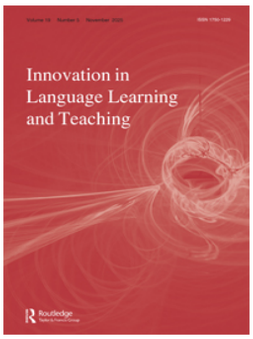

Innovation in Language Learning and Teaching

ISSN: 1750-1229 (Print) 1750-1237 (Online) Journal homepage: www.tandfonline.com/journals/rill20

The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study

Zhihui Zhang, Scott Aubrey, Xiaomeng Huang & Thomas K. F. Chiu

To cite this article: Zhihui Zhang, Scott Aubrey, Xiaomeng Huang & Thomas K. F. Chiu (13 May 2025): The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study, Innovation in Language Learning and Teaching, DOI: 10.1080/17501229.2025.2503890

To link to this article: https://doi.org/10.1080/17501229.2025.2503890

INNOVATION IN LANGUAGE LEARNING AND TEACHING
https://doi.org/10.1080/17501229.2025.2503890

Routledge
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study

Zhihui Zhang[a], Scott Aubrey [a], Xiaomeng Huang[b] and Thomas K. F. Chiu [c]

[a]Department of Curriculum and Instruction, Faculty of Education, The Chinese University of Hong Kong, Hong Kong SAR, China; [b]Alibaba Cloud, Alibaba Hangzhou Headquarters, Hangzhou, China; [c]Department of Curriculum and Instruction Faculty of Education and Centre for Learning Sciences and Technologies and, Centre for University and School Partnership, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

ABSTRACT

This study examines the impact of Generative AI (GenAI) – generated feedback and hybrid feedback (GenAI and human tutor input) on L2 academic writing skills. Over 12 weeks, 60 Chinese EFL students were divided into two groups: Group 1 received only GenAI feedback, while Group 2 received hybrid feedback. Writing performance was assessed using GRE writing rubrics. Results showed that GenAI feedback improved Group 1's writing performance, particularly in grammar and sentence variety. However, its impact on higher – order skills like critical thinking and organization was limited. In contrast, Group 2 demonstrated greater improvement, with significant gains in organization, critical thinking, and sentence variety. The hybrid feedback group also reported higher motivation and more positive feedback perceptions. These findings suggest that while GenAI feedback is effective for basic writing skills, hybrid feedback is more beneficial for developing complex academic writing skills. The study highlights the importance of combining AI and human feedback to meet diverse learning needs in L2 writing instruction. Future research should explore the long – term effects of hybrid feedback and its application in different writing contexts.

ARTICLE HISTORY
Received 14 January 2025
Accepted 3 May 2025

KEYWORDS
Generative AI; L2 writing; academic writing; feedback; critical thinking

1. Introduction

Writing feedback is a critical component in the development of writing skills, as it helps learners identify issues, enrich content, and align their work with expectations (Nunes et al. 2022). Historically, Automated Writing Evaluation (AWE) systems, such as Grammarly, have supported learners by offering automated feedback. However, these systems struggle to address the increasingly complex and dynamic needs of modern learners. The emergence of Generative AI (GenAI), exemplified by GPT models, represents a significant leap in artificial intelligence (AI) technology. With its advanced contextual understanding, GenAI can deliver nuanced, goal-oriented, and human-like feedback, providing the necessary conditions for language learners seeking personalized support (Li et al. 2024).

Despite the potential of GenAI, its application in L2 writing remains underexplored. Since 2011, research on AWE systems has primarily focused on language and writing courses in higher

CONTACT Thomas K. F. Chiu tchiu@cuhk.edu.hk Department of Curriculum and Instruction Faculty of Education and Centre for Learning Sciences and Technologies and, Centre for University and School Partnership, The Chinese University of Hong Kong, Shatin, Hong Kong SAR

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17501229.2025.2503890.

education, evaluating effectiveness through metrics like feedback accuracy, user engagement, and impact on writing performance (Nunes et al. 2022; Shi and Aryadoust 2024). However, while various AWE systems (e.g. Criterion, Grammarly) have been studied, there is a significant gap in research on the application of emerging GenAI technologies like ChatGPT. Empirical studies on the impact of GenAI on writing skills present mixed findings. Some research suggests that ChatGPT significantly enhances writing quality, reduces writing anxiety, and increases motivation (Escalante, Pack, and Barrett 2023; Song and Song 2023), while other studies (Chowdhury et al. 2024) highlight potential drawbacks, such as impaired creative writing abilities and lower scores in originality and content accuracy.

This mixed evidence underscores the need for further investigation into the effectiveness of GenAI-generated feedback, particularly in the context of L2 writing. In L2 settings, the effectiveness of GenAI-generated feedback is further complicated by the revision strategies employed by learners, such as translation-based revisions and reliance on templates modes. These strategies often result in a focus on linguistic form, such as grammar and spelling, while neglecting higher-order concerns like self-monitoring and content development (Mohsen 2022). This raises critical questions about how L2 learners can effectively utilize AI-generated feedback to improve their writing skills.

Moreover, most existing research examines single feedback modes, such as purely AI or human feedback, leaving a gap in understanding the potential of hybrid feedback models that combine AI and human input. While technological tools are most effective when they complement, rather than replace, human feedback (Chiu 2024; Mohsen 2022; Shi and Aryadoust 2024), research on hybrid feedback models remains limited. Hybrid feedback has the potential to enhance L2 learners' writing skills and self-regulated learning abilities by addressing both linguistic form and higher-order concerns. While GenAI offers timely, consistent, and objective feedback on surface-level issues, human tutors provide depth and nuance, particularly in areas like personalized guidance and motivational support. It can be extended to other disciplines and fields, promoting the deep integration of educational technology with teaching practices and improving overall educational quality.

To address these gaps, this study proposes an innovative approach by comparing the effects of GenAI-generated feedback with hybrid feedback (GenAI + human tutor) on L2 academic writing skills. Grounded in Self-Determination Theory (SDT) (Deci and Ryan, 2012) and Hattie and Timperley's Feedback Model (2007), the study examines how these feedback modes meet learners' psychological needs (autonomy, competence, and relatedness) and address different levels of feedback (task, process, self-regulation, and self-level). Specifically, the study seeks to answer two research questions:

1. To what extent can GenAI-generated feedback enhance L2 academic writing skills?
2. How does hybrid feedback – combining GenAI with human input – improve L2 academic writing skills compared to GenAI alone?

To investigate these questions, the study employs an experimental design involving 60 Chinese EFL students, divided into two groups: one receiving only GenAI feedback and the other receiving hybrid feedback. Writing performance is assessed using GRE writing rubrics, while feedback experiences are evaluated through questionnaires and semi-structured interviews. This study aims to make a unique contribution to the field of AI feedback research in L2 contexts. The findings are expected to offer concrete guidance for L2 writing instruction, enrich theoretical foundations of L2 writing skills, and provide practical strategies for educators and students in selecting the most effective feedback modes. Additionally, this approach is especially significant for learners with limited resources, such as those in remote areas, as it combines AI with the personalized touch of human instruction, ensuring a more equitable and effective learning experience.

2. Literature review

Writing feedback is crucial for the academic development of L2 learners, as academic writing requires a strong integration of vocabulary, grammar, and clear structure, which often poses

challenges for them (Cen and Zheng 2024). Feedback helps learners identify errors, enhance clarity, logical flow, and accuracy, and guides them in meeting academic standards, thereby improving their ability to express complex ideas effectively.

2.1. Automated writing feedback (AWF)

The emergence of AWF stems from the need for efficient writing assessment tools. Traditional manual evaluation methods are time-consuming and labor-intensive, particularly in large-scale writing assessments and daily teaching practices. To address these challenges, computer technologies were introduced into the field of writing assessment to enable automated scoring and feedback functionalities. Early systems like Project Essay Grade (PEG) were primarily designed for scoring based on basic features such as word count and sentence length, rather than interactive feedback. Advances in Natural Language Processing (NLP) technologies facilitated broader adoption, with tools such as ETS's e-rater, integrated into standardized testing (e.g. TOEFL), providing feedback on grammar, style, and organization. The integration of AWF into high-stakes tests solidified its role in standardized assessment.

In the 2010s, AWF gained widespread attention with tools like Grammarly and the Pigai system (popular in China), offering real-time corrective feedback to language learners in both classroom and self-study environments. Additionally, AWF began to develop personalized features, such as Grammarly's ability to provide targeted grammar and vocabulary suggestions based on students' language proficiency. Research during this period highlighted AWF's potential to scaffold writing processes, though its main focus remained on grammar and spelling (Nunes et al. 2022). Furthermore, AWF systems enhanced user experience through graphical interfaces and interactive tools, further promoting student motivation. Shi and Aryadoust (2024) noted the widespread application of AWF systems in language learning and writing instruction, particularly in ESL and EFL writing classrooms.

