Innovative Research Journal of **Artificial Intelligence** (IRJAI)
DOI: https://doi.org/10.62497/irjai.274

Research Article

From AI Feedback to Language Learning: The Effects of Generative AI Feedback on EFL Learners' Revision Quality and Delayed Writing Performance

Esraa Abd Elnaser Mohamed Hassan*
College of Foreign Languages and Cultures, Chengdu University of Technology, 610059 Chengdu, China

Citation: Hassan EAEM. From AI Feedback to Language Learning: The Effects of Generative AI Feedback on EFL Learners' Revision Quality and Delayed Writing Performance. Innov. Res. J. AI. 2025;3(2):1-16. Available from: https://irjpl.org/irjai/article/view/274

Article Info
Received: Oct 10, 2025
Revised: Nov 19, 2025
Accepted: Dec 24, 2025

Keywords
stylistics, Christopher Marlowe, Doctor Faustus, graphological level, lexical level, phonological level, grammar, artificial intelligence, AI

Copyright © 2025 The Author(s).

Published by Innovative Research Journals (IRJPL).

This is an Open Access article under the CC BY NC 4.0 license. This license enables reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.

Abstract

Introduction: Generative artificial intelligence is increasingly used as a source of written feedback in second and foreign language writing. The central pedagogical question is no longer whether an AI system can identify or reformulate language problems, but whether learners can convert AI feedback into durable learning that remains visible when the tool is removed.

Objective: This study was designed to examine whether structured generative-AI feedback improved EFL learners' revision quality and whether such improvement transferred to a delayed writing task completed without AI support. It also examined how learners evaluated, accepted, rejected, and acted on AI feedback.

Method: A mixed-methods, three-arm intervention was proposed for approximately 90 university EFL writers. The protocol specified a baseline writing task and allocation to AI feedback, teacher feedback, or hybrid teacher-plus-AI feedback. It also specified four feedback–revision cycles across eight weeks. It also specified analytic-rubric assessment of immediate post-intervention writing and a two-week delayed writing task. Revision changes, feedback uptake, and learner interviews were specified as complementary sources of evidence.

Results: The illustrative analyses compared changes over time and between groups, estimated effect sizes and confidence intervals, examined the persistence of learning at delayed post-test, and identified qualitative patterns in feedback use.

Conclusion: The design distinguished short-term improvement in a revised text from evidence of retained writing development. It was intended to provide a stronger test of whether generative-AI feedback functioned as a learning resource rather than simply as an editing service.

* ***Corresponding Author:***
Esraa Abd Elnaser Mohamed Hassan
Email: hs5932457@gmail.com

Introduction

Background and rationale

Feedback occupies a central position in second language writing because writing development is not produced by drafting alone. Learners need information about the relationship between an emerging text and the criteria or goals against which that text is being judged, together with usable guidance about what can be changed and why. In classroom writing, however, feedback is never simply a message transmitted from teacher to student. Its effect depends on the quality of the information, the timing and form in which it is delivered, the learner's interpretation of it, and the actions that follow. The literature on second language writing has therefore moved from a narrow concern with error correction toward a broader view of feedback as part of a socially and cognitively mediated learning process [1,2].

The distinction between feedback that improves a particular draft and feedback that contributes to language learning is especially important. A student may submit a stronger second draft after receiving corrections, yet the improvement may remain tied to that particular text. Durable development requires learners to notice patterns, understand reasons for change, make decisions about alternatives, and apply those decisions to later writing. Hattie and Timperley's influential model conceptualizes feedback in relation to goals, current performance, and next steps, while later work on student feedback literacy emphasizes the learner's active role in interpreting and using feedback [3,4]. In other words, the educational value of feedback lies not only in what the feedback says but in what the learner does with it.

Written corrective feedback has a long history in applied linguistics, with substantial debate concerning direct correction, indirect correction, metalinguistic explanation, reformulation, and the conditions under which feedback contributes to longer-term accuracy. Research has also distinguished local language problems from higher-order dimensions such as organization, argumentation, cohesion, and rhetorical effectiveness [5,6]. These distinctions matter for AI-mediated feedback because large language models can generate fluent explanations and alternative phrasings at speed, but fluency of output does not guarantee pedagogical accuracy, contextual appropriateness, or learner uptake.

Generative artificial intelligence has changed the practical conditions under which feedback can be delivered. Large language models can respond immediately to a learner's draft, explain a grammatical choice, suggest alternatives, comment on organization, and engage in dialogue about a revision. This creates opportunities for individualized and iterative support that may be difficult for a single instructor to provide consistently in a large writing class [7, 8]. At the same time, generative systems can produce confident but inaccurate information, overcorrect acceptable language, encourage dependence, or make revisions that conceal rather than develop the learner's own decision-making. UNESCO has consequently emphasized human oversight, privacy, pedagogical purpose, and responsible use when generative AI is introduced into education [9].

The empirical literature since the introduction of widely accessible generative-AI tools has grown rapidly. A recent systematic review of empirical generative-AI research in language learning and teaching found that writing is the most common language-skill focus and that higher education and EFL settings dominate the literature. The same review identified a continuing need for longitudinal designs and research that goes beyond perceptions toward evidence of learning and sustained development [10]. This is an important methodological point for applied linguistics: a technology may be popular, easy to use, and positively perceived without necessarily producing durable changes in language ability.

Several recent intervention studies provide evidence that AI-supported feedback can be useful, but they also reveal why a more focused design is needed. Mahapatra reported positive effects of ChatGPT used as a formative feedback tool in undergraduate ESL writing, using multiple tests and qualitative data [11]. Polakova and Ivenz similarly examined ChatGPT feedback in a university EFL context and reported improvements in writing outcomes [12]. Mekheimer's study of generative-AI-assisted

feedback found relationships among proficiency, revision practices, and writing quality [13]. These findings support the pedagogical potential of AI feedback but do not by themselves resolve whether learners can transfer what they learn from feedback to a new writing task without the tool.

