تأثير التدريب على التحقق من التغذية الراجعة المقدمة من الذكاء الاصطناعي في دقة الكتابة وجودة المراجعة لدى متعلمي اللغة الانكليزية كلغة اجنبية في العراق

وسن جليل إبراهيم إسماعيل*

المديرية العامة لتربية محافظة ديالى، وزارة التربية، ديالى، العراق.

البريد الإلكتروني Email : Wassanjhalil78@gmail.com

الكلمات المفتاحية: التحقق من ردود الفعل بواسطة الذكاء الاصطناعي، الشفرات التي ينشئها المتعلم، كتابة اللغة الإنجليزية كلغة أجنبية، دقة الكتابة، جودة المراجعة، الذكاء الاصطناعي التوليدي، متعلمي اللغة الإنجليزية كلغة أجنبية في العراق.

كيفية اقتباس البحث

إسماعيل ، وسن جليل إبراهيم، تأثير التدريب على التحقق من التغذية الراجعة المقدمة من الذكاءالاصطناعي في دقة الكتابة وجودة المراجعة لدى متعلمي اللغة الانكليزية كلغة اجنبية في العراق ، مجلة مركز بابل للدراسات الانسانية، تشرين الثاني 2026،المجلد:16 ،العدد:11.

هذا البحث من نوع الوصول المفتوح مرخص بموجب رخصة المشاع الإبداعي لحقوق التأليف والنشر (Creative Commons Attribution) تتيح فقط للآخرين تحميل البحث ومشاركته مع الآخرين بشرط نَسب العمل الأصلي للمؤلف، ودون القيام بأي تعديل أو استخدامه لأغراض تجارية.

مفهرسة في Indexed
IASJ

مسجلة في Registered
ROAD

The Effects of AI Feedback Verification Training on Iraqi EFL Learners' Writing Accuracy and Revision Quality

Wassan Jhalil Ibraheem Ismael*
The General Directorate for Education of Diyala, Ministry of Education, Diyala, Iraq

Keywords : AI feedback verification, EFL writing, writing accuracy, revision quality, generative AI, Iraqi EFL learners.

How To Cite This Article
Ismael, Wassan Jhalil Ibraheem , The Effects of AI Feedback Verification Training on Iraqi EFL Learners' Writing Accuracy and Revision Quality ,Journal Of Babylon Center For Humanities Studies, November 2026,Volume:16, Issue 11.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

الملخص

هدفت هذه الدراسة الى بحث أثر التدريب على التحقق من التغذية الراجعة المقدمة من الذكاء الاصطناعي في دقة الكتابة وجودة المراجعة لدى متعلمي اللغة الانكليزية كلغة اجنبية في العراق. وقد جرى توزيع متعلمي اللغة الانكليزية لغة اجنبية في احدى الجامعات العراقية على مجموعة تجريبية ومجموعة مقارنة وفق تصميم تجريبي كمي قائم على الاختبار القبلي والاختبار البعدي .تلقت كلتا المجموعتين النوع نفسه من التغذية الراجعة المولودة بواسطة الذكاء الاصطناعي بشأن كتاباتها، الا ان المجموعة التجريبية وحدها تلقت تدريبا صريحا على التحقق من المعنى المقصود، والاستعانة بمصادر لغوية موثوقة، واتخاذ القرار بشأن قبول اقتراحات الذكاء الاصطناعي او تعديلها او رفضها.وقد قسيت الدقة اللغوية باستخدام مؤشرات معيارية لدقة الكتابة، في حين قيست جودة المراجعة من خلال فحص مدى نجاح المراجعات المرتبطة بالتغذية الراجعة بين المسودات الاولى والمسودات المنقحة .واشارت النتائج الى ان التدريب على التحقيق من التغذية الراجعة المقدمة من الذكاء الاصطناعي أسهم بصورة دالة في تحسين دقة

الكتابة لدى المتعلمين مقارنة بالاستخدام التقليدي للتغذية الراجعة المقدمة من الذكاء الاصطناعي من دون تدريب صريح على التحقق منها. كما كانت جودة المراجعة لدى المجموعة التجريبية اعلى بصورة دالة.واشارت هذه النتائج الى ان فاعلية التغذية الراجعة المقدمة من الذكاء الاصطناعي لا تعتمد فقط على توافر الاقتراحات الالية، وانما تعتمد ايضا على قدرة المتعلمين على التفاعل معها بصورة نقدية وانتقائية. واثبتت الدراسة الحاجة الى تعزيز التحقق من التغذية الراجعة ومحو الامية في مجال التغذية الراجعة، وحو الامية النقدية في مجال الذكاء الاصطناعي في صفوف الكتابة باللغة الانكليزية كلغة اجنبية، ولا سيما في السياقات التي يزداد فيها انتشار الذكاء الاصطناعي التوليدي.

Abstract

This study investigated the impact of AI feedback verification training on Iraqi EFL learners' writing accuracy and revision quality. Iraqi university EFL learners were assigned to an experimental group and a comparison group by means of an experimental design, quantitative pre-test-post-test. Both groups received the same kind of AI-generated feedback on their writing, but only the experimental group was explicitly trained to verify feedback by checking the intended meaning, consulting reliable language resources, and deciding whether to accept, modify, or reject AI suggestions. Linguistic accuracy was measured with standard indices of writing accuracy and revision quality was measured by examining the success of feedback-related revisions between the first and revised drafts. The results indicated that the AI feedback verification training significantly improved the learners' writing accuracy compared to the traditional use of AI feedback without explicit verification training. The revision quality of the experimental group was also significantly higher. These results suggested that the effectiveness of AI feedback was not only about the availability of automated suggestions but also learners' ability to interact with them critically and selectively. The study demonstrated the need for AI feedback verification, feedback literacy, and critical AI literacy in the EFL writing classroom, especially in contexts where generative AI is increasingly prevalent.