Despite significant advancements, AWF systems still face several challenges. For instance, the accuracy of AWF scoring may be limited by training data, and it cannot fully replace human evaluators' contextualized assessment of writing (Link, Mehrzad, and Rahimi 2022). Additionally, AWF systems need improvement in depth and personalization of feedback to better meet diverse learner needs. AWF systems also have limitations in writing instruction. Ranalli (2018) argued that AWF lacks a human pedagogical element, overlooks individual learner differences, and sometimes results in excessive corrections. Although most research during this period affirmed AWF's efficacy in improving writing accuracy and student engagement, meta-analyses reported a medium effect size on writing performance, with mixed results (Ding and Zou 2024).

Existing studies lack in-depth exploration of the long-term effects of AWF system feedback. Most research focuses on short-term writing performance improvements, without adequately assessing the potential impact of AWF systems on students' long-term writing development (Ngo, Chen, and Lai 2024). Huang and Renandya (2020) found that Pigai's application among low-level EFL learners was ineffective, possibly due to students' unfamiliarity with the tool and the short duration of the study. Saricaoglu (2019) found that AWF did not significantly improve the writing performance of low-level EFL learners, possibly due to the short study duration (two weeks) and students' unfamiliarity with AWF tools. Most existing studies employ small-scale experiments or exploratory research designs, limiting the generalizability and reliability of the findings and making it difficult to apply them directly to broader educational practices. Therefore, this study will adopt a mixed-methods approach, combining quantitative data analysis with qualitative interviews, to comprehensively understand students' experiences and feedback during long-term use of AWF systems. Existing systems primarily focus on micro-level feedback (McCarthy et al. 2022), with fewer studies on other writing types (e.g. argumentative essays) and macro-level aspects (e.g. structure and rhetoric), as well as comparative studies on students' critical thinking and organization in writing.

2.2. Generative AI (GenAI) in education

GenAI brings the evolution of Automated Writing Feedback (AWF) systems. Tools such as ChatGPT and Microsoft Copilot are excel in delivering more accurate and comprehensive feedback. These systems not only enhance precision in feedback but also offer deeper insights into content, coherence, and organization, thereby aiding students in refining their writing skills (Ding and Zou 2024; Ngo, Chen, and Lai 2024). While research has begun to explore GenAI's utility as an AWF tool, its pedagogical implications remain underexamined. GenAI has demonstrated promise in assisting L2 learners with foundational skills such as grammar and sentence structure. Its advanced natural language generation capabilities enable it to manage nuanced expressions and multi-level semantics, making it particularly valuable for complex writing tasks. Some studies have even found that GenAI-generated feedback is comparable to human feedback in quality, offering both supportive and motivational responses (Cen and Zheng 2024; Steiss et al. 2024).

Despite its potential, GenAI's feedback quality remains inconsistent and limited in certain respects (Steiss et al. 2024). For example, AI-generated feedback may lack accuracy, particularly when evaluating high-quality essays, potentially leading students to internalize incorrect guidance and hindering their writing development.. Research by Chowdhury et al. (2024) highlights that the use of ChatGPT may negatively impact students' creative writing abilities. Their findings reveal that students relying on ChatGPT performed worse in areas such as content accuracy, presentation, expansion, and originality compared to those who did not use the tool. While ChatGPT-generated content is often grammatically and structurally sound, it tends to lack depth and uniqueness, potentially fostering over-reliance on template-based outputs and stifling independent thinking and creativity (Kohnke 2024). ChatGPT excels in content generation and feedback provision, it struggles with organizing writing structure and maintaining logical coherence. Students may overlook the overall structure and logical flow of their essays, resulting in content that lacks depth and cohesion – a limitation attributed to GenAI's superficial processing of language (Kohnke 2024). Moreover, although some studies suggest that GenAI can boost writing motivation, others indicate that prolonged use may diminish students' interest in writing, as over-reliance on AI tools may erode the sense of achievement and enjoyment derived from autonomous writing.

Another critical gap in the literature is the insufficient attention to the differential impact of GenAI across various language learning groups, particularly non-native English speakers (ELS) and disadvantaged learners. These shortcomings underscore the need for further research to better understand and optimize the educational effectiveness of GenAI in diverse language learning contexts. Addressing these gaps will be essential for harnessing the full potential of GenAI in enhancing writing instruction and learning outcomes.

2.3. Hybrid feedback

While students and instructors generally hold positive attitudes towards AWF, recognizing their potential to enhance writing skills and teaching efficiency, some research indicates a preference for feedback from human raters, whether teachers or peers (Shi and Aryadoust 2024). Specifically, compared to traditional feedback, AWF systems do not necessarily lead to better writing improvement. For instance, Luo and Liu (2017) found that while experimental group students included more relevant themes in their writing, they performed worse in organization, readability, and diversity compared to the teacher feedback group. Ware (2014) found that students using the Criterion system did not show significant improvement in writing quality, length, and mechanics, and the effectiveness of automated feedback was not superior to traditional feedback methods like teacher or peer feedback. Zhang and Hyland (2018) found that compared to teacher feedback, AWF was less effective in certain aspects (e.g. content relevance). Sari and Han (2022) found that Criterion performed poorly in consistency with human raters, with significant discrepancies between different raters.

Most studies examine AI feedback or human feedback in isolation, with few exploring the potential synergies of a hybrid model. Therefore, the current trend in AWF development requires the integration of more instructional strategies to better support students' writing development. There is also rising interest in hybrid feedback systems, which combine AI-generated insights with human expertise to offer a more balanced approach. For instance, preliminary research suggests that students value AI for its immediacy but still rely on instructors for nuanced, context-specific guidance. Tang and Rich (2017) found in a Chinese EFL context that while students' writing quality improved, the effect was not significant, and the combination with teacher feedback yielded better results. Hybrid feedback has proven particularly beneficial in high-stakes assessments like TOEFL, enhancing evaluation quality and student engagement (Shi and Aryadoust 2024). Despite these benefits, effectively integrating AI and human feedback remains challenging, and more research is needed to optimize this combination.

3. This study

This study aims to address the gaps in existing research by comparing the effects of Generative AI (ChatGPT) and hybrid feedback in improving L2 writing skills, providing a more comprehensive perspective for language education. By combining the immediate feedback of Generative AI with the personalized guidance of human instructors, this study will explore the hybrid feedback model's potential to enhance writing quality, motivation, and critical thinking and organization in writing, offering new practical guidance for language education.

3.1. Conceptual framework

This study is based on Hattie and Timperley's (2007) Feedback Model and Self-Determination Theory (SDT) (Deci and Ryan 2012). Hattie and Timperley's model highlights the role of task-level, process-level, and self-regulation feedback in learning, while SDT focuses on how feedback supports intrinsic motivation and psychological needs. This framework will explore how generative AI and hybrid feedback, through different feedback levels and psychological need fulfillment, enhance second language academic writing skills, particularly through the synergy of AI and human feedback.

3.2. Hattie and timperley feedback model

Hattie and Timperley (2007) define feedback as information that compares an individual's current performance with a desired standard, offering guidance to both students and teachers. Their model categorizes feedback into three functions – feed up, feedback, and feed forward – and four levels: task level, process level, self-regulation level, and self level. Feed up focuses on clarifying learning goals and setting clear objectives, which provides direction for improvement. Feedback assesses the student's current performance, highlighting strengths and areas needing enhancement. Feed forward offers actionable steps to help students achieve their learning goals, such as improving writing style or argument structure (Mandouit and Hattie 2023).

