Education and Information Technologies (2025) 30:26777–26799
https://doi.org/10.1007/s10639-025-13775-3

ChatGPT as an automated writing evaluation tool: how students perceive it and how it affects their writing

Linqian Ding[1] · **Di Zou**[2] · **Lucas Kohnke**[1]

Received: 24 November 2024 / Accepted: 31 August 2025 / Published online: 13 October 2025

Abstract

Generative artificial intelligence (GAI) language models, exemplified by ChatGPT, are significantly helping language learners in writing practice by providing immediate formative feedback. This study investigates whether ChatGPT's writing feedback influences graduate students' academic writing abilities and compares it with combined instructor and ChatGPT feedback. The students receiving only ChatGPT feedback (Group A) showed significant improvements in mechanics, tone, grammar, APA formatting, and overall writing quality, with lesser gains in organisation and no significant change in content. In contrast, the students who received combined feedback (Group B) exhibited significant enhancements in all of the parameters that were evaluated, including the areas in which Group A lagged. However, the only significant between-group difference was in grammar, although there was a marginal difference in organisation that suggested a trend towards greater improvement in Group B. The participants expressed positive attitudes about using ChatGPT as an automated writing evaluation tool, especially for correcting grammar and enriching vocabulary.

Keywords Automated writing evaluation · ChatGPT · Writing performance · Student perceptions

✉ Di Zou
daisy.zou@polyu.edu.hk

Linqian Ding
dinglinqian19951228@gmail.com

Lucas Kohnke
lucaskohnke@gmail.com; lmakohnke@eduhk.hk

[1] Department of English Language Education, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong SAR, China

[2] Department of English and Communication, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

1 Introduction

Over the past few years, the advancement of generative artificial intelligence (GAI) has rapidly changed educational practices (Baskara, 2023). ChatGPT, a large language model (LLM) that can generate human-like text and provide real-time language support, has gained significant attention in the field of education (Kohnke et al., 2023). It has shown great potential as an automated writing evaluation (AWE) tool as it can understand personalised instructions and generate nuanced, conversational feedback on students' writing (Dai et al., 2023; Mizumoto & Eguchi, 2023).

AWE tools are computer-based systems that can assess and provide feedback on written texts with the help of natural language processing (NLP), latent semantic analysis, and artificial intelligence (AI; Bai & Hu, 2017; Warschauer & Ware, 2006). They have two main functions: scoring student writing and offering formative corrective feedback (Page, 2003; Shermis et al., 2013). Traditional AWE tools, such as Criterion, e-rater, and Grammarly, have been widely used in large English writing classes to provide students with immediate and personalised feedback (Li, 2021; Guo et al., 2022). However, these tools cannot address higher-order writing concerns such as argumentation, organisation, and content (Barrot, 2021). In addition, their feedback formats and scoring rubrics are limited (Gao, 2021). These drawbacks mean that they cannot be used alone as a replacement for teacher feedback (Hyland & Hyland, 2019). GAI tools, which possess advanced natural language understanding and generation capabilities, offer a more interactive and nuanced form of feedback. Thus, they are expected to overcome the limitations of traditional AWE tools (Ding & Zou, 2024).

To better understand GAI's potential as an AWE tool, it is essential to investigate whether it can improve student writing performance. How students perceive GAI tools is equally important information as learners play a crucial role in the integration of educational technologies (Zhang et al., 2023). This study employs ChatGPT as a representative GAI tool to evaluate the quality of its feedback on key aspects of academic writing, including grammar, mechanics, content, organisation, and APA formatting, and assesses student perceptions of ChatGPT as an AWE tool. It also compares the effectiveness of ChatGPT with ChatGPT plus teacher feedback in improving student writing. By addressing these dimensions, this research aims to contribute to the growing body of research on the pedagogical integration of GAI in writing instruction. It also aims to identify the benefits and limitations of using such tools to support student writing development.

2 Literature review

2.1 Theoretical framework

The present study is grounded in activity theory (Engeström, 1987). Activity theory is a conceptual framework that analyses human practice by identifying the components of a specific activity system (Engeström, 2001). It is widely applied in educational research, especially technology-enhanced learning, to examine how tools mediate

learning activities and how interactions among learners and other components shape educational outcomes (Barab et al., 2004; Blin, 2004). According to Engeström (1987), an activity system has six components: subject, object, tools, rules, community, and division of labour. Chung et al. (2019) adapted the six components to educational context. They defined (a) the subjects as the participants in learning activities, including teachers and students; (b) the objects as the learning goals, such as enhancing L2 writing performance; (c) the tools as the supporting instruments and materials, such as computers; (d) the community as the setting of the activity, such as a classroom; (e) the rules as the strategies and methods applied in learning, such as feedback mechanisms; and (f) the division of labour as the roles and responsibilities within the learning activities, such as the roles in feedback provision.

Activity theory is frequently used as a conceptual framework to analyse writing processes accompanied by AWE interventions. Barrot (2021) applied this theory to investigate the effects of AWE feedback on the writing accuracy of college-level L2 learners. Similarly, Rahimi et al. (2024) employed it to assess the impact of AWE feedback on the academic writing skills of learners of English as a foreign language (EFL). In the present study, we applied activity theory to describe the interactive elements of the use of ChatGPT for academic writing feedback (Fig. 1). The subjects were the students who participated in an intervention in which they received feedback from ChatGPT. The primary tool used in this activity was the ChatGPT platform, which was freely available to the students through their university system. The rules included the English class regulations and the prompts given to students to elicit feedback from ChatGPT. The community consisted of the students, their instructors, and the ChatGPT system interacting within the conceptual framework. The labour was divided between the students engaging with feedback and the instructors managing the overall process and providing supplementary feedback. Ideally, engagement

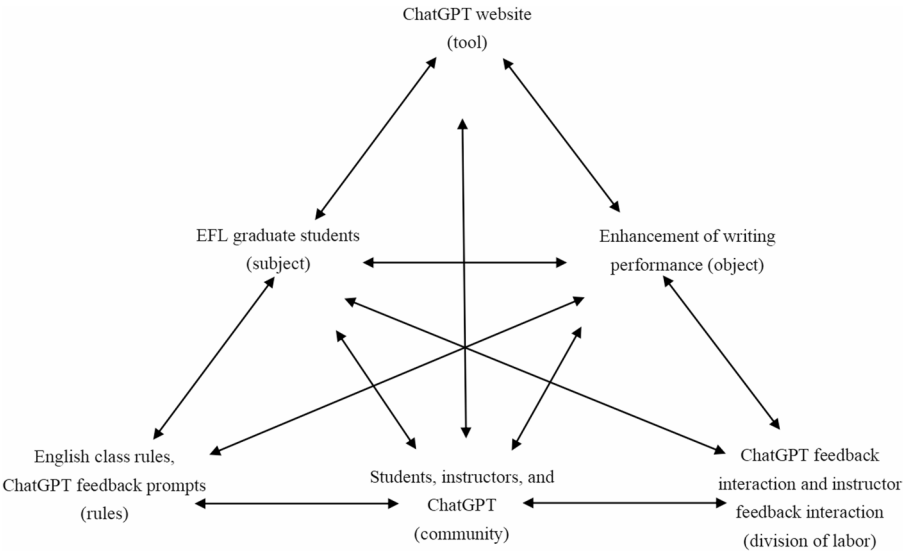

Fig. 1 Activity Theory Model for ChatGPT Feedback Intervention

in this ChatGPT-involved English writing activity would improve student writing performance.