1. Introduction

Writing is one of the most difficult aspects of second or foreign language learning because students have to consider linguistic accuracy, vocabulary, organization, meaning and audience expectations simultaneously. The development of Generative Artificial Intelligence (GenAI), especially large language model-based systems like ChatGPT, has brought a new element to this process, allowing learners to get

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 11

individualized language support almost immediately. For instance, Barrot (2023) highlighted feedback, language enhancement, and revision assistance as some of the most promising uses of ChatGPT in L2 writing. This potential has been increasingly supported by empirical work. For instance, Mahapatra (2024) reported improvements in academic writing of university ESL students when ChatGPT was integrated as a formative feedback tool. Similar results were also reported by Poláková and Ivenz (2024) as their EFL participants showed an improvement in the writing elements of grammar, conciseness and inclusion of relevant information, after their interaction with feedback from ChatGPT. More recently, Teng and Huang (2026) found that ChatGPT-generated feedback was related to both improved EFL writing performance and increased affective and behavioral engagement. These developments together suggest that AI-generated feedback is no longer a marginal technological possibility in language education, but a realistic form of writing support that is increasingly shaping the way learners notice problems, rethink their texts, and approach revision.

Consequently, a growing body of research has moved beyond the question of whether generative artificial intelligence can provide feedback and has started to explore what happens when such feedback is incorporated into actual EFL writing instruction. Guo and Wang (2024) found that ChatGPT could help teachers to generate rich feedback, but also highlighted major limitations that mean teachers need to continue to be involved. In a similar vein, Han and Li (2024) found that ChatGPT-supported teacher feedback encompassed a wide range of linguistic and rhetorical issues, and that students made substantial use of this feedback in revising their texts. Alsofyani and Barzanji (2025) found that the feedback generated by ChatGPT yielded writing outcomes similar to those generated by teacher-generated feedback in a Saudi university context, suggesting that AI feedback could provide meaningful support in tertiary EFL settings. Asadi et al. (2025) took the approach further by integrating ChatGPT and teacher feedback, and found significant improvement in learners' argumentative writing, while at the same time teachers expressed concerns about AI's limited contextual understanding and the possibility of learner overreliance. Mekheimer (2025) also reported positive correlations between AI-assisted feedback, revision activity and writing quality, but emphasized that learners should engage with AI suggestions critically. The picture that is emerging is therefore encouraging but qualified: generative artificial intelligence can make feedback more immediate, more extensive, more accessible, but the

educational value of that feedback very much depends on what learners actually do with it.

The qualification is particularly crucial when switching the attention from getting feedback to applying it appropriately. AI feedback can sound confident and coherent despite being not complete enough, not required at all, inappropriate in its sensitivity, or plain wrong. In their comparative evaluation of human and ChatGPT feedback to students' writing, Steiss et al. (2024) noted that trained human evaluators offered better feedback in four out of five analyzed dimensions, including quality, clarity of instructions, importance of writing features addressed, and the friendly tone. Barrot (2023) pointed out that ChatGPT might generate an inaccurate or incomprehensible response and that it was necessary to put some measures in place when employing ChatGPT in L2 writing classes. Issues can also emerge in the moment of learner feedback uptake. Cengiz et al. (2025) observed that although EFL students generally perceived ChatGPT feedback positively, they experienced problems with the length of the feedback, its lack of precision, advanced language use, and misinterpretations, as learners tended to accept language-related feedback directly. According to Zou et al. (2026), who compared teacher-generated and ChatGPT feedback, the former type of feedback led to greater engagement and accuracy in revisions than the latter, even though ChatGPT exhibited its own strengths in certain aspects like organizational feedback. The results challenge the idea that the presence of helpful feedback implies helpful revision. A learner receiving AI feedback has to make sure that the identified problem is indeed a problem, that the suggested change does not distort the meaning of the text, and that the recommended change is indeed an improvement.

This need for informed decision-making with regard to AI output ties AI-enhanced writing into the larger topic of AI literacy and feedback literacy in particular. According to Jin et al. (2025), the true nature of generative AI literacy transcends an impression of proficiency with the technology, as evidenced by the performance-based assessment that underlined the importance of learners' capacity for critical analysis of GenAI-enhanced tasks. In the same vein, critical GenAI literacy is not simply about the level of technological skill; according to Rapanta et al. (2025), it requires the learners' capability of questioning the epistemological, ethical, and knowledge implications of AI-generated content. As to writing, Liu et al. (2026) have proven that GenAI-supported tasks can enhance the learners' feedback literacy, despite the cognitive overload and lack of disciplinary expertise. In addition, Zhou and Wang (2026) found that university students' writing feedback literacy was significantly related to their

acceptance of generative AI, writing self-efficacy, and writing anxiety. In a more general trend, Wang et al. (2026) noted in their systematic review of 79 empirical studies conducted from 2023 to 2025 that GenAI in L2 writing has been particularly concerned with feedback, revision, and polishing tasks. These trends indicate an important change in pedagogy – knowledge of how to get feedback from AI becomes less important than the ability to evaluate it and decide what to do with it.