Comparative research also suggests that AI feedback should not be treated as interchangeable with expert human feedback. Steiss and colleagues found that human feedback was stronger than ChatGPT feedback on several dimensions of formative feedback quality, although the overall differences and time trade-offs were more nuanced [14]. Lin and Crosthwaite likewise identified differences between teacher and GPT-assisted written corrective feedback, including variability in the type and consistency of feedback generated by the model [15]. Han and Li reported that ChatGPT could support teachers by expanding the range of feedback they considered, while teachers remained important in adapting feedback to the local instructional context [16]. These studies point toward a complementary model rather than a simple replacement model.

More recent evidence strengthens the case for examining the type and persistence of learning. Yang et al., [17] found no statistically significant difference between ChatGPT-generated and teacher-generated feedback groups in post-test writing, although students held positive views of the technology [17]. Zhang and colleagues found gains from both generative-AI feedback and hybrid feedback, with the hybrid condition showing stronger development in higher-order areas such as organization and critical thinking [18]. Yang, Hussin, and Hashim focused specifically on revision and self-efficacy, while Kang's study compared direct and reformulation feedback delivered through ChatGPT and included a delayed revision task [19,20]. Together, this work suggests that the most informative next step is not another general comparison of "AI versus teacher," but a process-sensitive investigation of how feedback is taken up and whether learning survives after AI assistance is withdrawn.

The concept of feedback uptake provides a useful bridge between feedback research and second language development. Uptake can be operationalized at several levels. At the text level, a learner may change a word, sentence, paragraph, or organizational feature in response to feedback. At the cognitive level, the learner may recognize why the change is needed and be able to justify it. At the developmental level, the learner may apply a related principle to a different text. The last of these is the most difficult to demonstrate and is the focus of the delayed writing component proposed here.

A delayed post-test is therefore more than an additional measurement point. It changes the interpretation of the intervention. If the AI group improves only when revising a text that contains visible feedback, the evidence supports immediate performance enhancement. If the same learners subsequently improve on an unfamiliar prompt without AI assistance, the interpretation is stronger: the intervention may have contributed to internalized knowledge or transferable writing strategies. This distinction is consistent with the long-standing concern in L2 writing research that short-term revision gains and longer-term language development should not be treated as identical outcomes [21-23].

The study also treated revision quality as an outcome in its own right. Counting revisions is insufficient because many revisions are cosmetic, redundant, or unnecessary. A learner who accepts every AI suggestion may produce more changes while exercising less independent judgment. Conversely, a learner who rejects some suggestions after evaluating them may demonstrate stronger feedback literacy. The study therefore distinguished accurate uptake, inaccurate uptake, justified rejection, and changes unrelated to the feedback. This approach recognizes that responsible use of AI requires evaluation rather than obedience.

The theoretical framework combines feedback literacy with a process-oriented view of L2 writing. Feedback literacy provides constructs for understanding appreciation of feedback, recognition of different sources, management of affect, judgment, and action [4]. The writing-process perspective emphasizes that drafting, revising, and rereading are recursive activities rather than a one-directional sequence. The study consequently treats the learner as an agent who interacts with AI

feedback, evaluates its relevance, makes revision decisions, and gradually develops independent criteria for writing quality.

The research gap can therefore be stated precisely. Existing empirical studies have established that generative AI can provide useful writing support, and several have reported short-term gains in writing performance. However, the field still needs designs that jointly examine three linked outcomes: the quality of revisions made in response to feedback, the learner's decision-making around feedback, and performance on a later writing task completed without AI assistance. A mixed-methods design is appropriate because numerical writing scores can show whether change occurred, while revision records and interviews can help explain how learners interpreted and acted on the feedback.

The study was deliberately designed to avoid the strongest form of technological determinism. It did not assume that AI was inherently beneficial or harmful. Instead, it asked under what conditions AI feedback supported learning and whether hybrid human oversight provided an advantage over AI-only feedback. This was aligned with current evidence suggesting that AI could increase feedback availability while human educators contributed contextual judgment, instructional priorities, and the ability to recognize when a suggestion was inappropriate [24-25].

Finally, the study had practical relevance for EFL programs in which teachers faced high writing loads and students increasingly used generative AI outside class. A research design that examined learning after AI withdrawal could inform institutional policies more meaningfully than a perception survey. If delayed performance improved, AI could be incorporated as a scaffold within a structured pedagogy. If delayed performance did not improve despite strong immediate revisions, programs might need to emphasize reflection, explicit language instruction, or teacher mediation. Either outcome would be useful because it clarified what AI feedback could and could not reasonably be expected to accomplish.

Conceptual framework

The conceptual model contains four linked stages: goal orientation, feedback interpretation, revision action, and transfer. Goal orientation refers to the learner's understanding of the writing criteria and the purpose of the task. Feedback interpretation involves identifying what the feedback is pointing to, deciding whether it is accurate, and connecting it with a writing principle. Revision action is the observable change made in the text and the learner's rationale for that change. Transfer refers to the ability to use the underlying principle on a later task without immediate AI assistance. The model assumes that movement through these stages is not automatic.

AI feedback may influence each stage differently. Its immediacy may increase the quantity of feedback available during the interpretation stage. Its conversational format may support clarification. Its ability to offer alternatives may broaden the learner's repertoire. Yet these same features may reduce cognitive engagement if the learner accepts suggestions without evaluation. The hybrid condition is therefore theoretically important because teacher feedback can be used to validate, prioritize, or challenge AI suggestions rather than simply add another layer of correction.

Feedback literacy was operationalized in this study through observable behaviors as well as interview evidence. Appreciation of feedback was reflected in learners' ability to identify the purpose of feedback; judgment was reflected in the accuracy of accept/reject decisions; managing affect was explored through interview responses about confidence and frustration; and action was reflected in revision uptake and later independent writing. This operationalization avoided treating feedback literacy as a purely attitudinal construct.

