Discover Education

Research

Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality

Mohamed Mekheimer[1]

Received: 23 December 2024 / Accepted: 3 June 2025
Published online: 12 June 2025

Abstract
This study investigated the impact of generative AI-assisted writing feedback, specifically using Grammarly, on English as a Foreign Language (EFL) learners' writing proficiency, revision practices, and writing quality. Sixty postgraduate EFL students were randomly assigned to either an experimental group using Grammarly as an AI-enhanced feedback system or a control group receiving traditional instruction. Pre- and post-tests assessed overall writing proficiency, and semi-structured interviews were conducted with 25 students from the experimental group to gather qualitative data. Quantitative results indicated that the experimental group achieved significantly higher post-test scores than the control group. Additionally, positive correlations were found between the use of AI features, including generative AI, grammar, and vocabulary tools, and increased revision frequency, as well as improvements in content, organization, and cohesion. Thematic analysis of the qualitative data revealed that AI-assisted feedback promoted student engagement by reducing frustration, boosting confidence, and facilitating a greater sense of accomplishment. Participants emphasized the need to strategically integrate AI tools with traditional instruction, to foster critical thinking, and for AI tools to prioritize natural language fluency as well as to provide detailed feedback beyond basic grammar and mechanics. These results suggest that generative AI-assisted writing feedback, when used with pedagogical awareness, can effectively enhance EFL learners' writing skills and support active participation in the writing process. The study also highlights that critical thinking skills are essential and that students should avoid overreliance on AI tools.

Keywords Generative AI · AI-assisted writing feedback · EFL writing · Writing proficiency · Revision · Student engagement

1 Introduction

The integration of artificial intelligence (AI) into education is rapidly evolving, with significant potential recognized in language learning, particularly for developing writing skills [2, 11, 16]. Among the various AI applications, tools providing automated writing feedback, such as grammar and style checkers enhanced with AI capabilities (akin to Grammarly, the specific tool used in this study as detailed in the methods), have become widely accessible to students [12, 25]. These

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44217-025-00602-7.

✉ Mohamed Mekheimer, mohamed.abdelgawad3@edu.bsu.edu.eg | [1]Faculty of Education, Beni-Suef University, Beni Suef Governorate, Salah Salem Street, Beni-Suef 62511, Egypt.

tools offer immediate, individualized feedback primarily focused on surface-level aspects like grammar, punctuation, spelling, and style suggestions [1, 20].

While the accessibility and immediacy of such AI-assisted feedback are clear advantages over potentially delayed traditional teacher feedback [37], its specific impact on the development of English as a Foreign Language (EFL) writing proficiency requires more nuanced investigation. Much existing research either examines broader AI instructional systems [4, 51] or explores student perceptions [23], often lacking direct empirical comparison between the effects of these readily available AI feedback tools and traditional teacher feedback methods within authentic classroom settings. Furthermore, there is *a specific gap* in understanding how interaction with AI feedback focused predominantly on form and style influences not only the final writing product's quality (across different dimensions like content and organization, not just language use) but also the students' writing *process*, particularly their revision behaviors [39]. Understanding these differential impacts is crucial for educators seeking to effectively integrate AI tools into EFL writing pedagogy.

This study aims to address this gap by empirically comparing the effects of using an AI-assisted feedback tool (Grammarly) versus traditional teacher feedback on EFL students' writing development. We specifically investigate how these two feedback modalities influence overall writing proficiency, the frequency and nature of student revisions, and improvements in distinct components of writing quality, namely content, organization, and language use (cohesion/mechanics). By examining these aspects, the study seeks to provide clearer insights into the practical benefits and potential limitations of employing common AI feedback assistants in the complex context of EFL academic writing development, where feedback needs span from sentence-level accuracy to higher-order concerns like argumentation and structure.

2 Research questions

Therefore, this study investigates the comparative effects of AI-assisted feedback and traditional feedback on EFL writing by addressing the following research questions:

RQ1: How does using AI-assisted writing feedback (focused on grammar, style, mechanics via Grammarly) impact overall EFL writing proficiency compared to receiving traditional teacher feedback?

RQ2: How does access to AI-assisted writing feedback influence EFL students' revision frequency and engagement during writing tasks compared to traditional feedback methods?

RQ3: To what extent does AI-assisted writing feedback, compared to traditional teacher feedback, contribute to improvements in specific writing components: content, organization, and language use (cohesion/mechanics) in EFL writing?

2.1 Research hypotheses

Based on the research questions, this study proposes the following hypotheses:

H1: Students receiving AI-assisted writing feedback will demonstrate significantly higher overall writing proficiency scores on the post-test compared to students receiving traditional teacher feedback.

H2: Students utilizing AI-assisted writing feedback will exhibit a significantly higher frequency of revisions during the writing process compared to students receiving traditional teacher feedback.

H3: Students utilizing AI-assisted writing feedback will show significantly greater improvement in the language use (cohesion/mechanics) component of their writing, and potentially different patterns of improvement in content and organization, compared to students receiving traditional teacher feedback, as measured by the post-test scoring rubric.

2.2 Review of literature

The integration of Artificial Intelligence (AI) into English as a Foreign Language (EFL) instruction represents a rapidly expanding field [22, 31], leading to the widespread adoption of specialized tools designed to provide automated writing feedback [10, 25, 47]. This trend reflects a broader interest in leveraging AI for various aspects of language teaching and learning [1], with automated writing evaluation and feedback systems becoming increasingly prominent in educational settings, aiming to enhance aspects like writing skills and potentially even self-efficacy [17, 52].

Commonly available tools, such as grammar checkers, style assistants, and platforms incorporating AI features like Grammarly (the type of tool central to this study), primarily function by delivering immediate and accessible feedback [20]. This feedback typically targets surface-level aspects of writing, including grammatical accuracy, mechanics

(punctuation, spelling), vocabulary choice, and adherence to stylistic conventions [1, 16]. These systems analyze text and provide suggestions often aimed at error correction and improving clarity at the sentence level [30], representing a form of technology-mediated support often used alongside or in place of traditional teacher feedback cycles [34].

Proponents suggest a significant potential benefit of these AI tools lies in their capacity for personalization and fostering learner autonomy [21, 48]. AI systems can analyze patterns in student writing to offer tailored feedback or suggest learning pathways adapted to individual needs [19, 23, 27, 35, 38, 44, 45]. This potential for individualized support is seen as a key driver for enhancing student engagement and creating more adaptive learning experiences [32, 49], possibly even extending to recognizing learner affective states [28]. The expectation is that such targeted assistance could lead to more effective and efficient language acquisition [15], potentially transforming collaborative learning experiences as well [13], although the full extent and nature of these benefits require ongoing empirical investigation.

2.3 Examining the evidence and identifying the gaps

While the potential benefits of immediacy and accessibility are appealing, particularly compared to the constraints of traditional teacher feedback cycles, the actual impact of these common AI feedback tools on comprehensive EFL writing development requires critical examination. Several **key gaps and limitations** persist in the current literature:

1. *Lack of Direct Comparative Studies:* Much research investigates AI feedback tools in isolation or focuses on student perceptions [14, 24, 40], rather than conducting direct empirical comparisons with traditional, human teacher feedback within the same instructional context. Meta-analyses show positive trends for AI assistance generally [29, 41], but often aggregate diverse AI types and intervention designs, making it difficult to isolate the specific effect of common feedback tools like Grammarly compared directly to teacher feedback on multifaceted writing outcomes.
2. *Focus on Surface-Level versus Higher-Order Concerns:* Many readily available AI feedback tools excel at identifying grammatical errors and stylistic inconsistencies [16]. However, their effectiveness in promoting improvement in higher-order writing aspects, such as content development, logical organization, argumentation, and coherence—areas where teacher feedback is often considered crucial [3, 8]—remains less substantiated. Reviewer 1 correctly notes that classifying tools like Grammarly strictly as "generative AI" akin to ChatGPT might be misleading, their primary function is corrective/suggestive feedback, not necessarily deep content generation or holistic rhetorical analysis. This distinction is critical when evaluating their impact on overall writing quality.
3. *Impact on the Writing Process (Revision):* While AI feedback might prompt *more* frequent revisions due to its immediacy [6, 42], the *nature* and *depth* of these revisions compared to those prompted by teacher feedback are under-explored. Does AI feedback encourage deeper engagement with meaning and structure, or primarily superficial editing? How does this compare to the revision process guided by teacher comments that might focus more on meaning and global issues?
4. *Contextual Factors in EFL Writing:* EFL learners face specific challenges including L1 interference, lexical gaps, and navigating different rhetorical conventions [9, 33, 50]. The extent to which current AI feedback tools can provide culturally and linguistically sensitive feedback that addresses these nuances, compared to an experienced EFL teacher, is an open question requiring empirical investigation.