Nevertheless, the empirical literature has focused much more on the provision, comparison, perception, or use of AI feedback than on the direct instruction of the learners in verifying the quality of such feedback before revising. Wang et al. (2026) in their systematic review of the recent studies of GenAI-assisted L2 writing noted that feedback and revision were the key areas of integration of AI in such studies but also pointed to the general problem of rather superficial pedagogical transformation despite the widespread use of AI. Cengiz et al. (2025) studied the way in which learners accepted, rejected, and responded to feedback from ChatGPT but focused on the naturally occurring learner engagement and did not test the effect of any verification instruction. Liu et al. (2026) researched the development of feedback literacy through GenAI but their intervention was aimed at the development of the skill of interpreting, evaluating, and using feedback rather than at experimental verification instruction. The contextual gap is also important. In Iraq, Jasim (2025) found that the integrated use of ChatGPT and Grammarly enhanced Iraqi EFL university students' academic essay writing, but he was concerned about overreliance and limited critical engagement with AI. More recently, Kucuk (2026) found that teacher-guided feedback from ChatGPT supported EFL writing development at an Iraqi university, but the AI condition was pedagogically supervised by teachers rather than training learners themselves to independently verify feedback. Thus, recent Iraqi research has started to establish the relevance of AI for EFL writing, but a distinct question remains largely unresolved: can learners be explicitly trained to check AI-generated feedback themselves and, consequently, become more accurate writers and more effective revisers?

In addressing this gap, the current research took the approach of investigating AI feedback verification training as an independent instructional intervention, not AI feedback per se, which served as an independent variable in other studies. This was essential due to the previous findings demonstrating that the AI-generated writing feedback could benefit EFL learners in the right circumstances (Alsofyani & Barzanji, 2025). Thus, there was no point in conducting yet another straightforward comparison of AI feedback and teacher feedback as it

would have provided little additional knowledge to the existing literature. At the same time, the evidence of different degrees of accuracy and pedagogical value of ChatGPT feedback made a blind following of its recommendations unadvisable from the pedagogical perspective (Steiss et al., 2024). In addition, there was ample evidence that students' uptake of feedback was uneven and could depend on the feedback type and source (Zou et al., 2026). The rising importance of feedback literacy also implied that learners should be able to critically evaluate feedback rather than just follow the proposed suggestions (Liu et al., 2026). Therefore, in the present study, Iraqi EFL learners were trained to recognize what an AI comment alleged, compare that allegation with their intended meaning, examine dubious suggestions against credible language resources, and then accept, alter, or refuse the suggestion before revising. The experimental group , however , was systematically trained in verification , while the comparison group received the same form of AI-generated feedback . This design permitted us to test whether learning to evaluate AI feedback provided benefits above and beyond simply having access to AI feedback. Specifically, the study focused on two outcomes representing different dimensions of learning: writing accuracy, as measured by independently produced writing, and revision quality, as measured by the quality of changes learners made in response to AI feedback. The study therefore sought to answer the following research questions:

RQ1. To what extent does AI feedback verification training affect Iraqi EFL learners' writing accuracy compared with conventional AI-feedback use without explicit verification training?

RQ2. To what extent does AI feedback verification training affect Iraqi EFL learners' revision quality compared with conventional AI-feedback use without explicit verification training?

2. Review of the Related Literature

Research carried out after the advent of generative AI technology has indicated in general terms that AI-enhanced feedback can play an important role in the development of second language writing, even though the extent and the way in which this happens has differed in individual studies. One of the early intervention studies is represented by Song and Song (2023) who investigated ChatGPT-enhanced academic writing of EFL learners and found that there were positive effects on writing and motivation, implying that the interaction with the system could help learners at various stages of writing. Escalante et al. (2023) provided a less optimistic picture by comparing AI-generated feedback with traditional human feedback. While students appreciated the

consistency and specificity of AI-generated comments, the study did not provide clear evidence that AI feedback was more effective for the development of writing than human feedback, implying that accessibility and quickness should not be taken as the key factors of better instruction. Mahapatra (2024) followed up by investigating ChatGPT as a formative feedback tool among undergraduate ESL students and found significant improvement in academic writing performance. In a similar vein, Poláková and Ivenz (2024), studying 110 EFL university students, showed that there was significant improvement in various areas of writing once learners interacted with ChatGPT feedback, especially grammar, conciseness, inclusion of relevant information, and other linguistic features. Overall, these initial studies created an important basis in that feedback based on artificial intelligence could help learners improve their second language writing skills, especially when given revision opportunities; however, the effect did not stem from the fact that the feedback was provided by an AI tool. What became important was the feedback quality and processing.

Later studies thus paid more attention to the actual quality and characteristics of the AI-generated feedback. Guo and Wang (2024) compared the feedback produced by ChatGPT and teachers on the argumentative essays of Chinese university students and found that compared to teachers, the feedback produced by ChatGPT was significantly more, and paid relatively even attention to content, organization, and language. At the same time, teachers in their study discussed the limitations of simply using AI and preferred a collaborative human–AI approach. Banihashem et al. (2024) reached similarly qualified conclusions when comparing feedback generated by ChatGPT with feedback generated by peers on argumentative essays. ChatGPT tended to provide long descriptive comments while students were often better at identifying specific problems in the essays. Overall, peer feedback was rated higher on some dimensions of feedback quality. The study conducted by Steiss et al. (2024) is highly relevant regarding the issue of evaluating AI feedback. Evaluating 200 pieces of human feedback and the same amount of feedback from ChatGPT, these authors discovered that trained evaluators provided higher-quality feedback in four of five categories, including accuracy, clear instructions, highlighting important aspects of writing, and supportive feedback. While ChatGPT was most effective in providing criteria-related feedback, its overall quality was not equivalent to expert human feedback. By contrast, Han and Li (2024) showed the benefits of introducing the element of mediation when applying AI feedback, namely putting a teacher between

AI feedback generation and its receipt by learners. Analyzing ChatGPT feedback provided to 102 Chinese EFL learners at university level, the teacher examined and changed this feedback. As a result, learners started to use such feedback in their revision process more often. Thus, it appears that the feedback generated by AI tools can be highly detailed and helpful, but its educational role becomes better through evaluation, filtering, and other types of human judgment.