The Hattie and Timperley Feedback Model is chosen over other frameworks, such as Black & Wiliam's formative assessment model, due to its comprehensive and structured approach to feedback. While Black & Wiliam's model emphasizes the broader principles of formative assessment (Black and Wiliam 1998), Hattie and Timperley provide the classification system that focuses on future-oriented guidance (Hattie and Timperley 2007), particularly valuable in analyzing feedback in the context of generative AI and hybrid feedback scenarios (see in Figure 1). In the first stage, AI provides task-level feedback, offering prompt corrections that help students focus on higher-order writing skills. This task-level feedback from AI tools can be easily quantified and analyzed for immediate error correction. In the second stage, human teachers provide process-level feedback, refining writing strategies and addressing more complex writing issues, which deepens students'

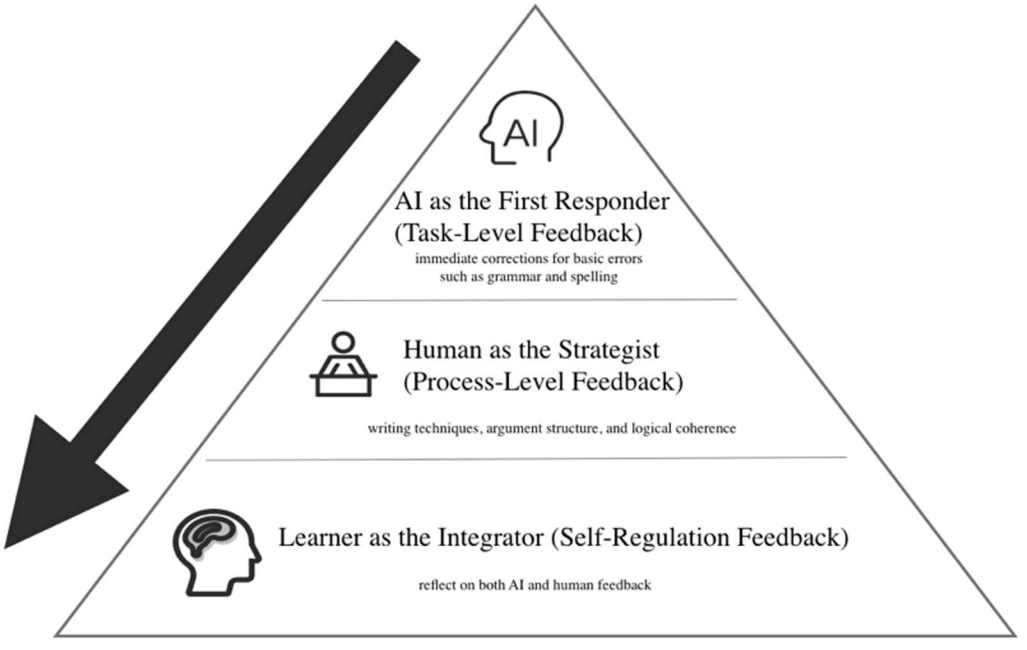

Figure 1. A Staged Model for Hybrid Feedback. A triangular diagram illustrating a staged model for hybrid feedback. The sections are labeled: 'AI as the First Responder,' 'Human as the Strategist,' and 'Learner as the Integrator.' Icons represent each stage of the feedback process.

understanding (Gombert et al. 2024). Furthermore, this process-level feedback can be tailored to the individual, supporting long-term writing development. In the final stage, learners reflect on both AI and human feedback, enhancing autonomy, metacognitive skills, and confidence, particularly in complex tasks. The inclusion of self-regulation feedback in this stage fosters learner autonomy and promotes long-term writing growth.

3.3. Self-Determination theory (SDT) and feedback

Self-Determination Theory (SDT), introduced by Deci and Ryan (2012), explores motivation, particularly the interaction between intrinsic and extrinsic motivators. SDT identifies three core psychological needs – autonomy, competence, and relatedness – that are essential for fostering intrinsic motivation (Deci and Ryan 2012). These needs are especially relevant in educational settings, where supporting students' psychological well-being is crucial for long-term engagement and success.

SDT provides a more suitable framework than Goal Setting Theory (GST) for understanding and promoting long-term motivation and psychological well-being in second language (L2) writing, particularly in the context of hybrid feedback. SDT emphasizes intrinsic motivation, which is vital for sustained engagement and development in L2 writing. Intrinsic motivation cultivates a genuine interest in learning, which is fundamental for long-term progress in language acquisition. In contrast, GST focuses primarily on extrinsic, goal-specific outcomes, which may improve short-term performance but often overlook the deeper psychological needs that drive long-term learning (Jansen et al. 2024).

In the context of L2 writing, SDT effectively explains how hybrid feedback better supports learners' psychological needs and writing development compared to AI-only feedback. Human teachers promote autonomy by providing delayed, personalized feedback that encourages students to reflect on their writing and make independent decisions regarding their improvements. This scaffolded feedback provided by human instructors is particularly critical in L2 writing. The collaborative

nature of hybrid feedback also strengthens relatedness by fostering a supportive learning environment where students feel connected to both AI and human instructors. Motivational feedback from human instructors can help reduce writing anxiety (Cen and Zheng 2024) and promote intrinsic interest through encouragement and reflective questioning. For example, human teachers can frame feedback in a way that highlights growth and effort rather than focusing solely on errors. While pure AI feedback is consistent and objective, it lacks the emotional and interpersonal elements that human feedback provides. Hybrid feedback enhances competence by offering detailed, multidimensional feedback on students' writing, particularly in foundational writing skills. In L2 writing, where learners often struggle to master both language and composition skills, this combination of AI and human feedback ensures a balanced approach to skill development (Mohsen 2022).

4. Methodology

4.1. Participants

This study involved 60 Chinese EFL students (23 males and 37 females), with an average age of 23.8 years. The participants were either undergraduates or graduates from universities in China or the United States, all with TOEFL scores above 90 and IELTS scores over 7, indicating intermediate English proficiency. All participants were native Chinese speakers and had been learning English for over ten years. Participants were recruited from language training institutions focused on Graduate Record Examinations (GRE) preparation. Only those actively preparing for the GRE and fully available for training were selected. Screening criteria included motivation to improve GRE writing scores, academic goals, GPA, and major field of study: humanities and social sciences (22), engineering (17), science (13), and medicine (8). Students without specific goals for improving GRE writing did not qualify. The average GPA was 3.51 (SD = 0.24).

The human tutor provided feedback based on GRE writing standards and held master's degrees in TESOL or English literature. The selection of tutors also considered experience, proficiency in online teaching, and technology readiness. All feedback during the course was provided by the human tutor.

4.2. Research design

This study employed an experimental method to investigate the effects of different feedback mechanisms on participants' writing performance. Figure 2 provides a flowchart of the experimental design.

Participants were divided into two groups based on their initial writing performance: Group 1 (N = 30) and Group 2 (N = 30), ensuring similar pre-writing scores. Group 1 received only GenAI feedback (mean score 2.88, SD = 0.79), while Group 2 received both GenAI and human feedback (mean score 2.87, SD = 0.87). All participants took a GRE training course in August 2023, which covered verbal reasoning, writing tasks, and quantitative reasoning, including ten 90-minute writing sessions. Writing practice occurred biweekly for a total of six exercises. Group 1 received immediate feedback from GenAI after each submission. Group 2 first received GenAI feedback, followed by additional human tutor feedback within three days.

To ensure consistency, accuracy, and relevance to GRE writing, a customized feedback system was developed using Coze, integrating GPT-4 Turbo (https://www.coze.com/store/agent/7473732560649404421?bot_id = true). The system delivered structured feedback through three components: Holistic Essay Assessment, Detailed Sentence-by-Sentence Review, and Encouraging Feedback. It evaluated essays across dimensions such as critical thinking, organization, development, sentence variety, grammar, and tone, while providing specific suggestions for improvement at the sentence level, including examples and supporting evidence. A GRE essay knowledge base was incorporated to align feedback with GRE standards and address EFL students' needs. Constraints were also applied to restrict feedback to GRE writing topics, ensuring focus and applicability.

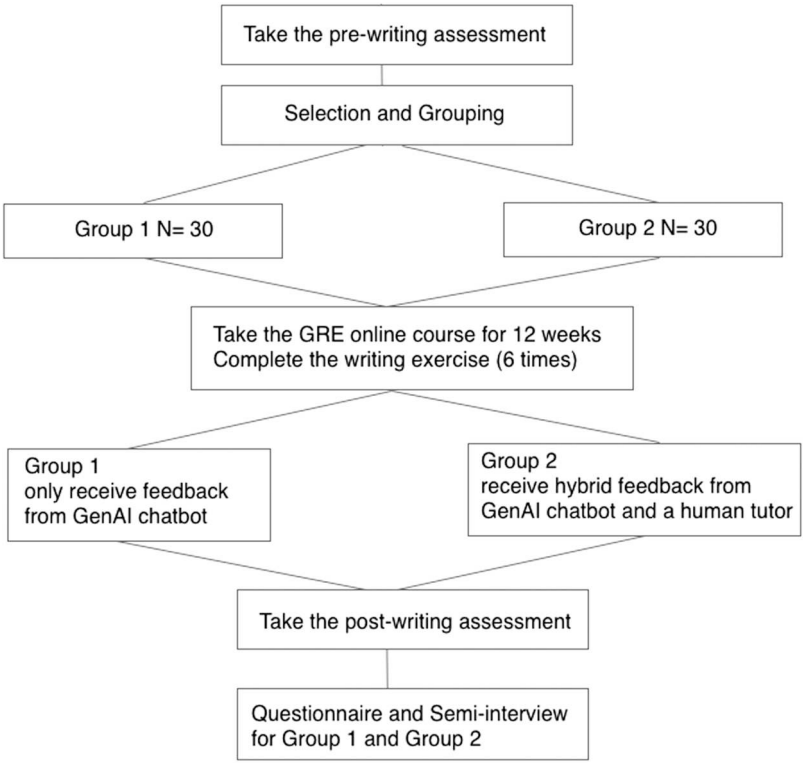

Figure 2. Flowchart of our study, showing how to compare two groups (Group 1 and Group 2) in a 12-week GRE online course.