2.2 The impact of AWE tools on writing quality

Numerous studies have found that AWE tools improve writing quality. Attali (2004) found that an AWE tool, Criterion, reduced writing errors by about 25 per cent in an analysis of over 9,000 essays from students in Grades 6 to 12. Likewise, Barrot (2021) observed that students who used Grammarly to obtain writing feedback, compared with those who did not, significantly improved their writing accuracy. Furthermore, EFL students who employed Grammarly's feedback to correct 85% of their mistakes significantly reduced their error scores (Guo et al., 2022). Shen et al. (2023) observed that the impact of AWE tools varied by proficiency level: less proficient learners improved more in grammatical accuracy, whereas more proficient learners made greater progress in lexical complexity.

Nevertheless, some studies have questioned the effectiveness of AWE tools. Huang and Renandya (2020) found that using Pigai did not significantly improve low-proficiency learners' overall writing performance. They attributed this to students' limited ability to detect or fully understand inaccurate feedback. Several studies have pointed out that AWE systems tend to focus on surface-level features, such as grammar and spelling, rather than deeper aspects of writing (Barrot, 2021; Gao, 2021). Han et al. (2021) found that Pigai improved students' overall writing performance but did not enhance their lexical diversity or sophistication. This was probably because the feedback emphasised lexical accuracy over complexity. Likewise, Xu and Zhang (2022) found that after 15 weeks of using Pigai, writing accuracy became more similar in learners across proficiency levels. Nevertheless, there was a slight improvement in higher-level skills, such as syntactic complexity and fluency.

2.3 Students' perceptions of AWE tools

Students tend to have mixed perceptions of AWE tools; most students recognise both the strengths and limitations of the tools. Yousofi (2022) reported that learners expressed favourable attitudes towards using Grammarly to revise their writing. They found that the tool improved writing quality, enhanced mechanical accuracy, saved time, and supported independent revision. However, some students said that its feedback was not always professional, and they criticised its inability to address content-related issues or meet the needs of lower-level learners. Gao (2021) found similar variations in student perceptions of Pigai. The students generally valued its word-level feedback but found it less effective in helping them improve syntactic complexity. Li et al. (2015) also reported that students were satisfied with AWE feedback on grammar and mechanics but considered its comments on organisation and rhetoric less effective than teacher feedback.

Some studies have compared students' perceptions of AWE tools with other sources of writing feedback, especially teacher feedback. Mohsen and Abdulaziz (2019) examined students' experiences with the AWE tool MyTutor and found that learners struggled to understand its feedback on content and organisation. They

found teacher feedback on content and organisation, received in addition to MyTutor feedback, clearer, more helpful, and more intelligible. Similarly, Thi and Nikolov (Thi, et al., 2022) found that students perceived that Grammarly feedback was useful for improving grammar and vocabulary but not for developing content or organisation. They considered combined AWE and teacher feedback more comprehensive and effective across all aspects of writing. This suggests that AWE tools help address lower-level writing concerns, but teacher input remains essential to support students in developing higher-order writing skills such as content elaboration and organisational coherence.

2.4 ChatGPT as an AWE tool

Compared with previous AWE tools, GAI models, such as ChatGPT, demonstrate a more advanced ability to understand context, generate coherent texts, and provide interactive feedback (Baskara, 2023; Han et al., 2021; Kohnke, 2024). These capabilities suggest that GAI may address some of the longstanding limitations of traditional AWE systems. A growing body of research has begun exploring the potential of ChatGPT in AWE. Many of them focused on comparing ChatGPT feedback with human feedback, and they generally found that ChatGPT can generate automated scores that closely align with human evaluations (e.g., Mizumoto & Eguchi, 2023; Naismith et al., 2023), or even provide more comprehensive and balanced feedback than teachers (e.g., Dai et al., 2023; Guo & Wang, 2023). Several studies focused on examining the effectiveness of ChatGPT on students' writing performance. For example, Meyer et al. (2024) found that EFL learners who received ChatGPT feedback shower greater improvement in writing compared to those who received no feedback. Similarly, Shi et al. (2025) reported that students who used ChatGPT outperformed those who used a traditional AWE tool (Pigai) in their overall writing scores.

However, these studies evaluated students' overall writing performance without analysing specific aspects of writing quality. This is a crucial gap, as existing research suggests that while AWE tools may enhance surface-level features such as grammar, they are often less effective in improving higher-order skills (Barrot, 2021). Therefore, to determine whether ChatGPT is more effective, it is essential to evaluate its feedback across distinct aspects of writing quality, rather than treating writing performance as a unified construct. Yang et al. (2025) examined the impact of ChatGPT feedback on lexical sophistication, syntactic complexity, and textual cohesion. They found significant improvements in the first two dimensions, but not in cohesion. Nevertheless, their study used pre-collected writing samples rather than data from real classroom contexts. This warrants research in authentic classroom settings, because students' real experiences may influence the effectiveness of the tool (Sanosi, 2022). Zhang et al. (2025) compared improvement in different aspects of writing between students who received only ChatGPT feedback with those who received hybrid feedback (ChatGPT and teacher feedback). ChatGPT helped students to improve their grammar and sentence variety but had limited effect on higher-order skills such as organisation and critical thinking. In contrast, the students who received hybrid feedback improved their higher-order skills. However, they did not explore students' per-

ceptions of ChatGPT as an AWE tool, although attitudes can significantly influence how students engage with and benefit from ChatGPT.

To address these gaps, the present study investigates the effectiveness of ChatGPT feedback on specific aspects of writing performance, including APA formatting, mechanics, tone, grammar, organisation, and content, in an authentic classroom setting. More importantly, we also explore students' perceptions of ChatGPT as an AWE tool. We aim to provide a more comprehensive understanding of ChatGPT's potential and limitations in real-world educational contexts, as well as to explore its more effective practices in future applications. Specifically, it seeks to answer the following research questions:

What is the impact of ChatGPT feedback on student writing quality?

How does combining ChatGPT and teacher feedback affect student writing compared to ChatGPT feedback alone?

What are students' perceptions of ChatGPT writing feedback?

3 Methodology

3.1 Participants and research context

This study included 77 graduate students at a university in Hong Kong. They were all enrolled in the Faculty of Education, with a specific focus on English Language Education. The majority of participants were female ($n=68$), which was attributed to the typical gender distribution in language education-related fields. Most of them aged between 18 and 34 ($n=71$), and six aged 35 or above. All of them had more than 10 years of English learning experience, but none of them used English as their first language or had lived in an English-speaking country for a long period. Moreover, all participants were native Mandarin speakers from mainland China and had taken the International English Language Testing System (IELTS) examination as their application for graduate study. Their most recent IELTS scores ranged from 6.5 to 8. The IELTS test is widely recognised as a reliable test for determining whether candidates are prepared to study at English-medium universities. According to IELTS guidelines, an overall band score of 6.5 is generally considered satisfactory and necessary for further English studies (IELTS, 2024). This suggests that the student participants were well-prepared to handle English writing tasks.

