Journal of Artificial Intelligence and Computing (JAIC)
ISSN (e): 3005-3358
Volume: 4, Number: 1, Pages: 8 - 18, Year: 2026

Generative AI as an Automated Written Corrective Feedback Provider in EFL Academic Writing: A Mixed-Methods Investigation of Accuracy Gains, Feedback Quality, and Learner Engagement

Mujtaba Kamal Pasha

Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan
Corresponding author: Mujtaba Kamal Pasha (Email: mujtaba.kamal@uet.edu.pk)

Received: 10/03/2026, Revised: 26/05/2026, Accepted: 15/06/2026

***Abstract*— The integration of generative artificial intelligence (GenAI) and large language models (LLMs) into language technology has reopened long-standing questions about the role of automated written corrective feedback (AWCF) in second-language (L2) writing development. While early enthusiasm has positioned conversational LLMs such as GPT-4 as scalable substitutes for the labour-intensive provision of teacher feedback, the empirical base remains fragmented, and few controlled studies examine accuracy outcomes, feedback quality, and learner engagement together. This study addresses that gap through a controlled investigation conducted in an under-represented South Asian English-as-a-foreign-language (EFL) setting. Adopting a sequential explanatory mixed-methods design, ninety undergraduate EFL learners were stratified and randomly assigned to three conditions—GPT-4 AWCF, expert teacher-written corrective feedback, and a self-revision control—across an eight-week, six-cycle argumentative writing programme.**

Quantitative strands comprised error-rate analysis, an analytic rubric, and a validated engagement questionnaire; qualitative strands comprised corpus coding of 1,860 feedback moves, stimulated-recall interviews, and reflective journals, integrated through a joint display. Analysis of covariance indicated that the GPT-4 group significantly outperformed the control on both accuracy and rubric scores, with gains statistically comparable to, and on local accuracy marginally exceeding, the teacher group, and with most gains retained at delayed post-testing. Feedback coding revealed that GPT-4 favoured metalinguistic explanation and reformulation but showed reduced precision on discourse-level concerns, where redundancy and occasional hallucinated corrections appeared. Engagement was high—particularly on the affective dimension—but cognitively demanding, and learners expressed concern about over-reliance and loss of authorial voice. The study contributes controlled, integrated evidence on the pedagogical value and limitations of LLM-based language technology and argues for a principled human-in-the-loop model of AI-assisted feedback rather than wholesale automation.

Keywords: **Generative artificial intelligence; large language models; automated written corrective feedback; EFL academic writing; learner engagement; natural language processing in education; computer-assisted language learning.**

I. INTRODUCTION

Writing is considered to be the most complex and the slowest of the language skills that English-as-a-foreign-language (EFL) students need to master. Writing in the academic world is very different from speaking, in that it demands that the idea be generated, the rhetorical organisation, the precision of the lexis and the accuracy of the grammar all happen at the same time and that the product be permanent and subject to evaluation. In the last few decades, the pedagogical pillar of corrective feedback on written production has been regarded as the means by which learners discover the discrepancy between their interlanguage and the norm of the target language in order to gradually revise their L2 writings.

However, the lack of timely, individualised and sustained feedback provision at scale has been an ongoing structural limitation in language-education systems around the world. In large programmes, and particularly in the high enrolment and high resource constraint environments common in much of South Asia, the hope of many such comments, for each student, each term, is largely unfeasible, as one teacher may be expected to mark the written work of more than one hundred learners. Historically, language technology has addressed this limitation with the development of automated writing evaluation (AWE) systems that promised to reduce teacher workload by providing fast scores and pre-defined corrective feedback. The learning outcomes of such systems, however, have been varied. Rule-based and statistical tools have been shown to be very good at identifying surface level flaws in spelling, punctuation and local grammar, but very poor at discourse level flaws, such as

argumentation, coherence and rhetorical appropriateness, that define good academic writing.

A recent systematic review of one of the most widely used commercial AWE tools found that although it is good for surface accuracy, it does not dive deep into other high-order writing processes [1]. This has restricted the perception of what is possible with language technology, and set the expectations for the current generation of new technology. It is that new generation of transformer driven large language models, specifically LLMs and conversational systems such as ChatGPT and GPT-4 that entered the world of education in late 2022. This release has been described as a milestone for the conception and application of automated written corrective feedback (AWCF) in L2 teaching and learning [2]. Unlike previous AWE tools that generate fixed templates, LLMs can offer contextual, explanatory, and dialogic responses, which can serve as a metalinguistic commentary from a human tutor and can be questioned, edited, and re-prompted by the learner [4].

In the realm of early intervention, positive outcomes on writing performance [5], [6], self-regulated learning and motivation [6] have been reported when LLMs are systematically integrated into instruction, while learners' perceptions of LLMs as supportive, non-threatening feedback companions have been largely positive [7], [8]. But this is an optimism in conflict with several realities. First, the accuracy and reliability of LLM output will remain unknown: Research on automated essay scoring and AI-generated feedback shows that validity and reliability will be highly dependent on the model, rubric, task, and population and that the apparent fluency of LLM responses can mask mistakes and even invented corrections [9], [10]. Secondly, scholars point out that AI feedback might act as a thinking inhibitor rather than a facilitator as students blindly follow the suggestions, and this will replace the productive cognitive effort on which durable learning depends [11]. Third, and most significantly for cumulative knowledge building, a recent scoping review of over fifty studies in the field indicates that the field is growing fast and that there is considerable methodological fragmentation, with relatively few controlled designs reporting on outcomes for accuracy, quality of feedback and learner engagement in a single study [12]. It's easy to overlook the methodological stakes of this fragmentation. Much of the available evidence is based on a single group, pre-post design, or perception-only design, and it is difficult to separate the impact of the AI feedback from the impact of practice, maturation, and an exciting gadget. Apparent gains are nothing but repetition of writing if no feedback control is used, and claims of similarity with human feedback are unjustified in the absence of an expert-teacher comparison.