In Group 2, participants submitted their writing tasks in Word format, which were processed through the system by researchers. The generated feedback was compiled and forwarded to human tutors. As shown in Figure 3, the hybrid feedback process began with GenAI-generated feedback, which was then reviewed and enhanced by a tutor. The tutor identified specific issues, particularly in critical thinking and structural coherence, and provided targeted strategies and personalized explanations tailored to each student's proficiency level. Additionally, the tutor offered comprehensive guidance, including plans for future tasks, time management strategies, and exam preparation tips. For example, a tutor might comment, 'This essay expresses your ideas well, but the prompt asks for a compelling reason that challenges your position. Include a counterargument.' This combined approach ensured participants received both automated and human-driven insights, enhancing the depth and applicability of the feedback.

After completing the GRE course, all participants underwent a final writing assessment, evaluated by two GRE instructors. A third instructor was consulted if score discrepancies exceeded one point. The assessment was conducted as a timed task, requiring participants to complete an argumentative essay in 60 minutes. To ensure authenticity, essays were submitted through a secure online platform that disabled copy–paste functions and were checked for plagiarism using Turnitin.

4.3. Data collection

Instrument 1: Writing Assessment

The primary data collection tool was a writing assessment simulating the GRE Issue Writing Task. Prompts were adapted from course materials, and scoring used GRE rubrics alongside the 6-Traits Writing Model (Culham 2003). Each of the seven categories – Critical Thinking, Organization,

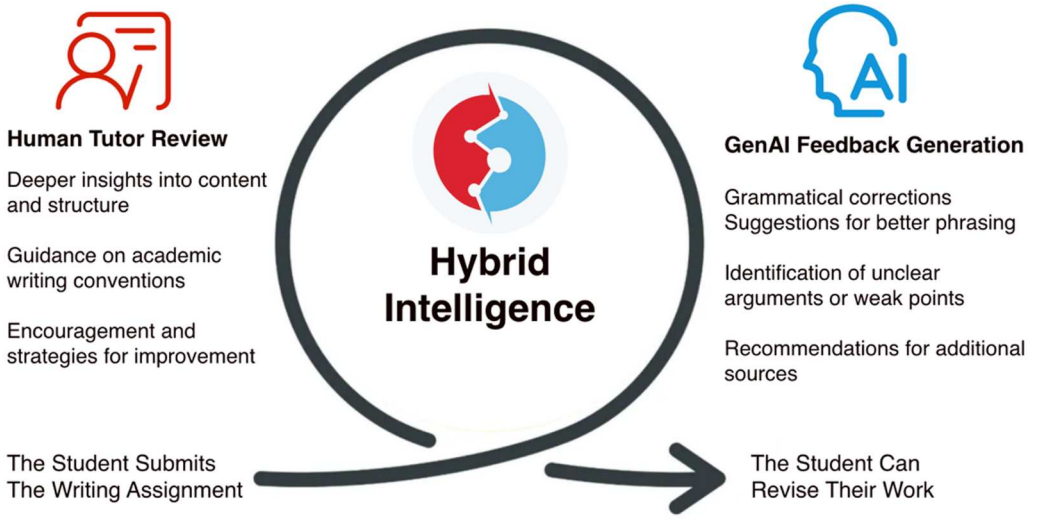

Figure 3. Diagram illustrating the hybrid feedback process, combining human tutor review and GenAI feedback generation.

Development, Sentence Variety, Grammar, Voice, and Tone – was scored on a 0–6 scale, with increments of 0.5 (see in appendix A).

Instrument 2: Questionnaires

The questionnaire (see Appendix B) was designed to gather participants' feedback experiences and motivations. It used a 1–5 Likert scale and was based on Hattie and Timperley's feedback model, as well as the self-determination theory (Deci and Ryan 2012; Hattie and Timperley 2007). It covers four levels of feedback. Task-level feedback assesses the clarity of feedback on specific tasks (e.g. 'How clear was the feedback you received on the specific tasks such as grammar corrections?'). Process-level feedback evaluates whether the feedback provided effective strategies to improve the writing process, (e.g. Did the feedback help you structure your arguments?). Self-regulation feedback evaluates whether the feedback help in regulating their learning independently. (e.g. Did the feedback give you enough confidence to make substantial changes independently?) Self-level feedback is often in the form of praise (e.g. How motivating was the personal praise you received in the feedback?) . Additionally, the questionnaire examines participants' motivations through questions on intrinsic motivation, such as 'Has your interest in writing increased after receiving feedback?'; extrinsic motivation, such as 'Does the feedback make you feel that writing is merely about fulfilling course requirements?'; and amotivation, such as 'Do you feel that writing is meaningless even after receiving feedback?'.

Instrument 3: Semi-structured Interviews

Two researchers conducted semi-structured interviews with ten students to gain deeper insights into their perceptions of and motivations related to writing feedback. The interviews explored how students improved their writing after feedback, and which feedback types were most beneficial, offering a comprehensive understanding of students' perspectives.

4.4. Data analysis

Data were analyzed using SPSS version 27. Descriptive statistics summarized the mean, standard deviation, and frequency. For each group, the Related-Samples Wilcoxon Signed Rank Test evaluated within-group differences in pre – and post-writing scores. The Mann–Whitney U test compared performance differences between groups, analyzing writing scores across skills and questionnaire responses (questionnaire reliability analysis, Cronbach's Alpha = 0.97). Effect sizes were calculated

using Cohen's d, with Hedges' g applied for adjustments due to the small sample size. Qualitative data from interviews and feedback content underwent thematic analysis to identify recurring themes and patterns in students' feedback experiences.

5. Results

5.1. Pre- and post-Writing assessments scores for two groups

After the writing practice, both groups showed improvements in their post-writing scores. The overall mean post-writing score increased to 3.76 (SD = 1.01), with a range of 2.0–6.0 (see Figure 4). Group 1's improvement from pre- to post-test was modest (Δmean = 0.38), whereas Group 2 exhibited a larger improvement (Δmean = 1.38). Most scores for Group 1 ranged between 2.375 and 3.5, while Group 2's scores were concentrated between 3.0 and 4.5. Overall, Group 2 demonstrated better outcomes compared to Group 1, with a higher mean score of 4.25 (SD = 0.95) compared to 3.26 (SD = 0.82).

To explore changes in writing skills within the experimental groups, the Wilcoxon signed-rank test was applied due to non-normal distribution. Table 1 shows that both Groups 1 and 2 achieved statistically significant improvements ($p < 0.001$) with standardized test statistics greater than 3, indicating strong effects.

Figure 5 shows the pre- and post-test changes across various dimensions of writing skills for both groups. For Group 1 (Figure 5a), the highest post-assessment score was in 'Voice and Tone' (mean = 3.93), followed by 'Grammar and Mechanics' (mean = 3.70), while 'Critical Thinking and Analysis' had the lowest score (mean = 2.44). Notably, 'Sentence Variety and Style' showed the largest mean improvement (0.52), whereas 'Critical Thinking and Analysis' had the smallest increase (0.25). Despite these improvements, the highest and lowest scoring skill areas remained consistent from pre- to post-assessment. This indicates that pure GenAI feedback, while showing some efficacy in specific aspects, falls short in 'Critical Thinking and Analysis.'

Figure 5b presents the results for Group 2, where 'Voice and Tone' achieved the highest post-assessment score (mean = 4.96), followed by 'Sentence Variety and Style' (mean = 4.60). Consistent with Group 1, 'Critical Thinking and Analysis' was the lowest scoring category (mean = 3.61). The greatest improvement for Group 2 was seen in 'Organization' (mean increase = 1.63) and Critical Thinking and Analysis (mean increase = 1.5), while 'Voice and Tone' showed more modest gains with 1.14, Overall, the performance of various functions in Group 2 showed a tendency towards balance in the post-test.

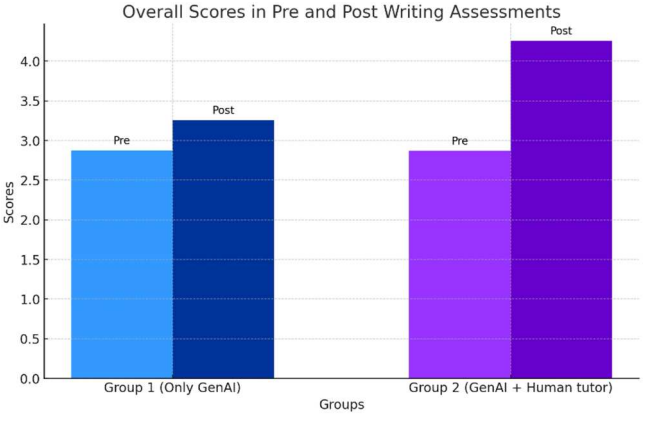

Figure 4. Bar chart comparing the overall scores in pre- and post-writing assessments for two groups. Students learning with GenAI and human tutor had better writing performance than those only learning with GenAI

Table 1. Related-samples wilcoxon signed rank test.