All of the participants were enrolled in two classes of Academic English Writing taught by the same instructor. The classes were part of the second semester of their one-year graduate program. One class was randomly assigned to be Group A ($n=30$) and the other to be Group B ($n=47$). Each course ran for 16 weeks, with one 120-minute session per week. Students completed three writing assignments during the course. The second assignment, a review of literature on digital literacy curricula and pedagogy, was selected as the focal assignment for this study as it was completed independently and aligned with academic writing standards.

3.2 Research design

The study employed a quasi-experimental design with two intact groups (Groups A and B). The procedure is presented in Fig. 2. In Week 5 of the course, both groups attended a 60-minute training session to learn how to use ChatGPT to obtain writing feedback. This session introduced the basic functions of ChatGPT in academic writing evaluation, including how to upload drafts, input effective prompts, and interpret the feedback.

After the training, the students were taught how to write a Literature review and asked to collect relevant materials. In Week 6, the students in both groups completed their Literature review writing task in class within 60 min. They uploaded their drafts to the university platform, which the instructor and the research team could access. During the second half of the Week 6 class, the students used ChatGPT independently

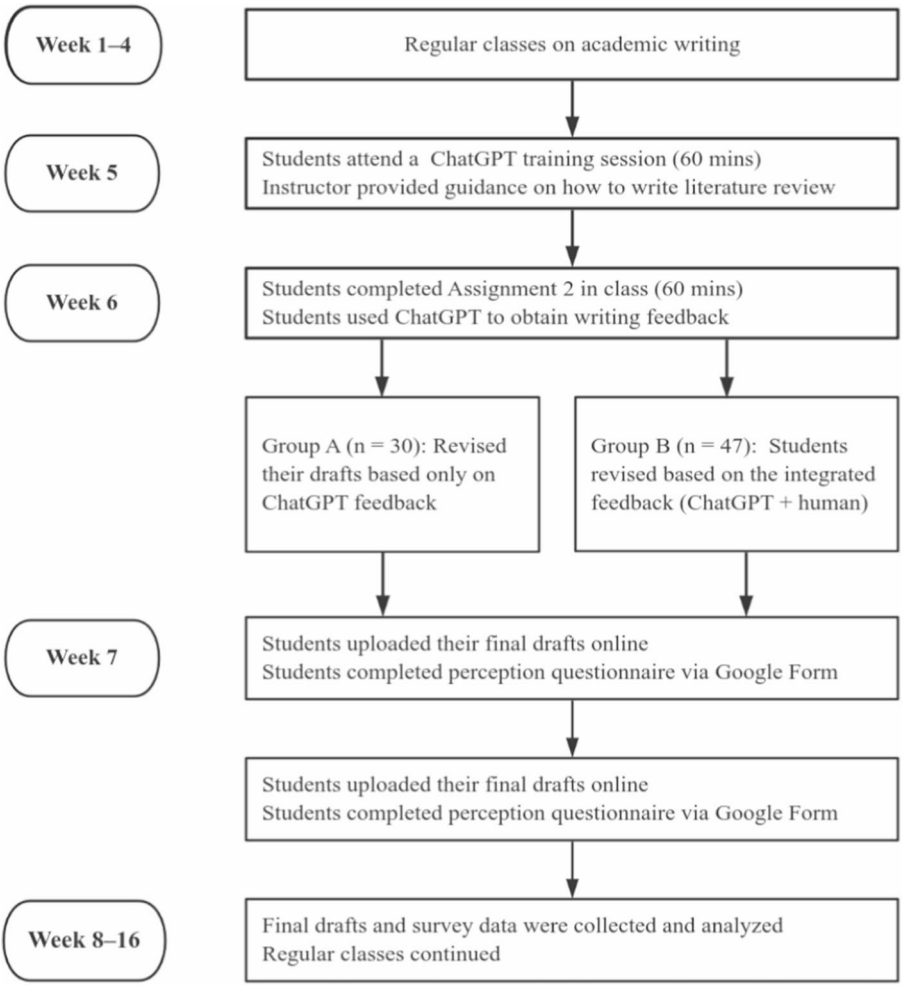

Fig. 2 Outline of experimental procedures

to obtain feedback on their drafts. To ensure that they received relevant and appropriate feedback, they were guided to use a standardised prompt. A prompt refers to the input a user gives to guide a GAI's response, and its quality affects the output (Wan & Chen, 2024). Our study adapted a prompt from Steiss et al. (2024), who developed it through multiple rounds of testing by experts in writing and AI-assisted feedback. This prompt instructed ChatGPT to provide specific, actionable feedback, highlight what had been done well and what could be improved, use a friendly and encouraging tone, and offer examples for improving the essay. After receiving ChatGPT feedback, the Group A participants were instructed to revise their drafts based solely on this feedback and then upload their final versions.

In contrast, the ChatGPT feedback from Group B was collected and forwarded to a human rater – a third researcher involved in the study – who was a native English speaker with over eight years of experience in teaching academic writing. The human rater carefully reviewed both students' original drafts and the ChatGPT feedback. He then refined it with a focus on structural coherence and topic relevance. In addition, he provided targeted revision strategies and personalised explanations tailored to each student's writing topic. Once the integrated feedback was returned to the students in Group B, they revised their drafts based on it and submitted their final drafts online.

In Week 7, students in both groups completed a perception questionnaire designed to gather their views on the feedback they received. The questionnaire included both closed-ended and open-ended items. The responses were collected via Google Forms and compiled by the researchers for further analysis.

3.3 Data collection

We collected quantitative and qualitative data to evaluate writing performance and student perceptions of ChatGPT feedback. To assess writing quality, we employed a comprehensive writing rubric (see **Appendix 1**). It assessed six aspects of academic writing: mechanics, tone, grammar, organisation, content, and APA formatting. These features of writing are commonly considered in university-level writing assessments (e.g. Brigham Young University, 2024; Utica University, 2024). Each aspect was scored on a five-point scale, with 1 being the lowest and 5 being the highest. After the students submitted their final drafts, they received independent scores for each aspect and an overall score calculated by summing the independent scores.

To investigate the students' perceptions of the ChatGPT feedback, we adapted Huang and Renandya's (2020) questionnaire (see **Appendix 2**). The questionnaire includes ten closed-ended and three open-ended questions. The closed questions were organised into three subsections: perceived comprehensibility of the feedback, perceived value of the feedback for revision, and perceived value of the feedback for English writing performance. We modified the original scale by replacing references to 'Pigai' with 'ChatGPT' and recalculated Cronbach's alpha using our data set. The Cronbach's alpha values were 0.81, 0.76, and 0.78, respectively, for the three subsections, indicating satisfactory reliability. The closed-ended items were collected through a Google Form. The open-ended items asked students to elaborate on their experience with the feedback they received. These questions focused on the

same three areas as the closed-ended items. The students were encouraged to provide examples to support their reflections. They recorded their responses and submitted them through the same Google Form.

3.4 Data analysis

To address the first research question, the first drafts and revisions from both groups were collected and evaluated blindly using the same writing rubric. Both researchers independently assessed six randomly selected compositions to ensure reliability and achieved strong inter-rater agreement. A normality test and independent *t*-test were conducted to check for group differences in the initial drafts. Paired samples *t*-tests were then used to compare writing scores between the first to final drafts within each group.

To address the second research question, we created delta variables for each writing aspect to represent the changes from the first to the final draft. A one-way analysis of variance (ANOVA) was then performed to compare these improvements between the two groups. This allowed us to assess the differential impact of ChatGPT feedback alone versus combined ChatGPT and human feedback on the development of academic writing skills among the participants.