2.4 Theoretical perspectives on feedback mechanisms

To understand the potential differential impacts of AI versus teacher feedback, certain theoretical frameworks are particularly relevant:

- *Sociocultural Theory (SCT):* Vygotsky's [46] theory highlights the role of mediation in learning. Both AI tools and teachers act as mediators providing feedback, but they do so differently. Teacher feedback is inherently social and dialogic, potentially offering tailored scaffolding based on inferred understanding and interaction [53]. AI feedback, while potentially interactive, is algorithmic and lacks genuine social-cognitive understanding. This study explores whether these different modes of mediation lead to different learning outcomes in terms of writing quality and revision practices. *While AI might connect students in collaborative platforms* [18]*, the focus here is on the individual student's interaction with the feedback source itself.*

- *Cognitive Load Theory (CLT):* Sweller's [43] theory suggests learning is optimized when cognitive load is managed effectively. Immediate, constant AI feedback could potentially overload learners with information (extraneous load), especially concerning surface errors, detracting from focus on higher-order concerns. Conversely, teacher feedback, often more synthesized and delivered later, might impose a different load profile, potentially facilitating deeper processing of key issues. This study implicitly examines how these different feedback modalities might influence cognitive load and subsequent writing performance.

General learning theories like Constructivism (e.g., [36]) and Situated Learning [26] inform our understanding of writing as a constructive and context-bound activity. However, to specifically dissect the potential differences arising directly from the *source and nature of the feedback*—the core comparison in this study—Sociocultural Theory (SCT) and Cognitive Load Theory (CLT) offer more targeted analytical frameworks. SCT prompts examination of how learning is mediated differently through interaction with technology versus interaction with a human expert [46], while CLT helps analyze how the information processing demands might differ between immediate, often fragmented AI feedback and holistic, potentially delayed teacher feedback [43]. Consequently, SCT and CLT are deemed more pertinent for analyzing the specific effects of the *feedback mechanism* (AI vs. human) under investigation.

2.5 Synthesizing the need for this study

Existing research acknowledges the potential of AI feedback tools but also highlights significant gaps in our understanding, particularly regarding their specific impact compared to traditional teacher feedback. There is a critical need for empirical studies that directly compare these two feedback modalities within authentic EFL classroom settings, focusing not just on overall proficiency scores but also on specific writing quality components (content, organization, language use) and the nature of the student revision process. Critiques regarding the over-reliance on frameworks like "SMART" without sufficient empirical backing for specific EFL contexts [21] and concerns about the depth of feedback from current tools further highlight this need.

This literature review reveals that while AI feedback tools offer promise, their effectiveness relative to established pedagogical practices (teacher feedback) requires rigorous investigation. By directly comparing AI-assisted feedback (via a tool like Grammarly) with traditional teacher feedback, this study aims to provide much-needed empirical evidence on their respective influences on EFL students' writing proficiency, revision frequency, and quality across different writing dimensions, thereby addressing the identified gaps and informing pedagogical choices regarding technology integration in EFL writing instruction.

3 Methods

3.1 Participants and setting

The study sample comprised 60 postgraduate students enrolled in a mandatory EAP (English for Academic Purposes) writing course during the second semester of the 2023–2024 academic year at a public university in Egypt. All participants were non-native English speakers pursuing Master's or PhD degrees across various disciplines (primarily education, humanities, and social sciences) and provided informed consent prior to participation.

- *Proficiency Assessment and Control:* Participants possessed varying English proficiency levels typical of postgraduate entrants at the institution, generally ranging from B2 to C1 on the Common European Framework of Reference for Languages (CEFR), as indicated by university placement records. While strict proficiency homogeneity was not feasible, the pre-test (detailed in 3.3) served as a baseline measure to statistically account for initial writing ability. Furthermore, the **random assignment** of participants to the experimental and control groups immediately following the pre-test aimed to distribute variations in proficiency and other individual differences evenly between the groups, minimizing systematic bias.
- *Instructor Details:* Both the experimental and control group sections of the course were taught by the **same instructor** (one of the researchers) to ensure consistency in core content delivery, teaching style, and overall task expectations, beyond the specific feedback intervention. To mitigate potential instructor bias, standardized lesson plans were used

for shared content, and distinct, detailed protocols were followed for feedback delivery in each group, as specified in Sect. 3.4. The instructor was experienced in teaching EAP writing at the postgraduate level.

4 Research design

This study employed a **quasi-experimental, pre-test/post-test control group design** to investigate the comparative effects of two feedback conditions on EFL postgraduate students' academic writing development.

- *Independent Variable (IV):* Type of writing feedback provided (AI-assisted feedback via Grammarly Premium versus traditional written teacher feedback).
- *Dependent Variables (DVs):*

  1. Overall academic writing proficiency (measured by pre-test/post-test essay scores).
  2. Changes in specific writing quality components: Content, Organization, and Language Use (measured by analytic scoring rubric, Appendix II).
  3. Revision frequency/engagement (measured quantitatively via counts of revisions [e.g., tracked via platform/manual count] and qualitatively via thematic analysis of semi-structured interviews).

- *Groups:* Participants were randomly assigned (using a random number generator) after the pre-test to either:

  - *Experimental Group (n = 30):* Received standard EAP instruction supplemented with AI-assisted feedback using Grammarly Premium.
  - *Control Group (n = 30):* Received the same standard EAP instruction supplemented with traditional written teacher feedback.

- *Duration:* The intervention period spanned 6 weeks (Weeks 2–7) within an 8-week study timeframe, with the course meeting twice weekly for 1.5 h per session (totaling 18 h of instruction plus assignment work). See Fig. 1 for a visual timeline.
- *Feedback Timing Consideration:* Acknowledging an inherent difference, the AI feedback (experimental group) was available immediately upon student use of the tool, while teacher feedback (control group) was provided with a typical delay (e.g., 1 week turnaround). This timing difference is intrinsic to the feedback types being compared and represents a potential limitation discussed in the limitations section.

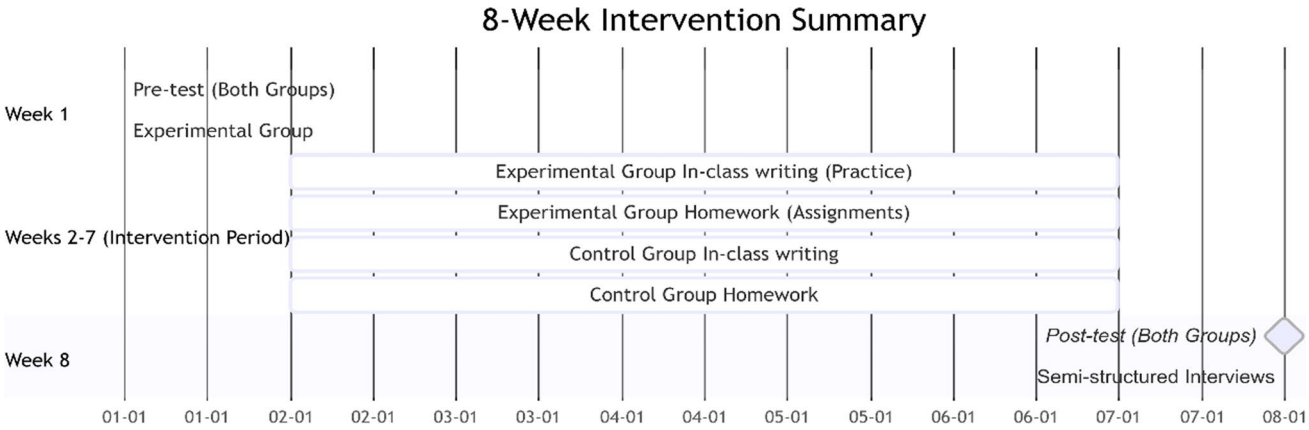

Fig. 1 Timeline of 8-Week Intervention of Experimental and Control groups

4.1 Materials and instruments

4.1.1 Pre-test and post-test

Design: Both tests were designed by the researchers, targeting EAP skills relevant to postgraduate academic writing. They required participants to write a short argumentative essay (approx. 300–400 words) on one of three provided topics (The challenges and opportunities of using technology in education; the role of cultural understanding in international communication; the impact of globalization on local communities), incorporating information from a short source text provided. The topics were counterbalanced across pre/post-tests and groups.

Time Limit: A time limit of 60 min was allocated for each test, determined through pilot testing with similar students to ensure sufficient time for planning, writing, and basic revision under timed conditions.

Writing Assessment Rubric: For scoring both pre-test and post-test essays, a detailed analytic rubric, developed by the researchers based on established EAP writing assessment criteria, was utilized (see Appendix II). This rubric assessed three key components of academic writing: Content (evaluating relevance, idea development, and argumentation; maximum 20 points), Organization (assessing overall structure, coherence, and paragraphing; maximum 20 points), and Language Use (covering grammatical accuracy, vocabulary range and accuracy, mechanics, and cohesion; maximum 20 points), yielding a total possible score of 60 per essay.