Pair	Test Statistic	Standard Error	Standardized Test Statistic	Asymptotic Sig. (2-sided test)
Group 1 Pre VS Group 1 Post	465.00	48.59	4.79	0.088
Group 2 Pre VS Group 2 Post	465.00	48.53	4.79	< 0.001

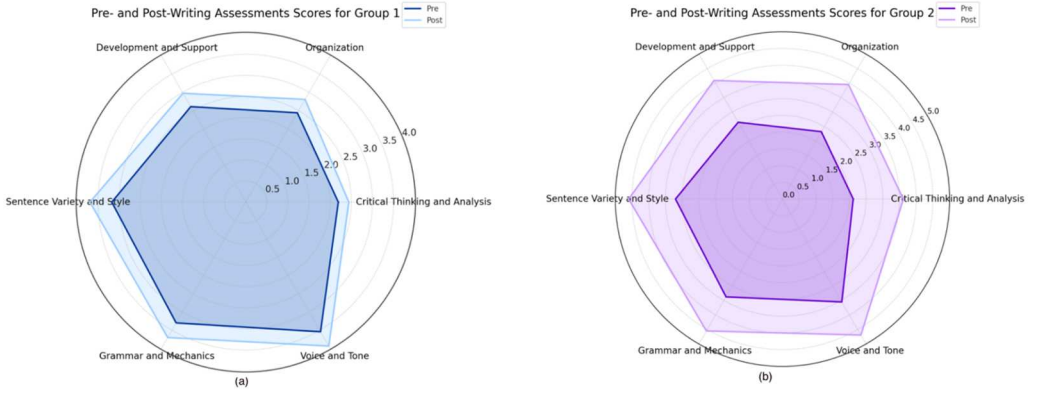

Figure 5. Two radar charts comparing pre- and post-writing assessment scores for Group 1 and Group 2 across five categories: Sentence Variety and Style, Grammar and Mechanics, Development and Support, Critical Thinking and Analysis, and Spelling and Tone. Group 2 improved all the assessments scores than Group 1. Learning with GenAI and human tutor has more benefits than learning with GenAI only.

Effect size analysis, as shown in Table 2, highlighted differences in the impact of interventions across groups. Group 1 demonstrated a moderate overall effect size (Hedge's $g = 0.45$), with the largest effect observed in 'Sentence Variety and Style' (Hedge's $g = 0.64$). In contrast, the intervention had a minimal impact on 'Critical Thinking and Analysis,' which showed the smallest effect size (Hedge's $g = 0.25$). Group 2, on the other hand, displayed a higher overall effect size (Hedge's $g = 1.45$), indicating more substantial improvements. All categories exhibited large effect sizes above 0.8, indicating that the hybrid feedback intervention led to a great improvement in various writing skills. The largest effect size was observed in 'Organization' (Hedge's $g = 1.62$), while 'Critical Thinking and Analysis' and 'Sentence Variety and Style' also showed strong improvements.

5.2. Comparison of writing scores between groups

To evaluate the effect of different feedback types on student performance, the Mann–Whitney U test was used to compare the two groups, as summarized in Table 3. Statistically significant differences were found between Group 2 and Group 1 ($p < 0.001$). Furthermore, a detailed analysis of specific writing skills showed statistically significant differences across all skill areas between the two groups, particularly in 'Critical Thinking and Analysis,' 'Organization,' and 'Development and Support' ($p < 0.001$).

To further quantify the differences, Hedge's g effect size was calculated for each comparison (Figure 6). Group 2 showed larger effect sizes in overall writing performance (g = 1.07 compared

Table 2. Hedge's g effect size for pre- and post-test results by group.

Groups	overall scores	Critical Thinking and Analysis	Organization	Development and Support	Sentence Variety and Style	Grammar and Mechanics	Voice and Tone
Group 1	0.45	0.25	0.42	0.39	0.64	0.52	0.52
Group 2	1.45	1.35	1.62	1.37	1.44	1.42	1.40

Table 3. Mann-Whitney U test results for different writing skills.

Statistic	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
G2 vs G1_Total Scores	189.00	654.00	−3.86	< 0.001
G2 vs G1_ Pair1	188.0	653.0	−3.887	< 0.001
G2 vs G1_ Pair2	173.0	638.0	−4.117	< 0.001
G2 vs G1_ Pair3	177.0	642.0	−4.055	< 0.001
G2 vs G1_ Pair4	216.5	681.5	−3.471	0.001
G2 vs G1_ Pair5	217.5	682.5	−3.451	0.001
G2 vs G1_ Pair6	217.0	682.0	−3.462	0.001

Note: Pairs 1–6 represent Critical Thinking and Analysis, Organization, Development and Support, Sentence Variety and Style, Grammar and Mechanics, and Voice and Tone respectively. Additionally, CG refers to the Control Group, G1 to Group 1, and G2 to Group 2.

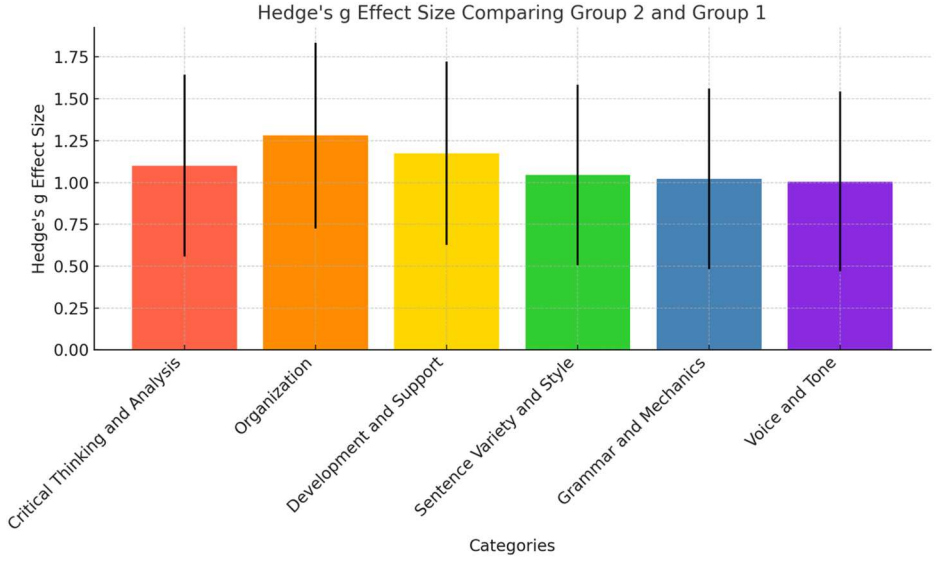

Figure 6. Bar chart showing Hedge's g effect size comparing Group 2 (GenAI + human tutor feedback) and Group 1 (GenAI-only feedback) across different writing assessment categories. Learning with GenAI and human tutor is more effective than learning only with Group 1 in Organization, Development and Support, and 'Critical thinking and analysis'.

to Group 1), indicating a clear benefit of the combined feedback approach. Additionally, Group 2's hybrid feedback showed significant improvements over Group 1, particularly in 'Organization' (g = 1.21), 'Development and Support' (g = 1.12), and 'Critical Thinking and Analysis' (g = 1.05). These results suggest that hybrid feedback provides a notable advantage over pure GenAI feedback in these writing skills.

5.3. Questionnaire analysis: perceptions and motivation

The questionnaire data show that, for Group 1, the highest score was for the 'Task-Level Feedback' category (mean = 3.68, SD = 0.98), while the lowest score was for the 'Self-Level Feedback' category (mean = 1.57, SD = 0.70). A similar pattern was observed in Group 2, where 'Task-Level Feedback' had the highest score (mean = 3.72, SD = 0.90) and 'Self-Level Feedback' received the lowest score (mean = 2.25, SD = 0.94). Additionally, the Mann–Whitney U Test (Table 4) was used to compare Group 1 and Group 2 with respect to different types of feedback. The results indicated no significant difference between the two groups for 'Task-Level Feedback' ($p > 0.05$). However, significant differences were found between the groups for 'Process-Level Feedback,' 'Self-Regulation Feedback,' and 'Self-Level Feedback' ($p < 0.05$).

Table 4. Mann-Whitney U test results for questionnaire.