Thus, a three arm design that simultaneously compares the AI feedback to a true control and to an expert human is not simply a luxury of the method, but a prerequisite for credible inference. The stakes are also quite high in terms of the context. In systems where enrolments are high and resources are scarce, the promise of scalable individualised feedback is greatest where teacher time is most limited, but it is here, too, that the potential inequity of access to devices and connectivity poses most strongly as a new source of disadvantage. Producing rigorous evidence, therefore, out of such a context, and not extrapolating from better resourced settings, is a substantive contribution in and of itself.

To tackle exactly that fragmentation is the present study. It is located in a Pakistani university's English department—a context that is important in terms of the quantity of EFL learners but not representative in international published literature on AWCF—and takes a controlled, three-arm, sequential explanatory mixed-methods design to test GPT-4 as an AWCF provider in comparison to expert human feedback and a self-revision control. In doing so it addresses directly the technological dimension of current language, literature and technology scholarship and in particular the interplay of natural language processing, e-learning, and language-education assessment that is becoming more and more important in the field of computer-assisted language learning. Instead of simply questioning whether or not AI feedback ‘works', the study breaks down the construct, exploring where in fact LLM feedback can be seen to be effective, where it is inferior to human feedback, and where and how students are actually using AI feedback.

A. Research Objectives

- To examine the effectiveness of GPT-4-generated automated written corrective feedback (AWCF) in improving the linguistic accuracy and overall quality of EFL learners' academic writing compared with expert teacher feedback and self-revision.
- To compare the characteristics and accuracy of GPT-4-generated feedback with expert teacher written corrective feedback in EFL academic writing.
- To explore EFL learners' behavioural, cognitive, and affective engagement with AI-mediated written corrective feedback and examine their perceptions of its pedagogical value and limitations.

B. Research Questions

1. To what extent does GPT-4-generated AWCF improve the linguistic accuracy and overall quality of EFL learners' academic writing relative to teacher feedback and a self-revision control?
2. How do the characteristics and judged accuracy of GPT-4 feedback moves compare with those of expert teacher feedback?
3. How do learners engage—behaviourally, cognitively, and affectively—with AI-mediated written feedback, and how do they interpret its value and limitations?

II. Literature Review

Written corrective feedback (WCF) involves responding to the learners' written errors to encourage noticing, uptake and longer term accuracy. The construct is in the middle of a long and sometimes contentious controversy in second-language-acquisition research about whether, and how, error correction can help the development of writing. Opinions about the viability of gains from correction span from doubt that such gains would last to strong assertions that gains resulting from focused feedback on discrete linguistic features are measurable and durable. They are distinguished with respect to whether the

feedback is direct (providing the correct form) or indirect (signaling an error without providing it); whether focused (providing feedback on a few features) or comprehensive (providing feedback on all errors); and whether metalinguistic (providing explicit rules/explanations) or reformulation (modelling target-like rephrasing). These distinctions are relevant for the current study because they provide the vocabulary for comparing human and AI feedback and because the balance a feedback source strikes between them is pedagogically relevant. The WCF tradition offers two additional insights into the automation debate. The first is that feedback is not effective when he or she gives it, it is effective when it is received, understood and acted upon; feedback that is not received, understood and acted upon is not effective, even if it is accurate. The second is that the provision of feedback is highly labour intensive, and that the cognitive and time commitment for teachers to provide rich, individualised feedback is not linear in class size. This is where automated systems have been looking to solve the scaling issue, and it is upon this dual measure of uptake and quality that their solutions should be judged. What makes these design differences possible is a collection of cognitive mechanisms believed to underlie the effect of feedback. One prominent theory of the effect of corrective feedback is that it causes noticing, defined as the conscious recording of a discrepancy between a learner's production and the desired form, and that noticing is a necessary precondition for acquisition. A complementary account is that language production under communicative stress brings the learner to the limits of his or her competence, thus enabling feedback to intervene most productively. Both accounts suggest that feedback is generative only to the extent that the feedback is cognitively processed, that is, if the feedback is only glanced at and copied, without being understood, the mechanism of generative feedback is not engaged. This puts the concept of engagement, which is elaborated below, at the heart of any consideration of feedback technology and is a prefiguration of the most important concern about LLMs: that their fluency and authority could lead to passive acceptance where active processing is needed.

A. From automated writing evaluation to generative AI

WCF automation is not new with generative AI, it goes back decades. First generation AWE systems included statistical scoring models along with natural language processing, and returned a set of templated comments and holistic or trait scores. Although these systems clearly decreased the teachers' workload and sped up the turnaround of feedback, their pedagogical scope was limited by their system: They were limited by the hand-designed features and the fixed bans of answers, and could not respond to the meaning and argument of a text or to a learner's individual error. This ceiling is reflected in the systematic review by [1], which found consistent evidence of improvements in the surface accuracy but little support for the rhetorical and ideational aspects of writing. But the advent of LLMs is not just an incremental improvement on this paradigm, it's a qualitative change. LLMs are based on transformer models that use self-attention mechanisms and trained on extensive text data sets, generating their output instead of retrieving it, which means that they can provide contextualised, explanatory, and conversational feedback, which is iteratively refined [4]. ChatGPT has demonstrated clear potential for use as a classroom WCF tool, but initial classroom studies have also found that feedback varies in quantity and quality over time [13].

