AWE and LLMs in L2 Writing Feedback: An Exploration of EFL Learners' Experiences

Ruyin Zheng[1,a,*]

[1]Department of Foreign Studies, Yangjiang Polytechnic, Yangjiang, 529500, China
[a]eunicecheng719@outlook.com
[*]Corresponding author

Abstract: *In second language (L2) writing instruction, providing real-time personalized feedback at scale within traditional classroom settings poses significant challenges for teachers. Technology tools such as the established Automated Writing Evaluation (AWE) and the emerging Large Language Models (LLMs), with their respective strengths, have emerged as viable solutions. To investigate leaners' experiences with these two distinct computer-generated feedback types, this study examined the cognitive load and perceptions of English as a Foreign Language (EFL) students when using AWE and LLMs for writing feedback. A between-group experiment with 76 Chinese university students revealed that both tools induced low to moderate cognitive load, indicating favourable technology acceptance of both tools. Students' perception ratings for LLMs were slightly higher than those for AWE across all three dimensions. Qualitative data also indicated a strong student preference for LLMs due to their detailed, adaptive, and interactive feedback, though some still favoured AWE for its simplicity. The findings supporting the previous research results that LLMs exhibit significant potential in providing multi-dimensional feedback, AWE remains relevant for certain learners. The study highlights the need for tool selection based on learner proficiency and the potential benefits of integrating Generative AI with traditional AWE methods in L2 writing instruction.*

Keywords: *Writing Feedback, Large Language Models, Automated Writing Evaluation, L2 Writing, EFL Learners*

1. Introduction and Literature Review

In second language (L2) writing, feedback refers to information provided by readers to help authors revise their compositions [1]. These suggestions act as scaffolding, effectively bridging the gap between learners' current cognitive levels and their potential cognitive levels, known as the Zone of Proximal Development (ZPD) [2]. High-quality writing feedback significantly improves the writing skills, motivation, and engagement of English as a Foreign Language (EFL) learners, making it an essential part of English writing instruction. Timely and personalized feedback, in particular, plays a decisive role in fostering students' writing development [3]. However, due to limited teaching resources and time, EFL teachers find it difficult to provide real-time individualized feedback on a large scale in traditional classroom settings, especially for complex writing tasks.

Since the 1960s, advancements in computer technology have expanded possibilities for language education. Researchers have begun exploring computer-assisted writing tools to address the limitations of manual feedback. Automated Writing Evaluation (AWE) is one of the earliest rule-based and statistical automated error-correction tools investigated by scholars. AWE employs natural language processing (NLP) technology to assess and provide feedback on essays through algorithms, encompassing grammar checks, vocabulary analysis, and sentence structure evaluation. Extensive empirical research has been conducted on the efficacy of AWE. Results indicate that AWE can deliver feedback to students swiftly and efficiently, aiding them in identifying and resolving issues in their writing [4][5][6][7][8]. It can also reduce teachers' workloads and offer students unlimited practice opportunities, significantly enhancing EFL students' writing skills, motivation, and engagement. Regarding feedback quality, AWE provides relatively basic feedback, such as grammar and vocabulary corrections. It is particularly stable and accurate in error detection and correction. Recent breakthroughs in generative artificial intelligence (GenAI) have led to the emergence of large language models (LLMs) such as ChatGPT, Claude, and DeepSeek, which hold great potential in language education [9]. Through unsupervised pre-training on human-generated text data, LLMs can understand complex linguistic structures and generate fluent, coherent text. Compared to AWE, LLMs offer more personalized writing feedback, diverse error classification, and comprehensive rhetorical feedback [10][11]. As writing feedback assistants, LLMs enhance feedback quality while reducing the burden on teachers, particularly in large-scale classrooms [12]. Banihashem et al. conducted a comparative analysis of LLMs-generated feedback and peer feedback, finding that LLMs feedback tends to be more descriptive, whereas peer feedback focuses more on problem identification [13]. Their study suggests that LLMs can serve as supplementary tools in peer feedback processes, improving overall feedback quality and depth. The quality of LLMs-generated feedback has also been empirically validated. Steiss et al. compared LLMs feedback with human teacher feedback across five dimensions and found that, even without pre-training, LLMs feedback was nearly on par with human feedback [14]. Notably, LLMs even outperformed human teachers in the “criterion-based” dimension. Similarly, Guo and Wang examined differences between LLMs-generated feedback and teacher feedback in the context of Chinese university students’ argumentative essays [15]. Their results showed that LLMs provided significantly more feedback than human teachers, with balanced coverage across content, organization, and language.