Statistic	Task-Level Feedback	Process-Level Feedback	Self-Regulation Feedback	Self-Level Feedback
Mann-Whitney U	433.0	230.0	274.0	236.0
Wilcoxon W	898.0	695.0	739.0	701.0
Z	−0.030	−3.31	−2.68	−3.27
Asymp. Sig. (2-tailed)	0.98	0.00	0.030	0.005

Group 2 consistently scored higher on all motivation measures compared to Group 1. Specifically, extrinsic motivation was greater in Group 2 (mean = 3.37, SD = 1.22) than in Group 1 (mean = 3.09, SD = 1.30). Similarly, intrinsic motivation was higher for Group 2 (mean = 3.10, SD = 0.99) compared to Group 1 (mean = 2.74, sd = 0.95). Scores for amotivation were similar between the groups, with Group 1 at a mean of 1.68 (SD = 0.60) and Group 2 at 1.57 (SD = 0.58).

5.4. Qualitative analysis

5.4.1. Feed up feedback

In the Feed Up dimension, human tutors tend to offer more frequent and broad-based advice on writing objectives and task selection. They often focus on helping students understand how to set achievable goals, providing guidance on the broader scope of their writing tasks. In contrast, GenAI feedback is more focused on aligning writing goals with academic standards, emphasizing how students can adjust their objectives to meet scholarly benchmarks.

Group 1 indicated that while they appreciated the clarity of GenAI's feedback, they found it somewhat overly academic. Participant 1: 'GenAI's feedback always reminded me of how to meet academic standards, but it didn't address the challenges I might encounter while writing or offer strategies to overcome them.' Group 2 expressed greater satisfaction with the feedback that combined both academic standards and broader goal-setting. Participant 2: 'Hybrid feedback not only helped me understand academic standards but also guided me on how to adjust my writing goals, providing a more holistic understanding of my writing process.'

5.4.2. Feedback

In the Feed Back dimension, human tutors focus on structural elements, grammar, and logical flow, providing specific, targeted corrections. Tutors often provide suggestions for improving the clarity of arguments and the overall structure of the text. On the other hand, GenAI emphasizes linguistic aspects, such as vocabulary choice, formality, and variety, offering more systematic suggestions related to word choices and tone.

Group 1 found GenAI's feedback particularly useful for grammar and vocabulary enhancement, but they noted its lack of focus on logical flow. Participant 3 'GenAI provides many vocabulary replacement suggestions, but there is little attention to the logical structure of my essay.' Group 2 appreciated the hybrid feedback for its balance between structural and linguistic improvements. Participant 4, 'My tutor gave me advice on the structure, while GenAI helped with word choice and grammar. The combination of the two was very beneficial for improving my writing.'

5.4.3. Feed forward feedback

In the Feed Forward dimension, human tutors provide more individualized suggestions, such as expanding arguments and incorporating specific examples, which are tailored to the student's needs and writing context. By contrast, GenAI feedback is less personalized, focusing primarily on standardized improvements such as enhancing formality and increasing sentence variety, often lacking direct engagement with the content of the student's arguments.

Group 1 felt that GenAI's feedback, while helpful in improving language formality, lacked personalized suggestions that could have deepened their arguments. Participant 5, 'GenAI gave me advice

on how to make my writing more formal and varied, but it didn't address the specific arguments I was trying to make.' Group 2, however, found the hybrid feedback more enriching, as it combined personalized advice from the tutor with linguistic adjustments from GenAI. Participant 6, 'My tutor encouraged me to expand on my ideas and provided specific examples, while GenAI helped me refine the wording and improve the sentence structure.'

5.4.4. Self-Regulation feedback

In the Self-Regulation dimension, human tutors provide a moderate level of emotional support and motivational feedback, encouraging students to reflect on their writing process. Tutors often include affirmations, helping students build confidence in their abilities. In contrast, GenAI feedback is primarily procedural, focusing on writing mechanics such as paragraph analysis and self-editing, but it lacks the emotional support that is present in human tutor feedback.

Group 1 expressed that GenAI's feedback was helpful in terms of technical writing aspects but lacked motivational elements. Participant 7, 'GenAI helped me with language issues, but it didn't encourage me to reflect on my writing or build my confidence.' Group 2 found that hybrid feedback was more holistic, addressing both writing mechanics and emotional support. Participant 9, 'My tutor not only gave me advice on the technical aspects of my writing but also encouraged me to stay confident and focused, which was really helpful during the stressful exam period.'

6. Finding and discussion

The findings aim to provide insights into the strengths and limitations of GenAI feedback, the added value of human input in hybrid feedback, and the implications for L2 writing instruction.

6.1. Finding 1: the effectiveness of GenAI feedback on L2 writing skills

The study revealed that GenAI feedback significantly improved L2 learners' writing skills, particularly in low-order aspects such as sentence variety, style, grammar, and vocabulary accuracy. However, its impact on higher-order writing skills, such as critical thinking and organizational structure, was limited. While the effect size of GenAI on L2 writing was lower than the overall effect size of AWE ($g = 0.67$) as reported in meta-analyses, it exceeded the effect size of surface-level feedback (Scherer, Graham, and Busse 2024). Additionally, when compared to other AWE tools, GenAI's effect size was higher than that of Pigai ($g = 0.09$) and Criterion ($g = 0.34$) but lower than that of Grammarly ($g = 1.04$; Ngo, Chen, and Lai 2024).

6.2. Finding 2: superiority of hybrid feedback over GenAI feedback

Hybrid feedback demonstrated greater effectiveness than GenAI feedback alone, particularly in enhancing L2 learners' ability to address complex writing issues. Moreover, the effect size of hybrid feedback significantly surpassed that of human tutor-only feedback ($g = 0.27$; Cen and Zheng 2024), peer tutoring ($g = 0.84$), and computer-mediated corrective feedback ($g = 1.21$; Mohsen 2022). This study highlights the superior impact of hybrid feedback over other feedback modalities.

6.3. Finding 3: motivation and feedback perception in the hybrid feedback group

The Hybrid Feedback group (Group 2) demonstrated significantly higher levels of both intrinsic and extrinsic motivation compared to the GenAI Feedback group (Group 1). Additionally, Group 2 reported significantly higher levels of task-level, process-level, self-regulation, and self-level feedback perception, particularly excelling in process-level and self-regulation feedback compared to Group 1.

This study validates previous research (e.g. Liu et al. 2024; Wang 2024) on the significant role of AI, such as ChatGPT, in enhancing students' writing quality, particularly in grammatical and structural revisions. Unlike prior studies that focused on the integration of GenAI into specific writing tasks (e.g. Liu et al. 2024, integrating GenAI into Cognitive Academic Language Learning Model on EFL writing instruction; Mahapatra, 2024, on GenAI as a writing assistance tool), this study expands the scope by examining the utility of AI-generated feedback in improving writing outcomes. While earlier research primarily explored AI applications in text continuation (e.g. continuation writing in Wang 2024; creative writing in Chowdhury, 2024; ENL essay evaluations in Escalante, Pack, and Barrett 2023), this study extends the application of GenAI to L2 academic writing, specifically for argumentative writing. However, the limited effect of GenAI contrasts with the work of Borge et al. (2024) and Lee & Low (2024), who found that interactive chat-based GenAI models, like OpenAI, support higher-order thinking when used in a more interactive manner. In this study, GenAI was used to provide feedback rather than directly interact with learners, which may explain the difference.

The findings confirm that hybrid feedback, which combines human and AI-generated feedback, is more effective than AI-generated feedback alone (Escalante, Pack, and Barrett 2023; Li, Zhou, and Chiu 2024; Meyer et al., 2024). This contrasts with earlier research (Escalante, Pack, and Barrett 2023), which compared automated feedback with teacher feedback and found no significant difference between ChatGPT-generated feedback and human tutor feedback. Additionally, this study diverges from research such as Meyer et al. (2024) and Liu et al. (2024), which highlighted the limitations of AI-generated feedback, indirectly underscoring the efficacy of combining AI with human feedback. Beyond analyzing the effects of AI and hybrid feedback, this study extends the evaluation by quantifying learners' perceptions of feedback and their writing motivation. Unlike Kao and Reynolds (2024), which examined perceived feedback quality, this study emphasizes the type of feedback, thereby expanding on previous qualitative research on students' motivation regarding AI.

The effectiveness of GenAI feedback in improving lower-order writing skills can be explained by the Hattie and Timperley's (2007) feedback model. Task-level feedback, which focuses on the correctness or accuracy of specific tasks, is particularly effective in helping students improve basic skills, such as spelling. The feedback perceptions of Group 1 students predominantly focused on task-level feedback, reflecting GPT's ability to identify grammatical errors and provide specific improvement suggestions (Escalante, Pack, and Barrett 2023; Kao and Reynolds 2024; Meyer et al., 2024; Steiss et al. 2024). This is further supported by students' high level of trust in AWE tools, particularly in areas such as grammar and vocabulary selection.