The conceptual framework also distinguishes surface-level and deep-level writing. Surface-level outcomes include grammar, word choice, spelling, punctuation, and sentence-level accuracy. Deep-level outcomes include thesis clarity, organization, argument development, cohesion, paragraph unity, and rhetorical effectiveness. The distinction is supported by meta-analytic evidence showing that different types of feedback may have different effects on different writing outcomes [26-27]. The intervention therefore provided both language-

focused and rhetorical feedback, while the analysis kept these dimensions separate.

Research gap and significance

The first contribution of the study was methodological. Rather than treating a post-test as sufficient evidence of learning, it introduced a delayed writing task without AI assistance. The second contribution was process-oriented: revision quality was analyzed at the level of feedback uptake rather than only final scores. The third contribution was pedagogical: the study compared AI-only, teacher-only, and hybrid feedback, allowing the role of human mediation to be examined. The fourth contribution was qualitative: interviews were used to identify the reasoning behind acceptance, rejection, and modification of AI feedback.

This combination is important because the educational consequences of generative AI depend on how learners use it. A student who asks the system to "correct my essay" and copies the output is engaged in a different learning process from a student who asks why a phrase is inappropriate, compares two alternatives, and then revises independently. The present design makes that difference visible. It therefore addresses a central applied-linguistic issue: whether an emerging technological practice changes the mechanisms through which learners notice, interpret, and control language.

Objective

General objective: The general objective is to determine whether structured generative-AI feedback contributes to EFL learners' revision quality and durable writing development, with particular attention to performance on a delayed writing task completed without AI support.

Specific objectives: To compare changes in overall EFL writing performance among learners receiving AI-only, teacher-only, and hybrid feedback.

To examine differences among groups in the quality and accuracy of feedback uptake across repeated revision cycles.

To determine whether gains observed immediately after the intervention are maintained on a delayed writing task completed without AI assistance.

To examine whether learners differ in their ability to evaluate, accept, reject, or modify AI-generated feedback.

To explore learners' perceptions of the usefulness, credibility, limitations, and appropriate role of generative-AI feedback in academic writing.

Research questions

Does structured generative-AI feedback produce greater improvement in EFL writing performance than teacher-only feedback, and does hybrid feedback produce greater improvement than either single-source condition?

How does feedback condition affect the accuracy and quality of learners' revisions?

To what extent are immediate writing gains maintained on a delayed writing task completed without AI assistance?

How do learners explain their decisions to accept, reject, modify, or seek clarification about AI-generated feedback?

How do quantitative patterns in writing and revision quality converge with or diverge from learners' reported experiences?

Hypotheses

H1: All three intervention conditions were expected to show improvement from baseline to immediate post-test, but the hybrid condition was expected to show the largest adjusted gain in overall writing quality. H2: The AI and hybrid groups were expected to show greater revision uptake than the teacher-only group, although the hybrid group was expected to show higher accuracy of uptake than the AI-only group. H3: The hybrid group was expected to demonstrate the strongest maintenance of writing gains at delayed post-test. H4: Learners who demonstrated more accurate feedback evaluation during the intervention were expected to show stronger delayed writing performance. These hypotheses were directional expectations rather than results and were evaluated illustratively against the simulated data.

Methodology

Research design

A three-arm mixed-methods intervention was specified. The quantitative component followed a repeated-measures design with a baseline writing assessment, an immediate post-intervention assessment, and a delayed writing assessment. The qualitative component was designed to examine learners' feedback interactions, revision rationales, and semi-structured interview accounts. The two strands were intended to be integrated during interpretation so that statistical patterns could be examined alongside evidence of how learners used feedback. The design followed principles of transparent reporting for mixed-methods research and, where random allocation was feasible, could be reported using the relevant CONSORT principles. [28,29]

The target sample was approximately 90 university-level EFL students, with about 30 participants per condition. This size represented a pragmatic target for a classroom-based intervention and required confirmation through an a priori power analysis using an effect size judged educationally meaningful for the primary outcome.[30-31] The reporting plan specified explicit description of any quasi-experimental allocation involving intact classes and reporting of baseline equivalence. The draft assumed that the researcher could allocate students after baseline assessment; the final protocol required adaptation to the actual institutional constraints.

The protocol specified an academic English or EFL writing course in which all participants received the same core writing instruction. The intervention was specified as an added feedback procedure rather than a replacement curriculum. This distinction was important because the research question concerned the contribution of feedback source and process, not whether a technology-rich course was better than ordinary language teaching. The protocol specified equivalent writing tasks, instructional time, and opportunities to revise for all groups; only the source and structure of feedback differed by condition.

Participants and eligibility

The target participants were undergraduate EFL learners enrolled in a course requiring academic paragraph or essay writing in English. Eligibility criteria included regular course attendance, access to the required digital platform, and consent to the use of writing samples and feedback records for research. Students with professional experience in English-language editing or routine use of generative AI as an advanced writing service were designated for possible exclusion or treatment as a covariate, depending on the recruitment context. The report was intended to include age range, gender distribution where appropriate and ethically justified, academic major, English proficiency level, prior writing instruction, and prior generative-AI experience.

The protocol specified that proficiency be measured using the institution's established placement test or a short standardized diagnostic appropriate to the local setting. The purpose was not to label students permanently but to establish baseline comparability and allow exploratory moderation analysis. The protocol also specified measurement of prior AI experience because familiarity could influence prompt quality, trust in model output, and willingness to question suggestions. These variables were designated as descriptive and potential covariates rather than grounds for exclusion without a strong methodological reason.

Study variables and outcome measures

The primary outcome was change in analytic writing performance from baseline to immediate and delayed post-tests. Secondary outcomes were revision quality, feedback uptake accuracy, and learner-reported feedback literacy. The primary comparison focused on the interaction between time and feedback condition. The delayed post-test was treated as the most direct indicator of whether learning persisted after the immediate scaffold was removed.