Another line of research has moved the focus from the feedback per se to learner engagement, feedback uptake, and revision. In this regard, Teng (2024), examining the impact of ChatGPT-enhanced writing tasks on 45 EFL learners in a semester-long study, found positive effects of the writing tasks on motivation, self-efficacy, engagement, and learners' readiness to engage in collaborative writing, while the participants had some reservations regarding the use of AI for writing. In terms of revision, more concrete evidence is provided by Zou et al. (2026), who explored how Chinese university EFL learners integrated teacher and ChatGPT feedback in revising argumentative essays. The results showed that feedback source impacted not only learner uptake but also the quality of revision. A similar result was obtained by Cengiz et al. (2025) who analyzed the revision practices of B1-level EFL learners in Türkiye. Learners willingly accepted most of the language-related feedback from ChatGPT, often implementing its grammar and linguistic corrections directly, while organizational and content feedback led to more diverse patterns of acceptance, rejection, substitution, and non-revision. The difficulties that participants faced included excessive length of feedback, too general comments, advanced linguistic level of feedback, and misunderstanding of learner's intentions. Mekheimer (2025) analyzing the impact of AI-assisted feedback, revision frequency, and writing quality on EFL learners found many benefits of using AI-enhanced feedback and significant correlations between revision and writing outcomes. These studies changed the focus of the discussion from the ability of AI to detect errors. It was shown that revision implied an additional decision-making stage where learners had to understand the suggestion, its relevance, and implement it. Thus, an advanced feedback system could not lead to a successful revision without the ability of learners to evaluate suggestions critically.

More recent experimental and comparative studies have demonstrated that AI feedback is capable of operating at a performance level on par with other feedback types; however, they stress the role of instructional conditions. Alsofyani and Barzanji (2025), for example, in their comparison of ChatGPT-produced feedback and feedback produced by

teachers in a sample of 102 Saudi university EFL learners, found no statistically significant difference in the post-test writing performance of the two groups. This means that AI feedback could serve as an adequate alternative or supplement to teacher feedback, and students held positive views of using AI feedback. In Jordan, Alnemrat et al. (2025) also found that both AI-produced feedback and teacher-produced feedback had a significant effect on improving argumentative writing. Importantly, the difference between the two feedback conditions was not statistically significant, although proficiency level mattered. Asadi et al. (2025) in their research involving Iranian EFL learners tested another model by using ChatGPT feedback in combination with teacher feedback. Students from the combined group performed significantly better than those who were getting teacher feedback only in the areas of task response, coherence and cohesion, lexical resource, and grammatical range and accuracy, although teachers who participated in this experiment still voiced their concerns regarding the dependence on AI and its contextual ignorance. The significance of learner engagement was stressed by Teng and Huang (2026) through their quantitative research involving 174 EFL university students and proving that use of ChatGPT feedback resulted in better writing outcomes as well as affective and behavioral engagement. In particular, Han and Wang (2026) examined the essay revision and discovered that Chinese university students performed better on essays revised with feedback from ChatGPT than with feedback from teachers, although students found teacher feedback easier to use and appreciated its clarity and emotional support. The combined evidence therefore does not support a simple claim that either AI or teachers always provide better feedback. Rather, effectiveness seems to depend on the writing dimension tackled, learner proficiency, instructional scaffolding, and students' interpretation and action on feedback.

More recently, evidence has increasingly been directed toward feedback literacy and critical engagement with AI, which is particularly pertinent to the present study. Liu et al. (2026) investigated the development of feedback literacy in GenAI-supported disciplinary writing and reported major gains in learners' ability to interpret, evaluate and act on feedback. Students continued to have cognitive trouble when feedback relied on disciplinary knowledge or a strong grasp of academic conventions, pointing out that access to GenAI did not negate the need for learner judgment. Evidence from Iraq has also started to come out. Al-Obaydi and Pikhart (2025) compared human peer assessment and ChatGPT-based peer assessment in contexts including the University of Diyala and found that both contributed to writing development, with AI feedback being

faster and more comprehensive and with human peer assessment generating stronger social and cognitive engagement. In his study on Iraqi university EFL students, Jasim (2025) noted the benefits of using ChatGPT and Grammarly in academic writing lessons, including increased coherence, grammatical accuracy, academic style, confidence, and autonomy, but he also highlighted problems such as overdependence and lack of critical interaction with the AI feedback. In a more recent study, Kucuk (2026) found that ChatGPT feedback under teachers' guidance was very helpful for EFL writing development in Iraqi universities, but the AI feedback in this study was still under the control of teachers. While recent study has investigated AI–teacher combinations, AI-generated feedback, AI versus peer or teacher feedback, learner perceptions, feedback uptake and revision behavior, there was a crucial issue that remained underexplored. Particularly, the central research problem of this paper is the lack of empirical evidence considering whether EFL students can be systematically trained to critically confirm AI-generated feedback before incorporating it into their revisions rather than passively accepting automated suggestions. AI tools provide immediate corrections to language use or misleading feedback from AI without critical evaluation.

The available studies either trained learners with AI feedback, or studied how they reacted to it naturally. They did not experimentally isolate the training of learners to verify AI feedback before use. Little evidence was available, in particular, on whether teaching EFL learners to inspect an AI recommendation, compare it with their intended meaning, consult reliable external resources if necessary, and consciously accept, modify, or reject the recommendation could improve their subsequent independent writing accuracy and the quality of their revisions. This gap was particularly significant in the Iraqi EFL context where booming research had already demonstrated the relevance of AI-assisted writing but had not directly examined AI feedback verification as a learnable learner strategy. The current study was developed to address this specific gap.