According to Hattie and Timperley's model, feedback is most effective when it addresses three key levels: feedup (clarifying learning goals), feedback (providing specific guidance for improvement), and feedforward (helping students plan for future learning). Hybrid feedback excels in addressing all three levels, making it more effective than AI-only feedback, which often lacks the depth and contextual understanding required to support complex learning needs. While AI feedback can incorporate new perspectives into revisions, it frequently suffers from inconsistencies in quality. For instance, ChatGPT performs well in providing feedback on lower-quality writing, but its accuracy diminishes with higher-quality texts (Steiss et al. 2024). The quality of AI feedback can fluctuate depending on the student's text and the context in which the feedback is generated, potentially resulting in feedback that does not fully address the student's needs (Meyer et al., 2024). In contrast, human feedback can address issues that AI feedback may overlook, offering more targeted and nuanced suggestions for improvement (Weber et al., 2024).

Hybrid feedback provides process-level guidance that not only informs students about what to improve but also explains why and how to do so. AI feedback, on the other hand, tends to be more generalized and may fail to address specific student needs, such as clarifying complex causal relationships or ensuring overall coherence in their work (Li, Zhou, and Chiu 2024). Human tutors are better equipped to guide students in organizing essay structure and logical flow, thereby helping them produce more coherent and well-structured writing (Li, Zhou, and Chiu

2024). However, students sometimes find GenAI feedback too indirect or difficult to understand, leading to reluctance in accepting its suggestions (Liu et al. 2024). Additionally, students using AI feedback must verify the accuracy of the information, especially when dealing with unfamiliar topics (Niloy et al. 2024). Effective use of AI feedback requires a certain level of AI literacy and an understanding of its limitations (Niloy et al. 2024). Furthermore, a single instance GenAI feedback has limited long-term impact on students' writing proficiency, and multiple feedback iterations may be necessary to achieve more significant learning outcomes (Meyer et al., 2024).

Self-regulation feedback is essential for fostering critical thinking, metacognitive skills, and organizational abilities (Hattie and Timperley 2007). Such feedback goes beyond identifying problems; it encourages students to reflect on their learning process through questions or prompts, thereby promoting self-regulatory skills. Human tutors can guide students toward independent thinking through interaction and discussion, reducing over-reliance on technological tools and enhancing students' autonomous writing abilities (Liu et al. 2024; Song and Song 2023).

Emotional support is another critical advantage of hybrid feedback, as it positively influences students' writing motivation. According to SDT, the personalized guidance provided by human tutors enhances students' sense of autonomy and competence, thereby strengthening their intrinsic motivation (Chiu et al. 2024). While AI feedback can contribute to positive emotions and motivation, its impact varies depending on individual student differences (Meyer et al., 2024; Yang & Li, 2024). Unlike human tutors, AI feedback cannot provide the same level of emotional support and encouragement, which are crucial for sustaining student motivation and engagement throughout the learning process. Consequently, AI feedback alone is insufficient for helping students build confidence in their writing.

6.4. Theoretical and practical implications

Hattie and Timperley's (2007) feedback model emphasizes the role of feedback at different levels in enhancing learning. However, it does not explicitly address the source of feedback (human vs. AI) or the integration of multiple sources. To address this gap, an updated model could introduce a new dimension that distinguishes between feedback sources, such as AI-only, human-only, and hybrid feedback. Additionally, the model could incorporate a skill-level dimension, differentiating between feedback for lower-order skills (e.g. task-level feedback, such as spelling corrections) and higher-order skills (e.g. process-level and self-regulation feedback, such as critical thinking and coherence).

Furthermore, Hattie and Timperley's model does not explicitly incorporate motivation as a mediating factor in feedback effectiveness. To address this, the model could be integrated with SDT. For instance, AI feedback could be categorized as more effective for task-level feedback, which supports competence, but less effective for process-level and self-regulation feedback, which are crucial for fostering autonomy and relatedness. This integration would provide a more comprehensive framework for understanding feedback's role in learning and motivation.

To maximize the benefits of hybrid feedback, instructors can use GenAI tools to provide immediate, automated feedback on lower-order skills, freeing up valuable time for educators to focus on delivering process-level feedback. This includes guiding students through critical stages of writing, such as brainstorming, drafting, and revising. Additionally, educators should actively teach students how to interpret and apply GenAI feedback effectively. This can be achieved by encouraging goal-setting, progress monitoring, and reflective practices, which foster self-regulation and a sense of ownership over the learning process. Scaffolding students' use of AI feedback – through guided practice or structured reflection exercises – can further enhance their ability to understand and act on automated recommendations (Chiu and Chai 2020).

To address potential challenges, such as the misinterpretation of AI-generated feedback, educators can organize peer review sessions where students collaboratively discuss and critique AI

feedback on each other's work. This not only improves feedback literacy but also promotes collaborative learning. Furthermore, providing targeted training and resources for both educators and students is essential to ensure a deeper understanding of how to effectively integrate GenAI tools into writing instruction. This approach helps address the diverse needs of L2 learners and creates a more progressive and inclusive learning experience. Furthermore, to quantify the impact of these strategies, educators can implement pre- and post-intervention assessments to measure improvements in feedback literacy, writing quality, and self-regulation skills. This data-driven approach ensures that the integration of GenAI tools and hybrid feedback models is both effective and aligned with students' developmental needs.

7. Conclusion

This study shows that while GenAI-generated feedback can enhance L2 academic writing by improving lower-order skills like grammar and sentence variety, its impact on higher-order skills such as critical thinking, organization, and argumentation is limited. The hybrid feedback model, combining GenAI with human input, is more effective, promoting academic proficiency and improving various writing skills by providing balanced feedback at multiple levels, including process-level, self-regulation, and motivational support.

This study highlights the value of integrating GenAI tools with traditional teaching methods for L2 writing instruction, addressing a critical gap in research on the long-term impact of AWE (Ngo, Chen, and Lai 2024). By providing empirical evidence for the sustained benefits of hybrid feedback, this study not only demonstrates its effectiveness in enhancing writing skills but also underscores its ability to address students' psychological needs. Grounded in SDT, the findings reveal that hybrid feedback better meets students' needs for autonomy, competence, and relatedness, supporting both cognitive development and emotional growth. Additionally, the study demonstrates the importance of training educators in the effective use of GenAI tools. Educators need to understand AI's capabilities and limitations to effectively integrate these tools into their teaching, addressing the diverse needs of L2 learners.

This study has several limitations. First, the quality of AI-generated feedback can vary significantly depending on prompt design and customization, which may influence the consistency and effectiveness of the feedback provided. Additionally, it remains unclear whether similar results would be obtained using other AI models (e.g. LLaMA), as this study focused exclusively on a single AI system. Finally, the effects of the intervention are short term, and the development of writing skills are developed over time. It is necessary to see if the effects are long term.

Future research should explore the performance and reliability of alternative AI models in generating feedback for argumentative writing. Additionally, this study primarily adopts the integrated hybrid feedback model. Future research directions could explore the effectiveness of various hybrid feedback models, such as parallel hybrid feedback (where both technology and human instructors provide feedback simultaneously) and peer-assisted hybrid feedback. Future studies could adopt a longitudinal design to investigate the long-term effects, for example, adding delay posttests.

Ethics approval and consent to participate

This study got ethical clearance from the first author's university and got the consents from the participants to collect their data.

Consent for publication

The participant has consented to the submission of the manuscript to journals and conferences.

Availability of data and materials

The datasets used for the current study are available from the corresponding author on reasonable request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Zhihui Zhang is a PhD student at Department of Curriculum and Instruction, The Chinese University of Hong Kong, and her research areas include AI and language education.

Professor Scott Aubrey is an associate professor at Department of Curriculum and Instruction, The Chinese University of Hong Kong and his main research areas include second language motivation, learner engagement, and task-based language teaching.

Xiaomeng Huang is an industrial researcher.

Professor Thomas K. F. Chiu is an assistant professor at Department of Curriculum and Instruction, The Chinese University of Hong Kong and his main research areas include AI and Education.

ORCID

Scott Aubrey http://orcid.org/0000-0003-4365-0516
Thomas K. F. Chiu http://orcid.org/0000-0003-2887-5477

References

Black, P., and D. Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5 (1): 7–74.

Cen, Y., and Y. Zheng. 2024. "The Motivational Aspect of Feedback: A Meta-Analysis on the Effect of Different Feedback Practices on L2 Learners' Writing Motivation." *Assessing Writing* 59:100802. https://doi.org/10.1016/j.asw.2023.100802.