However, both computer-assisted writing tools have their limitations. First, AWE relies on predefined rules and lacks the ability to evaluate deeper logical and creative aspects of writing. This means AWE systems may fail to provide sufficient feedback on certain dimensions of writing, such as discourse coherence and content development. Additionally, AWE only offers standardized feedback that is difficult to personalize, which may introduce bias and limit its adaptability to diverse writing styles and cultural contexts. As a result, AWE often struggles with specific writing tasks, particularly open-ended questions or complex arguments [16][17]. Furthermore, since AWE primarily provides surface-level feedback, its effectiveness may vary depending on students’ proficiency levels. When learners progress to higher stages of their ZPD, they may no longer benefit from AWE, suggesting that AWE may not support long-term writing improvement [18][19]. On the other hand, as cutting-edge AI products, LLMs also face challenges in practical applications. These include issues like hallucinations, which refers to generating inaccurate information and inconsistency in scoring due to variations in model versions and prompts [20]. Additionally, due to their inherent design, LLMs may generate different feedback responses even when given the same prompt, raising concerns about consistency and reliability. Moreover, LLMs-generated feedback may contain redundant or overly verbose explanations, which could be difficult for lower-proficiency learners to understand, potentially increasing their anxiety. Furthermore, Su et al. found that most students, due to limited proficiency, primarily use LLMs for error correction rather than substantive revision, which restricts their effectiveness in improving learning outcomes [21]. This observation was supported by Koltovskaia et al., who noted that students mainly relied on LLMs to fix low-level errors such as grammar, spelling, and punctuation rather than engaging in deeper writing improvement [22].

2. The Present Study

Given the respective strengths and limitations of both tools, it is essential to return to the writing classroom context to gain deeper insights into how students experience and perceive computer-generated feedback. This study focuses on two computer-assisted writing tools: rule-based AWE, which is most commonly used in English writing classrooms, and the more advanced GenAI-powered LLMs. The research particularly examines how these tools affect learners’ cognitive load and how learners perceive them.

Cognitive load refers to the burden placed on working memory when processing information [23]. It reflects the amount of information learners must simultaneously handle and the cognitive effort required to process that information during learning. Cognitive Load Theory provides a crucial theoretical lens for evaluating the effectiveness of technology-enhanced learning environments. As a key indicator of tool usability, the intensity of cognitive load directly influences students’ acceptance, as per the UTAUT model and long-term willingness to use such tools [24]. By measuring cognitive load, we can assess the suitability of AI-generated feedback in writing instruction, identify potential issues in feedback design, and predict learners’ acceptance and usage behaviors—ultimately guiding the effective integration of computer-assisted tools in writing classrooms.

Feedback does not automatically lead to improvement, while its impact depends on learners’ active engagement with the feedback content or the feedback process itself [25][26]. Therefore, how students perceive feedback directly influences their willingness to accept it, interpret it, and apply it to improve their writing. Unlike traditional teacher or peer feedback, computer-generated feedback may differ in nature and presentation, potentially affecting student perceptions. Factors such as AI’ s “non-human” identity, potential hallucination issues, and the automated nature of feedback could shape learners' attitudes and engagement.

By investigating students' perceptions of AWE and LLMs feedback, this study aims to uncover the strengths and limitations of these tools. The findings will contribute to optimizing feedback design and implementation, as well as informing strategies for integrating computer-generated feedback with other feedback forms to build a more effective writing feedback ecosystem. Specifically, this exploratory study seeks to address the following questions:

RQ1: How are students' cognitive loads while utilizing AWE and LLMs for essay revision?

RQ2: What are students' perceptions and preferences of the writing feedback provided by AWE and LLMs?

3. Methods

3.1 Participants

The participants in this study were 76 second-year students from two classes in a Chinese college. Each class had 38 students. All participants were from the English Education program, and their native language was Chinese. According to the results of a standardized college English test, the participants' English proficiency ranged from low to intermediate levels. The students were randomly divided into two groups, with one class using AWE and the other using an LLMs tool. This study employed *Pigai* (http://www.pigai.org/) as the AWE system. As one of the most widely utilized AWE platforms in China's English writing instruction, its automated feedback features on grammar and vocabulary can effectively meet the writing training needs of students. All the participants in this study were experience AWE users according to records from *Pigai*. Due to access restrictions to ChatGPT in Chinese mainland, this research selected DeepSeek as the LLMs tool, one of the most powerful LLMs, offering functions similar to ChatGPT and providing free services for educational users.