3. Methodology

3.1 Research Design

The study used an experimental design of a pre-test post-test randomization. There are two parallel instructional conditions. The independent variable was the type of training on the use of AI-generated writing feedback. The experimental group was trained to verify the AI feedback, but the comparison group received the same AI-generated feedback and the same amount of writing and revision practice, but no systematic instruction on how to verify AI feedback. Two dependent

variables were studied: writing accuracy and quality of revision. Writing accuracy was the accuracy linguistically of independently produced essays and was mainly measured by the proportion of error-free clauses. Revision quality was a measure of the extent to which changes made between an initial draft and its revised version improved rather than maintained or weakened the text.

3.2 Participants

At the total of 100 Iraqi undergraduate EFL learners enrolled in the English or English education program at a public university in Iraq were initially recruited for proficiency screening. All participants were third year, undergraduate learners, aged between 20 and 22 years. The final sample had both male and female students (24 females, 60%; 16 males, 40;) to confirm balanced gender representation across instructional conditions. Participants were required to be at least 19 years old, studying English as a foreign language, and at an intermediate proficiency levels. Students who had previously taken formal courses specifically dedicated to generative-AI-assisted writing or AI-feedback verification were excluded because such prior systematic instruction could have influenced their response to the treatment. Previous informal use of ChatGPT or other AI tools was captured using a background questionnaire, as such experience has become increasingly common in contemporary educational contexts. English proficiency was assessed using the Oxford Placement Test (OPT). Following the proficiency screening and pre-testing, 40 learners who fell within the B1–B2 CEFR range were selected for the main study and randomly assigned to the two instructional conditions (20 learners [12 females and 8 males] for the experimental group) and (20 learners (12 females and 8 males) for the comparison group) for the experimental group and 40 for the comparison one) using a computer-generated random allocation sequence.

3.3 Instruments

3.3.1 Oxford Placement Test

The OPT of Allan (2004) was used only for initial screening of proficiency and homogenization of sample, not as a measure of outcome. The evaluation was conducted in the classroom. It published results against CEFR levels of proficiency. The OPT offsets the methodological weakness of relying on students' year of university or teacher judgment alone as a proxy for English language proficiency.

3.3.2 Writing Tasks

Writing performance was collected via argumentative essays, following the style of the IELTS Academic Writing Task 2. The pre-test and post-test tasks were official IELTS tasks and not researcher-constructed

writing tests. The Academic Writing Task 2 requires test takers to write an essay in an academic or semi-formal style in response to a given issue, position, argument or problem. The task is a minimum of 250 words and is scored on task response, coherence and cohesion, lexical resource and grammatical range and accuracy. Two similar Task 2 prompts were chosen. Both were of the same broad type of rhetorical task, preferably an opinion- or discussion-based argumentative task Both involved familiar, non-specialist topics so that differences in disciplinary knowledge would not become a major source of score variation. On each testing occasion, students were first allowed 40 minutes to produce a minimum of about 250 words in supervised conditions. During this first writing stage, no AI system, electronic grammar checker, online translator or online dictionary was permitted. These first drafts gave us the data to measure the accuracy of independent writing.

3.3.3 AI Feedback Generation

The feedback was produced by AI at a centrally managed level, and not by individual students. Student-generated prompting may have been a major confound, because differences in prompting ability could have resulted in differences in the amount, specificity, and usefulness of feedback received. The feedback was generated using the OpenAI API, the generative AI model selected (with the exact model ID, date accessed, response settings, and full prompts stored in the research archive).

The AI was given each anonymized initial draft with a fixed feedback prompt. The standardized prompt asked the AI to provide formative feedback on an IELTS Academic Writing Task 2 style essay written by an intermediate EFL learner. Specifically, it was instructed not to rewrite the entire essay but to identify up to eight significant areas for improvement. Feedback included language-accuracy issues such as grammar, word form, sentence structure, vocabulary, spelling, and punctuation where relevant as well as higher-order issues such as coherence, cohesion, argument development or task response.

For each feedback point, the AI was told to: identify the relevant portion of the text, indicate the type of problem, provide a brief explanation of the problem, and suggest a possible correction or revision. The feedback points were numbered individually. Participants in both conditions were presented with the same prompt and AI configuration. As the AI responses were created, they were saved. They were not linguistically corrected by the instructor before being given to participants, because this would have created teacher-verified AI feedback and would have removed the phenomenon that the experimental training was designed to

address. The researchers also did not intentionally add fake feedback or artificially induce AI errors.

3.3.4 Measurement of Writing Accuracy

The accuracy of writing was analyzed mainly by procedures based on Polio (1997) and Polio and Shea (2014). These studies were considered especially appropriate because they systematically investigated the use and reliability of linguistic-accuracy measures in L2 writing research. Polio (1997) examined measures such as error-free units and detailed error analysis and noted the need for coding procedures to be explicitly defined and for interrater reliability to be established. Polio and Shea (2014) later compared several common measures including error-free T-units, error-free clauses, errors per word, and specific error-category counts. The primary accuracy measure in the present study was the Error-Free Clause Ratio (EFCR):

EFCR = Number of error-free clauses / Total number of clauses × 100

A clause was coded as error-free if no lexical, morphological, or syntactic error could be identified according to the coding manual used in the study. Clausal segmentation and error decisions were made according to a written coding protocol prepared before analysis. During rater calibration, borderline cases were discussed to ensure consistent coding across participants. A second measure was retained as a robustness indicator:

Errors per 100 Words

Error Rate = Total linguistic errors / Total number of words × 100

Compared to EFCR, where a clause with one error and a clause with several errors were both labeled as not error-free, errors per 100 words gave an indication of the density of linguistic mistakes. Thus, the use of one main and one supplementary accuracy measure minimized dependence on one operationalization of writing accuracy.