4.2 AI system: grammarly premium

The experimental group utilized Grammarly Premium, accessed through university-provided licenses. The selection of Grammarly was deliberate and aligns directly with the study's focus on automated *feedback* rather than AI-driven text *generation*. While Grammarly is recognized primarily as an AI-powered writing assistant offering corrective feedback, rather than a comprehensive generative AI (GAI) system like ChatGPT, this specific functional profile was considered a methodological advantage for this research and to avoid fake authorship of students' writing assignments [7].

Utilizing tools with more extensive generative capabilities would have introduced a significant confounding variable, making it difficult to isolate the effects of feedback itself and posing a considerable risk of students substituting AI-generated content for their own composition. Choosing Grammarly allowed for greater control over the intervention, ensuring the focus remained squarely on how students utilize automated feedback to revise and improve their *own* writing.

While it incorporates AI and offers some limited GAI features, such as sentence rewrite suggestions, its core function and primary application within this intervention was explicitly as a feedback and revision support tool. The specific functionalities leveraged in this study therefore included its advanced checks for grammar, mechanics, spelling, punctuation, clarity, conciseness, style consistency, vocabulary enhancement (suggesting synonyms, identifying repetition), and plagiarism detection.

Correspondingly, training and usage guidance directed students to focus on these core feedback features related to correctness, clarity, and engagement (vocabulary), alongside the plagiarism checker. Students were explicitly instructed to critically evaluate all suggestions provided by the tool and use them to inform their *own* revision decisions, rather than passively accepting changes or relying heavily on its limited generative capabilities.

4.3 Semi-structured interviews

Following the intervention period, semi-structured interviews were conducted during Week 8 with a volunteer sample of **25** students from the experimental group. The purpose of these interviews was to gather in-depth qualitative data on their experiences using Grammarly Premium. Specifically, the interviews explored participants' views on the tool's usability, the perceived helpfulness and limitations of specific feedback types, how they integrated Grammarly into their writing and revision processes, any resulting changes in engagement or writing habits, comparisons with traditional feedback, and their overall impressions regarding its impact on their writing quality and confidence. The interview guide detailing the specific questions used is provided in Appendix III.

4.4 Procedure

1. *Ethical Approval and Consent:*
   Approval was obtained from the university's Institutional Review Board. Participants received information sheets and signed consent forms.
2. *Week 1: Baseline and Preparation:*
   During the first week of the study, all 60 participants completed the timed argumentative essay pre-test under invigilated classroom conditions to establish baseline writing proficiency. Following the completion of this pre-test, participants were randomly assigned to either the experimental or control group.
   Subsequently, the experimental group alone received a dedicated 60-minute training session focused on the effective use of Grammarly Premium; this session covered essential aspects such as logging in, navigating the interface, submitting text, understanding the various feedback categories, interpreting suggestions critically, utilizing the plagiarism check appropriately, and implementing strategies for incorporating the tool's feedback into their revision process.
3. *Weeks 2–7: Intervention Phase:*
   During the 6-week intervention period (Weeks 2–7), both the experimental and control groups followed the identical core EAP curriculum content and completed the same writing tasks, all facilitated by the same instructor to ensure consistency. These tasks comprised a mix of shorter in-class exercises aimed at specific skills ('Practice Activities') and longer homework assignments requiring more substantial writing, like drafting essay sections ('Assignment Activities'). The crucial distinction between the groups lay entirely in the feedback mechanism employed.
   Participants in the experimental group were required to use Grammarly Premium to check drafts of all their practice and assignment tasks, benefiting from immediate, 24/7 access to automated feedback. Their writing process involved initiating drafts based on prompts and their own ideas, then iteratively using Grammarly during drafting and revision, specifically focusing on the feedback provided for grammar, vocabulary, style, and clarity to inform their editing decisions.
   The instructor's role for this group shifted to that of a facilitator, guiding students on how to interpret and critically evaluate Grammarly's suggestions, leading discussions on writing issues flagged by the tool, and concentrating direct instruction on higher-order concerns like argumentation and structure, which Grammarly addresses less directly.
   Importantly, the instructor did not provide separate written corrective feedback on drafts checked by Grammarly; the process emphasized student-AI interaction for feedback, mediated by instructor guidance on effective tool usage. In contrast, participants in the control group submitted drafts of the same tasks directly to the instructor and received traditional written feedback within approximately 1 week. This feedback consisted of annotations and end comments addressing a balance of global issues (content, organization) and patterns of local errors (grammar, vocabulary, mechanics), consistent with standard EAP pedagogy. The instructor's role for the control group remained traditional, involving the provision of direct corrective feedback and guidance via written comments on student work.
4. *Week 8: Post-Intervention Assessment:*
   In Week 8, concluding the intervention phase, all participants from both groups completed the timed argumentative essay post-test, administered under conditions identical to those of the pre-test. Additionally, during this final week, the planned semi-structured interviews were conducted with the volunteer participants from the experimental group to gather qualitative data on their experiences.
5. *Marking and Reliability:*
   To ensure objective assessment, all pre-test and post-test essays were first anonymized using codes. Subsequently, they were scored independently by two trained EAP raters—the course instructor and another qualified EAP teacher who was blind to group allocation—utilizing the analytic rubric detailed in Appendix II. Prior to scoring, both raters participated in a training session using anchor scripts to standardize their application of the rubric.

To establish inter-rater reliability, a subset comprising 20% of the essays was double-marked. The resulting Cohen's Kappa statistic exceeded $\kappa = 0.75$ for all rubric components, indicating substantial agreement between the raters. Any scoring discrepancies identified on the double-marked scripts were resolved through discussion, and the final scores for these essays were averaged for analysis. The remaining essays were marked by one rater each, with batches assigned randomly between the two trained raters.

4.5 Data analysis

Both quantitative and qualitative data analyses were conducted to comprehensively address the research questions. Initially, baseline equivalence between the experimental and control groups was confirmed by conducting an independent samples *t*-test on the pre-test scores, with Levene's test confirming the assumption of equal variances for this initial comparison, indicating no statistically significant difference between the groups at the outset.

To evaluate the intervention's impact on writing proficiency (RQ1 and RQ3), post-test scores were compared between the groups using Welch's *t*-test, which was specifically chosen because Levene's test indicated unequal variances between the groups on the post-test measures ($p = 0.003$). The Mann–Whitney U test was also performed on the post-test scores as a non-parametric confirmation of these findings.

Furthermore, to assess changes in specific writing components (Content, Organization, Language Use), separate Welch's *t*-tests (or Mann–Whitney U tests where appropriate for specific sub-scores) were conducted on the respective post-test sub-scores. To investigate the relationship between student engagement with the AI tool and writing outcomes within the experimental group (RQ2), correlational analyses (using Pearson's r or Spearman's rho, depending on data distribution) were performed.

This analysis focused specifically on three key learning behaviors deemed most relevant to the study's focus on AI feedback and revision: the frequency of using Grammarly's generative AI features (e.g., rewrite suggestions), measured using student self-report surveys administered after each major assignment asking participants to estimate usage frequency; the frequency of using corrective features (grammar and vocabulary checks), similarly measured via the self-report surveys; and the total number of revisions made between drafts, quantified by manually comparing sequential draft versions submitted by participants via the course learning management system.

A regression analysis was subsequently conducted using these three usage/revision metrics to examine their predictive power on post-test scores within the experimental group.

Complementing the quantitative results, thematic analysis, following the procedures outlined by Braun and Clarke [5], was conducted on the transcribed data from the semi-structured interviews with the experimental group participants. This qualitative analysis aimed to identify recurring themes regarding students' engagement levels, revision processes, interactions with AI feedback, perceived benefits and challenges, and overall experiences. All statistical analyses were performed using SPSS [Version 25].

5 Findings

This section presents the results of the quantitative data analyses conducted to address the research questions regarding writing proficiency (RQ1 and RQ3) and the relationship between AI tool usage and outcomes (RQ2 and RQ3). Preliminary analyses verifying baseline group comparability are presented first, followed by the main analyses comparing post-test performance between the control and experimental groups, and finally, correlational and regression analyses exploring relationships between learning behaviors and outcomes within the experimental group. Qualitative findings are presented in the subsequent Sect. 4.5.

5.1 Preliminary analysis: baseline comparability

To ensure the experimental and control groups were comparable regarding initial writing proficiency, scores from the pre-test administered before the intervention were compared. Levene's test for equality of variances indicated that the variances were equal between the control ($n = 30$) and experimental ($n = 30$) groups, $F(1, 58) = 0.51$, $p = 0.479$. Subsequently, an independent-samples *t*-test revealed no statistically significant difference in mean pre-test scores between the control group ($M = 27.13$, $SD = 4.45$) and the experimental group ($M = 28.60$, $SD = 3.96$), $t(58) = -1.35$, $p = 0.183$, 95% CI for the difference [−3.64, 0.71]. These results, detailed in Table 1, support the assumption that the groups possessed equivalent baseline academic writing proficiency prior to the intervention.