3.2 Instruments

3.2.1 Cognitive Load Questionnaire

The Cognitive Load Questionnaire (CLQ) employed in this study was adapted from Hwang et al.'s validated instrument, which was originally developed based on the measures of Paas and Sweller et al. [27][28][29]. Modifications were made to better align with the current research context. CLQ consists of 8 items across two categories: mental load (5 items) and mental effort (3 items), using a 6-point Likert scale (1 = *strongly disagree*, 6 = *strongly agree*). The measurement tool demonstrated good reliability, with a Cronbach's α of 0.844, indicating high internal consistency and ensuring the validity of the engagement metrics.

3.2.2 Feedback Perceptions Questionnaire

The study employed the Feedback Perceptions Questionnaire (FPQ) to measure students' perception of feedback, developed by Strijbos et al. [30]. The original version focused on how learners perceive, interpret, and utilize peer feedback. To align with the distinctive nature of AI-mediated feedback environments, our study implemented modifications [31]. Specifically, the Acceptance dimension was removed due to insufficient reliability ($\alpha < 0.70$ in pilot testing). The fairness dimension was also excluded because this construct does not align with the human-machine interaction context in this study. The finalized instrument comprised three dimensions and 9 items: usefulness (3 items), willingness to improve (2 item) and affect (4 item), rated on a 6-point Likert scale (1 = *strongly disagree*, 6 = *strongly agree*). The overall Cronbach's alpha was 0.93 indicating high internal consistency. To capture qualitative insights, the questionnaire concluded with an open-ended item: "Which computer-assisted writing tool would you prefer for future writing revisions, and why?" This supplementary question was designed to triangulate quantitative findings with learners' tool selection rationale.

3.3 Research Procedure

The experimental procedure is illustrated in figure 1. This study first employs a between-group experimental design to compare the differences in cognitive load between students using an AWE system and a LLMs tool for essay revision. This design effectively controls for potential interference from learning effects and fatigue on the research outcomes by preventing participants from undergoing repeated experimental treatments. When comparing students' perceptions of the two computer-assisted writing tools, a single-group delayed recall design was employed. Specifically, students of the LLMs group evaluated their perceptions of both the LLMs and AWE tools one week after completing the revision task with LLMs feedback. The delayed recall design ensured that assessments of both computer-assisted writing tools were based on retained memory, reflecting lasting impressions of the tools. Although the single-group experiment does not enable direct comparison between the two tools, this design simulates the real-world process of students transitioning naturally from AWE to LLMs, providing user insights to inform decisions on educational technology adoption.

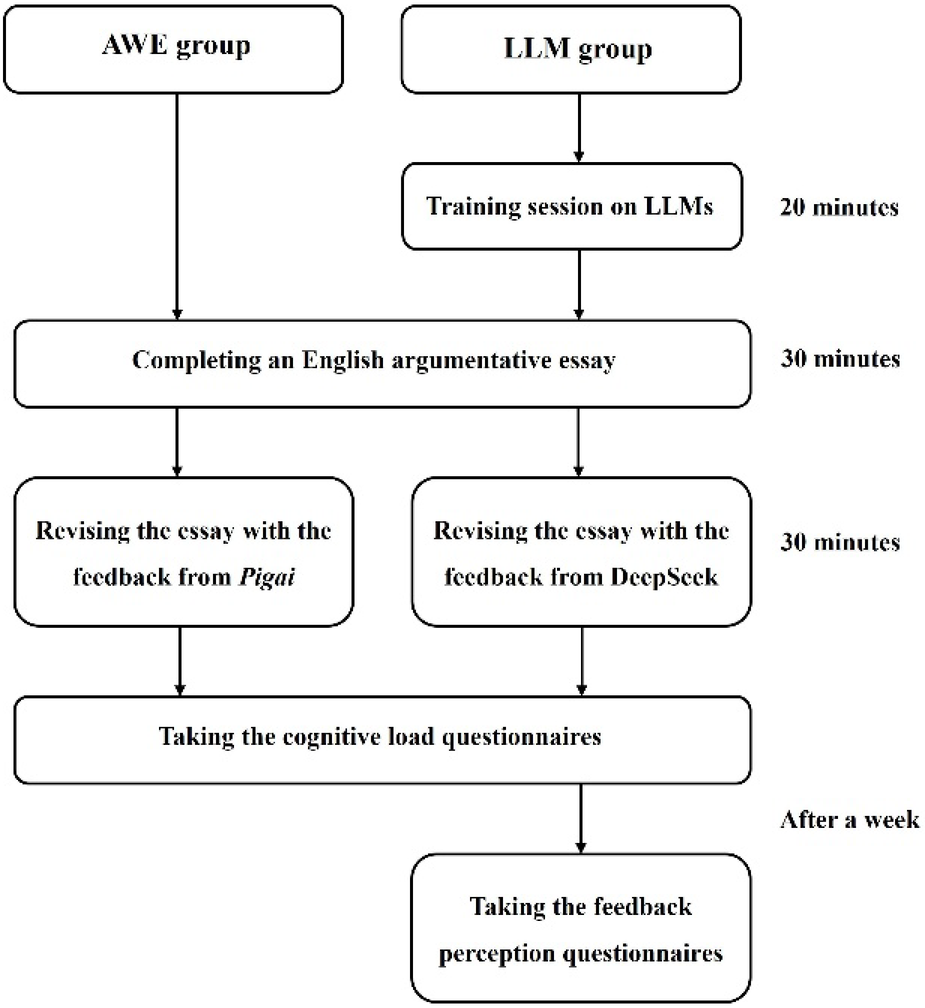
AWE group
LLM group
Training session on LLMs
20 minutes
Completing an English argumentative essay
30 minutes
Revising the essay with the feedback from *Pigai*
Revising the essay with the feedback from DeepSeek
30 minutes
Taking the cognitive load questionnaires
After a week
Taking the feedback perception questionnaires