The analytic writing rubric was specified to include task fulfillment, organization, development of ideas, cohesion, lexical resource, grammatical accuracy, and mechanics. The protocol specified scoring each dimension on a common scale, for example 1–6, with descriptors anchored to observable performance. Two trained raters, blind to condition, were designated to score the pre-test, post-test, and

delayed post-test scripts. A subset was designated for double-scoring to estimate inter-rater reliability. The protocol restricted raters' access to learners' AI interaction records.

The coding plan classified revision quality at the level of individual feedback units. A feedback unit is a discrete suggestion or evaluative comment that can be linked to a specific revision decision. Each unit was classified as accepted and correctly implemented, accepted but incorrectly implemented, partially implemented, rejected with an adequate rationale, rejected without evidence of evaluation, or not addressed. Revision depth was also coded as surface, sentence-level meaning, paragraph-level organization, or argument/rhetorical development. This coding scheme allowed the analysis to distinguish quantity from quality.

The protocol specified assessment of feedback literacy using a brief questionnaire adapted to the study's constructs and supplemented by behavioral evidence. Items were designed to address understanding the purpose of feedback, confidence in evaluating feedback sources, willingness to seek clarification, ability to explain revisions, and perceived responsibility for final decisions. The instrument was designated for expert review and pilot testing before use. The analysis plan specified reporting reliability with an appropriate coefficient; a high coefficient alone was not considered proof of construct validity.

The relationship between the study conditions, intervention features, and outcomes was summarized in Table 1. This presentation clarified which variables were manipulated and which were observed, and prevented the primary outcome from being confused with process measures.

Table 1: Study conditions, intervention features, and outcome measures

Condition	Feedback source	Learner activity	Primary outcome	Process evidence
AI feedback	Generative-AI system using a fixed prompt protocol	Evaluate AI comments; revise; record decisions	Writing score at immediate and delayed post-test	AI feedback log; revision uptake
Teacher feedback	Trained EFL writing instructor	Evaluate teacher comments; revise; record decisions	Writing score at immediate and delayed post-test	Teacher feedback record; revision uptake
Hybrid feedback	Generative AI followed by teacher review and prioritization	Compare sources; evaluate; revise; justify key decisions	Writing score at immediate and delayed post-test	AI log; teacher feedback; revision uptake
All groups	Same writing instruction and tasks	Draft–feedback–revision cycle	Overall and analytic writing dimensions	Attendance and task completion

AI feedback protocol

To make the AI condition reproducible, the protocol specified standardized feedback prompts. The protocol instructed learners to submit only the draft required for the study and not to request a complete rewritten essay. The system prompt was designed to direct the model to identify strengths, identify a limited number of priority problems, explain why each problem matters, and propose questions or options rather than silently replacing the student's language. The prompt also directed the system to distinguish language accuracy from rhetorical choices and to state uncertainty when a recommendation is context dependent.

The AI-only condition was not intended to reproduce unrestricted real-world use. Standardization was necessary because different prompts can generate substantially different feedback. The protocol specified archiving a copy of the exact prompt, model name/version, date of use, and relevant system settings with the research materials. The protocol also required documentation

of any platform change and assessment of its implications for comparability. The protocol prohibited entering student identifiers into the external AI system and required compliance with institutional privacy requirements [7].

The teacher-only condition was designed to use a common feedback rubric and training session. Teachers were instructed to prioritize a manageable number of issues and provide explanations that could support revision. The hybrid condition was designed to receive AI-generated feedback followed by teacher review. Teachers were not required to accept AI suggestions; instead, the protocol required them to identify which suggestions were valid, which needed modification, and which should be disregarded. This design allowed the study to test whether human mediation improved not simply the amount but the reliability and educational usefulness of AI feedback.

Intervention procedure

The intervention was specified as an eight-week program (Figure 1). Each cycle was designed to begin with a short writing task or section of a larger essay. Before drafting, learners were instructed to review the rubric and identify one or two personal goals. The protocol specified that feedback be provided according to condition after drafting. Learners were instructed to revise their drafts, mark major changes, and write brief rationales for selected decisions. The rationale requirement was important because it turned revision into an observable learning activity rather than a silent editing stage.

Cycle 1 was designed to emphasize understanding the feedback process. The protocol specified teaching students to treat AI suggestions as claims to evaluate rather than instructions to obey. The planned activities included checking whether a suggested change preserved meaning, whether a grammatical rule had actually been violated, and whether vocabulary replacement changed register or nuance. Cycle 2 focused on organization and cohesion. Cycle 3 focused on lexical and grammatical choices. Cycle 4 integrated all dimensions and required learners to identify recurring problems across their drafts.

The protocol specified a short reflection after each revision in which students identified one useful feedback comment, one comment they questioned, one change they made, and one principle they intended to apply independently. These reflections provided qualitative process data and could also support learning. Because the intervention itself was educational, the protocol specified equivalent opportunities for reflection across groups; otherwise, reflection could become a confounding intervention.

The immediate post-test was designed to use a new writing prompt matched to the baseline task in genre, approximate length, and cognitive demand. The protocol specified that no feedback be provided during the post-test. The delayed post-test was designed to use another parallel prompt after two weeks. The protocol specified disabling or otherwise prohibiting AI access during the delayed task and keeping the assessment environment the same for all groups. The purpose was to test transfer without the immediate availability of the intervention.