For the third research question, closed-ended questionnaire responses were coded using a Likert scale (1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Somewhat Disagree*, 4 = *Somewhat Agree*, 5 = *Agree*, 6 = *Strongly Agree*). The open-ended responses were first coded according to the three dimensions used in the questionnaire. Based on this initial coding, an inductive thematic analysis was conducted to identify sub-themes and patterns in participants' feedback experiences. Two researchers independently coded a subset of the data to ensure reliability and resolved any discrepancies through discussion. Representative responses were selected to illustrate the key themes.

4 Results

4.1 ChatGPT's influence on student writing outcomes

The results of the independent *t*-test showed that there were no significant differences between the first drafts of the students in Groups A and B in mechanics ($p = .52$), tone ($p = .58$), grammar ($p = .63$), organisation ($p = .25$), APA formatting ($p = .179$), and overall ($p = .25$) scores. Only the content category showed a possible trend towards significance ($p = .08$). The results suggest that the participants in the two groups had similar writing abilities before the intervention and would not affect the validity of the subsequent analysis.

p values less than 0.05 indicate statistical significance.

To determine whether the students improved their writing after the ChatGPT intervention, we independently conducted paired-sample *t*-tests using the Group A and Group B data. As shown in Table 1, the Group A students who received feedback from ChatGPT significantly improved their overall writing scores ($t = -6.22$, $p < .001$, $d = 2.88$). The Cohen's *d* is a measure of effect size that quantifies differences

Table 1 Writing improvements from first to final drafts (Group A)

Aspects	Scores of 1 st version		Scores of the final version		t	*p*	Cohen's d
	M	SD	M	SD			
APA formatting	2.90	1.03	3.67	0.84	−5.43	<0.001	0.77
Mechanics	3.03	0.81	3.67	0.61	−4.83	<0.001	0.72
Tone	3.10	0.80	3.70	0.70	−3.84	<0.001	0.86
Grammar	3.00	0.87	3.83	0.75	−6.11	<0.001	0.75
Organisation	2.87	0.94	3.13	0.97	−2.11	0.043	0.69
Content	2.77	0.94	2.93	0.98	−1.72	0.096	0.53
Overall	17.67	4.43	20.93	3.90	−6.22	<0.001	2.88

Note: M mean, *SD* standard deviation, *t* t-value, *p* p-value, *Cohen's d* effect size

p values less than .05 indicate statistical significance

Table 2 Writing improvements from first to final drafts (Group B)

Aspects	Scores of 1 st version		Scores of the final version		t	*p*	Cohen's d
	M	SD	M	SD			
APA formatting	3.23	1.07	3.94	0.85	−5.45	<0.001	0.88
Mechanics	3.15	0.72	3.85	0.59	−6.99	<0.001	0.69
Tone	3.19	0.65	3.89	0.60	−6.19	<0.001	0.78
Grammar	2.91	0.65	4.09	0.62	−11.44	<0.001	0.70
Organisation	3.13	0.97	3.74	0.77	−4.84	<0.001	0.87
Content	3.13	0.74	3.43	0.74	−3.48	0.001	0.59
Overall	18.74	3.66	22.94	3.21	−10.03	<0.001	2.86

between two means in standard deviation units, and a d value above 0.80 is usually considered a large effect (Cohen, 2013). Therefore, a d of 2.88 indicates a strong improvement in writing performance. More specifically, the students demonstrated improved skills in APA formatting ($t = -5.43$, $p < .001$, $d = 0.77$), mechanics ($t = -4.83$, $p < .001$, $d = 0.72$), tone ($t = -3.84$, $p < .001$, $d = 0.86$), and grammar ($t = -6.11$, $p < .001$, $d = 0.75$). These medium-to-large effect sizes suggest that the improvements in these areas were noticeable. As for the improvement in organisation, even though the differences was statistically significant ($t = -2.11$, $p = .043$), the effect size was relatively smaller ($d = 0.69$). Therefore, we can conclude that students demonstrated only modest improvements in organisation in their final drafts. Moreover, no significant differences were found in content between the two drafts ($t = -1.72$, $p = .096$, $d = 0.53$). This suggests that the ChatGPT intervention had minimal impact on the quality of the content. It was effective in helping students revise surface-level aspects of writing but had limited impact on deeper-level issues, for which students may require guidance from teachers.

Our analysis revealed significant improvements across all aspects of writing for the students in Group B, who received both ChatGPT and human feedback (Table 2). The overall writing score increased markedly ($t = -10.03$, $p < .001$, $d = 2.86$). To be more specific, the most significant improvement was in grammar ($t = -11.44$, $p < .001$, $d = 0.70$), followed by mechanics ($t = -6.99$, $p < .001$, $d = 0.69$), tone ($t = -6.19$, $p < .001$, $d = 0.78$), and APA formatting ($t = -5.45$, $p < .001$, $d = 0.88$). Notably,

higher-order writing skills, including organisation ($t = -4.84$, $p < .001$, $d = 0.87$) and content ($t = -3.48$, $p = .001$, $d = 0.59$), also improved significantly. While the effect size for content ($d = 0.59$) was smaller than those associated with the other aspects of writing, it still represented a moderate improvement. This might be because content feedback was relatively more difficult to understand and apply. For example, one student wrote a broad literature review on digital tools supporting homework completion. Her first draft included tools from different fields and education levels. The instructor commented,

You need to narrow your focus and state the context of your paper in the first introduction section. For example, university students, primary, high school? Where?

In the final draft, the student added that the paper focused on university students. However, the scope was still too wide, so the improvement in content was small. In other cases, some students received feedback like 'Any pros and cons of the tools?' In the revision, they only added the pros and forgot to include the cons. Without full revisions, their content scores improved only slightly. Therefore, when feedback was not direct or specific enough, students might not know how to revise, which explained why content improvement was weaker than in other areas.

A one-way ANOVA was conducted to compare the writing improvements between Groups A and B. The results in Table 3 show that there were no significant between-group differences in APA formatting ($F(1, 75) = 0.11$, $p = .744$, $\eta^2 = 0.001$), mechanics ($F(1, 75) = 0.18$, $p = .675$, $\eta^2 = 0.002$), tone ($F(1, 75) = 0.29$, $p = .590$, $\eta^2 = 0.004$), or content ($F(1, 75) = 0.99$, $p = .324$, $\eta^2 = 0.013$). The Eta squared (η^2) is a measure of effect size that indicates how much of the variance in the outcome is explained by group differences, and η^2-values of 0.01, 0.06, and 0.14 are typically considered as small, moderate, and large effects respectively (Cohen, 2013). These very small effect sizes in the results suggest that ChatGPT feedback and combined feedback were similarly effective in promoting improvements in these aspects of writing. With regard to grammar, Group B showed significantly greater improvement in grammatical accuracy than Group A ($F(1, 75) = 4.02$, $p = .049$, $\eta^2 = 0.051$). This means that the combined feedback was more effective in helping students improve their grammatical accuracy than ChatGPT alone. For organisation, the result was not statistically significant at the conventional level ($F(1, 75) = 3.44$, $p = .067$). However, the effect size was moderate ($\eta^2 = 0.044$), which indicates that teacher feedback may offer an