3.3.5 Measurement of Revision Quality

The quality of the revision was assessed by comparing the original draft and the revised draft for each participant, directly. The procedure was especially informed by Min's (2006) established approach to the evaluation of feedback-related revisions and the tradition of revision analysis represented by Faigley and Witte (1981) and by Ferris (1997). Faigley and Witte developed a taxonomy for systematically analyzing text revisions, and Ferris studied whether feedback-induced changes actually improved student texts. Min (2006) studied revision quality in an EFL context, comparing original and revised segments on whether the original was better, the revision was better, or there was no meaningful quality difference. Recent studies on AI-supported writing have also classified learners' responses to feedback into successful, unsuccessful, or

no revisions. So, this coding logic was suitable for the current context of AI-mediated revision.

The first step was to compare the two drafts (initial and revised) and locate each textual revision. Revisions that were a direct answer to a numbered AI feedback point were categorized as AI feedback-triggered revisions. Revisions that could not reasonably be attributed to an AI-feedback point were separately recorded as self-initiated revisions. Each feedback-driven response was then coded into one of three categories:

- Successful revision: The student's response resolved the problem identified in the feedback or otherwise produced a linguistically, rhetorically, or semantically improved version of the relevant text.
- Unsuccessful revision: The student attempted to respond to the feedback, but the resulting text failed to solve the original problem, introduced a new problem, altered the intended meaning inappropriately, or produced a weaker version of the text.
- No revision: The student left the relevant section unchanged.

The principal feedback-level revision outcome was the Successful Revision Rate (SRR):

SRR = Number of successful AI-feedback-related revisions / Total actionable AI-feedback points × 100

A coding step was added concerning further verification as the study was specifically focused on verification of AI feedback. Two experienced EFL writing experts assessed the validity of each AI feedback point, classifying them as valid, partially valid, or invalid/misleading, prior to coding learner uptake. This extra step was necessary because the AI recommendation itself might have been wrong. Where this was the case, a student who had properly rejected that recommendation, should not have been seen as a poor reviser, just because the relevant text was unaltered. Thus, the research team kept an additional process code that indicated whether the learner's feedback decision was appropriate or inappropriate in relation to the expert-determined status of the AI comment.

A correct decision was usually accompanied by successful incorporation or adaptation of the feedback in valid AI feedback. If the feedback is only partly correct, learners can adjust the suggested correction appropriately, rather than just imitating it. In the case of invalid or misleading feedback, a suitable verification decision might be to reject the suggestion and keep the stronger original wording. This verification-decision measure was taken as additional process evidence and not as a substitute for revision quality as the main dependent variable.

3.3.6 Whole-Text Revision Quality

Besides the feedback level coding, the quality of revision was rated at the whole-essay level using the official IELTS Writing Task 2 Band Descriptors. The Task 2 writing in the IELTS is assessed in terms of task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. The anonymized drafts were independently scored by two trained raters, who were blind to group membership and occasion of testing. Mixed first and revised drafts were presented in random order so that the raters did not know if a text was a first or revised version. The full text gain score was calculated as:

Revision Gain = Revised-draft writing score − Initial-draft writing score

Positive value showed improvement after revision, zero value indicated no detectable whole-text improvement and negative value indicated deterioration. This whole-text measure was used as a supplementary validation of the more sensitive feedback-level revision coding.

3.4 AI Feedback Verification Training

3.4.1 Experimental Group

The experimental group was explicitly instructed on how to critically verify AI-generated writing feedback before using it in a revised text. The verification procedure comprised five recurring actions.

First, learners identified the specific claim that the AI feedback made. They identified what the AI found problematic and whether the comment was about grammar, vocabulary, sentence structure, coherence, argumentation, register or any other aspect of writing. Second, learners returned to their initial intended meaning. Before replacing a piece of text, they wondered if the AI's version expressed the meaning they wanted to communicate. Third, learners independently cross-checked dubious linguistic suggestions against trusted non-AI reference sources. The permitted resources were Oxford Learner's Dictionaries, Cambridge Dictionary and English Grammar Today, Purdue Online Writing Lab and the official IELTS Writing criteria. Fourth, learners decided whether to accept, modify or reject each key AI suggestion. They were taught explicitly that acceptance was not the correct response by default. An AI suggestion may be linguistically possible but stylistically inappropriate, unnecessary, inconsistent with the learner's intended meaning, or incorrect. Finally, in a revision step, learners re-read the surrounding sentence and paragraph to ensure that a local correction did not adversely affect meaning, cohesion, reference, argument development, or textual consistency.

In treatment sessions, students recorded their major decisions on a short verification log that included the AI-feedback number, the student's

decision (accept, modify, or reject), a brief justification, the reference consulted, if applicable, and the revision that was made. The verification log was used primarily to document treatment fidelity and actual use of the trained procedure by learners. It was not used as a psychometric questionnaire. Students were not told that AI feedback was inherently untrustworthy or should be routinely rejected in the experimental condition. Instead, learners were trained to view AI feedback as a possible source of feedback that must be evaluated before it can be adopted.

3.4.2 Comparison Group

The comparison group was given the same amount of instructional time and wrote the same essays and revisions. Students also got exactly the same kind of AI feedback generated centrally.