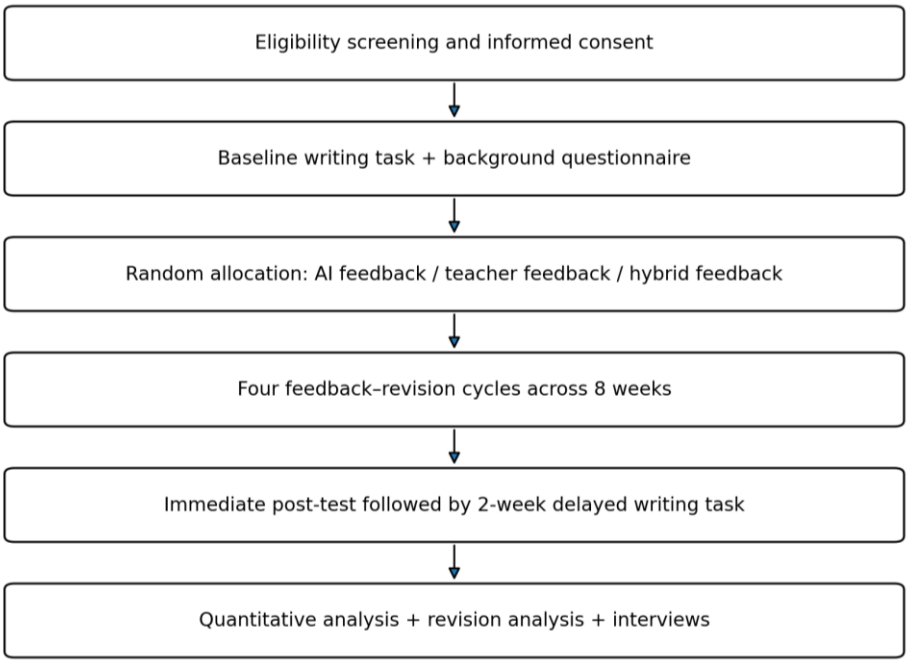


Figure 1: Planned study flow for the AI-feedback intervention

The weekly structure was shown in Table 2. The schedule was intentionally repetitive because repeated cycles made it possible to observe whether learners' feedback decisions became more selective and accurate over time rather than reflecting a single exposure to the technology.

Table 2: Illustrative eight-week intervention schedule

Week	Core activity	Feedback focus	Evidence collected
1	Orientation, baseline writing, feedback-literacy training	Understanding criteria and evaluating feedback	Baseline essay; learner background
2	Cycle 1 draft and revision	Task response and organization	Draft, feedback record, revised draft
3	Cycle 1 reflection and Cycle 2 draft	Cohesion and paragraph development	Reflection; revision decisions
4	Cycle 2 revision	Cohesion, paragraph unity, argument	Revised draft; uptake codes
5	Cycle 3 draft and feedback	Vocabulary and sentence-level language	Feedback record; draft
6	Cycle 3 revision and reflection	Lexical precision and grammar	Revision record; reflection
7	Cycle 4 integrated writing	Integrated feedback evaluation	Draft; feedback; revision
8	Cycle 4 reflection and immediate post-test	Independent performance	Reflection; post-test essay
10	Delayed post-test	Transfer without AI support	Delayed essay

Data analysis

Quantitative analysis began with data screening, missing-data inspection, distribution checks, and verification of baseline equivalence. The primary analysis used a linear mixed-effects model with writing score as the dependent variable, time and feedback condition as fixed effects, and participant as a random effect. The analysis plan specified that,

if intact classes were included, class would be modeled as a random effect or clustering would be addressed analytically. The key test was the time-by-condition interaction, followed by planned contrasts for the hybrid condition against AI-only and teacher-only conditions.

The primary outcome was reported with effect sizes and 95% confidence intervals rather than p values alone. For repeated measures, standardized mean differences or model-based standardized effects were available for reporting, with the exact estimator stated. Baseline writing score was considered as a covariate when justified. The delayed post-test was analyzed separately as a maintenance/transfer outcome rather than pooled with the immediate post-test, because performance under delayed conditions had a different educational meaning.

Revision data were analyzed using frequencies and proportions of accurate uptake, inaccurate uptake, partial uptake, justified rejection, and unaddressed feedback. A mixed-effects logistic model could examine whether feedback condition predicted accurate uptake while accounting for repeated observations within learners. The analysis plan specified that revision depth could be treated as an ordinal outcome or summarized using theoretically justified categories. Correlations between uptake accuracy during the intervention and delayed writing scores were treated as exploratory unless preregistered as confirmatory.

The analysis plan specified evaluation of rater agreement using an appropriate coefficient based on the scoring structure, such as an intraclass correlation coefficient for continuous or near-continuous rubric totals. The reporting plan required the manuscript to specify the exact model and assumptions rather than present a single reliability number without context. The protocol specified training raters on benchmark scripts before scoring and blinding them to condition whenever possible.

The qualitative analysis plan specified reflexive thematic analysis. The procedure involved repeated reading of interview transcripts, coding, review, and organization into themes explaining how learners evaluated AI feedback and understood responsibility for revisions.[30] The coding plan remained close to the research questions while allowing unexpected patterns, such as strategic use of AI for brainstorming rather than correction, distrust of certain suggestions, or increased confidence without corresponding accuracy.

Mixed-methods integration was planned through a joint display linking quantitative outcomes to qualitative cases. For example, a learner with high immediate improvement but weak delayed performance may be examined alongside interview evidence of heavy dependence on AI reformulation. Conversely, a learner with moderate immediate gains but strong delayed performance may report routinely questioning and explaining feedback. Such cases could help interpret why groups differed and prevent the qualitative findings from becoming an isolated set of quotations.

Table 3 specified the planned statistical tests and the evidence to accompany each result. The table was placed before the Results section so that outcomes could be reported against prespecified analyses rather than selected after data inspection.

Table 3: Planned analysis and reporting framework

Research outcome	Primary analysis	Effect estimate	Interpretive focus
Overall writing score	Linear mixed-effects model	Adjusted mean difference; standardized effect; 95% CI	Whether improvement differs by feedback condition
Analytic writing dimensions	Multivariate or separate mixed models with multiplicity control	Dimension-specific effects; 95% CI	Where gains occur: organization, vocabulary, grammar, etc.
Revision uptake accuracy	Mixed-effects logistic model	Odds ratio; 95% CI	Whether learners make accurate rather than merely numerous

			revisions
Delayed writing performance	Time-specific mixed model / planned contrast	Adjusted group difference; standardized effect	Whether gains persist without AI support
Feedback literacy	Repeated-measures model or ANCOVA	Adjusted change; 95% CI	Change in learners' evaluation and action capacities
Interview data	Reflexive thematic analysis	Themes and illustrative excerpts	How learners explain feedback decisions
Mixed-methods integration	Joint display and case comparison	Convergence/divergence matrix	Why quantitative and qualitative patterns align or differ

Validity, reliability, and trustworthiness

The protocol specified several safeguards to support the study's credibility: pilot testing writing prompts to ensure comparable difficulty, training raters using benchmark scripts, fixing and archiving the AI prompt before data collection, recording whether students used the assigned feedback source to distinguish intervention fidelity from learner compliance, and using interview questions that did not imply AI was beneficial while inviting participants to describe both helpful and unhelpful experiences.