5.2 Post-test performance comparison (RQ1 and RQ3)

To examine the impact of the intervention, post-test writing scores were compared between the groups. First, the total post-test score (maximum 60 points) was analyzed. Levene's test indicated unequal variances between the groups for the total

Table 1 Independent-Samples *t*-test for Pre-test Scores by Group

Group	*N*	*M*	*SD*	*t*	*df*	*p*	MD [95% CI]
Control	30	27.13	4.45	− 1.35	58	0.183	− 1.47 [− 3.64, 0.71]
Experimental	30	28.60	3.96				

Table 2 Welch's *t*-test comparison of total post-test scores between control and experimental groups

Group	*M*	*SD*	Levene's *F* (*p*)	t-Statistic	*p*	MD [95% CI]
CTRL ($n = 30$)	45.50	5.32	9.51 (0.003)*	$t(47.35) = 10.89$	< 0.001	8.87 [7.23, 10.50]
Ex. ($n = 30$)	54.37	5.54				

*$p < 0.05$ for Levene's test indicates unequal variances assumed

Table 3 Mann–Whitney U test comparison of total post-test score ranks between control and experimental groups

Group	Mean rank	*U*	*W*	*z*	*p*
CTRL ($n = 30$)	16.05	16.50	481.50	6.42	< 0.001
Ex. ($n = 30$)	44.95				

Table 4 Independent-Samples *t*-test Comparison for the Content Component Post-Test Score

Component	Group	*M*	*SD*	Levene's *F* (*p*)	Test Statistic	*p*	MD [95% CI]
Content (Max 20)	CTRL ($n = 30$)	16.10	2.30	1.23 (0.272)	$t(58) = 1.68$	.0.098	0.95 [− 0.18, 2.08]
	EX ($n = 30$)	17.05	2.15				

MD Mean Difference; *CI* Confidence Interval. Mean difference reported as EX—CTRL. Max score = 20 points

score, $F(1, 58) = 9.51$, $p = 0.003$. Given that the assumption of homogeneity of variances was violated, Welch's *t*-test, which does not assume equal variances, was deemed the appropriate parametric test for comparing the group means.

As shown in Table 2, the results revealed a statistically significant difference, with the experimental group ($M = 54.37$, $SD = 5.54$) scoring significantly higher than the control group ($M = 45.50$, $SD = 5.32$), Welch's $t(47.35) = 10.89$, $p < 0.001$, 95% CI for the difference [7.23, 10.50].

Furthermore, a Mann–Whitney *U* test was conducted as a robust non-parametric confirmation, often used when assumptions for parametric tests are violated or data may not be perfectly normally distributed. As shown in Table 3, this test confirmed a statistically significant difference in the score distributions ($U = 16.50$, $z = 6.42$, $p < 0.001$), with the experimental group exhibiting a higher mean rank (44.95) compared to the control group (16.05). Both parametric and non-parametric tests consistently indicate superior post-test performance by the experimental group, supporting Hypothesis 1.

Next, the intervention's effect on specific writing components (Content, Organization, Language Use; each scored out of 20 points) was examined. For the Content sub-score, Levene's test indicated equal variances ($F(1, 58) = 1.23$, $p = 0.272$); an independent-samples *t*-test showed no significant difference between the experimental group ($M = 17.05$, $SD = 2.15$) and the control group ($M = 16.10$, $SD = 2.30$), $t(58) = 1.68$, $p = 0.098$. This comparison is shown in Table 4.

For the Organization and Language Use sub-scores, Levene's tests indicated unequal variances (Organization: $F(1, 58) = 5.15$, $p = 0.027$; Language Use: $F(1, 58) = 8.82$, $p = 0.004$). Therefore, Welch's *t*-tests were used. For Organization, a significant difference was found, with the experimental group ($M = 18.12$, $SD = 1.95$) outperforming the control group ($M = 15.65$, $SD = 2.55$), Welch's $t(51.88) = 4.31$, $p < 0.001$. For Language Use, a large significant difference was observed, with the experimental group ($M = 19.20$, $SD = 1.40$) scoring substantially higher than the control group ($M = 13.75$, $SD = 2.80$), Welch's $t(40.15) = 9.55$, $p < 0.001$. These component score comparisons requiring Welch's *t*-test are detailed in Table 5.

Table 5 Welch's *t*-test comparisons for organization and language use component post-test scores

Component	Group	*M*	*SD*	Levene's *F* (*p*)	Test Statistic	*p*	MD [95% CI]
Organization	CTRL ($n=30$)	15.65	2.55	5.15 (0.027)*	$t(51.88)=4.31$	< 0.001	2.47 [1.32, 3.62]
(Max 20)	EX ($n=30$)	18.12	1.95				
Language Use	CTRL ($n=30$)	13.75	2.80	8.82 (0.004)*	$t(40.15)=9.55$	< 0.001	5.45 [4.29, 6.61]
(Max 20)	EX ($n=30$)	19.20	1.40				

MD Mean Difference; *CI* Confidence Interval. Mean difference reported as EX—CTRL. Max score per component = 20 points. *$p < 0.05$ for Levene's test indicates unequal variances assumed

5.3 Correlations between learning behaviors and writing outcomes (experimental group only) (RQ2 and RQ3)

Within the experimental group ($n=30$), Pearson correlation coefficients (r) were calculated to explore relationships between the frequency of specific Grammarly usage behaviors (e.g., measured via platform logs or self-reports) and writing outcomes.

First, correlations with the total post-test score were examined (Table 6). Significant positive correlations were found between the total score and frequency of using limited generative AI features (practice: $r=0.346$, $p=0.049$; assignment: $r=0.373$, $p=0.032$), frequency of using grammar/vocabulary corrective features (practice: $r=0.374$, $p=0.032$; assignment: $r=0.348$, $p=0.047$), and total number of revisions (practice: $r=0.359$, $p=0.040$; assignment: $r=0.375$, $p=0.032$).

Next, correlations between usage behaviors and specific writing component scores were explored (Table 7). Use of limited Generative AI features correlated significantly with Content scores (practice and assignment) and Organization scores (practice and assignment), as well as Cohesion and Consistency (assignment only). Use of Grammar and Vocabulary corrective features correlated significantly only with the Grammar and Vocabulary component score (practice and assignment).

Finally, the relationship between using limited generative AI features and revision frequency was examined as shown in Table 8. Significant positive correlations indicated that more frequent use of these features was associated with a higher total number of revisions during both practice ($r=0.421$, $p=0.012$) and assignment activities ($r=0.398$, $p=0.018$).

Table 6 Correlations between learning behaviors and total post-test score (Ex. Group)

Learning Behaviour (N = 30)	Correlation with total post-test score (*r*)
Total Use of Generative AI Features (Practice Activity)	0.346*
Total Use of Generative AI Features (Assignment Activity)	0.373*
Total Use of Grammar and Vocabulary Features (Practice)	0.374*
Total Use of Grammar and Vocabulary Features (Assignment)	0.348*
Total Revisions (Practice Activity)	0.359*
Total Revisions (Assignment Activity)	0.375*

*$p < 0.05$ (2-tailed)

Table 7 Correlations between learning behaviors and specific writing aspect scores (experimental group)

Learning Behaviour	Content (*r*)	Organization (*r*)	Grammar and Vocabulary (*r*)	Cohesion and Consistency (*r*)
GAI Features (Practice Activity)	0.360*	0.346*	0.285	0.325
Generative AI Features (Assignment Activity)	0.369*	0.389*	0.312	0.369*
Grammar and Vocabulary Features (Practice Activity)	0.253	0.211	0.415*	0.305
Grammar and Vocabulary Features (Assignment Activity)	0.298	0.265	0.395*	0.342

$n=30$. Cohesion and Consistency score was likely part of the Language Use component but separated for this specific analysis in the original text. *$p < 0.05$ (2-tailed)

Table 8 Correlations between generative AI feature use and total revisions (Ex. Group)

Activity	Correlation (*r*)
Practice	0.421*
Assignment	0.398*

*$p < 0.05$ (2-tailed)

5.4 Regression analysis: predicting post-test performance (ex. group only)

To assess the predictive value of assignment performance and specific AI-related behaviors on post-test scores within the experimental group ($n = 30$), a hierarchical multiple regression was conducted. Predictors were entered sequentially: Model 1 contained total assignment score; Model 2 added frequency of generative AI use (assignment); Model 3 added total revisions (practice). Multicollinearity diagnostics were within acceptable limits (Tolerance > 0.86, VIF < 1.16).

Table 9 indicates that assignment score was a significant initial predictor ($R^2 = 0.162$). The model significantly improved when generative AI usage was added ($\Delta R^2 = 0.113$), resulting in a model where both assignment score ($\beta = 0.38$) and GenAI usage ($\beta = 0.34$) were significant unique predictors ($R^2 = 0.275$). Revision frequency did not significantly add to the prediction ($\Delta R^2 = 0.010$).