Table 3 ANOVA results comparing writing improvements between group A and group B

Aspects	Group A		Group B		F (1, 75)	*p*	η2
	M	SD	M	SD			
DeltaAPA formatting	0.77	0.77	0.70	0.88	0.11	0.744	0.001
DeltaMechanics	0.63	0.72	0.70	0.69	0.18	0.675	0.002
Delta Tone	0.60	0.86	0.70	0.78	0.29	0.590	0.004
DeltaGrammar	0.83	0.75	1.17	0.70	4.02	0.049	0.051
DeltaOrganisation	0.27	0.69	0.62	0.87	3.44	0.067	0.044
DeltaContent	0.17	0.53	0.30	0.59	0.99	0.324	0.013
DeltaOverall	3.27	2.88	4.19	2.86	1.90	0.172	0.025

Note: *Delta* refers to the value calculated by subtracting the first draft score from the final draft score (i.e., *DeltaOverall* Final draft overall scores minus First draft overall scores), *F* F-ratio, η^2 eta squared (effect size)

advantage. There was no significant difference in the overall improvement in writing quality between Groups A and B ($F(1, 75) = 1.90$, $p = .172$, $\eta^2 = 0.025$); hence, integrating teacher feedback did not notably enhance the overall effectiveness of ChatGPT as an AWE tool.

4.2 Perceptions of using ChatGPT as an AWE tool

4.2.1 Perceived comprehensibility of ChatGPT feedback

Participants' perceptions of the comprehensibility of ChatGPT feedback can be seen in Table 4. Most participants considered the ChatGPT feedback on their writing comprehensible ($M = 5.04$). The statement 'I think the feedback by ChatGPT is clear' received the highest average score ($M = 5.23$). The qualitative data further support this finding. One student noted, 'I can understand the feedback.. the way it outputs responses is quite similar to human dialogue.' Another student added, 'It presents the feedback in a table, and I find that very clear.' Other students did not mention this table format. This is likely because they did not follow the recommended prompt, and ChatGPT generated different feedback based on the their instructions or earlier input.

A handful of students perceived ChatGPT feedback as unclear ($n = 7$). One participant noted that feedback on broader questions tended to be vague. She said, 'When I asked how to improve my essay but didn't specify what aspect, it gave general advice like "accumulate more vocabulary", which was not specific to the essay.' Another student complained that ChatGPT sometimes presented scores as ranges rather than specific numbers. This highlights the importance of giving clear and focused instructions to ChatGPT. Vague or general prompts may result in less actionable or overgeneral feedback.

Almost all of the students (93.5%) agreed with the statement 'I can understand feedback by ChatGPT'. Nevertheless, some students expressed concerns about the appropriateness of ChatGPT feedback. One student shared, 'I often doubt whether the revised sentences are natural or sound strange to native speakers.' Similarly, another commented, 'I suspect that native speakers might find some of the sentences ChatGPT revises quite odd.' These concerns suggest that the student participants in the study, who were at an intermediate to upper-intermediate level of English proficiency, tended to evaluate ChatGPT feedback critically rather than accepting it blindly.

The average score for students' confidence in revising their work was slightly lower than the average score of the overall perceived comprehensibility ($M = 4.81$).

Table 4 Perceived comprehensibility of feedback

Item	StD	D	SoD	SoA	A	StA	M
I can understand feedback from ChatGPT.	1	1	3	10	30	32	5.09
I know how to revise the composition based on feedback I receive from ChatGPT.	1	3	4	18	27	24	4.81
I think the feedback by ChatGPT is clear.	1	3	3	8	33	29	5.23
Average							5.04

Note: *StD* Strongly Disagree, *D* Disagree, *N* Neutral, *A* Agree, *StA* Strongly Agree, *M* average response

In total, 10.39% of students expressed uncertainty about how to apply the feedback to their writing. Some students pointed out that the feedback was often vague or insufficient in detail. One participant noted, 'When I asked for feedback on the whole essay, the suggestions were relatively few and not as helpful as I expected.' Another student mentioned that the feedback could be 'abstract', and she found it difficult to know how to adjust her writing. In such cases, these students found that providing more specific instructions to ChatGPT, such as asking for examples or detailed suggestions, led to more precise and more actionable feedback. One student explained, 'I had to ask it to give me examples to support my argument, and once I did, the feedback became much clearer.' This further emphasises the importance of designing effective prompts. It is necessary to teach students how to formulate specific and well-structured prompts to obtain more accurate, relevant, and actionable feedback.

4.2.2 Perceived usefulness of ChatGPT feedback for composition revision

As shown in Table 5, most students considered ChatGPT feedback useful for revising the composition they were working on. The majority agreed that the feedback helped them correct grammar mistakes ($n=73$, 94.81%) and believed it could help them achieve a higher score on their composition ($n=73$, 94.81%). In addition, most students agreed that the feedback could help improve the overall quality of their composition ($n=72$, 93.51%).

The interviews confirmed that most students found ChatGPT feedback beneficial, particularly in refining grammar, sentence structure, vocabulary, and APA formatting. Typical comments included 'It corrected my spelling errors, inconsistent tenses, and subject-verb agreement issues'; 'It helps me with vocabulary, like suggesting better words to express certain ideas'; 'It pointed out my mistakes in APA formatting and taught me the correct rules' and 'It points out the grammatical mistakes I made, explains the correct rules, and provides a refined version'. Most students expressed satisfaction with ChatGPT because it helped them correct their surface-level writing issues.

Some students also reported that ChatGPT feedback helped make their writing more thoughtful and persuasive. One participant noted,

> I think it improved the content of my writing. I initially used shallow examples, but ChatGPT's suggestions were more in-depth. It pointed out that my use of 'I believe' was too subjective and recommended changing it to 'It is believed,'

Table 5 Perceived usefulness of ChatGPT feedback for composition revision

Item	StD	D	SoD	SoA	A	StA	M
The feedback can help me correct grammar mistakes in this composition.	0	1	3	7	29	37	5.27
I think it can help me improve the quality of this composition.	0	1	4	15	23	34	5.10
It can help me get a higher score. For this composition.	0	2	2	8	23	42	5.31
Average							5.23

Note: *StD* Strongly Disagree, *D* Disagree, *N* Neutral, *A* = Agree; *StA* Strongly Agree, *M* average response

and adding citations to make my argument more convincing. That was something I hadn't considered before.

Moreover, several students observed improvements in the organisation of their essays. They commented that ChatGPT 'provided clear guidance on how to divide my essay into specific paragraphs' and 'suggested that I should incorporate some authentic references to strengthen my argument instead of writing purely subjective thoughts'. However, not all students were convinced that ChatGPT significantly enhanced the overall content and coherence of their essays. One student claimed, 'It rewrites beautiful sentences, but when I put them together in my essay, they don't flow smoothly.' Another student noted that the improvements in structure or content were 'less noticeable compared to those in grammar and vocabulary'. Similarly, another student added, 'I don't think the content has changed much. It mostly sticks to what I originally wrote. The logic hasn't changed either. It just refined what I had already said. So, I feel the main improvements are in grammar and vocabulary, not so much in content or structure.' Although some students appreciated ChatGPT's support in dealing with deeper-level aspects of writing, many perceived its impact as limited in that domain.

