaei, M. Tashkovska, P. Mejia-Domenzain, T. Wambsganss, T. Käser, User-centric reflective writing assistance: Leveraging rag for enhanced personalized support, in: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–8.
- [34] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al., Autogen: Enabling next-gen llm applications via multi-agent conversations, in: *First Conference on Language Modeling*, 2024.
- [35] Z. Hu, H. P. Chan, J. Li, Y. Yin, Debate-to-write: A persona-driven multi-agent framework for diverse argument generation, in: *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4689–4703.
- [36] J. Jordan, X. Yin, M. Fabros, G. Ranade, N. Norouzi, Magic: Multi-agent argumentation and grammar integrated critiquer, *arXiv preprint arXiv:2506.13037* (2025).
- [37] Z. Liu, S. X. Yin, N. F. Chen, Optimizing code-switching in conversational tutoring systems: A pedagogical framework and evaluation, in: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2024, pp. 500–515.
- [38] A. Collins, J. S. Brown, A. Holum, et al., Cognitive apprenticeship: Making thinking visible, *American educator* 15 (1991) 6–11.
- [39] Z. Liu, S. X. Yin, G. Lin, N. F. Chen, Personality-aware student simulation for conversational intelligent tutoring systems, in: *Proceedings of the 2024 Conference on Empirical Methods in*

Natural Language Processing, 2024, pp. 626–642.

- [40] C. Gerald, M. Joseph, Paragraphing strategies in academic writing: A systematic review of structured approaches, *Asian Journal of Education and Social Studies* 50 (2024) 160–174.
- [41] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, *Learning and individual differences* 103 (2023) 102274.
- [42] P. R. Pintrich, E. V. De Groot, Motivational and self-regulated learning components of classroom academic performance., *Journal of educational psychology* 82 (1990) 33.
- [43] J. A. Fredricks, P. Blumenfeld, J. Friedel, A. Paris, School engagement, What do children need to flourish? Conceptualizing and measuring indicators of positive development (2005) 305–321.
- [44] A. Sher, Assessing the relationship of student-instructor and student-student interaction to student learning and satisfaction in web-based online learning environment., *Journal of Interactive Online Learning* 8 (2009).
- [45] S. X. Yin, Z. Liu, D. H.-L. Goh, C. L. Quek, N. F. Chen, Scaling up collaborative dialogue analysis: An ai-driven approach to understanding dialogue patterns in computational thinking education, in: *Proceedings of the 15th International Conference on Learning Analytics and Knowledge (LAK '25)*, 2025, pp. 47–57.
- [46] E. L. Deci, R. M. Ryan, Self-determination theory, *Handbook of theories of social psychology* 1 (2012) 416–436.
- [47] A. Glassner, M. Weinstock, Y. Neuman, Pupils' evaluation and generation of evidence and explanation in argumentation, *British Journal of Educational Psychology* 75 (2005) 105–118.
- [48] K. L. McNeill, Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year, *Journal of research in science teaching* 48 (2011) 793–823.

A. Full Related Work

A.1. Argumentation Skill and Writing Practice in Elementary Education

Children at ages 7-11 demonstrate emerging abilities to engage in abstract reasoning and perspective-taking [14]. During this concrete operational stage, students begin to develop logical thinking skills that form the foundation for argumentative reasoning [15]. Research demonstrates that early development of argumentative reasoning significantly impacts students' critical thinking, evidence-based reasoning, and persuasive communication across disciplines [17]. For example, Kuhn & Moore [2015] found that students who received argumentation training in grades 4-5 demonstrated superior reasoning abilities and metacognitive awareness. Similarly, Newell et al. [2011] found that argument writing instruction significantly improved elementary students' comprehension and analytical thinking skills, with effects extending beyond language arts to science and social studies domains.

Traditional approaches to teaching argumentation skills have primarily relied on writing practice through teacher-centered instruction. Usually, teachers taught argument structures through direct instruction, followed by individual writing assignments. In this way, students practiced argumentative writing independently, and teachers provided written feedback on completed drafts in one-on-one assessment cycles [19, 20]. A meta-analysis [21] found that while explicit teaching of writing strategies produced positive effects, it often resulted in limited student engagement and delayed feedback that hindered iterative improvement [22]. With the recent development of educational technology, argumentative writing incorporates digital tools and collaborative platforms that enable real-time peer interaction and scaffolded argumentation development. Studies by Zhang et al. [2020] and Hwang & Fu [2019] demonstrated that online argumentation platforms with structured sentence starters, argument mapping tools, and peer review features significantly improved elementary students' argument quality and complexity. Automated writing evaluation systems, such as those examined by Wilson & Roscoe [2020], Guo et al. [2022] and Su et al. [2023], provide immediate formative feedback on argument

structure, evidence usage, and counterargument development, allowing students to revise iteratively.