B. Empirical effects of LLM feedback on writing

An increasing number of studies have begun to measure the impact LLM feedback can have on writing. Liu in [14] found that, in the context of EFL writing education, integrating an LLM-based model into the learning process led to substantial improvements in writing, self-regulated learning strategy use, and writing motivation in a randomized study, highlighting the crucial role of human teachers in the cycle. In a mixed-methods intervention with tertiary ESL learners, Mahapatra [6] reported that ChatGPT as a formative feedback tool significantly improves academic writing skills and that it was overwhelmingly well-received by the students. The perception findings are striking because they all point to the affective experience of LLM feedback: learners often see LLM feedback as a companion instead of evaluative threat, as Teng [7] reported lower writing apprehension among EFL learners, and as Du and Alm [8] found in their self-determination theory interpretation of LLM feedback. However, a systematic review of GenAI in the language classroom warns of the varying quality of implementation and the ongoing strong reliance on perception data and under-controlled designs [15]. The strength and persistence of effects are quite variable, however, in this growing body of literature, and the drivers of this heterogeneity are not yet well known. Research studies vary in the type of model/version used, control of prompts (learner or researcher), type of writing genre addressed, level of writing proficiencies addressed, duration and intensity of the intervention, and time of point of measure (immediate or delayed). Thus, the very same tool may be transformative in one study and marginal in another, not because it has different capacities, but because the circumstances of its use are different. This diversity in itself argues for controlled designs in which such conditions are deliberately controlled, enabling the contribution of the feedback source to be separated from the contribution of the confounds of prompt skill, genre and dosage. It also warns against interpreting any single estimate of the effect, such as the one presented here, as an absolute property of the technology and not one generated during the situation [15], [12].

C. Feedback quality, accuracy, and assessment validity

A second strand of research is not about the liking of LLM feedback or about gains, but whether it is good, accurate, appropriate & trustworthy. Comparing the feedback of humans and AI has been particularly informative. Lin and Crosthwaite [4] discovered that GPT provided a higher proportion of metalinguistic explanation and reformulation than teachers, while demonstrating less attention to discourse-level issues, and that although the feedback provided by GPT was frequently beneficial, it also featured inaccuracies and inconsistencies. This profile—strength on local, explainable correction, weakness and less reliability on global, rhetorical correction—can be found throughout the comparative literature. Optimism is qualified with the parallel literature on automated essay scoring (AES). Complementary studies demonstrated performance to vary as a function of model and prompt when

using LLMs to evaluate ELL writing [9], non-native Japanese writing [16], and LLMs in general for automated writing evaluation [5].

D. Learner engagement with AI-mediated feedback

The value of feedback is manifested in uptake, therefore engagement of the learners cannot be viewed as a peripheral variable but a determinant of what they will achieve. The most prominent conceptualisation is one that identifies three interdependent aspects that can be considered to be aspects of engagement, namely behavioural engagement (what learners see, hear and do to receive feedback), cognitive engagement (how deeply learners process the feedback and what metacognitive and cognitive strategies they take to process it), and affective engagement (what emotions do learners feel towards the feedback and feedback provider). Yan and Zhang [2] conducted a multiple case study of ChatGPT-generated AWCF using this model, which discovered that ChatGPT created a competence and time-intensive learning environment that also exposed learners to an emotionally supportive setting where they were not threatened by the faces of others and felt free to request feedback several times. Process-oriented research validated the mediating role of engagement: Schiller [3] demonstrated that the effectiveness of automated feedback is not simply because it is available, but actually because of a material engagement with it, using keystroke logging. There are two underlying concerns to these affordances. The first is over-reliance: Guo [11] cautions that uncritically accepted AI feedback can inhibit the independent thinking that learning requires. The second is the displacement of teacher expertise, and thus the suggestion in literature is for complementary configurations where AI is used to complement rather than replace teacher feedback [17].

E. Theoretical framework

The study is explained by two complementary perspectives. The first is sociocultural theory and its concept of the zone of proximal development, in which feedback serves as a tool of mediation to allow a learner to do what he or she is unable to do independently, and to internalise this mediated performance into the ability to do it alone. From this point of view, it is not so much whether AI feedback looks like teacher feedback but whether it offers graduated and contingent feedback that can help with internalisation, or whether, as the over-reliance critique suggests, it actually short-circuits this process and gives away the very cognitive product that constitutes learning. Empirically, the second perspective is provided by the so-called engagement framework given above, which provides the behavioural, cognitive, and affective aspects operationalised in the empirical strands. These lenses inspire a design that assesses outcomes as well as the quality of feedback and how learners engage with this feedback.

F. The research gap

There are three gaps from this review. Controlled designs that incorporate accuracy results, feedback-quality analysis, and engagement in a single study are scarce, even with the growth of research in the field, as recognized in the recent scoping review [12]. Conceptually, the evidence base consists largely of East Asian and western context settings, and the representation of the large EFL populations in South Asia is significantly under-represented. In methodological terms, the two most common designs are perception and case-study: there are quite few three-armed controlled comparisons both with expert human feedback and a true control.

III. Methodology

A. Research design and rationale

The research design and the rationale for the study. A sequential explanatory mixed-methods design was used, where a quantitative quasi-experimental strand was followed by a qualitative strand, and the two strands were integrated using a joint display at the interpretation stage. This design was chosen because the three research questions are of different kinds: RQ1 is a comparative-effects question, for which controlled measurement is appropriate; RQ2 is a question about the qualities of feedback, which can only be addressed by fine-grained analysis of the corpus; and RQ3 is a question about engagement and meaning, which requires the learners' own accounts. A mixed-methods architecture enables the strengths of each strand to fill in the shortcomings of the other: the explanatory power of learner testimony is disciplined by the quantitative pattern; the quantitative pattern is given explanatory depth by the learner testimony. The entire process is summarised in Fig. 1.

B. Context and participants

The research took place in the compulsory undergraduate section of academic writing course in the English department of a university in Lahore, Pakistan. As with much high enrolment EFL provision in the region, the setting has large classes, wide ability levels, limited teacher time with individual students and varying, albeit increasing, access to digital devices and connectivity. Using purposive sampling, 90 students were selected from a larger cohort of students in a course. Students were excluded from the study if they were not in the B1–B2 range of the Common European Framework of Reference according to the Oxford Placement Test and if they had previous experience with the use of AI writing tools. The final sample consisted of learners from a variety of disciplinary backgrounds, and a fairly even gender distribution with all learners saying that Urdu/Punjabi was their native language and English was the language of instruction used. Each proficiency band and gender was then randomly allocated to one of three groups: a GPT-4 AWCF group, a teacher WCF group, and a self-revision control group, which consisted of thirty students in each group. Simple randomisation was not used because it was important to balance the conditions on the two characteristics most likely to affect writing-gain comparisons, which was achieved using stratified random assignment. Baseline equivalence was then statistically demonstrated by the absence of significant between-group differences on pre-test accuracy, and rubric scores.