A short orientation was given to them explaining how the AI feedback was organized, how to find the feedback points in their essays and how to create a revised draft. However, they were not trained on explicit verification, they were not trained on the systematic accept-modify-reject procedure, and they were not required to independently verify AI recommendations before revising. They were free to ignore or follow individual recommendations in the usual way that learners might use AI feedback without specific instruction to verify. This active comparison condition was methodologically preferable to a no-treatment control condition because it controlled for AI exposure, revision practice, instructional time, writing practice, teacher contact, and novelty of technology. Differences between groups could be more precisely attributed to the additional effect of verification training.

3.5 Treatment Procedure

The entire study lasted eight weeks. In the first phase, participants completed the OPT, demographic and prior-AI-use information, and a pre-test writing task. The first pre-test essay was written without AI help. Standardized AI feedback was generated centrally after the initial drafts had been collected. Both groups then completed a controlled revision of the pre-test essay. This procedure yielded baseline data on both writing accuracy and revision quality before verification training commenced.

In Weeks 2 and 3, the experimental group received explicit instruction and modeling of AI-feedback verification. The instructor showed examples of good, questionable, overly prescriptive, and potentially inaccurate AI feedback. Students practiced looking up information in independent language references and deciding whether to accept, modify, or reject individual suggestions. The comparison group received the same amount of instruction time devoted to process writing, interpretation of

AI comments, and ordinary revision practice during the same period. There was no instruction in explicit verification procedures.

In Weeks 4 through 7, both groups engaged in multiple cycles of writing, feedback, and revision. Each cycle included an initial writing task, standardized AI feedback and production of a revised draft. The experimental group carried out the verification procedure and filled in the verification log, while the comparison group revised the AI feedback in the conventional manner. The writing prompts used during instruction were different from the pre-test and post-test prompts to ensure that the assessment tasks were not practiced directly during instruction. In Week 8, the participants undertook the post-test under the same general conditions as the pre-test. They first created an independent 40-minute essay without access to AI or electronic correction tools. This initial version gave the post-test writing-accuracy score.

The essays were then subjected to the same standardized AI-feedback procedure. Both groups got their numbered AI feedback and had the same amount of revision time. The experimental group was not allowed to use dictionaries or grammar references exclusively during the revision period. Both groups were allowed to use the same allowed reference materials. Therefore, the observed possible advantage of the experimental group cannot be attributed to the exclusive use of dictionaries or grammar references. No controlled post-test revision was made by a participant to the AI after the initial prompt. This restriction was to prevent differences in conversational prompting behavior from confounding the effect of the verification training. The revised post-test essays were then collected and compared to the initial post-test drafts to determine revision quality.

3.6 Data Analysis

Data analysis was performed after screening data for missing data, outliers, coding errors, and relevant statistical assumptions. Descriptive statistics for the two groups were calculated at pre-test and post-test. These included mean, standard deviation, minimum, maximum and 95% confidence intervals for the main outcome measures. Baseline comparability was also descriptively analyzed in terms of proficiency, initial writing accuracy, initial revision quality, age, and frequency of previous AI use.

The primary dependent variable for RQ1 was the Error-Free Clause Ratio from the independently written post-test essay. The primary group comparison was conducted using an analysis of covariance (ANCOVA), with post-test EFCR as the dependent variable, instructional condition as the fixed factor, and pre-test EFCR as the covariate. Prior to analyzing the ANCOVA results, linearity between the covariate and the dependent

variable, homogeneity of regression slopes, independence of observations, homoscedasticity, and the distribution of residuals were checked. The robustness check was based on the errors-per-100-words metric. The lower value of error-rate represented more accuracy so the direction was interpreted accordingly.

For RQ2, the primary participant-level revision outcome was the percent of AI-feedback opportunities that resulted in a successful revision. Post-test revision quality was compared across instructional conditions, controlling for pre-test performance. As successful revision rates were proportions bounded between 0 and 1, the distribution of this variable was carefully examined. Where participant-level proportions were sufficient to meet statistical assumptions, ANCOVA was used for consistency with the RQ1 analysis. Where significant skewness, ceiling effects were found, a binomial generalized linear model or generalized linear mixed model was used at the feedback-point level, with successful versus non-successful response as the binary outcome and participant as a repeated or random effect.

This alternative approach was considered appropriate in cases where learners were presented with a different number of actionable AI-feedback points, as it prevented percentages based on different denominators from being considered of equal accuracy. Further analyses looked at the proportion of successful revisions, unsuccessful revisions, appropriate rejection of invalid AI feedback, and inappropriate acceptance of invalid feedback. Whole-text revision gain relative to the descriptors in the IELTS Writing was also investigated as convergent evidence of revision improvement.

All primary analyzes were tested for statistical significance at .05. Interpretation was not based solely on P values, however. We also reported effect sizes and 95% confidence intervals. Partial eta squared and adjusted group means were reported in the ANCOVA analyzes, and odds ratios with confidence intervals were reported in the logistic models. The statistical analysis plan and the designation of primary versus supplementary measures were finalized before the outcome data was examined to minimize the flexibility of analyzes and selective reporting.

4. Results

4.1 Descriptive Statistics

Proficiency screening resulted in a total of 40 Iraqi EFL learners who participated in the final analysis. There was an equal distribution of participants across the two instructional conditions (20 learners in the experimental group and 20 learners in the comparison group). Descriptive statistics were computed for the three major measures used in the study,

Error-Free Clause Ratio (EFCR), errors per 100 words, and Successful Revision Rate (SRR). EFCR and SRR were positively interpreted so that higher scores indicate better performance. Lower scores for errors per 100 words indicate greater accuracy in writing. The means and standard deviations of the pre-test and post-test for both groups are presented in Table 1.