The hierarchical regression analysis sought to identify significant predictors of post-test writing performance within the experimental group. The results indicate that performance on prior assignments (Total Assignment Score) served as a significant baseline predictor, explaining approximately 16% of the variance in post-test scores (Model 1). Importantly, adding the frequency of using Grammarly's generative AI features during assignment activities (Model 2) significantly improved the model's explanatory power, accounting for an additional 11.3% of the variance. In this final significant model (Model 2), both prior assignment performance ($\beta = 0.38$) and the use of generative AI features during assignments ($\beta = 0.34$) emerged as significant, independent predictors of higher post-test scores.

Interestingly, adding the total number of revisions made during practice activities (Model 3) did not significantly contribute further to predicting post-test scores once assignment performance and generative AI usage were already accounted for. This suggests that, within this group and model, assignment performance and engagement with the tool's generative features during assignments were the most salient factors predicting success on the final writing assessment.

5.5 Qualitative findings: student experiences with AI-assisted feedback (RQ1, RQ2, and RQ3)

To gain deeper insights into student experiences and perceptions regarding the use of Grammarly Premium, thematic analysis was conducted on semi-structured interview data collected from 25 participants in the experimental group. The analysis revealed three overarching themes related to the participants' interactions with the AI-assisted feedback tool: (1) Enhanced Efficiency and Confidence in the Writing Process, (2) Perceived Improvements in Accuracy and Language Awareness, and (3) The Necessity of Critical Engagement and Desire for Broader Language Support. These themes are summarized in Table 10 and elaborated below with illustrative participant comments.

Table 9 Hierarchical multiple regression analysis predicting post-test scores (Ex. Group)

Model	Predictor(s)	*B*	*SE B*	β	R^2	ΔR^2	*FΔ*	*pΔ*
1	(Constant)				0.162	0.162	5.98*	0.02*
Total Assignment Score		0.59	0.24	0.40*				
2	(Constant)				0.275	0.113	4.69*	0.04*
Total Assignment Score		0.57	0.23	0.38*				
Total Use of Generative AI (Assignment)		0.13	0.06	0.34*				
3	(Constant)				0.285	0.01	0.42	0.52

β represents standardized regression coefficients from the final significant model step (Model 2). Δ indicates the change introduced at each step; ΔR^2 quantifies the added explained variance, while FΔ and pΔ test its significance. Model 3 predictors are omitted because the change in variance (ΔR^2) was non-significant. n = 30. *p < 01

Theme 1: Enhanced Efficiency and Confidence in the Writing Process

A predominant sentiment among participants was that using Grammarly significantly improved the efficiency of their revision process. Many described how the tool allowed them to quickly identify and address surface-level errors, freeing up cognitive resources to concentrate on more substantial aspects of their writing. As one participant stated, *"It just made fixing mistakes much faster… I could spend more time thinking about my ideas instead of worrying about every comma."*

This perceived streamlining of revisions was often linked to an increase in writing confidence. Several participants expressed feeling more assured about the quality of their work after using the tool, knowing that potential errors had been flagged. One student commented, *"I felt more sure submitting my work, knowing Grammarly had checked it. It reduced my anxiety about making silly mistakes."* Furthermore, for some, interacting with the tool's suggestions prompted further learning, with several participants mentioning they were encouraged to consult external grammar resources to better understand the feedback, suggesting a move towards more autonomous learning exploration.

Theme 2 Perceived Improvements in Accuracy and Language Awareness

Participants widely perceived Grammarly as effective in improving the technical quality of their writing. Many highlighted how the tool helped them produce texts with greater grammatical accuracy and overall clarity. *"Grammarly caught so many small errors I would have missed,"* remarked one participant, adding, *"My writing definitely became clearer after using it."*

Beyond simple error correction, a number of students felt that the tool contributed to their language learning process. They appreciated the explanations often accompanying suggestions, which they felt enhanced their understanding of specific grammar rules. For example, a student explained, *"Seeing the explanations sometimes helped me understand why it was wrong, not just that it was wrong. That helps me learn for next time."* This suggests that, for these participants, the tool served not just as a proofreader but also as a supplemental learning resource.

Theme 3 The Necessity of Critical Engagement and Desire for Broader Language Support

While acknowledging the tool's benefits, participants also articulated a more nuanced perspective, emphasizing the crucial role of their own judgment. A significant number stressed the importance of critically evaluating Grammarly's suggestions rather than accepting them passively. *"You can't just accept everything,"* cautioned one participant, *"sometimes the suggestion changes your meaning, so you have to think carefully."* Concerns about potential over-reliance were also voiced, with students recognizing the need to develop their own editing skills independently.

Alongside this need for critical engagement, participants expressed desires for AI support that extends beyond grammatical mechanics. Several students highlighted the limitations of the tool in assisting with aspects like achieving a natural tone or appropriate style, particularly relevant for EFL learners navigating different communicative contexts. One participant noted, *"Sometimes the suggestions sound too formal or robotic, not natural for the way I want to write. It needs help with fluency too."* This indicates a perceived gap between the tool's capabilities and the broader linguistic needs of advanced EFL writers.

5.6 Integration of quantitative and qualitative findings

This section synthesizes the quantitative results comparing the experimental (AI-assisted feedback) and control (traditional feedback) groups with the qualitative insights gathered from interviews with experimental group participants. The integration aims to provide a more comprehensive understanding of the impact of AI-assisted feedback on EFL writing, addressing the key areas investigated by the research questions. A joint display summarizing this integration is presented in Table 11.

Regarding Overall Writing Proficiency (RQ1):

The quantitative analysis demonstrated a statistically significant advantage for the experimental group on the total post-test score (Table 2), supporting H1. This finding is complemented by the qualitative data (Theme 1, Table 10), where a large majority of participants reported feeling significantly more confident about their writing after using Grammarly. As one student noted, the tool *"reduced my anxiety about making silly mistakes."* The quantitative improvement in proficiency may, therefore, be linked not only to direct error correction but also to the affective benefits of increased confidence and potentially reduced writing apprehension reported qualitatively.

Regarding Specific Writing Components (RQ3): The quantitative results showed a differentiated impact across writing components. The experimental group made significantly larger gains in Language Use and Organization compared to the control group, but the groups did not differ significantly on Content scores (Tables 3, 4 and 5), partially supporting H3. This pattern aligns well with the qualitative findings. Participants widely perceived improvements in accuracy and clarity (Theme 2, Table 10), directly relating to Language Use, with one commenting, *"My sentences are definitely clearer*

Table 10 Thematic analysis of student interviews on experiences with grammarly premium (n = 25)

Theme	Codes	Illustrative aspects/key sentiments reported	Example quotes	Freq	(%)
1. Enhanced Efficiency and Confidence in the Writing Process	INV-I-1	Streamlined revision process, allowing focus on higher-order concerns	*"It made fixing small mistakes so much quicker, so I could actually think about my argument."*	18	72%
	INV-I-2	Boosted confidence due to error identification and perceived improvement	*"I felt less nervous submitting the paper because I knew Grammarly caught things I might miss."*	20	80%
	INV-I-3	Encouraged exploration of external grammar resources to understand suggestions	*"Sometimes I didn't understand a suggestion, so I looked it up online. It made me check the rules."*	15	60%
2. Perceived Improvements in Accuracy and Language Awareness	INV-II-1	Improved grammatical accuracy and clarity of writing	*"My sentences are definitely clearer now. It catches run-ons and awkward phrasing."*	22	88%
	INV-II-2	Enhanced understanding and learning of specific grammar rules via feedback/explanations	*"Seeing the explanation for why a comma was needed helped me learn the rule, not just fix it."*	19	76%
	INV-II-3	Streamlined revision process, contributing to overall perceived quality improvement (mentioned again in this context)	*"The whole editing process felt smoother and less overwhelming."*	17	68%
3. Critical Engagement and Need for Broader Language Support	INV-III-1	Importance of critical thinking, evaluating suggestions, and avoiding over-reliance	*"You have to be careful and read the suggestions; sometimes they change your meaning or sound wrong."*	14	56%
	INV-III-2	Desire for AI support beyond mechanics, focusing on natural language use, fluency, and appropriate style for EFL context	*"It's good for grammar, but I wish it helped more with sounding natural, not so formal or simple."*	12	48%

Themes derived from thematic analysis of semi-structured interviews with 25 experimental group participants. Frequency (n) indicates the number of participants whose responses contained evidence related to the specific code/aspect. Percentages calculated based on N = 25. INV = Interview