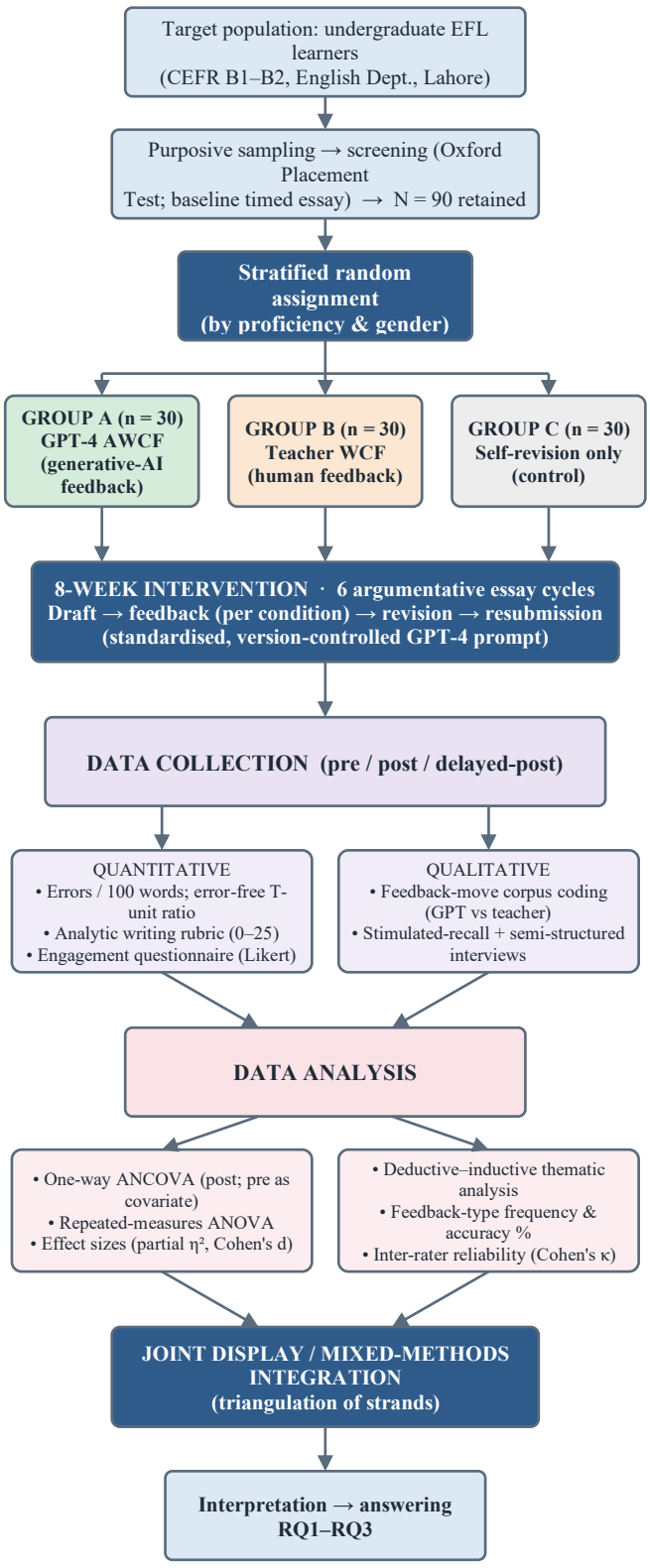

Fig. 1: Sequential explanatory mixed-methods research design.

C. *The intervention*

Each of the three groups went through the same set of six cycles in which they wrote argumentative essays, all of which were matched for genre, length, and cognitive demand, and all were completed over a period of eight weeks. Each cycle involved a four-step draft–feedback–revision–resubmission process: the learners gave a first draft, the feedback was pertinent to the situation, they revised based on feedback, and they resubmitted a second draft. Time-on-task, number of cycles, prompts, and submission infrastructure were kept constant across conditions, thus limiting the experimental contrast to the source and nature of the feedback. The GPT-4 condition involved the use of a standardised, version-controlled prompt template for all learner drafts to create feedback. The template asked the model to detect mistakes and weaknesses, to explain metalinguistically instead of silently rewriting, to comment on the organisation and argument in addition to the language used, and specifically not to copy entire rewritten texts that learners could copy. We chose to keep the elicitation prompt fixed as a methodological decision: LLM output is very sensitive to prompt formulation and to let the learners formulate their own elicitation prompt would have made it difficult to disambiguate the feedback source from the ability of the prompts to elicit the right answer. The model's output was given to the learners but the elicitation was controlled. In the teacher condition, two trained teachers gave feedback according to an agreed upon, pre-standardised annotation protocol to achieve a focus and form which were consistent. In the control condition, students amended their work on their own based on a general self-editing checklist, which is similar to the real-world situation where there was no individualised feedback. The design of the GPT-4 prompt template deserves explanation as it represents a number of the methodological commitments of the study. The template was developed in an iterative fashion over the pilot phase and fixed throughout the intervention to ensure that it is comparable for learners and cycles of the intervention. It had been instructed that it could not rewrite more than a certain percentage of any passage, thereby allowing the learner to retain ownership of their writing; to give detailed explanations for every correction, but only provide a short and easily understood rationale; to highlight uncertainty instead of making confident substitutions; and to score the comments in order of importance, starting with comments on task response and organisation and ending with comments on sentence level. The model was provided with learner identifiers and drafts in a standardised wrapper and outputs were delivered to learners without editing by the researcher. This controlled-elicitation technique sacrifices a bit of ecological validity (real learners creating their own, often weaker, stimuli) for a measure of internal validity to ensure that any outcome is the result of the feedback source and not the variation of the prompting skill.