4.2.3 Perceived usefulness of ChatGPT feedback in enhancing writing performance

The students' perceptions of the usefulness of ChatGPT feedback in enhancing writing performance are summarised in Table 6. On average, students rated the tool positively across all items (M=4.82). The highest-rated item was 'It can help me enlarge my vocabulary' (M=5.01), followed by 'It can help me improve my grammar' (M=4.95). Many students thought highly of the immediate and targeted feedback you provided on language usage. They claimed that ChatGPT 'offers suggestions that are more personalised and tailored to my current level', 'gives me specific, relevant feedback on my writing', and 'teaches me how to use transitional words to make my writing smoother'. These results indicate that the students valued ChatGPT feedback on vocabulary and grammar. Both grammar and vocabulary are key indicators of writing ability as vocabulary improvement enhances writing performance (Johnson et al., 2016), and grammatical accuracy is critical for conveying meaning (Miranty & Widiati, 2021).

The scores for the statements 'I think it can help me enhance my writing performance' (M = 4.70) and 'The feedback can help me realise my writing problems' (M = 4.62) was slightly lower. This suggests that the students recognised the tool's

Table 6 Perceived usefulness of ChatGPT feedback in enhancing writing performance

Item	StD	D	SoD	SoA	A	StA	M
The feedback can help me realise my writing problems.	2	7	4	10	36	18	4.62
It can help me improve my grammar.	2	4	4	10	23	34	4.95
It can help me enlarge my vocabulary.	1	3	3	8	33	29	5.01
I think it can help me enhance my writing performance.	3	3	6	13	29	23	4.70
Average							4.82

Note: *StD* Strongly Disagree, *D* Disagree, *N* Neutral, *A* Agree, *StA* Strongly Agree, *M* average response

potential to improve their writing but viewed its impact on higher-order writing skills as less significant. One student explained,

> It helps me polish my essays, but improving my writing ability is a long-term process that requires more than just fixing grammar and vocabulary issues. When it comes to the overall content, the changes are minimal. The tool doesn't necessarily enhance my ideas or the logic behind them.

Several students pointed out that ChatGPT was helpful in refining surface-level aspects of their writing but did not significantly enhance the coherence or depth of their arguments. Furthermore, some students expressed concerns about relying too much on it. As one student said, 'Once I'm on my own, I'm still unsure of the words, and I go back to old habits.' Some students suggested that the effectiveness of ChatGPT largely depended on the user: 'If you don't actively learn from it, your writing won't improve in the long run. It's a tool, not a teacher.' These students recognised that meaningful improvement required autonomy and conscious learning.

5 Discussion

5.1 ChatGPT's potential as an AWE tool

This study, grounded in Engeström's (1987) activity theory, examined the impact of ChatGPT feedback on students' writing outcomes to determine whether it was an effective AWE tool. Comparisons between students' initial and final writing scores indicated that ChatGPT's feedback improved their overall writing performance, which supports its potential as a viable AWE tool. Significant improvements were observed in surface-level writing features, such as APA formatting, mechanics, tone, and grammar. However, little to no improvements were found in higher-order writing skills, including organisation and content development. This echoes the performance of standard AWE tools, such as Pigai and Grammarly, which are also widely recognised for addressing surface-level writing issues but not the deeper aspects (Barrot, 2021; Gao, 2021). While traditional AWE tools depend on fixed rubrics to evaluate writing, ChatGPT uses deep learning algorithms to generate feedback that is more natural and varied (Hong, 2023). We expected that ChatGPT could outperform traditional AWE systems in writing evaluation. However, this expectation was not achieved.

This aligned with findings from Yang et al. (2025) and Zhang et al. (2025), who also reported that ChatGPT helped students improve surface-level features but had limited effect on higher-order writing skills. This was potentially because, although ChatGPT can generate human-like responses, it lacks a true understanding of authorial intent. Therefore, it was unable to offer feedback on content and logic like an experienced instructor (Roumeliotis & Tselikas, 2023; Steiss et al., 2024). Even if such feedback could be provided, students may be reluctant to take it because content and logic revisions require much more time and cognitive effort (Black & Nanni, 2016). Improvements in higher-order writing skills were therefore not observed.

We also evaluated the effects of both ChatGPT feedback and combined ChatGPT and teacher feedback on students' writing quality. Compared to feedback from ChatGPT alone, the combined feedback led to greater improvements in higher-level writing aspects, including content and organisation. Although the statistical differences between the groups were insignificant, the trend indicated the pedagogical value of incorporating teacher input that addresses complex writing challenges. This was consistent with earlier studies on standard AWE tools, which have shown that combining automated and teacher feedback elicits better outcomes than using AWE alone (Ebadi et al., 2023; Thi & Nikolov, 2022).

Unlike studies that have suggested ChatGPT can revolutionise writing instruction and become a more reliable writing feedback provider (e.g., Meyer et al., 2024; Shi et al., 2025), the current study found that while ChatGPT has the potential to be used as an AWE tool, it did not present clear advantages over other methods. Instead, they performed at a similar level. It remains unclear whether ChatGPT can replace other AWE tools, and more detailed comparative studies are needed to evaluate their effectiveness across different aspects of writing. Regardless, ChatGPT is not a replacement for human feedback, but a supplementary tool that can assist with immediate surface-level revisions.

5.2 Students' acceptance of ChatGPT as an AWE tool

This study also explored students' acceptance of the feedback. Their perceptions were examined through both questionnaires and semi-structured interviews. Overall, most students described ChatGPT as helpful, convenient, and accessible, which aligned with previous studies reporting generally positive student attitudes towards AWE feedback (Fu et al., 2024). This positive acceptance is significant, as student willingness to engage with feedback is a key factor contributing to its effectiveness.

The qualitative findings also provided possible explanations for differential effectiveness of ChatGPT feedback across various aspects of writing. Students generally perceived feedback related to grammar or vocabulary revisions as easier to understand and more directly applicable. As a result, they were more willing to revise based on this type of feedback. This higher uptake led to more significant improvement in these areas. In contrast, students expressed uncertainty and scepticism toward ChatGPT's suggestions regarding higher-order writing aspects. Some felt that the feedback lacked relevance to their writing content or was too vague to implement effectively. Consequently, they were less likely to engage with it. This may have resulted in a lack of significant improvement in these areas.

Moreover, the results indicated the importance of prompts in ChatGPT's ability to intervene in writing instruction. This is an aspect that has received limited attention in existing research (Wu et al., 2025). Unlike standard AWE tools, the quality and relevance of ChatGPT's feedback heavily depend on the prompt input. In addition, ChatGPT allows for unlimited user interactions so that learners can refine their prompts and ask follow-up questions as needed. Students' percep-

tions of the usefulness and comprehensibility of the feedback varied depending on the quality of the prompts. If prompts were too brief or vague, the generated feedback tended to also be general, which may have discouraged students from engaging in further revision. Moreover, students who were more willing to engage in multiple rounds of interaction with ChatGPT were also more likely to find the feedback helpful.

This study has offered a new perspective on evaluating the effectiveness of AWE tools in the GAI era. Its effectiveness should be evaluated not only by feedback accuracy, but also by prompt quality, student interaction, and learner perceptions. This broader approach allows better understandings of the mechanisms behind the varied impact of ChatGPT's feedback and can inform strategies for optimizing the use of GAI tools in writing instruction.