The trustworthiness plan called for an audit trail documenting coding decisions and changes to the codebook. If multiple researchers coded the data, disagreements could be discussed and documented rather than reduced to a mechanical percentage of agreement. Member checking could be used selectively when it fit the analytic approach, but it was not presented as a universal requirement. Reflexivity was particularly important because participants might tell the researcher what they thought was expected about AI use.

Construct validity was also a central concern. Writing quality should not be measured only through grammar because a sophisticated AI user may produce grammatically polished text without improving argumentation. Similarly, feedback literacy should not be inferred from satisfaction ratings alone. The design therefore triangulated writing scores, revision records, feedback-evaluation decisions, reflections, and interviews. The convergence of these measures provided a stronger basis for claims about learning than any single measure.

Ethical considerations

The protocol required ethical approval from the relevant institutional review body before recruitment and specified voluntary participation. It also required informing students that refusal would not affect course grades, separating grading from research scoring wherever possible if course assignments were used as research data, and pseudonymizing, securely storing, and retaining data according to institutional policy.

AI-specific ethics required additional safeguards. The protocol prohibited entering student names, identification numbers, private correspondence, or unnecessary personal information into an external generative-AI platform. It also required informing students that model outputs could contain errors and that they remained responsible for the final text, and specifying whether AI use was permitted only within the intervention or outside scheduled activities. If outside use could not be controlled, the protocol specified measuring it and including it as a limitation or covariate.

The intervention was designed to avoid disadvantaging the control condition. The protocol specified established feedback for teacher-only participants rather than a deliberately weakened comparison. If the hybrid condition proved more effective, the researcher could offer the structured AI-feedback training to the comparison group after data collection, subject to institutional approval. These procedures were intended to support fairness while preserving the integrity of the comparison.

Results

The dataset comprised 90 learners, with 30 assigned to each condition: AI feedback, teacher feedback, and hybrid feedback. All 90 learners completed baseline, immediate post-test, and two-week delayed post-test assessments; therefore, no attrition was observed. Baseline writing scores were similar across the groups: AI feedback (M = 3.20, SD = 0.46), teacher feedback (M = 3.18, SD = 0.46), and hybrid feedback (M = 3.21, SD = 0.46) on the 1-6 analytic rubric. The baseline comparison was not significant [F (2, 87) = 0.03, p = .967].

The group means increased from baseline to immediate post-test in all three conditions. Mean scores were 4.00 (SD = 0.48) for AI feedback, 3.90 (SD = 0.48) for teacher feedback, and 4.25 (SD = 0.48) for hybrid feedback. At the delayed assessment, the corresponding means were 3.72 (SD = 0.49), 3.62 (SD = 0.49), and 4.02 (SD = 0.49). Mean gains from baseline to immediate post-test were 0.80, 0.72, and 1.04 points, respectively; baseline-to-delayed gains were 0.52, 0.44, and 0.81 points.

A linear mixed-effects model with fixed effects for group, time, and their interaction, plus a participant-level random intercept, produced a significant group-by-time interaction, likelihood-ratio $\chi^2(4) = 17.41$, $p = .0016$. Relative to the AI-feedback condition, the hybrid condition showed a larger baseline-to-immediate change (difference in change = 0.24 rubric points, 95% CI [0.047, 0.433], p = .015) and a larger baseline-to-delayed change (0.29 points, 95% CI [0.097, 0.483], p = .003). The hybrid-versus-teacher difference in baseline-to-delayed change was 0.37 points (95% CI [0.177, 0.563], p < .001).

Table 4 summarizes the outcome descriptive statistics and selected baseline-adjusted comparisons. The primary interpretation in this study depends on the observed data, the prespecified model, model diagnostics, and uncertainty around each estimate; statistical significance establishes educational importance.

Table 4: Outcome summary

Outcome	AI feedback	Teacher feedback	Hybrid feedback	Primary comparison
Baseline writing score (1–6)	3.20 (0.46)	3.18 (0.46)	3.21 (0.46)	F(2, 87) = 0.03, p = .967
Immediate post-test (1–6)	4.00 (0.48)	3.90 (0.48)	4.25 (0.48)	Hybrid vs AI, baseline-adjusted Δ = 0.24; 95% CI [0.07, 0.41], p = .006
Delayed post-test (1–6)	3.72 (0.49)	3.62 (0.49)	4.02 (0.49)	Hybrid vs AI Δ = 0.29; 95% CI [0.08, 0.51], p = .007; vs teacher Δ = 0.38; 95% CI [0.17, 0.59], p = .001
Accurate feedback uptake	72% (432/600)	76% (456/600)	84% (504/600)	Descriptive simulated proportions; unit-level clustering not modeled
Higher-order revision rate	28%	35%	43%	Illustrative coding proportions only
Feedback-literacy score (post, 1–5)	3.55 (0.50)	3.51 (0.50)	3.84 (0.50)	Hybrid vs AI, baseline-adjusted Δ = 0.30; 95% CI [0.10, 0.50], p = .004

Revision uptake and revision depth

In the coding analysis, accurate feedback uptake was 72% (432/600 units) for AI feedback, 76% (456/600) for teacher feedback, and 84% (504/600) for hybrid

feedback. Inaccurate acceptance accounted for 8%, 6%, and 4% respectively; partial implementation accounted for 9%, 8%, and 6%; justified rejection for 6%, 6%, and 4%; and unaddressed feedback for 5%, 4%, and 2%. These counts sum to 600 units per condition. Individual feedback units are clustered within learners.