Figure 1: Diagram of research procedure.

Before the experiment, the researchers provided all participants with a detailed explanation of the study's purpose, procedures, and their rights (including the principle of voluntary participation and confidentiality clauses). The experiment commenced only after obtaining written informed consent. To ensure standardized use of the AI tools, the LLMs group received a 20-minute training session on tool operation, prompt optimization and ethical considerations. The AWE group followed a standard operating procedure without additional training, as the tool's operation was relatively simple and participants already had basic experience using it.

The experimental tasks were conducted in a standardized classroom setting. All participants first completed an English argumentative essay within 30 minutes. The difficulty of the writing task was equivalent to that of the CET-4 writing section, aligning with current teaching requirements and students' proficiency levels. After completing the initial draft, the AWE group uploaded their texts to the *Pigai* website to receive writing feedback, while the LLMs group interacted with DeepSeek to obtain revision suggestions. Subsequently, participants revised their essays based on the AI-generated feedback within 30 minutes. Immediately after the revision task, participants filled out the CLQ to measure their cognitive load during the task. Additionally, the LLMs group completed the FEQ, which combined quantitative and qualitative methods to assess students' usage experiences and preferences for AWE (based on prior use) and LLMs (current use). Participants were also asked to provide examples for their choices. Given that the participants' native language was Chinese, all questionnaires were administered in Chinese to ensure the comprehension of the questions and enable them to express genuine opinions in open-ended responses. All data collection is completed through *Wenjuanxing* platform.

3.4 Data analysis

Data analysis was conducted using the statistical software SPSS 27. For Research Question #1 (RQ1), the study first used descriptive statistical methods to analyze the cognitive load of participants when using AWE and LLMs to revise their essays, followed by t-tests to compare the means between the two groups. Addressing the Research Question #2 (RQ2), the study also conducted descriptive statistical methods to analyze participants' perceptions of the two tools. T-tests was also used to examine the differences in students' perceptions of the two tools across three dimensions.

To further explore students' preferences of computer-assisted writing tools, this study coded the textual responses to open-ended questions to identify high-frequency themes. The coding tool used was NVivo 15. First, NVivo 15 was employed to perform word frequency analysis on the text to understand the preference proportions for the two tools. Then, thematic analysis was conducted on the text. The specific operational procedure followed the thematic analysis steps proposed by Braun and Clarke [32]. The researchers repeatedly read the textual data, used an inductive approach to derive themes from the initial codes, and refined them through multiple revisions to finalize the theme names. To ensure reliability, an experienced EFL writing teacher reviewed the coding scheme to enhance content validity. The scheme was revised iteratively by both parties until consensus was reached. Cohen's Kappa coefficient was 0.93.

4. Results

4.1 Cognitive load

The descriptive statistical results of cognitive load indicate that the distribution of cognitive load in both groups of participants follows a pattern dominated by moderate and low load, with no individuals exhibiting high load (see Table 1). Specifically, in the AWE group, 60.5% (n=23) of the participants were in a state of moderate load, 39.5% (n=15) were in a state of low load, and the proportion of high load was 0%. The distribution in the LLMs group was similar, with 55.3% (n=21) at moderate load and 44.7% (n=17) at low load, also showing no high-load individuals (0%). The results indicate that the overall cognitive load levels in both groups were relatively low, with a balanced distribution of cognitive demands within each group and no extreme polarization observed.