Chiu, T. K. F. 2024. "A Classification Tool to Foster Self-Regulated Learning with ChatGPT by Applying Self-Determination Theory from a Teacher Perspective." *Educational Technology Research & Development* 72:2401–2416. https://doi.org/10.1007/s11423-024-10366-w.

Chiu, T. K. F., and C. S. Chai. 2020. "Sustainable Curriculum Planning for Artificial Intelligence Education: A Self-Determination Theory Perspective." *Sustainability* 12 (14): 5568. https://doi.org/10.3390/su12145568.

Chiu, T. K. F., B. L. Moorhouse, C. S. Chai, and M. Ismailov. 2024. "Teacher Support and Student Motivation to Learn with Artificial Intelligence (AI) Based Chatbot." *Interactive Learning Environments* 32 (7): 3240–3256. https://doi.org/10.1080/10494820.2023.2172044.

Chowdhury, A. G., M. M. Islam, V. Kumar, F. H. Shezan, V. Jain, and A. Chadha. 2024. Breaking Down The Defenses: A Comparative Survey of Attacks on Large Language Models. *arXiv preprint arXiv:2403.04786*.

Culham, R. 2003. *6+1 Traits of Writing: The Complete Guide*. New York, NY: Scholastic Inc.

Deci, E. L., and R. M. Ryan. 2012. "Self-determination Theory." *Handbook of Theories of Social Psychology* 1 (20): 416–436.

Ding, L., and D. Zou. 2024. "Automated Writing Evaluation Systems: A Systematic Review of Grammarly, Pigai, and Criterion with a Perspective on Future Directions in the age of Generative Artificial Intelligence." *Education and Information Technologies* 29 (11): 14151–14203. https://doi.org/10.1007/s10639-023-12402-3.

Escalante, J., A. Pack, and A. Barrett. 2023. "AI-generated Feedback on Writing: Insights Into Efficacy and ENL Student Preference." *International Journal of Educational Technology in Higher Education* 20: 57. https://doi.org/10.1186/s41239-023-00425-2.

Gombert, S., A. Fink, T. Giorgashvili, I. Jivet, D. Di Mitri, J. Yau, and H. Drachsler. 2024. "From the Automated Assessment of Student Essay Content to Highly Informative Feedback: A Case Study." *International Journal of Artificial Intelligence in Education* 34 (4): 1378–1416. https://doi.org/10.1007/s40593-023-00387-6.

Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112. https://doi.org/10.3102/003465430298487.

Huang, S., and W. A. Renandya. 2020. "Exploring the Integration of Automated Feedback among Lower-Proficiency EFL Learners." *Innovation in Language Learning and Teaching* 14 (1): 15–26. https://doi.org/10.1080/17501229.2018.1471083.

Jansen, T., J. Meyer, J. Fleckenstein, A. Horbach, S. Keller, and J. Möller. 2024. "Individualizing Goal-Setting Interventions Using Automated Writing Evaluation to Support Secondary School Students' Text Revisions." *Learning and Instruction* 89:101847. https://doi.org/10.1016/j.learninstruc.2023.101847.

Kao, C. W., and B. L. Reynolds. 2024. "Timed Second Language Writing Performance: Effects of Perceived Teacher vs Perceived Automated Feedback." *Humanities and Social Sciences Communications* 11:1012. https://doi.org/10.1057/s41599-024-03522-3.

Kohnke, L. 2024. "Exploring EAP Students' Perceptions of GenAI and Traditional Grammar-Checking Tools for lan Guage Learning." *Computers and Education: Artificial Intelligence* 7:100279. https://doi.org/10.1016/j.caeai.2024.100279.

Li, B., V. L. Lowell, C. Wang, and X. Li. 2024. "A Systematic Review of the First Year of Publications on ChatGPT and Language Education: Examining Research on ChatGPT's use in Language Learning and Teaching." *Computers and Education: Artificial Intelligence* 7:100266. https://doi.org/10.1016/j.caeai.2024.100266.

Li, Y., X. Zhou, and T. K. F. Chiu. 2024. "Systematics Review on Artificial Intelligence Chatbots and ChatGPT for Language Learning and Research from Self-Determination Theory (SDT): What are the Roles of Teachers?" *Interactive Learning Environments* 33 (3): 1850–1864. https://doi.org/10.1080/10494820.2024.2400090.

Link, S., M. Mehrzad, and M. Rahimi. 2022. "Impact of Automated Writing Evaluation on Teacher Feedback, Student Revision, and Writing Improvement." *Computer Assisted Language Learning* 35 (4): 605–634. https://doi.org/10.1080/09588221.2020.1743323.

Liu, Z. M., G. J. Hwang, C. Q. Chen, X. D. Chen, and X. D. Ye. 2024. "Integrating Large Language Models Into EFL Writing Instruction: Effects on Performance, Self-Regulated Learning Strategies, and Motivation." *Computer Assisted Language Learning* 36 (2): 187–209. https://doi.org/10.1017/S0958344023000265.

Luo, Y., and Y. Liu. 2017. "Comparison Between Peer Feedback and Automated Feedback in College English Writing: A Case Study." *Open Journal of Modern Linguistics* 7 (4): 197–215. https://doi.org/10.4236/ojml.2017.74015.

Mandouit, L., and J. Hattie. 2023. "Revisiting "The Power of Feedback" from the Perspective of the Learner." *Learning and Instruction* 84:101718. https://doi.org/10.1016/j.learninstruc.2022.101718.

McCarthy, K. S., R. D. Roscoe, L. K. Allen, A. D. Likens, and D. S. McNamara. 2022. "Automated Writing Evaluation: Does Spelling and Grammar Feedback Support High-Quality Writing and Revision?" *Assessing Writing* 52:100608. https://doi.org/10.1016/j.asw.2022.100608.

Mohsen, M. A. 2022. "Computer-mediated Corrective Feedback to Improve L2 Writing Skills: A Meta-Analysis." *Journal of Educational Computing Research* 60 (5): 1253–1276. https://doi.org/10.1177/07356331211064066.

Ngo, T. T. N., H. H. J. Chen, and K. K. W. Lai. 2024. "The Effectiveness of Automated Writing Evaluation in EFL/ESL Writing: A Three-Level Meta-Analysis." *Interactive Learning Environments* 32 (2): 727–744. https://doi.org/10.1080/10494820.2022.2096642.

Niloy, A. C., S. Akter, N. Sultana, J. Sultana, and S. I. U. Rahman. 2024. "Is ChatGPT a Menace for Creative Writing Ability?" *An Experiment. Journal of Computer Assisted Learning* 40 (2): 919–930. https://doi.org/10.1111/jcal.12929.

Nunes, A., C. Cordeiro, T. Limpo, and S. L. Castro. 2022. "Effectiveness of Automated Writing Evaluation Systems in School Settings: A Systematic Review of Studies from 2000 to 2020." *Journal of Computer Assisted Learning* 38 (2): 599–620. https://doi.org/10.1111/jcal.12635.

Ranalli, J. 2018. "Automated Written Corrective Feedback: How Well Can Students Make use of it?" *Computer Assisted Language Learning* 31 (7): 653–674. https://doi.org/10.1080/09588221.2018.1428994.

Saricaoglu, A. 2019. "The Impact of Automated Feedback on L2 Learners' Written Causal Explanations." *ReCALL* 31 (2): 189–203. https://doi.org/10.1017/S095834401800006X.

Scherer, S., S. Graham, and V. Busse. 2024. "How Effective is Feedback for L1, L2, and FL Learners' Writing? A Meta-Analysis." *Learning and Instruction* 93:101961. https://doi.org/10.1016/j.learninstruc.2024.101961.

Shi, H., and V. Aryadoust. 2024. "A Systematic Review of AI-Based Automated Written Feedback Research." *ReCALL* 36 (2): 187–209. https://doi.org/10.1017/S0958344023000265.

Song, C., and Y. Song. 2023. "Enhancing Academic Writing Skills and Motivation: Assessing the Efficacy of ChatGPT in AI-Assisted Language Learning for EFL Students." *Frontiers in Psychology* 14:1260843. https://doi.org/10.3389/fpsyg.2023.1260843.

Steiss, J., T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, and C. B. Olson. 2024. "Comparing the Quality of Human and ChatGPT Feedback of Students' Writing." *Learning and Instruction* 91:101894. https://doi.org/10.1016/j.learninstruc.2024.101894.

Tang, J., and C. S. Rich. 2017. "Automated Writing Evaluation in an EFL Setting: Lessons from China." *The JALT CALL Journal* 13 (2): 117–146. https://doi.org/10.29140/jaltcall.v13n2.j215.

Wang, J. 2024. "An Empirical Study on Continuation Writing in Senior High School Under the Assessment and Feedback of ChatGPT." *Journal of Theory and Practice of Contemporary Education* 4 (05): 7–22. https://doi.org/10.53469/jtpce.2024.04(05).02.