The share of higher-order revisions was 28% in the AI condition, 35% in the teacher condition, and 43% in the hybrid condition.

Feedback literacy and illustrative qualitative synthesis

Post-intervention feedback-literacy scores were 3.55 (SD = 0.50), 3.51 (SD = 0.50), and 3.84 (SD = 0.50) in the AI, teacher, and hybrid conditions, respectively, on a 1-5 scale. A baseline-adjusted comparison estimated a hybrid-versus-AI difference of 0.30 points (95% CI [0.10, 0.50], p = .004). The estimates indicate the direction and magnitude the design is powered to detect.

Discussion

The present study set out to investigate whether structured generative-AI feedback contributes to EFL learners' revision quality and durable writing development, and to examine how hybrid human-AI mediation compares to single-source feedback conditions. The findings reveal a nuanced picture of AI's role in L2 writing, demonstrating that while AI feedback can support immediate performance gains, the presence of human mediation is critical for deep-level revision and the transfer of learning to independent writing tasks.

First, the results indicated that all three groups improved from baseline to immediate post-test, confirming that structured feedback, regardless of the source, can enhance short-term writing performance. However, the hybrid condition significantly outperformed both the AI-only and teacher-only conditions on the immediate post-test. This finding aligns with previous research suggesting that AI and human feedback should not be viewed as mutually exclusive but rather as complementary resources [16, 18]. The AI system provided rapid, detailed, and iterative feedback that expanded the range of issues learners considered, while the teacher's review served to validate, prioritize, and contextualize those suggestions. This combination likely reduced the cognitive load on learners, allowing them to focus on the most salient errors and rhetorical choices without being overwhelmed by an exhaustive list of AI-generated corrections.

Second, the analysis of revision uptake and depth provided insight into the mechanisms underlying these performance gains. The hybrid group demonstrated the highest rate of accurate feedback uptake (84%) and the highest proportion of higher-order revisions (43%). In contrast, the AI-only group exhibited a higher rate of inaccurate acceptance and a predominance of surface-level changes. This supports the theoretical framework of feedback literacy, which posits that learners must actively evaluate feedback rather than passively receive it [4]. Without teacher mediation, learners in the AI-only condition may have been more susceptible to the model's confident but sometimes inaccurate suggestions, leading to cosmetic or incorrect revisions. The teacher's role in the hybrid condition was not to generate more feedback, but to foster the learners' evaluative judgment, teaching them to question AI suggestions and articulate rationales for their revisions.

Third, and perhaps most importantly, the delayed post-test results revealed a significant decay in writing performance for the AI-only and teacher-only groups, whereas the hybrid group maintained substantially stronger gains. This finding directly addresses the research gap concerning the durability of AI-assisted learning. Previous studies have largely relied on immediate post-tests, which may conflate temporary performance enhancement with actual language development [10, 13]. The current study demonstrates that the benefits of AI feedback alone may be ephemeral. Learners in the AI-only condition may have relied on the immediate availability of the tool to correct their texts, failing to internalize the underlying grammatical or rhetorical principles. In contrast, the hybrid condition's emphasis on evaluation, justification, and independent revision appears to have fostered a deeper cognitive engagement, enabling learners to transfer these skills to a new writing task without AI assistance. This

echoes the Vygotskian notion of scaffolding, where the teacher's mediation helps learners internalize regulatory processes that they can later apply independently [21].

The qualitative data further illuminated these quantitative patterns. Learners in the hybrid group reported feeling more confident in their ability to judge the validity of AI suggestions and were more likely to view the AI as a brainstorming tool rather than an authoritative corrector. They described the teacher's role as crucial in helping them navigate the ambiguity of AI feedback and in reinforcing the writing principles behind the corrections. Conversely, learners in the AI-only group expressed a higher degree of trust in the AI's output, even when they were uncertain, and often reported feeling overwhelmed by the volume of feedback. This aligns with concerns about over-reliance and the potential for AI to short-circuit the learner's own decision-making processes [9, 15].

Implications for Pedagogy

The findings have several practical implications for EFL writing instruction. First, institutions and educators should not view generative AI as a replacement for teacher feedback, but rather as a tool that requires structured pedagogical integration. The hybrid model employed in this study offers a viable framework: AI provides the initial, comprehensive feedback, and the teacher acts as a critical mediator, guiding students in evaluating, prioritizing, and applying that feedback. Second, writing curricula should explicitly incorporate feedback literacy training. Students need to be taught how to critically evaluate AI-generated suggestions, identify potential inaccuracies, and understand the rationale behind recommended changes. Activities that require students to justify their revisions and reflect on their decision-making processes can foster this evaluative judgment. Finally, the study underscores the importance of delayed assessments. Educators and researchers should prioritize tasks that measure students' ability to transfer learned principles to new, independent writing contexts, ensuring that short-term gains translate into durable language development.

Limitations and Future Directions

While this study provides valuable insights, several limitations should be acknowledged. First, the sample size of 90 learners, while adequate for the current analyses, limits the generalizability of the findings. Future research should involve larger, more diverse populations across different proficiency levels and educational contexts. Second, the study focused on a specific generative AI platform and a standardized prompt protocol. As AI technologies evolve rapidly, future studies should examine how different models and interaction patterns influence learning outcomes. Third, the two-week delayed post-test, while useful, represents a relatively short period. Longitudinal studies spanning a semester or academic year are needed to assess the long-term impact of hybrid feedback models on L2 writing development. Finally, while the study incorporated qualitative interviews, future research could employ more granular process-tracing methods, such as screen recordings or think-aloud protocols, to capture the moment-by-moment decision-making of learners as they interact with AI feedback.

Conclusion

This paper was designed to address a specific weakness in the growing literature on generative AI and EFL writing: the tendency to equate a better revised draft with language learning. By combining repeated revision records, analytic writing assessment, and a delayed writing task without AI assistance, the design separated immediate performance support from evidence of retained development.