A.2. LLM-based Writing System

The advancement of LLMs has catalyzed the development of writing support systems, which can be categorized into four types: prompt-based writing systems, domain-specific systems, retrieval-augmented generation (RAG) systems, and multi-agent writing systems.

Prompt-based systems integrate pedagogical frameworks into the interaction architecture to guide students through structured writing processes. These systems employ prompt engineering techniques—carefully designed input templates that shape LLM responses to align with instructional goals [28]. For example, the Essay Scoring system [29] uses multi-stage prompting where students first brainstorm claims, then the LLM suggests evidence sources (without generating complete arguments), and finally students synthesize information into their own arguments.

Domain-specific systems trained or fine-tuned on educational datasets to provide more targeted instructional support. These systems build upon base models (e.g., GPT-3.5, LLaMA-2) through supervised fine-tuning on corpora of student writing, teacher feedback, and exemplar texts [30, 31]. The automated writing evaluation system [31] fine-tuned GPT-3.5 on a corpus of 480 argumentative essays with ground truth scores, achieving 84.72% accuracy in evaluation.

RAG Systems represent an emerging architecture that combines LLMs with external knowledge retrieval to enhance factual accuracy and source attribution—critical for argumentative writing instruction [32]. For example, Neshaei et al. [33] implements RAG architecture for argumentative writing support, automatically retrieving credible sources for student claims and generating properly cited evidence.

Multi-Agent Writing Systems employ multiple specialized LLM instances that collaborate to support complex writing tasks. This architecture draws inspiration from cognitive apprenticeship theory, where different agents simulate expert roles such as brainstormer, critic, editor, and mentor [34]. Hu et al. [35] and Jordan et al. [36] implement a four-agent architecture: a Claim Generator that helps students articulate positions, an Evidence Advisor that suggests supporting data, a Counterargument Analyst that identifies weaknesses, and a Critic Agent that guides improvement.

While these systems demonstrated effectiveness in supporting argumentative writing performance, existing research has mainly focused on secondary and higher education contexts [7, 10], leaving elementary school populations under-explored. Unlike adults who can process comprehensive abstract instructions, children aged 7-11 operate primarily within the concrete operational stage, relying heavily on tangible examples [14]. These learners require substantial scaffolding to bridge the gap between concrete experiences and abstract reasoning [38], suggesting that system designed for adults may not address their developmental needs.

B. PEEL Writing Rubric

C. Experimental Methodology

C.1. Study Context and Participants

Following ethical approval from the university Institutional Review Board [Anonymized for double-blind review], we recruited 130 fifth-grade elementary students (66 female, 64 male; Mean age = 10.79 years). Participants were randomly assigned to three groups. Control group (N = 39) students attended traditional classes using paper-based practice. Teachers collected their work after class and returned it with feedback in the next lesson. To ensure comparability with the AI-assisted conditions, teachers structured their written feedback according to the PEEL framework. Experimental group 1 students (N = 39) used an LLM-based writing system without PEEL guidance. Feedback was presented in plain text format. Experimental group 2 students (N = 52) used a PEEL-guided writing system where feedback was decomposed into four components and presented in different colors. The writing practice was integrated into students' regular English lessons. During each class, students were required to

Table 2
PEEL Writing Rubric for Elementary Students

Component	1-Point	2-Point	3-Point
Point	The main idea is unclear or does not directly answer the question.	The main idea is stated but could be clearer in answering the question.	The main idea clearly and directly answers the question.
Evidence	Few or no details are given to explain the main idea.	Some details are given to explain the main idea. The explanation makes sense but could be clearer.	Clear and logical details are given that fully explain the main idea.
Explanation	No example is given or the example does not connect to the main idea.	An example is given that connects to the main idea.	Specific and convincing examples are given that strongly supports the main idea.
Link	No concluding sentence or the sentence does not connect back to the main idea.	A concluding sentence is included but simply repeats the main idea.	A concluding sentence summarizes the main idea.

compose an argumentative paragraph following the PEEL structure in response to a given topic within 25 minutes. All writing topics have been reviewed and approved by both the school principal and supervising teachers.

C.2. Experimental Design and Ethical Consideration

We conducted a four-week experiment. Before starting, students completed a pre-test to establish a baseline of argumentative writing proficiency, as well as a survey measuring engagement, self-efficacy, and satisfaction. Next, students attended four weekly lessons that provided PEEL structure instruction and argumentative writing practice. At the end of the experiment, students completed a post-test to assess their argumentative writing performance and the same survey instrument measuring engagement, self-efficacy, and satisfaction. To protect student privacy and ensure confidentiality, no personally identifiable information was collected during writing tasks or surveys, and all data were coded using student ID numbers. The system was carefully designed to prevent harmful inputs and outputs throughout the intervention. Participation in the study was entirely voluntary, and students were informed of their right to withdraw at any time without consequence or penalty. Students who chose to withdraw during the experiment were able to continue attending regular classes and received alternative instruction to ensure no disruption to their learning.