Table 1 Descriptive statistical results of AWE and LLMs group on cognitive load.

Group	Low CL(n/%)	Mid CL(n/%)	High CL(n/%)	Total
AWE	15 (35.9%)	23 (60.5%)	0 (0%)	38
LLMs	17 (44.7%)	21 (55.3%)	0 (0%)	38

*Cognitive Load=CL

The independent samples t-test was used to compare the differences in cognitive load between the LLMs group and the AWE group. As shown in Table 2, the cognitive load score of the LLMs group (M=2.88, SD=0.63) was slightly lower than that of the AWE group (M=3.02, SD=0.74). The results of the independent samples t-test indicated that the difference in cognitive load ratings between the two groups did not reach statistical significance (*p*=0.38). This suggests that there was no significant difference in the cognitive load experienced by students in the two groups when using different tools for essay revision under the conditions of this study.

Table 2 Independent-sample T-test result of the cognitive load on the questionnaire scores of AWE and LLMs group.

Group	N	Mean	SD	*t*	*p*	Cohen's d
AWE	38	3.02	0.74	0.88	0.38	0.69
LLMs	38	2.88	0.63			

4.2 Feedback perception

The second research question explores students' perceptions in using two tools for essay revision, covering three dimensions: usefulness, willingness to improve, and affect. Table 3 presents the descriptive statistics of the single group of students (n=38) regarding their perceptions of using AWE and LLMs. The descriptive statistics show that students' perception rates for LLMs were slightly higher than those for AWE across all three dimensions. In terms of Usefulness, LLMs (M=4.46, SD=0.66) scored higher than AWE (M=4.28, SD=0.75). Similarly, in the Willingness to Improve dimension, LLMs (M=4.51, SD=0.73) outperformed AWE (M=4.29, SD=0.84). The same trend was observed in the Affect dimension (LLMs: M=4.33, SD=0.47, AWE: M=4.24, SD=0.55). Independent samples t-tests revealed that these differences did not reach statistical significance (Usefulness *p*=0.29, Willingness to Improve *p*=0.22, Affect *p*=0.47). Considering that even minor differences may have practical implications in education, we further explored the manifestations of these differences through qualitative data.

Table 3 Independent-sample t-test result of perception of AWE and LLMs on the questionnaire scores.

	AWE		LLMs				
	38		38				
	Mean	SD	Mean	SD	*t*	*p*	Cohen's d
Usefulness	4.28	0.75	4.46	0.66	-1.08	0.29	0.71
Willingness to Improve	4.29	0.84	4.51	0.73	-1.24	0.22	0.79
Affect	4.24	0.55	4.33	0.47	-0.73	0.47	0.51

To gain a deeper understanding of the slight preference for LLMs among students as indicated by the quantitative data, we conducted a thematic analysis to explore the underlying reasons. A total of 32 valid responses were collected from the open-ended questionnaires. The word frequency analysis showed that among the 93.75% (n=30) of participants exhibited a preference for LLMs-generated writing feedback, which is consistent with the quantitative results. The reasons students described for their preference for LLMs primarily revolved around two core dimensions: (1) the features of LLMs feedback, and (2) the interactive experience with LLMs (table 4).

Regarding feedback features, participants emphasized four key advantages of LLMs. Students pointed out that LLMs feedback is highly specific and detailed, capable of accurately identifying grammatical errors and providing concrete improvement suggestions. Second, students considered the clarity of LLMs feedback to be one of its notable strengths, as its expressions are easy to understand. Additionally, the comprehensiveness of LLMs feedback was also acknowledged by students, as it gave suggestion from multiple dimensions. Most significantly, students particularly valued the personalized nature of LLMs feedback, emphasizing its adaptive capacity to generate tailored suggestions based on individual writing patterns and needs. In terms of interactive experience, students recognized the user-friendly interface of LLMs, noting the process as convenient and efficient. Particularly noteworthy was the widespread acclaim for LLMs' follow-up questioning feature. Students mentioned that this function allows them to delve deeper into the feedback through continuous interaction until they fully grasp the revision suggestions.

Table 4 Students' reasons for preferring LLMs feedback.