D. *Instruments and measures*

There were two complementary measures of quantitative writing outcomes. Errors per 100 words and percentage of error free T-units were used to measure linguistic accuracy, double coded by two trained raters using an explicit error taxonomy, and discrepancies discussed. Overall writing quality was assessed using an analytic rubric that was double scored and assessed from zero to twenty-five across four equally weighted dimensions: task response, coherence and cohesion, lexical range and accuracy, and grammatical range and accuracy. A validated questionnaire was used to measure engagement, which included Likert-scaled items in three behavioural, cognitive and affective subscales, and the two feedback groups

were asked to fill out the questionnaire after the intervention. The qualitative data consisted of three sources. Firstly, a set of feedback moves (those that occurred in the learner drafts in the stratified sample across both feedback groups) were collected and coded according to the type of feedback (direct correction, indirect signalling of error, metalinguistic explanation, reformulation, and content/organisation comment) and independently rated by accuracy for each feedback move by two raters. Secondly, stimulated-recall interviews and semi-structured interviews were conducted with a purposive subsample of learners, based on their own annotated versions of their drafts. Third, learners kept reflective journals during the intervention that detailed their experiences of seeking, interpreting and responding to feedback.

E. Data analysis

The quantitative analysis was carried out in three stages. Prior to this, baseline equivalence was established by one-way analysis of variance on the pre-test measures. Then, one-way analysis of covariance was used in the primary comparative analyses, with the respective pre-test score entered as a covariate to account for any pre-test residual variation and to increase statistical power, and partial eta-squared was reported as the effect-size index, with appropriate adjustments for multiple testing in post hoc pairwise comparisons. Repeated-measures analysis of variance was used to analyze retention of gains across pre-, post- and delayed-post-test points. Independent-samples t-tests were conducted on the engagement subscales for the two feedback groups with Cohen's d. A combined deductive-inductive thematic analysis was used with the feedback corpus to code the data according to the move taxonomy and the interview and journal data to identify common themes which were then connected back to the engagement framework. The Cohen's kappa to determine interrater reliability for the coding of feedback moves was calculated and was found to be substantial. Lastly, the strands were combined into an integrated display which contrasted the quantitative findings with the qualitative themes that accounted for them.

F. Trustworthiness and ethics

There were some steps taken that supported the trustworthiness of the study. Double rating ensured the reliability of quantitative coding; member checking of interview interpretation and triangulation of interview, journal and feedback corpus ensured credibility of qualitative claims; and the controlled, 3-arm design ensured internal validity of comparisons. Ethical clearance for the study was obtained. The following steps have been implemented: all participants were informed about the purpose of the study and their right to withdraw without facing any penalty; data was anonymised and securely stored; and, after the study, all participants were given access to the GPT-4 feedback workflow to avoid any disadvantage to the control group.

G. Pilot, positionality, and limitations of the design

A two-cycle pilot study was first implemented with another group of similar learners to fine-tune the prompt structure of GPT-4, the teacher annotation protocol, and to test the coding scheme as well as the understandability of the questionnaire for engagement and the interview guide. The pilot identified some problems with the first, namely that it tended to overwrite the learner texts in some instances and that two of the questionnaire items were ambiguous, as well as informing the training of the raters to an acceptable level of agreement before live coding started. Like in any mixed-methods research, the researchers' positionality was reflexively acknowledged: In studying, they were also members of the teaching community and had commitments to the value of human feedback that might, if unexamined, bias their interpretation to the conclusion that AI cannot match teachers. This was addressed through pre-registering the analytic plan for the quantitative strand, by double-coding with another rater outside of the plan, and by member-checking qualitative interpretations with participants. These do not remove the fundamental problems of a single-context, time-bounded, single-model study (directly discussed in the concluding section), but they do help to confidence that the pattern reported is indeed a product of the data and not the investigations' priors.

IV. Findings

A. Writing accuracy and quality gains

The learner is consistently able to write with accuracy and good quality. Pre- and post intervention writing measures for each group are shown in Table 1. Pre-test measures did not show any significant differences between groups for accuracy or rubric scores, indicating baseline equivalence. For the post-test accuracy, a covariate analysis of variance with the pre-test accuracy as covariate found a significant effect of condition, $F(2, 86) = 21.4$, $p < .001$, with a large effect size (partial $\eta^2 = .33$). Similarly, a parallel analysis of rubric scores showed that the effect of condition was significant, $F(2, 86) = 17.8$, $p < .001$, partial $\eta^2 = .29$. Both feedback groups significantly exceeded the control on both measures ($p < .001$) in pairwise comparisons. No significant difference was found between the GPT-4 and the teacher groups on rubric scores, which means that these groups performed similarly in relation to the holistic quality of their work, while on local accuracy GPT-4 group showed marginally larger gains (the average error rate per hundred words decreased from 9.84 to 5.12), likely owing to the dense local metalinguistic correction they received from GPT-4. Repeated-measures analysis for the delayed post-test showed that both feedback groups maintained most of the gains they made, and that there was no significant drop in accuracy in each group, which meant that gains were not just temporary effects of the immediate scaffolding.

Table I provides a pattern to answer RQ1 directly. GPT-4 AWCF generated statistically significant and held improvements in accuracy and overall quality scores, comparable to those of the expert teacher feedback for the holistic quality scores and slightly higher than for the local accuracy scores, and much higher than for the unsupported self-revision on both. The slight improvement in the control group was likely due to maturation and familiarity with the task over the six writing cycles, highlighting the significance of individualised feedback, whether human or AI.