5.3 Using ChatGPT for AWE: best practices

We recommend several strategies to integrate ChatGPT more effectively into writing instruction. First, it is essential to combine teacher feedback with ChatGPT feedback. Human instructors are better equipped than ChatGPT to provide feedback on higher-order concerns such as content development, coherence, and argumentation (Barrot, 2021). Human feedback can also focus on issues that GAI may overlook and provide suggestions that are more appropriate to the content (Zhang et al., 2025). Furthermore, human tutors encourage engagement and foster independent thinking through real-world interactions (Liu et al., 2024). This reduces students' over-reliance on GAI tools and enhances their autonomous writing abilities.

Second, the effectiveness of ChatGPT's feedback is highly dependent on the quality of the prompts it receives. Therefore, providing prompt training to both teachers and students is essential. Teachers can instruct students about how to design clear and focused prompts, help them revise and improve their wording, and demonstrates how variations in input can produce significantly different feedback. This study adapted a prompt from Steiss et al. (2024), and the students were typically pleased with the feedback they received from ChatGPT. However, they sometimes received irrelevant feedback when they asked follow-up questions using their own prompts, which led to confusion or reluctance to revise. This indicated an ongoing need to teach students how to write effective prompts to maximise the pedagogical value of ChatGPT as an AWE tool.

Finally, critical engagement tasks must be integrated into the writing process. Teachers can design activities in which students critically evaluate ChatGPT feedback: for example, by comparing multiple ChatGPT-generated responses, contrasting ChatGPT feedback with teacher or peer feedback, identifying inaccuracies, or justifying whether to accept a suggested revision. Teachers can also guide students to experiment with different prompt strategies to observe how prompt quality influences the usefulness of the feedback. Encouraging students to engage in multiple rounds of interaction with ChatGPT can further enhance the feedback's effectiveness. Through these tasks, ChatGPT can be used not just as a one-way

feedback provider, but as an interactive tool that supports reflection, experimentation, and deeper involvement in writing.

6 Conclusion and limitations

This study employed a quasi-experimental design and a mixed-methods approach to investigate the impact of ChatGPT as an AWE tool on students' writing quality. It also compared the effects of ChatGPT-only feedback with combined ChatGPT–human feedback and explored students' perceptions of ChatGPT feedback. The students who received only ChatGPT feedback (Group A) and those who received combined feedback (Group B) both demonstrated significant improvements in several important aspects of writing: grammar, tone, mechanics, and APA formatting. However, the students in Group B showed greater improvement than those in Group A in content development and organisation. Despite these improvements, no statistically significant differences in overall writing performance were found between the two groups, though grammar and organisation were slightly better in Group B.

Most participants found the ChatGPT feedback clear and helpful, particularly regarding surface-level issues with grammar and vocabulary. They appreciated the immediate, precise feedback, especially when they were given appropriate prompts to elicit it. However, some students noted limitations in ChatGPT's ability to address the more complex aspects of writing, including content, coherence, and argumentation. Some expressed concerns about becoming over-reliant on ChatGPT and suspected that it might not foster long-term improvement in their writing on its own. Overall, the study demonstrated ChatGPT's potential as a new-generation AWE tool when used appropriately. It also indicated the importance of human feedback, the quality of prompts, and students' perceptions of the tool.

The study had two major Limitations. First, the experiment compared the first and final drafts of a single assignment, which allowed the identification of improvements but could not provide evidence of long-term writing development. Future research could adopt a longitudinal design to examine whether ChatGPT usage can contribute to long-term writing progress. For example, researchers might follow students over a 16-week course and track their writing across multiple assignments to investigate how students improve over time. Second, ChatGPT was not compared with standard AWE tools, so whether it offers unique advantages is unknown. Future studies could compare ChatGPT with other AWE systems across specific writing aspects, such as surface-level and in-depth feedback, to better understand its relative effectiveness and pedagogical value. To examine which approach more effectively enhances students' writing outcomes, researchers could even compare human–machine feedback combinations, such as teacher feedback combined with standard AWE tools versus teacher feedback combined with ChatGPT.

Appendix 1

Table 7 Writing Task Rubric

Criteria	5	4	3	2	1
APA Formatting	Perfect adherence to APA formatting in all elements including citations, references, headings, spacing, and font.	Minor errors in APA formatting that do not substantially detract from the readability or professional appearance of the document.	Noticeable APA formatting errors; however, attempts to follow guidelines are evident.	Multiple APA formatting errors, indicating a lack of understanding or attention to APA style guidelines.	Minimal or no application of APA formatting guidelines.
Mechanics	No spelling or punctuation errors.	Few minor spelling or punctuation errors that do not impact readability.	Some spelling or punctuation errors that slightly distract the reader.	Frequent spelling or punctuation errors; reader's understanding is occasionally impeded.	Persistent errors in spelling and punctuation; difficult to understand.
Tone	Consistently maintains an unbiased, scientific, and appropriate academic tone throughout.	Generally maintains an appropriate tone, with minor lapses in objectivity.	Mostly appropriate tone but some sections may not seem completely unbiased or scientific.	Frequently strays from an unbiased or scientific tone; some sections may be subjective or informal.	Tone is not appropriate for an academic or scientific audience; highly subjective or biased.
Grammar	Grammar is consistently correct and sentences are well-constructed.	Occasional grammatical errors but they do not impede understanding.	Some grammatical errors that may distract or occasionally confuse the reader.	Frequent grammatical errors, making comprehension difficult at times.	Grammar is consistently incorrect, significantly impeding reader comprehension.
Organisation	Exceptionally well-organized, logical flow with seamless transitions between sections and ideas; arguments are fully developed and supported.	Well-organized and logical with effective transitions; most arguments are developed and adequately supported.	Generally organized but some sections may lack logical flow or effective transitions; some arguments are underdeveloped.	Poor organisation and weak transitions make the text difficult to follow; several arguments are insufficiently developed.	Lacks coherent structure and transitions; arguments are fragmented and unsupported.

Table 7 (continued)

Criteria	5	4	3	2	1
Content	Topics are thoroughly examined with comprehensive, well-supported assertions; demonstrates deep understanding.	Topics are adequately examined and assertions are generally supported with evidence; demonstrates clear understanding.	Covers most topics, but some areas are superficially treated; support for assertions is somewhat inconsistent.	Inadequate examination of topics and insufficient support for assertions; shows limited understanding.	Minimal content relevance and very little or no evidence supporting assertions; fails to demonstrate understanding of the subject.

Appendix 2. questionnaire and interview of students' perceptions of ChatGPT feedback

Part 1: Please indicate how much you agree or disagree with the statements. Put a tick (√) in the relevant box. 1 = Strongly Disagree; 2 = Disagree; 3 = Somewhat Disagree; 4 = Somewhat Agree; 5 = Agree; 6 = Strongly Agree.

1. The feedback can help me correct grammar mistakes in this composition.
2. I can understand feedback by ChatGPT.
3. The feedback can help me realize my writing problems.
4. I know how to revise the composition based on feedback I receive from ChatGPT.
5. I think it can help me improve the quality of this composition.
6. I think the feedback by ChatGPT is clear.
7. It can help me improve my grammar.
8. It can help me enlarge my vocabulary.
9. It can help me get a higher score for this composition.
10. I think it can help me enhance my writing performance.