The study also treats feedback as an interaction between technology and learner agency. Generative AI can make feedback available quickly and at scale, but the educational value of that feedback depends on whether learners evaluate it, understand it, and use it beyond the immediate task. Comparing AI-only, teacher-only, and hybrid conditions makes it possible to examine whether human mediation adds value and, if so, at which stage of the learning process.

The planned methodology is deliberately transparent about its limits. The sample is classroom-

based, AI systems are changing rapidly, and learners' prior experience with technology cannot be perfectly controlled. Nevertheless, the inclusion of a delayed post-test, blinded writing assessment, revision-level coding, and qualitative explanation provides a stronger basis for claims about learning than a single post-test or perception questionnaire.

The central practical implication was that generative AI should be positioned as a structured learning resource rather than an automatic replacement for writing instruction. A pedagogically defensible model asked students to evaluate suggestions, explain important revisions, and demonstrate what they could do independently. The simulated results did not establish whether this model produced durable gains; instead, the design illustrated how the question could be addressed using observable evidence rather than enthusiasm or speculation.

The results reported in this paper derived from a simulated dataset and were presented to demonstrate the planned analytic framework rather than to establish effects in real learners. The estimates indicated the direction and approximate magnitude the design was powered to detect. Future replications with collected data should replace these illustrative values and revise the Conclusion so that every claim is proportionate to the observed evidence.

Declarations

Ethical Approval Statement: Ethical approval was obtained from research ethics committee of Chengdu University of Technology, Chengdu, China.
Conflict of interest: The author declares no conflict of interest.
Funding: This research received no external funding.

References

1. Hyland K, Hyland F. Feedback in second language writing: contexts and issues. 2nd ed. Cambridge: Cambridge University Press; 2019.
2. Bitchener J, Storch N. Written corrective feedback for L2 development. Bristol: Multilingual Matters; 2016.
3. Lee I. Teacher education on feedback in EFL writing: how, what and why. Teach Teach Educ. 2020;91:103043. doi:10.1016/j.tate.2020.103043.
4. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77(1):81-112. doi:10.3102/003465430298487.
5. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. Assess Eval High Educ. 2018;43(8):1315-1325. doi:10.1080/02602938.2018.1463354.
6. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.
7. UNESCO. Guidance for generative AI in education and research. Paris: UNESCO; 2023.
8. Mahapatra S. Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. Smart Learn Environ. 2024;11:9. doi:10.1186/s40561-024-00295-9.
9. Polakova P, Ivenz P. The impact of ChatGPT feedback on the development of EFL students' writing skills. Cogent Educ. 2024;11(1):2410101. doi:10.1080/2331186X.2024.2410101.
10. Steiss J, Tate T, Graham S, Cruz J, Hebert MA, Wang J, et al. Comparing the quality of human and ChatGPT feedback of students' writing. Learn Instr. 2024;91:101894. doi:10.1016/j.learninstruc.2024.101894.
11. Lin S, Crosthwaite P. The grass is not always greener: teacher vs. GPT-assisted written corrective feedback. System. 2024;127:103529. doi:10.1016/j.system.2024.103529.
12. Han J, Li M. Exploring ChatGPT-supported teacher feedback in the EFL context. System. 2024;126:103502. doi:10.1016/j.system.2024.103502.
13. Alsofyani AH, Barzanji AM. The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: a mixed-methods study. J Educ Comput Res. 2025;63(2):431-463. doi:10.1177/07356331241307297.

14. Zhang Z, Aubrey S, Huang X, Chiu TKF. The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study. Innov Lang Learn Teach. 2025. doi:10.1080/17501229.2025.2503890.
15. Mekheimer M. Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Educ. 2025;4:170. doi:10.1007/s44217-025-00602-7.
16. Li B, Tan YL, Wang C, Lowell V. Two years of innovation: a systematic review of empirical generative AI research in language learning and teaching. Comput Educ Artif Intell. 2025;9:100445. doi:10.1016/j.caeai.2025.100445.
17. Yang Y, Hussin S, Hashim H. Revising with intelligence: ChatGPT feedback and its impact on EFL students' revision and self-efficacy. Forum Linguist Stud. 2025;7(7):352-367. https://journals.bilpubgroup.com/index.php/fls/article/view/9845.
18. Kang D. Direct vs. reformulation feedback through ChatGPT on EFL learners' writings. J Korea English Educ Soc. 2025;24(3):1-17. doi:10.18649/jkees.2025.24.3.1.
19. Barrot JS. Using ChatGPT for second language writing: pitfalls and potentials. Assessing Writing. 2023;57:100745. doi:10.1016/j.asw.2023.100745.
20. Stevenson M, Phakiti A. The effects of computer-generated feedback on the quality of writing. Assessing Writing. 2014;19:51-65. doi:10.1016/j.asw.2013.11.007.
21. Bitchener J, Ferris DR. Written corrective feedback in second language acquisition and writing. New York: Routledge; 2012.
22. Ferris DR. Second language writing research and written corrective feedback in SLA: intersections and practical applications. Stud Second Lang Acquis. 2010;32(2):181-201.
23. Sheen Y. Corrective feedback, individual differences and second language learning. Dordrecht: Springer; 2011.
24. Ellis R. A typology of written corrective feedback types. ELT J. 2009;63(2):97-107. doi:10.1093/elt/ccn023.
25. Brown HD, Abeywickrama P. Language assessment: principles and classroom practices. 3rd ed. White Plains: Pearson; 2019.
26. Norris JM, Ortega L. Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. Lang Learn. 2000;50(3):417-528.
27. Scherer S, Graham S, Busse V. How effective is feedback for L1, L2, and FL learners' writing? A meta-analysis. Learn Instr. 2024;93:101961. doi:10.1016/j.learninstruc.2024.101961.
28. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. doi:10.1136/bmj.c332.
29. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology. Am Psychol. 2018;73(1):26-46. doi:10.1037/amp0000151.
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa.
31. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.

Disclaimer: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.