Table I — *Pre- and post-intervention writing measures by group (M, with SD in parentheses)*

Measure	Time	GPT-4 (n=30)	Teacher (n=30)	Control (n=30)
Errors per 100 words	Pre	9.84 (2.31)	9.91 (2.45)	9.78 (2.40)
	Post	5.12 (1.88)	5.64 (1.97)	8.41 (2.22)
	Delayed	5.39 (1.93)	5.81 (2.02)	8.35 (2.27)
Error-free T-unit ratio	Pre	0.41 (0.09)	0.40 (0.10)	0.41 (0.09)
	Post	0.63 (0.08)	0.60 (0.09)	0.45 (0.10)
Analytic rubric (0–25)	Pre	14.2 (2.6)	14.0 (2.7)	14.3 (2.5)
	Post	19.6 (2.2)	18.9 (2.4)	15.1 (2.6)

Note. The larger the error-free T-unit ratio and rubric scores and the lower the error rate, the more successful the performance. The difference between the two groups receiving feedback was small compared to the common benefit in comparison to the control. Post-test was conducted four weeks after the intervention.

B. Characteristics and accuracy of GPT-4 versus teacher feedback

The distribution and accuracy profile in Table 2 resulted from coding 1,860 feedback moves over the two feedback groups. The feedback signatures for the two sources were quite distinct. Teachers focused more on direct correction and spent a significantly greater percentage of their moves on discourse level comments on content and organisation, while GPT-4 gave the highest proportion of its moves to metalinguistic explanation and reformulation, which amounted to more than half of its feedback. In terms of accuracy, both sources were highly reliable with respect to local grammar & lexis but in terms of discourse, teacher accuracy was high across all categories, while GPT-4's accuracy decreased with respect to content & organisation, where a minority of its moves were redundant, generic, or—in a smaller but significant number of cases—hallucinated, leading to the identification of non-existent problems or the suggestion of changes that created new problems. This profile validates and builds upon previous comparative results [4] and the cautions to validity found in the essay-scoring literature [9], [18].

Table II — *Distribution and judged accuracy of feedback moves by source.*

Feedback move	GPT-4 (% moves)	Teacher (% moves)	GPT-4 accuracy (%)	Teacher accuracy (%)
Direct correction	21.4	38.2	94.1	97.0
Indirect (location only)	9.8	17.6	90.5	95.8
Metalinguistic explanation	34.7	19.1	88.3	96.2
Reformulation	23.5	11.4	85.6	94.7
Content / organisation	10.6	13.7	71.2	93.4
Total moves (n)	1,012	848	—	—

Note. Accuracy is the percentage of moves rated as appropriate by two trained raters (Cohen's $\kappa = .81$). The accuracy of GPT-4 on the content/organisation is lower, with some suggestions being redundant or hallucinated.

Table II answers RQ2. GPT-4's relative advantage was how much it gave in the way of local feedback: it could give more explanations for the errors made than teachers could, and it could give more target-like reformulations of the errors made than teachers, under time pressure, could. It did not compare well in precisely those areas where human expertise still outpaced it: in the higher order, context-dependent judgement that is necessary to be able to comment reliably on argument and structure. The asymmetry has pedagogical consequences: It does not imply that one source of strength is necessarily superior on a global scale, but on the contrary, that both sources are complementary, each of which is strongest in the places where the other is weakest.

C. Learner engagement with AI-mediated feedback

The results of the engagement questionnaire are shown in Table III. Overall, engagement was high in both feedback groups, with the GPT-4 group reporting significantly more affective engagement than the teacher group, representing the largest gap in the data. There were no significant differences in behavioural and cognitive engagement across the groups, but the subjective processing load was higher in the GPT-4 group. The presence of affective advantage reflects previous studies on engagement and perception [2], [7], and the higher cognitive load aligns with the fact that the effectiveness of feedback is actually mediated by the quality of engagement [3].

Table III— *Learner engagement by dimension (5-point scale; M, with SD in parentheses.)*

Engagement dimension	GPT-4 (n=30)	Teacher (n=30)	t	p
Behavioural	4.31 (0.52)	4.18 (0.57)	0.92	.36
Cognitive	4.05 (0.61)	3.97 (0.64)	0.50	.62
Affective	4.46 (0.49)	3.88 (0.66)	3.85	<.001
Overall	4.27 (0.46)	4.01 (0.55)	1.98	.05

Note. Independent-samples t-tests (df = 58). The biggest difference between the two conditions was the affective advantage of GPT-4 feedback.

The behavioural data supporting these scores revealed that learners revising in the GPT-4 condition were more likely to revise and iterate, asking the model for clarification as they could do so as many times as their time allowed, whereas the single-pass nature of teacher feedback did not allow for this. The affective advantage was attributed to two factors repeatedly pointed out by the learners: the feedback was immediate, and there was no evaluative face threat. Meanwhile, the similar cognitive engagement scores—along with the more extensive reported processing load—suggest that the productive interaction with the model was itself demanding—both in terms of having to prompt the model and in terms of the often extensive nature of the model's responses, and having to determine which to follow.

D. Qualitative themes

To elaborate and explain the quantitative pattern, the interviews and reflective journals have been thematically analysed resulting in the four themes summarised in Table IV. The first theme was immediacy of feedback on demand, which was the value that learners attributed to feedback being available whenever they needed it and could be repeated and developed as many times as they wanted. The second, over-reliance and ownership, expressed a concern in the opposite direction: some students feared that they were blindly copying ideas and that the authorial voice of the model was eroding their own – a fear which explicitly mirrors the thinking-inhibitor critique [11]. The third, discourse-level unevenness, represented learners' impression that the model was good on grammar and wording but weak, generic or repetitive on argument and structure—the experiential side of the accuracy gap recorded in Table II. The fourth, a preference for blended human–AI feedback, framed the resolution that the learners themselves proposed: AI for surface corrections (high volume), teachers as higher-order guidance and the kind of relational, trustful mentoring that the learners did not give AI. This preference leans towards augmentation, as opposed to replacement models of AI integration [17], [19].

The qualitative results collectively address RQ3, revealing enthusiastic yet discriminating responses to AI feedback. Learners were not uncritical adopters, nor were they sceptics, but rather they developed a nuanced, experience-based evaluation of the tool's utility, and independently arrived at a complementary configuration which the quality analysis suggests.