Part 2:

1. Can you understand the feedback by ChatGPT? Is there anywhere that you find not clear? Please explain in detail.
2. To what extent has the feedback by ChatGPT helped you improve the quality of this composition? In what way? Please give examples to illustrate your point.
3. Do you think ChatGPT can help you improve your writing performance? To what extent? In what aspect? Please give examples to illustrate your point.

Acknowledgements The work described in this paper was partially supported by the Start-up Fund for New Recruits of The Hong Kong Polytechnic University (Project No. P0056518). Additionally, support was received from the TDG at the Education University of Hong Kong (Project No. T0293).

Funding Open access funding provided by The Hong Kong Polytechnic University

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics declarations I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Conflict of interest None.

References

Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal Of Second Language Writing, 14*(3), 191–205.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barab, S. A., Evans, M. A., & Baek, E. O. (2004). Activity theory as a lens for characterizing the participatory unit. *Handbook of research on educational communications and technology* (2nd ed., pp. 199–213). Lawrence Erlbaum Associates.

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*(4). https://doi.org/10.1080/09588221.2021.1936071

Baskara, R. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics, 7*(2), 343–358. https://doi.org/10.21093/ijeltal.v7i2.1387

Black, D. A., & Nanni, A. (2016). Written corrective feedback: Preferences and justifications of teachers and students in a Thai context. *GEMA Online Journal of Language Studies*, *16*(3), 99–114.

Blin, F. (2004). CALL and the development of learner autonomy: Towards an activity-theoretical perspective. *ReCALL, 16*(2), 377–395. https://doi.org/10.1017/S0958344004000928

Brigham Young University (2024). Retrieved April 24, 2024, from https://fhsswriting.byu.edu/https:/brightspotcdn.byu.edu/ea/78/afb1465a48b286b8259e80e006af/literature-review-rubric.doc

Chung, C. J., Hwang, G. J., & Lai, C. L. (2019). A review of experimental mobile learning research in 2010–2016 based on the activity theory framework. *Computers & Education, 129*, 1–13. https://doi.org/10.1016/j.compedu.2018.10.010

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023, July). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). IEEE. https://doi.org/10.35542/osf.io/hcgzj

Ding, L., & Zou, D. (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,* (11), 53. https://doi.org/10.1007/s10639-023-12402-3

Ebadi, S., Gholami, M., & Vakili, S. (2023). Investigating the effects of using grammarly in EFL writing: The case of articles. *Computers in the Schools*, *40*(1), 85–105. https://doi.org/10.1080/07380569.2022.2150067

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit.

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, *14*(1), 133–156.

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system Pigai. *International Journal of Emerging Technologies in Learning, 16*(11), 322–330. https://doi.org/10.3991/ijet.v16i11.19657

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies,* 1–29. https://doi.org/10.1007/s10639-023-12146-0

Guo, Q., Feng, R., & Hua, Y. (2022). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, *35*(9), 2312–2331. https://doi.org/10.1080/09588221.2021.1879161

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hong, W. C. H. (2023). The impact of ChatGPT on foreign Language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation*, *5*(1). https://doi.org/10.61414/jeti.v5i1.103

Huang, S., & Renandya, W. A. (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

Hyland, K., & Hyland, F. (Eds.). (2019). *Feedback in second Language writing: Contexts and issues* (2nd ed.). Cambridge University Press.

IELTS (2024). Test taker performance 2023. Retrieved from https://www.ielts.org/teaching-and-research/test-taker-performance

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal, 7*(3), 700–715. https://doi.org/10.1002/tesj.238

Kohnke, L. (2024). Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning. *Computers & Education: Artificial Intelligence, 7*, Article 100279. https://doi.org/10.1016/j.caeai.2024.100279

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of second language writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System*, *99*, 102505. https://doi.org/10.1016/j.system.2021.102505

Liu, Z. M., Hwang, G. J., Chen, C. Q., Chen, X. D., & Ye, X. D. (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

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using llms to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199.

Miranty, D., & Widiati, U. (2021). An automated writing evaluation (AWE) in higher education. *Pegem Journal of Education and Instruction, 11*(4), 126–137. https://doi.org/10.47750/pegegog.11.04.12

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mohsen, M. A., & Abdulaziz, A. (2019). THE EFFECTIVENESS OF USING A HYBRID MODE OF AUTOMATED WRITING EVALUATION SYSTEM ON EFL STUDENTS'WRITING. *Teaching English with Technology*, *19*(1), 118–131.

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403).

Page, E. B. (2003). *Project essay grade: PEG. Automated essay scoring: A cross-disciplinary perspective*. /Lawrence Erlbaum Associates.

Rahimi, M., Fathi, J., & Zou, D. (2024). Exploring the impact of automated written corrective feedback on the academic writing skills of EFL learners: An activity theory perspective. *Education And Information Technologies*. https://doi.org/10.1007/s10639-024-12896-5

Roumeliotis, K. I., & Tselikas, N. D. (2023). Chatgpt and open-ai models: A preliminary review. *Future Internet, 15*(6), Article 192. https://doi.org/10.3390/fi15060192

Sanosi, A. B. (2022). The impact of automated written corrective feedback on EFL learners' academic writing accuracy. *Journal of Teaching English for Specific and Academic Purposes*, 301–307. https://doi.org/10.22190/jtesap2202301s

Shen, C., Shi, P., Guo, J., Xu, S., & Tian, J. (2023). From process to product: Writing engagement and performance of EFL learners under computer-generated feedback instruction. *Frontiers in Psychology*, *14*, 1258286. https://doi.org/10.3389/fpsyg.2023.1258286

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. *Handbook of automated essay evaluation* (pp. 1–15). Routledge.

Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, 1–28. https://doi.org/10.1080/09588221.2025.2454541

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., & Olson, C. B. (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

Thi, N. K., Nikolov, M., & Simon, K. (2022). Higher-proficiency students' engagement with and uptake of teacher and grammarly feedback in an EFL writing course. *Innovation in Language Learning and Teaching*, *0*(0), 1–16. https://doi.org/10.1080/17501229.2022.2122476

Utica University (2024). Retrieved April 24, 2024, from https://www.utica.edu/academic/Assessment/new/LitReview.doc

Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, *20*(1), 010152.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157–180.

Wu, J., Li, J., Ge, Z., Xu, M., Lin, L., & Zhang, R. (2025). Effectiveness of generative AI in automated written corrective feedback with prompting. *Journal of Educational Computing Research,* , Article 07356331251359430. https://doi.org/10.1177/07356331251359430

Xu, J., & Zhang, S. (2022). Understanding AWE feedback and English writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher, 31*(4), 357–367. https://doi.org/10.1007/s40299-021-00577-7

Yang, K., Raković, M., Liang, Z., Yan, L., Zeng, Z., Fan, Y., & Chen, G. (2025, March). Modifying AI, enhancing essays: How active engagement with generative AI boosts writing quality. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 568–578).

Yousofi, R. (2022). Grammarly deployment (in)efficacy within EFL academic writing classrooms: An attitudinal report from Afghanistan. *Cogent Education*, *9*(1), 2142446. https://doi.org/10.1080/2331186X.2022.2142446

Zhang, R., Zou, D., & Cheng, G. (2023). Learner engagement in digital game-based vocabulary learning and its effects on EFL vocabulary development. *System*, *119*, 103173.

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study. *Innovation in Language Learning and Teaching, 1–19*. https://doi.org/10.1080/17501229.2025.2503890