Theme	Sub-theme	Definition	Example
LLMs feedback features	Personalization	Student thinks LLMs feedback is tailored to individual needs	The LLMs provides feedback that better aligns with my personal needs. (Student 35)
	Specificity	Student thinks LLMs feedback is specific and detailed	The LLMs can identify my grammar errors, explain the reasons behind them, and provide specific suggestions for improvement. (Student 34)
	Clarity	Student thinks LLMs feedback is easy to understand	The LLMs explains things in a way that's easy to follow. (Student 12)
	Comprehensiveness	Student thinks LLMs feedback covers multiple aspects	Compared to AWE, the LLMs can analyze my essay from multiple perspectives. (Student 26)
Interaction with LLMs	User-friendly	Student think using LLMs to generate feedback is easy	The LLMs is very convenient to use. (Student 28)
	Follow-up Capability	Students think the follow-up capability of LLMs is useful	I can keep asking the LLMs follow-up questions about my revisions until I fully understand. (Student 17)

It is worth noting that a small number of students (n=2) still insisted on choosing AWE system. One reason was that students found Pigai simpler to use. For instance, one student remarked: *"Pigai is very convenient to use, because you can see the score and corrections immediately after submitting the essay, with no additional steps required."* Additionally, students felt that understanding the extensive writing feedback generated by AWE required less effort than LLMs. One student noted: *"The writing feedback generated by DeepSeek is sometimes incoherent and hard to understand."*

5. Discussion

This research measured the cognitive load of two groups of students using two different tools for essay revision. Students using AWE and those using LLMs both exhibited predominantly low to moderate levels of cognitive load, with neither group showing individuals with high cognitive load (0% high load). This indicates that neither AI tool imposed excessive cognitive pressure on students. From the tool perspective, it suggests that current mainstream AI writing assistance tools (including AWE and LLMs) do not exceed learners' cognitive processing capacity thresholds (Sweller et al., 2019) in terms of functionality, making them suitable for learners' proficiency levels. This finding confirms the fundamental advantage of AI writing assistance tools over traditional manual feedback methods. Previous research has shown that students receiving teacher feedback often experience high cognitive load potentially due to the authoritative nature of teacher feedback (Hyland & Hyland, 2006) whereas the "low-threat correction environment" created by AI tools in this study may be the key to avoiding high cognitive load (Xiao et al., 2024). While cognitive load in AI-assisted writing remains understudied, our findings demonstrate that both AWE and LLMs maintain manageable cognitive demands, supporting their practical adoption. As Xiao et al. (2024) noted, AI agents provide timely "learning scaffolds" to help learners break through their ZPD, thereby reducing cognitive load during tasks. This shared characteristic provides strong evidence for the feasibility of integrating intelligent educational tools into writing instruction, enhancing efficiency while maintaining cognitive safety.

Additionally, this study found that even as a new tool for participants, LLMs exhibited slightly lower cognitive load compared to the well-experiences AWE. First, this suggests that generative AI has a low learning curve, as students only needed short-term training to use LLMs effectively, indicating its interaction design is more intuitive. Second, it implies that the interaction style of generative AI may be more consistent with students' cognitive habits than traditional AI methods. Compared to AWE's rule-based feedback (e.g., error marking), LLMs' natural language generation (e.g., directly offering rewritten suggestions) may reduce students' cognitive burden. This suggests that LLM's generative feedback model may offer greater cognitive efficiency advantages over traditional rule-based AWE, particularly in reducing learners' cognitive load. However, whether this advantage translates to long-term learning improvements requires further validation.

Refer to student's perception of AWE and LLMs, analysis of student questionnaire responses indicates that while observed differences remain statistically modest, LLMs tools demonstrate significantly higher student preference across three key dimensions: usefulness, willingness to improve, and affect. This finding suggests students' recognition of LLMs-generated feedback as being more corresponding to their writing revision requirements. Qualitative data analysis provides evidence for this interpretation. Specifically, LLMs-generated feedback demonstrates four distinctive characteristics: specificity, clarity, comprehensiveness, and personalization. These findings supporting the previous research results that LLMs exhibit significant potential in providing personalized and multi-dimensional feedback. These attributes collectively facilitate students' ability to comprehend and implement suggested writing improvements with greater efficacy. A particular advantage lies in the personalization dimension, where LLMs exhibit the capacity to deliver customized recommendations by analyzing individual writing patterns and needs. AWE is unable to replicate this feature. Furthermore, LLMs possess a critical advantage through their contextual follow-up capability. This permits iterative inquiry, allowing learners to pursue in-depth exploration of specific writing issues and receive progressively detailed explanations. Such dynamic interaction fosters deeper cognitive engagement with the feedback, ultimately leading to more thorough internalization and application of revision strategies. Conversely, AWE systems are constrained by their corpus, which typically limits users to passive reception of predetermined feedback without opportunities for clarification or extended dialogue. These functional differences underscore LLMs' potential for developing writing competence.