Table IV — *Qualitative themes from interviews and reflective journals (paraphrased)*

Theme	Sub-themes	Summary of learner meaning (paraphrased)	Sources (n)
On-demand immediacy	Speed; iteration; autonomy	Learners revised more often because feedback was instant and could be requested repeatedly on demand.	27
Over-reliance & ownership	Dependence; loss of voice	Some feared accepting AI suggestions uncritically and losing their own style and judgement.	19
Discourse-level unevenness	Shallow content cues; redundancy	The model helped with grammar but gave thin or repetitive advice on argument and structure.	22
Preference for blended feedback	AI plus teacher; trust	Learners wanted AI for surface accuracy and teachers for higher-order, relational guidance.	24

Note. n = number of participants (60 in two feedback groups) who had interview/journal data that supported the theme.

V. Discussion

The results provide a nuanced, not a one-size-fits-all, response to the question of whether generative AI can be an effective written corrective feedback provider in EFL writing for academic purposes. GPT-4 AWCF yielded highly significant improvements in both holistic writing quality and local accuracy, which were not significantly different from expert teacher feedback, and was significantly and markedly better than unsupported self-revision, in response to RQ1. This finding aligns with intervention studies that have found performance, self-regulation and motivational benefits from

systematic integration of LLMs in writing instruction [5], [6], and extends this research in two ways: it provides a comparison with a legitimate expert-teacher comparison, not just a no-feedback baseline alone, and it shows this effect in a controlled three-arm study in an under-represented South Asian EFL context. Theory suggests that retention of gains at delayed post testing is important, since it indicates that gains are not simply short term, scaffolded performance but involve a certain level of internalisation, in line with a sociocultural view of contingent mediation that enables gradual restructuring. When it comes to RQ2, however, the feedback-quality analysis calls into question any simple celebration of automation. GPT-4's power was in providing more frequent, detailed and explanatory local feedback, explaining errors and modelling reformulations more often than time-constrained teachers did, plausibly accounting for the marginal accuracy benefit. However, its biggest weakness was exactly at the discourse level, where it was less precise and redundancy and occasional hallucination became apparent—the region where human expertise was still consistently superior. This asymmetry aligns with previous comparative results [4] and the general trend in the automated-scoring literature, that the validity of LLMs is determined by model, rubric, task, and population and not by surface fluency [9], [16], [5]. It also highlights the need to validate rather than simply take for granted assessment-oriented applications of GenAI and the defensibility of framing such applications as a way to reinvent, not just accelerate, assessment [18], [19], [10]. Regarding RQ3, engagement with the AI feedback was high, effortful, and discriminating. The most evident behavioural footprint was the affective benefit of AI feedback—its evaluating-free nature and its immediacy—which confirmed previous case study and perception research [2], [7], [8] and the self-determination account of LLM appeal. However, the higher cognitive demands of prompting and processing, along with learners' expressed concerns of over-reliance and loss of voice, suggest that exposure to AI feedback does not guarantee uptake; rather, it is the quality of engagement that determines uptake [3], [11]. The qualitative finding that learners converged independently on a blended configuration, which is AI for surface accuracy and teachers for higher-order and relational guidance, is striking because it is what the quality analysis would suggest on its own merits. Overall, the study suggests a principled approach to human-in-loop AI assisted feedback, instead of full automation. In such a model, well-prompted LLMs can learn the extensive, explainable, local correction, which takes up a disproportionate amount of teacher time, so that teachers can focus their skills on argument, organization, voice, and relational aspects of mentoring, while learners acquire prompting and critical uptake skills to work productively with AI output. This setting is the most consistently suggested in the literature of the review and comparison [17], [20], [15], [13], [12]. The study suggests that contemporary LLMs are not yet ready to replace human feedback in writing education but might offer a means of redistributing feedback labour, thereby, if pedagogically managed, expanding the scope of individualised feedback to students who are not currently receiving it.

A. Pedagogical implications

There are three implications of this fact. First, the use of GenAI should be clearly framed as a complement to human input, not a replacement for it, based on this comparative analysis of strengths: local accuracy by AI, discourse/relational guidance by teacher. Second, since productive interaction with LLM feedback is also challenging and requires skills, prompting and critical uptake should be taught as competences, so that learners learn to interrogate, verify and selectively accept AI suggestions, rather than blindly accepting them – this reduces the risk of over-reliance as identified by learners. Third, with a known rate of inaccuracy and hallucination at the discourse level, institutions engaging in AI feedback at scale should include a verification and oversight process and not rely on the model's output as authoritative.

B. Methodological reflection

Finally, the study's controlled, integrated design is a methodological contribution, revealing the feasibility and benefit of studying accuracy, feedback quality, and engagement simultaneously, as opposed to separately. The finding that the various strands—quantitative outcomes, corpus-based quality analysis, and learner testimony—converged on the same complementary configuration lends the finding a strength that is not possible with any one strand alone, and thus shows the explanatory power of the integration of mixed methods in this field.

C. Contribution to theory

The study has two theoretical implications, in addition to its practical implications. First, it sets the boundary conditions for LLM feedback efficacy very precisely. Instead of 'AI feedback' as a general one-size-fits-all, the point of efficacy is more likely to be found in the local, explainable, rules-based act of correction, while the fragility of such feedback is more likely to be found in the global, context-dependent act of rhetorical judgement. This level-specific account nuanced the sociocultural reading of feedback as mediation: LLMs seem to be able to provide contingent and explicit feedback that can help to internalise local form, but not so much higher order, dialogic feedback to scaffold argument and voice. Second, the study helps to better understand the moderating role of engagement, as opposed to engagement being just correlated with outcomes. Results indicate that high affective engagement on the part of learners and high reported cognitive load together with explicit over-reliance anxiety, suggest that the relationship between AI feedback and learning may not be monotonic: beyond a certain point ease and authority of AI feedback may diminish the depth of processing, on which uptake depends, of learners. This makes the quality of engagement, and the metacognitive skill of critical uptake, the key ones that pedagogy needs to develop if the affordances of the technology are to be realised, and not wasted.