Discover Education (2025) 4:170 | https://doi.org/10.1007/s44217-025-00602-7

Table 11 Joint display integrating quantitative and qualitative findings

Research question	Quantitative findings	Qualitative findings summary	Integrated inference
Overall Writing Proficiency (RQ1)	EX group significantly higher total post-test score vs CTRL group (Welch's *t*, M-W *U*). (Table 2)	Theme 1: Boosted confidence reported by majority (80%)	AI-assisted feedback associated with statistically higher proficiency scores, potentially supported by increased student confidence reported qualitatively
Specific Writing Components (RQ3)	EX group significantly higher scores in Language Use (large diff.) and Organization; No significant difference in Content versus CTRL group (Indep./Welch's *t*). (Tables 3, 4, 5)	Theme 2: Perceived improvements in accuracy/clarity (88%). Theme 3: Desire for support with natural language/fluency (48%)	AI tool effectively improved aspects it directly targets (language, structure), aligning with its function; minimal impact on content confirmed by both QUAN results and QUAL reports of unmet needs (fluency)
Revision Process and Frequency (RQ2)	Higher usage frequency and revision frequency correlated positively with EX group post-test scores (Table 6). GenAI usage correlated with higher revisions (Table 8)	Theme 1: Revision process perceived as streamlined/faster (72%)	AI feedback appears to facilitate more frequent revisions, which is linked to better outcomes. Qualitative data suggests this may be due to increased efficiency/reduced burden of the revision task
Engagement and Learning (RQ2 and Qualitative Themes)	GenAI feature usage during assignments significantly predicted post-test scores in EX group, controlling for prior performance (Regression, Table 9). Usage correlated with scores (Table 6)	Theme 1: Encouraged exploration of resources (60%). Theme 2: Enhanced grammar learning perceived (76%). Theme 3: Critical thinking needed (56%)	Engaging with the tool (esp. generative features) predicts better scores. Qualitatively, this engagement involves learning and efficiency gains, but students recognize the concurrent need for critical evaluation
Tool Limitations and User Perspective (Qualitative)	N/A (Focus is on QUAL nuances)	Theme 3: Importance of critical thinking/avoiding over-reliance (56%); Desire for broader support (fluency/style) (48%)	While generally beneficial, students perceive limitations in the tool's scope (beyond mechanics) and emphasize the crucial role of user judgment and critical thinking when interacting with AI feedback

EX Experimental Group; *CTRL* Control Group; *QUAN* Quantitative; *QUAL* Qualitative

now." They also frequently mentioned how streamlining revisions (Theme 1) allowed focus on higher-level aspects, potentially aiding Organization. However, the qualitative data also highlighted limitations (Theme 3), specifically the desire for support with fluency and natural language use, and implicitly, less perceived impact on idea generation or argumentation (Content). The combined data suggests the AI tool was highly effective for form and structure but less so for content development, consistent with Grammarly's primary functions.

Regarding Revision Frequency and Engagement (RQ2):

Quantitative correlations within the experimental group indicated that higher frequency of using Grammarly's features (both corrective and limited generative) and a higher number of revisions were positively associated with better post-test scores (Table 6). Furthermore, using generative features correlated positively with making more revisions (Table 8), and generative feature usage during assignments significantly predicted post-test scores even after accounting for prior assignment performance (Table 9), suggesting a link between engagement and outcomes (supporting H2 implicitly via correlation).

The qualitative data provides explanatory depth: participants described the revision process as "streamlined" and "smoother" (Themes 1 and 2, Table 10), suggesting the tool made engaging in revision less burdensome. While the quantitative data points to the *frequency* of interaction being beneficial, the qualitative data highlights the *perceived ease and confidence* associated with this interaction, potentially driving engagement.

However, the qualitative data also introduces a critical nuance (Theme 3) regarding the need for thoughtful engagement, as captured by one student's caution: *"You can't just accept everything… you have to think carefully."* This suggests that while frequency of use correlates with better scores, the *quality* of engagement (critical evaluation) is also crucial, a factor not fully captured by the quantitative usage metrics.

Overall, integrating the findings presents a cohesive picture: AI-assisted feedback via Grammarly significantly enhanced overall writing proficiency, primarily through improvements in language accuracy and organization, aligning with the tool's strengths. This quantitative improvement is supported by qualitative reports of a more efficient revision process, increased confidence, and enhanced grammar awareness.

However, both datasets point to limitations regarding content development and the critical need for users to actively evaluate suggestions rather than relying passively on the technology.

In summary, the integration of quantitative and qualitative findings, as displayed in Table 11, provides a robust and multifaceted understanding of the impact of AI-assisted feedback via Grammarly in this EFL context. The statistically significant improvements in overall writing proficiency, particularly in Language Use and Organization, observed in the experimental group are strongly corroborated by participant reports of enhanced accuracy, clarity, confidence, and a more streamlined revision process.

Concurrently, the lack of significant quantitative improvement in Content scores resonates with qualitative feedback highlighting the tool's limitations beyond mechanics and the expressed need for support with natural language fluency. Furthermore, while quantitative data links higher tool usage and revision frequency to better outcomes, the qualitative insights underscore that this engagement is facilitated by perceived efficiency gains but must be accompanied by critical evaluation from the learner to be most effective.

Collectively, the evidence suggests that while AI feedback tools like Grammarly offer significant advantages for improving specific aspects of EFL writing and boosting student confidence, their optimal use requires critical engagement from students, and their capabilities may not fully address all dimensions of advanced academic writing, particularly concerning content generation and nuanced language use.

6 Discussion

This study aimed to provide empirical insights into the comparative effects of AI-assisted writing feedback (via Grammarly Premium) and traditional teacher feedback on EFL postgraduate students' writing proficiency, specific writing components, and revision processes. The integrated quantitative and qualitative findings offer a nuanced perspective, largely supporting the hypotheses while also highlighting important considerations for pedagogy and technology development.

6.1 Impact on writing proficiency and specific components (RQ1 and RQ3)

The central finding confirmed H1: students using AI-assisted feedback demonstrated significantly greater improvements in overall post-test writing proficiency compared to those receiving traditional teacher feedback (Table 2). This aligns

with meta-analytic trends suggesting the general efficacy of technology-mediated feedback [29, 41] and specific studies showing positive impacts of AI tools on writing outcomes (e.g., [25, 40]). The qualitative findings provide potential affective explanations, suggesting the AI tool's ability to streamline revisions and identify errors boosted user confidence (Theme 1, Table 10), potentially reducing writing anxiety and fostering greater engagement, factors known to influence performance.

However, the analysis of specific writing components revealed a more differentiated picture, partially supporting H3. The largest gains for the AI group were in Language Use, followed by significant improvements in Organization, but no significant difference was found in Content scores compared to the control group (Tables 3, 4 and 5). This strongly supports the literature distinguishing the capabilities of tools like Grammarly—primarily focused on surface-level correctness and clarity [1, 16]—from the more holistic feedback often required for higher-order concerns [8].

The significant improvement in Organization suggests that feedback on clarity and conciseness may indirectly aid structural development. Conversely, the lack of significant improvement in Content aligns with participant desires for support beyond mechanics (Theme 3, Table 10) and reinforces the idea that feedback on argumentation, idea development, and substance often requires nuanced human judgment [3], which this type of AI tool currently provides less effectively. This finding also underscores the importance of accurately classifying the AI tool, interpreting these results requires acknowledging Grammarly's primary role as a corrective feedback assistant, not a comprehensive content generator or evaluator.

6.2 Influence on revision frequency and engagement (RQ2)

The study explored AI's influence on revision and engagement. Quantitative correlations within the experimental group showed positive associations between the frequency of using AI features (both corrective and limited generative) and total revisions, and between these behaviors and post-test scores (Tables 6, 8 and 9). This suggests that increased interaction with the tool and more frequent revisions are linked to better outcomes, implicitly supporting H2. The qualitative data (Theme 1, Table 10) helps explain this linkage, with participants describing the AI-facilitated revision process as "streamlined" and "faster," reducing the perceived burden and potentially encouraging more iterative work, consistent with suggestions by Song and Song [42] regarding immediacy.

However, simply correlating frequency with outcomes doesn't capture the complexity of engagement (as noted by R1). The qualitative findings (Theme 3, Table 10) introduce a crucial counterpoint: the necessity of *critical* engagement. Participants recognized the pitfalls of over-reliance and the need to evaluate suggestions thoughtfully. This implies that while the *quantity* of interaction may correlate with improvement (perhaps reflecting overall time-on-task or attention to feedback), the *quality* of that interaction—the critical appraisal of AI feedback—is vital for meaningful learning, a point echoed in concerns raised by Dux Speltz [12]. The regression analysis further nuances this, showing that generative feature usage during assignments (a specific type of engagement) significantly predicted scores, while sheer revision frequency (during practice) did not add predictive power once assignment performance and GenAI usage were controlled (Table 9).

6.3 Theoretical implications

The findings can be interpreted through the lenses of Sociocultural Theory (SCT) and Cognitive Load Theory (CLT). From an SCT perspective [46, 53], Grammarly functioned as an effective mediational tool for aspects of writing involving explicit rules (grammar, mechanics) and structure (clarity, conciseness leading to better Organization). It successfully scaffolded learning in these domains where algorithmic feedback suffices. However, its limitations in improving Content scores suggest it is less effective in mediating the complex, dialogic negotiation of meaning inherent in developing argumentation and substance, where human interaction and feedback remain paramount. The qualitative emphasis on critical thinking further highlights that the AI tool serves primarily as an external resource requiring learner agency for effective internalization, rather than a fully reciprocal learning partner.