C.3. Data Collection and Analysis

Data were collected at two time points: pre-intervention (Week 0) and post-intervention (Week 4). At both time points, students completed argumentative writing tests and surveys measuring self-efficacy, engagement, and satisfaction. The system automatically logged student writing samples and interaction data throughout the four-week intervention. All data were collected digitally through the learning platform and stored securely using student ID numbers to maintain anonymity. Multiple validated instruments were employed to assess the study's constructs. **Self-efficacy** was assessed through adapted scales from [42], measuring participants' confidence in their writing abilities and task-specific competence. **Engagement** was adapted from Fredricks et al. [2005]. This instrument assessed three engagement dimensions: behavioral ($\alpha = .82$), emotional ($\alpha = .90$), and cognitive ($\alpha = .79$) on 5-point Likert scales. **Satisfaction** was measured using Sher [2009] instrument to assess students' overall satisfaction with collaborative processes ($\alpha = .90$). In addition to multiple-choice questions, we also included two open-ended questions at the end of the survey to understand students' perceptions of the learning approach: (1) Please describe your learning experience in 1-2 sentences, and (2) How did the

learning approach impact your writing process?

To verify that random assignment produced comparable groups at baseline, we conducted one-way ANOVA on pre-test writing scores and pre-intervention survey measures (self-efficacy, engagement, satisfaction) across the three conditions. No significant differences were found across groups on pre-test writing scores ($F(2, 127) = 0.74, p = .481$) or any affective measure (all $p > .05$).

To address the first research question regarding the extent to which the PEEL-guided writing system improves students' argumentative writing skills, paired-samples t-tests were conducted to compare pre-test and post-test scores on argumentative writing performance. Writing samples were scored using a rubric aligned with argumentative writing standards, and effect sizes (Cohen's d) were calculated to determine the magnitude of improvement. For the second research question, we analyzed system logs to track word count growth, text similarity between consecutive revisions, and the feedback adoption rate to evaluate the depth and consistency of students' iterative writing processes. To address the third research question concerning how the system enhances engagement, self-efficacy, and satisfaction, paired-samples t-tests were employed to compare pre- and post-intervention scores on each construct. Additionally, students' responses to open-ended questions were analyzed through thematic analysis to capture qualitative perceptions of the tool's utility and learning impact.

D. Comparison of Writing Performance and User Experience Metrics

Table 3

Comparison of Writing Performance and User Experience Metrics (Mean $\pm$ SD) across Three Groups at Pre- and Post-Test Intervals.

Category	Time	Control ($n = 39$)	System_base ($n = 39$)	PEEL-guided LLM ($n = 52$)
<i>Writing Performance Scores</i>				
Total	Pre	8.51 ± 1.34	8.77 ± 1.48	8.43 ± 1.22
	Post	9.10 ± 1.05	9.26 ± 1.20	8.93 ± 1.23
-Point	Pre	2.87 ± 0.34	2.89 ± 0.32	2.80 ± 0.44
	Post	2.92 ± 0.27	2.91 ± 0.28	2.84 ± 0.37
-Evidence	Pre	1.74 ± 0.44	1.89 ± 0.32	1.84 ± 0.37
	Post	1.90 ± 0.31	1.91 ± 0.28	1.98 ± 0.23
-Explanation	Pre	1.82 ± 0.39	1.89 ± 0.47	1.91 ± 0.35
	Post	2.05 ± 0.32	2.17 ± 0.45	2.07 ± 0.42
-Link	Pre	2.08 ± 0.62	2.11 ± 0.76	1.88 ± 0.72
	Post	2.23 ± 0.67	2.26 ± 0.78	2.04 ± 0.74
<i>User Experience Metrics</i>				
Self-efficacy	Pre	3.74 ± 0.58	3.43 ± 0.81	3.36 ± 0.70
	Post	3.73 ± 0.56	3.65 ± 0.80	3.63 ± 0.70
Engagement	Pre	3.54 ± 0.65	3.53 ± 0.79	3.38 ± 0.64
	Post	3.44 ± 0.58	3.54 ± 0.86	3.57 ± 0.69
Satisfaction	Pre	3.64 ± 0.80	3.48 ± 1.05	3.34 ± 0.89
	Post	3.49 ± 0.67	3.59 ± 0.89	3.59 ± 0.98