VI. Conclusion

This study aimed to investigate the effectiveness of GPT-4 as an automated written corrective feedback (WCF) tool for EFL academic writing within an under-represented South Asian context using an explanatory mixed-methods study with a

controlled, sequential design. Three conclusions are noteworthy. LLM-based AWCF can lead to substantial and largely permanent improvements in writing accuracy and quality that are comparable with expert human feedback, and much higher than the gains obtained with unsupported self-revision, thus demonstrating a real pedagogical promise of modern LLM tools.

Second, this potential is uneven – GPT-4 was very good at high-volume and explainable, local correction but also less reliable at the discourse level, where it generated redundancies and, sometimes, hallucination, and where human judgement could not be dispensed with. Thirdly, learners were very positive about interacting with AI feedback, particularly from an affective perspective, but they were also cognitively challenged and concerned about potential over-reliance and loss of authorial voice, preferring a balanced blend between human and AI feedback. The study's most significant contribution is to shift the debate from the question of "Does AI feedback work?" to the more relevant question of "Where, for what, and under what conditions does AI feedback add value? It's main takeaway is that generative AI should be used as an assistant to human feedback in a regulated, human-in-the-loop approach, where the LLM handles mundane tasks like surface correction that take too long for teachers, the teacher handles argumentation, organisation, voice and mentoring, and the learner is specifically trained in prompting and how to critically engage with AI feedback. These claims had a number of restrictions. The study was conducted in a single context, in an 8 week period, with a single model, a controlled prompt configuration, and partially relied on self-reported engagement; there was also only a single cohort of students reported, which limits generalisation.

Future work needs to test other models and disciplines, expand the length of the time horizon to investigate the effects of durability and transfer, and focus directly on engagement through process data like keystroke and revision logs, as well as the equity of access to such language technology in resource-limited contexts where it is most scalable and where the digital divide is most profound. As LLMs continue to develop, evidence of this type will be key to their responsible implementation in language learning to guarantee that the technology is used to support the study of writing, not replace it, and to ensure that the evidence is rigorous, integrated, and contextually diverse.

FUNDING STATEMENT

The author(s) received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, validation, writing—original draft preparation, writing—review and editing, E.K., I.A.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

REFERENCES

[1] Dizon, G., & Gayed, J. M., A systematic review of Grammarly in L2 English writing contexts. *Cogent Education, 11*(1), 2024, 2397882. https://doi.org/10.1080/2331186X.2024.2397882

[2] Yan, D., & Zhang, S., L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications, 11*(1), 2024, 1086. https://doi.org/10.1057/s41599-024-03543-y

[3] Schiller, R., Fleckenstein, J., Höft, L., Horbach, A., & Meyer, J., On the role of engagement in automated feedback effectiveness: Insights from keystroke logging. *Computers & Education, 238*, 2025, 105386. https://doi.org/10.1016/j.compedu.2025.105386

[4] Lin, S., & Crosthwaite, P., The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*, 2024, 103529. https://doi.org/10.1016/j.system.2024.103529

[5] Liu, Z. M., Hwang, G. J., Chen, C. Q., Chen, X. D., & Ye, X. D., Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning, 39*(3), 2026, 466–490. https://doi.org/10.1080/09588221.2024.2389923

[6] Mahapatra, S., Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), 2024, 9. https://doi.org/10.1186/s40561-024-00295-9

[7] Teng, M. F., "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*, 2024, 100270. https://doi.org/10.1016/j.caeai.2024.100270

[8] Du, J., & Alm, A., The impact of ChatGPT on English for academic purposes (EAP) students' language learning experience: A self-determination theory perspective. *Education Sciences, 14*(7), 2024, 726. https://doi.org/10.3390/educsci14070726

[9] Pack, A., Barrett, A., & Escalante, J., Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, 2024, 100234. https://doi.org/10.1016/j.caeai.2024.100234

[10] Pack, A., Hartshorn, K. J., Escalante, J., & Gillette, N., How well can GenAI (GPT-4) provide written corrective feedback on English-language learners' writing? *International Journal of English for Academic Purposes: Research and Practice, 5*(1), 2025, 7–26. https://doi.org/10.3828/ijeap.2025.2

[11] Guo, X., Facilitator or thinking inhibitor: Understanding the role of ChatGPT-generated written corrective feedback in language learning. *Interactive Learning Environments. 2024,* Advance online publication. https://doi.org/10.1080/10494820.2024.2445177

[12] Crosthwaite, P., & Sun, S., Generative AI and L2 written feedback studies: A scoping review. *RELC Journal.* 2026, Advance online publication. https://doi.org/10.1177/00336882251386530

[13] Long, H. S., Exploring the use of ChatGPT as a tool for written corrective feedback in an EFL classroom. *The Journal of AsiaTEFL, 21*(2), 2024, 397–412. https://doi.org/10.18823/asiatefl.2024.21.2.8.397

[14] Liu, Y., Lu, X., & Qi, H., Comparing GPT-based approaches in automated writing evaluation. *Assessing Writing, 66*, 2025, 100961. https://doi.org/10.1016/j.asw.2025.100961

[15] Lee, S., Choe, H., Zou, D., & Jeon, J., Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments. 2025,* Advance online publication. https://doi.org/10.1080/10494820.2025.2498537

[16] Li, W., & Liu, H., Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social*

Sciences Communications, 11(1), 2024, 723. https://doi.org/10.1057/s41599-024-03209-9

[17] Guo, K., & Wang, D., To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 2024, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

[18] Bui, N. M., & Barrot, J. S., ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies, 30*(2), 2025, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

[19] Li, J., Huang, J., Wu, W., & Whipple, P. B., Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications, 11*(1), 2024, 1268. https://doi.org/10.1057/s41599-024-03755-2

[20] Foung, D., Lin, L., & Chen, J., Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence, 6*, 2024, 100250. https://doi.org/10.1016/j.caeai.2024.100250