From a CLT perspective [43], Grammarly may have reduced extraneous cognitive load associated with identifying and correcting surface errors, freeing learners' cognitive resources to focus on Organization, potentially explaining the significant gains in that area. The immediate, constant feedback could facilitate learning by addressing errors promptly. However, the potential for overload exists if students do not engage critically, passively accepting numerous suggestions without processing them, which could hinder deeper learning. The lack of Content improvement might also suggest the tool did not effectively manage the intrinsic load associated with this complex task or provide appropriate scaffolding.

6.4 Limitations of the study

Several limitations temper the interpretation of these results. **Contextual limitations** include the specific sample (Egyptian postgraduates) and the single AI tool (Grammarly Premium), restricting generalizability. **Temporal limitations** arise from the short 6-week intervention, requiring longitudinal research for long-term effects. **Methodological limitations** involve the reliance on subjective/manual usage and revision metrics, potential instructor-researcher bias, and the intrinsic confound of differing feedback delivery times between conditions.

6.5 Implications for pedagogy and future research

Despite limitations, the study offers practical implications. AI feedback tools like Grammarly can be valuable pedagogical assets, particularly for improving language accuracy and organization and potentially boosting student confidence. They can efficiently handle lower-order concerns, allowing instructors to focus feedback and class time on higher-order thinking, content development, and argumentation. However, successful integration requires explicit instruction on *how* to use these tools critically—evaluating suggestions, understanding their limitations, and maintaining ownership of the writing. Educators should design tasks that necessitate higher-order thinking beyond what the AI can easily address and foster metacognitive awareness about the appropriate use of such tools.

Future research should explore the long-term effects of different AI feedback tools across diverse learner populations and proficiency levels. Comparative studies involving more advanced GAI are needed, carefully distinguishing feedback functions from text generation capabilities. Investigating specific pedagogical strategies for integrating AI feedback effectively and promoting critical engagement is crucial. Furthermore, developing and evaluating AI tools designed with specific EFL learner needs in mind—particularly regarding natural language use, fluency, and culturally relevant feedback—remains an important avenue for research and development, ideally through collaboration between educators, linguists, and AI developers. Addressing the ethical dimensions, including plagiarism, bias, and the impact on learner autonomy and critical thinking, must remain central to this ongoing research agenda.

7 Conclusion

This study set out to empirically compare the impact of AI-assisted writing feedback, specifically using Grammarly Premium, with traditional teacher feedback on EFL postgraduate students' writing proficiency, revision processes, and specific writing quality components. Integrating quantitative and qualitative findings provides a synthesized understanding: the AI-assisted feedback significantly enhanced overall writing proficiency compared to traditional feedback, primarily driven by substantial improvements in Language Use and Organization. However, no significant gains were observed in the Content dimension. Qualitatively, students using the AI tool reported increased confidence, perceived improvements in accuracy and clarity, and a more efficient revision process, corroborating the quantitative gains in form and structure. Crucially, though, they also emphasized the necessity of critical engagement with AI suggestions and highlighted limitations concerning support for natural language fluency.

The nuanced results underscore that while AI feedback tools like Grammarly can be effective for improving targeted aspects of writing, particularly surface-level accuracy and organization, they are not a panacea for comprehensive writing development in EFL contexts. The lack of significant improvement in Content scores, coupled with students' expressed need for support beyond mechanics, points to the limitations of relying solely on such tools for fostering higher-order thinking skills like argumentation and nuanced expression. This study contributes valuable empirical evidence differentiating the effects of readily available AI corrective feedback tools from traditional methods, highlighting both their potential benefits (efficiency, confidence, specific skill improvement) and their current shortcomings (limited content/fluency support).

These findings yield actionable insights for pedagogy. AI writing assistants can serve as valuable supplements, particularly for managing surface errors and improving organization, thereby freeing instructor time for higher-order concerns. However, effective integration demands more than simply providing access; educators must actively teach students how to engage critically with AI feedback, evaluating suggestions for relevance and accuracy rather than

passively accepting them. Direct instruction should continue to prioritize content development, critical argumentation, and nuanced stylistic choices—areas where current AI tools offer less support. Emphasizing learner autonomy involves not only leveraging AI for efficiency but also cultivating the judgment needed to use it wisely.

Furthermore, this research informs future AI tool development and research directions. There is a clear need, articulated by learners, for AI tools tailored more specifically to EFL writers, potentially incorporating features that better support natural language use, idiomatic expression, and stylistic appropriateness. Future research should employ longitudinal designs to track long-term effects, investigate a wider array of AI tools (including more advanced generative models, while carefully controlling for their distinct functionalities), and explore diverse pedagogical integration models. Such research must also continue to address ethical considerations, including mitigating risks of over-reliance that could stifle independent skill development, addressing potential algorithmic bias, and ensuring AI tools are used to augment, not replace, the crucial elements of human interaction and critical thinking in education.

In conclusion, AI-assisted writing feedback tools like Grammarly offer demonstrable benefits for enhancing specific EFL writing skills and improving the revision experience. However, their effective and ethical integration into educational practice requires a balanced pedagogical approach, mindful of their current limitations and dedicated to fostering critical, autonomous learners who can leverage technology judiciously as one component of their writing development toolkit.

Author contributions M.M. conceived and designed the study, collected and analyzed the data, and wrote the entire manuscript, including all text, tables, figures, and appendices. M.M. is solely responsible for all aspects of this research project.

Funding This research received no external funding. However, if the manuscript is accepted, the Science and Technology Development Fund (STDF) in Egypt will cover the open access publication fees.

Data availability The data that supports the findings of this study is available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate "This study was conducted in accordance with the Institutional Research Board Guidelines of Beni-Suef University, which adheres to the directives of the Supreme Council of Universities in Egypt (Correspondence dated 28/03/2023, Article 25). This research is exempt from formal ethics approval per Article 25, as it falls under the category of humanities and social sciences research that does not involve sensitive human subject treatments."

Informed consent While informed consent is routinely obtained from participants in research conducted by the authors, the Supreme Council of Universities in Egypt has waived the requirement for formal Research Ethics Board (REB) approval for humanities and social sciences research that does not involve interventions or treatments of sensitive human subjects (correspondence dated 28/03/2023, Article 25). This study falls under this category. However, informed consent was obtained from all participants before their participation in this study.

Competing interests The authors declare no competing interests.

References

1. Ali Z. Artificial intelligence (AI): a review of its uses in language teaching and learning. IOP Conf Ser Mater Sci Eng. 2020;769: 012043. https://doi.org/10.1088/1757-899x/769/1/012043.
2. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics. 2020;52(4):200–2. https://doi.org/10.1152/physiolgenomics.00029.2020.
3. Alharbi MS. The effect of the process-oriented approach and the use of corrective feedback in English writing (Order No. 30314643) [Doctoral dissertation, University of Arizona]. ProQuest Dissertations and Theses Global. 2023. https://www.proquest.com/dissertations-theses/effect-process-oriented-approach-use-corrective/docview/2836737480/se-2.
4. Bernacki ML, Greene MJ, Lobczowski NG. A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? Educ Psychol Rev. 2021;33(4):1387–429. https://doi.org/10.1007/s10648-021-09615-8.

5. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77–101. https://doi.org/10.1191/1478088706qp063oa.
6. Campello B, Agostinho R, Roazzi A. ChatGPT, the cognitive mediation networks theory and the emergence of sophotechnic thinking: how natural language AIs will bring a new step in collective cognitive evolution. SSRN Preprint. 2023. https://doi.org/10.2139/ssrn.4405254.
7. Caprioglio A, Paglia L. Fake academic writing: ethics during chatbot era. Eur J Paediatr Dent. 2023;24(2):88–9. https://doi.org/10.23804/ejpd.2023.24.02.01.
8. Cheng X, Zhang LJ. Sustaining university English as a foreign language learners' writing performance through provision of comprehensive written corrective feedback. Sustainability. 2021;13(15):8192. https://doi.org/10.3390/su13158192.
9. Crompton H, Edmett A, Ichaporia N, Burke D. AI and English language teaching: affordances and challenges. In: British journal of educational technology. Advance online publication. 2024. https://doi.org/10.1111/bjet.13460
10. Diaz Maldonado YL. Aprendizaje en inglés en línea: efectos sobre la elaboración de textos escritos y relación con la percepción de los docentes sobre la integración de tecnología (Order No. 29060217) [Doctoral dissertation, Universidad De Puerto Rico]. ProQuest Dissertations and Theses Global. 2022. https://www.proquest.com/dissertations-theses/aprendizaje-en-inglés-línea-efectos-sobre-la/docview/2679795489/se-2.
11. Dodigovic M. Artificial intelligence in second language learning: raising error awareness. Multilingual Matters. 2005.
12. Dux Speltz E. Developing and evaluating an approach to individualized, automated, process-focused feedback for university writing instruction (Order No. 30693231) [Doctoral dissertation, University of Pittsburgh]. ProQuest Dissertations and Theses Global. 2023 https://www.proquest.com/dissertations-theses/developing-evaluating-approach-individualized/docview/2919582151/se-2.
13. Emerson N. AI-enhanced collaborative story writing in Japanese university EFL classes. Technol Lang Teach Learn. 2024;6(3):1764. https://doi.org/10.29140/tltl.v6n3.1764.
14. Escalante N, Yang Y, Chen L. Student perceptions of AI-generated feedback versus teacher feedback in EFL writing. Comput Assist Lang Learn. 2023;36(5–6):815–35. https://doi.org/10.1080/09588221.2022.2159781.
15. Faisal E. Unlock the potential for Saudi Arabian higher education: a systematic review of the benefits of ChatGPT. Front Edu. 2024;9:1325601. https://doi.org/10.3389/feduc.2024.1325601.
16. Gayed JM, Carlon MKJ, Oriola AM, Cross JS. Exploring an AI-based writing assistant's impact on English language learners. Comput Edu Artif Intell. 2022;3: 100055. https://doi.org/10.1016/j.caeai.2022.100055.
17. Ghafouri M, Hassaskhah J, Mahdavi-Zafarghandi A. From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. In: Language Teaching Research. Advance online publication. 2024. https://doi.org/10.1177/13621688241239764.
18. Gibson D, Kovanovic V, Ifenthaler D, Dexter S, Feng S. Learning theories for artificial intelligence promoting learning processes. Br J Edu Technol. 2023;54(5):1135–55. https://doi.org/10.1111/bjet.13341.
19. Halkiopoulos C, Gkintoni E. Leveraging AI in E-learning: personalized learning and adaptive assessment through cognitive neuropsychology—a systematic analysis. Electronics. 2024;13(18):3762. https://doi.org/10.3390/electronics13183762.
20. Hutson M. Robo-writers: the rise and risks of language-generating AI. Nature. 2021;591(7848):22–5. https://doi.org/10.1038/d41586-021-00530-0.
21. Hwang WY, Nurtantyana R, Purba SWD, Hariyanti U, Indrihapsari Y, Surjono HD. AI and recognition technologies to facilitate English as foreign language writing for supporting personalization and contextualization in authentic contexts. J Edu Comput Res. 2023;61(5):1008–35. https://doi.org/10.1177/07356331221137253.
22. Kartal G. Transforming the language teaching experience in the age of AI. IGI Global. 2023.
23. Khor ET, Mutthulakshmi K. A systematic review of the role of learning analytics in supporting personalized learning. Edu Sci. 2024;14(1):51. https://doi.org/10.3390/educsci14010051.
24. Kim J, Lee S. Actionable insights or generic comments? EFL learners' evaluation of AI writing feedback specificity. System. 2024;119: 103178. https://doi.org/10.1016/j.system.2023.103178.
25. Koltovskaia S. Automated writing evaluation for formative second language assessment: exploring performance, teacher use, and student engagement (Order No. 29257311) [Doctoral dissertation, Indiana University]. ProQuest Dissertations and Theses Global. 2022. https://www.proquest.com/dissertations-theses/automated-writing-evaluation-formative-second/docview/2771052729/se-2.
26. Lave J, Wenger E. Situated learning: legitimate peripheral participation. Cambridge: Cambridge University Press; 1991. https://doi.org/10.1017/CBO9780511815355.
27. Li C. Age of AI: explore the role of AI in personalized learning. Commun Hum Res. 2025;53(1):1–7. https://doi.org/10.54254/2753-7064/2025.lc20612.
28. Li J, Washington P. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: machine learning study. JMIR AI. 2024;3: e52171. https://doi.org/10.2196/52171.
29. Li Z, Wang L. Effectiveness of automated writing evaluation feedback on EFL writing: a meta-analysis. Comput Assist Lang Learn. 2022;35(9):2148–74. https://doi.org/10.1080/09588221.2021.1904478. (**Note: Assumed this is the correct reference based on context**).
30. Lin CJ, Hwang GJ, Fu QK, Cao YH. Facilitating EFL students' English grammar learning performance and behaviors: a contextual gaming approach. Comput Educ. 2020;152: 103876. https://doi.org/10.1016/j.compedu.2020.103876.
31. Loncar M, Schams W, Liang J-S. Multiple technologies, multiple sources: trends and analyses of the literature on technology-mediated feedback for L2 English writing published from 2015–2019. Comput Assist Lang Learn. 2021;34(sup1):1–63. https://doi.org/10.1080/09588221.2021.1943452.
32. Nasser M. Personalized learning through AI: enhancing student engagement and teacher effectiveness. Int J Teach Learn Edu. 2024;3(6):23–6. https://doi.org/10.22161/ijtle.3.6.4.
33. Nguyen T-H, Hwang W-Y, Pham X-L, Pham T. Self-experienced storytelling in an authentic context to facilitate EFL writing. Comput Assist Lang Learn. 2020;35(4):666–95. https://doi.org/10.1080/09588221.2020.1744665.
34. Octavio MM, González V, Pujolà J-T. ChatGPT as an AI L2 teaching support: a case study of an EFL teacher. Technol Lang Teach Learn. 2024;6(1):1–25. https://doi.org/10.29140/tltl.v6n1.1142.
35. Pan F. AI in language teaching, learning, and assessment. IGI Global. 2024.

36. Piaget J. Science of education and the psychology of the child (D. Coltman, Trans.). Orion Press. 1970.
37. Qiao H, Zhao A. Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. Front Psychol. 2023;14:1255594. https://doi.org/10.3389/fpsyg.2023.1255594.
38. Rane N, Choudhary S, Rane J. Education 4.0 and 5.0: integrating artificial intelligence (AI) for personalized and adaptive learning. SSRN Electr J. 2023. https://doi.org/10.2139/ssrn.4638365.
39. Schmidt T, Strassner T. Artificial intelligence in foreign language learning and teaching. Anglistik. 2022;33(1):165–84. https://doi.org/10.33675/angl/2022/1/14.
40. Selim M. The transformative impact of AI-powered tools on academic writing: perspectives of EFL university students. Int J Engl Linguist. 2024;14(1):14. https://doi.org/10.5539/ijel.v14n1p14.
41. Smith JA, Jones B, Williams R. The impact of AI feedback tools on L2 writing proficiency: a meta-analytic review. J Second Lang Writ. 2023;60: 101015. https://doi.org/10.1016/j.jslw.2023.101015.
42. Song C, Song Y. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Front Psychol. 2023;14:1260843. https://doi.org/10.3389/fpsyg.2023.1260843.
43. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci. 1988;12(2):257–85. https://doi.org/10.1207/s15516709cog1202_4.
44. Taylor DL, Yeung M, Bashet AZ. Personalized and adaptive learning. In: Hmelo-Silver CE, Chinn C, Chan M, O'Donnell A, editors. Innovative learning environments in STEM higher education: Opportunities, challenges, and looking forward. Springer. 2021. p. 17–34. https://doi.org/10.1007/978-3-030-58948-6_2
45. Thimmanna AVNSS. Personalized learning paths: adapting education with AI-driven curriculum. Eur Econ Lett (EEL). 2024;14(1):31–40. https://doi.org/10.52783/eel.v14i1.993.
46. Vygotsky LS. Mind in society: the development of higher psychological processes. In: Cole M, John-Steiner V, Scribner S, Souberman E, Eds. Harvard University Press. 1978.
47. Wang Z. Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. Library Hi Tech. 2020;40(1):80–97. https://doi.org/10.1108/lht-05-2020-0113.
48. Ward R. Personalised learning for the learning person. Emerald Publishing Limited. 2020.
49. Wong KM. "A design framework for enhancing engagement in student-centered learning: own it, learn it, and share it" by Lee and Hannafin (2016): an international perspective. Edu Tech Res Dev. 2020;69(1):93–6. https://doi.org/10.1007/s11423-020-09842-w.
50. Xu C, Xia J. Scaffolding process knowledge in L2 writing development: insights from computer keystroke log and process graph. Comput Assist Lang Learn. 2021;34(4):583–608. https://doi.org/10.1080/09588221.2019.1632901.
51. Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, Li X, Jin Y, Gašević D. Practical and ethical challenges of large language models in education: a systematic scoping review. Br J Edu Technol. 2023;54(5):1156–81. https://doi.org/10.1111/bjet.13370.
52. Zhang P, Tur G. A systematic review of ChatGPT use in K-12 education. Eur J Educ. 2023;59(2):281–303. https://doi.org/10.1111/ejed.12599.
53. Zhou W. Sociocultural theory and its implications in college English teaching and learning in the age of artificial intelligence. J Phys: Conf Ser. 2020;1646: 012142. https://doi.org/10.1088/1742-6596/1646/1/012142.