However, a small proportion of students continue to prefer AWE systems, primarily valuing their user-friendly interface and straightforward feedback. This preference suggests that AWE systems maintain distinct advantages, including immediate score displays, clear correction results, and concise feedback that is easy to interpret. Importantly, while group-level data may show minimal differences, individual experiences vary, and the preferences of these students should not be disregarded. From a tool-design perspective, AWE systems offer ease of use and feedback clarity, making them especially suitable for students who are less comfortable with new technologies or who prefer immediate, simplified responses. From a learner perspective, some students may face challenges in effectively using LLMs due to limited self-regulated learning skills or language barriers. When presented with the detailed and interactive feedback produced by LLMs, these learners may experience cognitive overload or difficulty applying the suggestions. Forcing them to use LLMs would cause frustration and confusion, negatively impacting their motivation and grit for writing.

6. Implication, Limitation and Future Directions

This study explores the cognitive load and student perceptions of using AWE and LLMs for English writing revision. The findings indicate that neither AWE nor LLMs imposed an excessively high cognitive load on students, with LLMs scoring slightly lower than AWE in terms of cognitive load. Meanwhile, LLMs held a slight advantage in student perceptions, primarily reflected in the specificity, clarity, comprehensiveness, and personalization of feedback. This aligns with existing empirical research conclusions, which suggest that LLMs have the potential to provide personalized and interactive feedback. However, some students still preferred using AWE systems, indicating that traditional AWE tools retain their relevance in EFL writing classrooms.

The findings offer insights into EFL learners' cognitive load and perceptions of using computer-assisted writing tools for writing revision. Educators should select appropriate tools based on individual student differences to ensure effective writing learning. Particularly for students with weak writing foundations or limited metacognitive awareness, AWE tools can provide real-time feedback on grammar and word usage, enabling rapid improvement in language expression while keeping cognitive load low. For intermediate-level students, GenAI tools like LLMs can help optimize logical coherence. Since AI tools still have certain limitations, teachers are advised to combine them with human feedback in actual writing instruction to build a more comprehensive and effective writing feedback support system. Additionally, writing teachers should enhance their own AI literacy, including the ability to apply AI tools and understand AI ethics. Only by fully grasping the capabilities and limitations of AI tools can they effectively integrate them into teaching practices.

The study has several limitations, which also point to directions for future research. First, the study primarily adopts a student perspective. Future research should also explore teachers' perspectives on the AWE and LLMs to provide additional insights. As organizers and guides of teaching activities, teachers' opinions and experiences with these tools will significantly impact teaching practices. Understanding their perspectives can offer a more comprehensive assessment of the strengths and limitations of these tools. Second, the experimental duration in this study was relatively short, future research should extend the experimental period to several weeks or months to examine changes in various dimensions after prolonged use of AWE or LLMs tools. Periodic evaluations should also be conducted, with data collected at different time points to analyze the adaptability of tool use and the sustainability of learning effects. Finally, the study relied on self-report scales, which carry the risk of subjective bias. Future research should incorporate behavioral data tracking, such as classroom observations and feedback engagement assessments, to enhance objectivity.

Computer-assisted writing tools hold immense potential in English writing education. Teachers need to understand the strengths and limitations of rule-based AWE and generative LLMs and effectively integrate them into writing instruction to help students improve their writing skills. Ultimately, this will lead to more efficient and personalized English writing instruction.

Acknowledgements

The work was supported by the Project of Quality Engineering in Yangjiang Polytechnic [Grant No. 2024JG013], Project title: Research on innovative teaching practices in English writing based on Large Language Models

References

[1] Hyland K. Teaching and researching writing[M]. Routledge, 2015.

[2] Vygotsky L S. Mind in society: The development of higher psychological processes[M]. Harvard university press, 1978.

[3] Lee I. Utility of focused/comprehensive written corrective feedback research for authentic L2 writing classrooms[J]. Journal of Second Language Writing, 2020, 49: 100734.

[4] Attali Y. Exploring the feedback and revision features of Criterion[J]. Journal of Second Language Writing, 2004, 14(3): 1-20.

[5] Bai L, Hu G. In the face of fallible AWE feedback: How do students respond?[J]. Educational Psychology, 2017, 37(1): 67-81.

[6] Barrot J S. Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy[J]. Computer Assisted Language Learning, 2023, 36(4): 584-607.

[7] Chen M, Cui Y. The effects of AWE and peer feedback on cohesion and coherence in continuation writing[J]. Journal of Second Language Writing, 2022, 57: 100915.

[8] Zhang Z V, Hyland K. Fostering student engagement with feedback: An integrated approach[J]. Assessing Writing, 2022, 51: 100586.

[9] Kohnke L, Moorhouse B L, Zou D. ChatGPT for language teaching and learning[J]. Relc Journal, 2023, 54(2): 537-550.

[10] Gillani N, Eynon R, Chiabaut C, et al. Unpacking the "Black Box" of AI in education[J]. Educational Technology & Society, 2023, 26(1): 99-111.

[11] Roumeliotis K I, Tselikas N D. Chatgpt and open-ai models: A preliminary review[J]. Future Internet, 2023, 15(6): 192.

[12] Han J, Yoo H, Myung J, et al. LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction[J]. arXiv preprint arXiv:2310.05191, 2023.

[13] Banihashem S K, Kerman N T, Noroozi O, et al. Feedback sources in essay writing: peer-generated or AI-generated feedback?[J]. International Journal of Educational Technology in Higher Education, 2024, 21(1): 23.

[14] Steiss J, Tate T, Graham S, et al. Comparing the quality of human and ChatGPT feedback of students' writing[J]. Learning and Instruction, 2024, 91: 101894.

[15] Guo K, Wang D. To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing[J]. Education and Information Technologies, 2024, 29(7): 8435-8463.

[16] Ranalli J. L2 student engagement with automated feedback on writing: Potential for learning and issues of trust[J]. Journal of Second Language Writing, 2021, 52: 100816.

[17] Meyer J, Jansen T, Schiller R, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100199.

[18] Koltovskaia S. Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study[J]. Assessing Writing, 2020, 44: 100450.

[19] Zhang Z V. Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student

perceptions and revisions[J]. Assessing Writing, 2020, 43: 100439.

[20] Guo K. EvaluMate: Using AI to support students' feedback provision in peer assessment for writing[J]. Assessing Writing, 2024, 61: 100864.

[21] Su Y, Lin Y, Lai C. Collaborating with ChatGPT in argumentative writing classrooms[J]. Assessing Writing, 2023, 57: 100752.

[22] Koltovskaia S, Rahmati P, Saeli H. Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement[J]. Journal of Second Language Writing, 2024, 65: 101130.

[23] Sweller J. Cognitive load theory, learning difficulty, and instructional design[J]. Learning and instruction, 1994, 4(4): 295-312.

[24] Hoch E, Sidi Y, Ackerman R, et al. Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective[J]. Educational Psychology Review, 2023, 35(2): 61.

[25] Hattie J, Timperley H. The power of feedback[J]. Review of educational research, 2007, 77(1): 81-112.

[26] Winstone N E, Nash R A, Rowntree J, et al. 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience[J]. Studies in Higher Education, 2017, 42(11): 2026-2041.

[27] Hwang G J, Yang L H, Wang S Y. A concept map-embedded educational computer game for improving students' learning performance in natural science courses[J]. Computers & Education, 2013, 69: 121-130.

[28] Paas F G W C. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach [J]. Journal of educational psychology, 1992, 84(4): 429.

[29] Sweller J, Van Merrienboer J J G, Paas F G W C. Cognitive architecture and instructional design[J]. Educational psychology review, 1998, 10: 251-296.

[30] Strijbos J W, Pat-El R, Narciss S. Structural validity and invariance of the feedback perceptions questionnaire [J]. Studies in Educational Evaluation, 2021, 68: 100980.

[31] Hancock P A, Billings D R, Schaefer K E. Can you trust your robot?[J]. Ergonomics in Design, 2011, 19(3): 24-29.

[32] Braun V, Clarke V. Using thematic analysis in psychology[J]. Qualitative research in psychology, 2006, 3(2): 77-101.

Appendix. Questionnaires.

Cognitive load

Mental load ML1 The writing feedback was difficult for me to understand. ML2 I had to put a lot of effort into revision with the writing feedback. ML3 Revising my essay based on the writing feedback was very troublesome for me ML4 I felt very frustrated when revising essay with the writing feedback. ML5 I need to spend a lot of time digesting and applying all the provided feedback and suggestions.
Mental effort ME1 The way this tool presents writing feedback required a lot of mental effort from me. ME2 I had to work very hard to complete the essay revision. ME3 It was difficult to keep up with the guidance and suggestion of the tool

Feedback perception

Usefulness US1 I think this type of feedback useful. US2 I think this type of feedback is very helpful for my writing. US3 I think this type of feedback to be very constructive.
Willingness to improve WI1 I am willing to revise my essay based on this type of feedback. WI2 I am willing to apply this type of feedback to my future writing.
Affect AF1 I was very dissatisfied with this type of feedback. AF2 I was very satisfied with this type of feedback. AF3 Receiving this type of feedback made me feel frustrated. AF4 Receiving this type of feedback made me feel confident.