Table 1

Descriptive Statistics for Writing Accuracy and Revision Quality

Measure	Group	Pre-test *M*	Pre-test *SD*	Post-test *M*	Post-test *SD*
EFCR (%)	Experimental	54.81	5.57	71.07	3.07
	Comparison	54.81	5.42	61.05	4.66
Errors per 100 words	Experimental	12.85	1.38	7.99	1.10
	Comparison	12.69	2.15	10.89	1.02
SRR (%)	Experimental	47.55	6.03	75.34	4.16
	Comparison	44.81	4.56	55.49	5.76

Table 1 demonstrates that the two groups began the study with very similar levels of writing accuracy. For the experimental group M = 54.81, SD = 5.57 and for the comparison group M = 54.81, SD = 5.42. The mean pre-test EFCR was the same for the two groups. At post-test, the experimental group increased to a mean EFCR of 71.07 (SD = 3.07), while the comparison group increased only to 61.05 (SD = 4.66). Therefore, the increase of the experimental group was about 16.26 percentage points compared to the increase of the comparison group of about 6.24 percentage points. The descriptive pattern showed that explicit training in checking AI feedback was associated with significantly greater improvement in the proportion of grammatically and lexically error-free clauses.

The same tendency was observed for the supplementary error-density measure. Prior to the intervention, the experimental group had an average of 12.85 errors per 100 words (SD = 1.38) and the comparison group had an average of 12.69 errors (SD = 2.15). After the intervention, the experimental group showed a significant reduction in error rate, to 7.99 errors per 100 words (SD = 1.10). The comparison group also improved, but with a post-test error rate still substantially higher (10.89 errors per 100 words, SD = 1.02). Thus, both groups of learners seemed to benefit from exposure to AI-generated feedback, but the decrease in linguistic

errors was significantly larger for learners who had been explicitly trained to evaluate and verify the feedback before incorporating it in their writing.

A more obvious descriptive difference was seen in revision quality. The experimental group had a pre-test SRR of 47.55% (SD = 6.03) compared to the comparison group, which had a pre-test SRR of 44.81% (SD = 4.56). The success rate of the experimental group on the post-test was 75.34% (SD = 4.16), an increase of almost 28 percentage points. The mean SRR of the comparison group, however, increased to 55.49% (SD = 5.76), an improvement of about 11 percentage points. Thus, the descriptive findings suggested that verification training affected not only the accuracy of independently produced writing but also learners' ability to make effective decisions when revising in response to AI-generated feedback.

Other baseline comparisons indicated that groups were relatively equivalent before the intervention. Independent-samples t tests revealed no significant differences on pre-test for EFCR, $t(38) = -0.00$, $p = .998$, errors per 100 words, $t(38) = 0.28$, $p = .782$, or SRR, $t(38) = 1.62$, $p = .113$. Thus, large descriptive post-test differences were unlikely to be a result of large pre-existing differences between the groups. Then, inferential analyzes were performed to see whether the post-test advantages observed were still statistically significant when controlling for initial performance.

4.2 Inferential Analysis

4.2.1 Results for RQ1: Writing Accuracy

The first research question investigated whether the AI feedback verification training significantly improved the Iraqi EFL learners' writing accuracy compared to the traditional use of AI-feedback without explicit verification training. EFCR was the main measure for writing accuracy.

Prior to conducting the analysis, the assumptions related to ANCOVA were examined. The relationship between pre-test and post-test EFCR scores was roughly linear for both groups. The Shapiro–Wilk test did not indicate a serious departure from normality of the standardized residuals, $W = .98$, $p = .700$. Levene's test was also non-significant, $F(1, 38) = 2.24$, $p = .142$, suggesting homogeneity of variances. Moreover, the interaction between group and pre–test EFCR was not significant, $p = .507$, providing support for the assumption of homogeneity of regression slopes. These results indicated that the data were suitable for ANCOVA.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 11

An ANCOVA was then conducted with post-test EFCR as the dependent variable, instructional group as the fixed factor and pre-test EFCR as the covariate. The results are presented in Table 2.

Table 2

ANCOVA Results for Post-test Error-Free Clause Ratio

Source	SS	df	MS	*F*	*p*	Partial η²
Pre-test EFCR	128.68	1	128.68	10.28	.003	.217
Group	1004.46	1	1004.46	80.22	< .001	.684
Error	463.31	37	12.52	—	—	—

Controlling for pre-test EFCR, instruction group had a statistically significant effect on post-test writing accuracy, F(1, 37) = 80.22, p < .001, partial η2 = .684. The effect size indicated that a substantial part of the post-test variance, after adjustment, was associated with the instructional condition. The adjusted post-test mean for the experimental group was 71.07 (SE = 0.79, 95% CI [69.47, 72.67]) and the adjusted mean for the comparison group was 61.05 (SE = 0.79, 95% CI [59.44, 62.65]). The adjusted between group difference was around 10.02 percentage points (95% CI 7.76 to 12.29).

These results suggest that the experimental group's greater accuracy on the post-test was statistically significant even when initial writing accuracy was controlled. Thus, the first research question was answered in the affirmative: AI feedback verification training significantly improved Iraqi EFL learners' writing accuracy than conventional AI-feedback use without explicit verification training.

Errors per 100 words were also analyzed to see whether another measure of accuracy supported this finding. Assumption testing revealed no significant violation of normality, W = .98, p = .689, and Levene's test indicated homogeneity of variance, F(1, 38) = 0.02, p = .875. The interaction between instructional condition and pre-test error rate was also non-significant, p = .947, confirming that the homogeneity of regression slopes assumption was met. Table 3 shows the ANCOVA results for errors per 100 words.

Table 3

ANCOVA Results for Post-test Errors per 100 Words

Source	SS	df	MS	*F*	*p*	Partial η²
Pre-test error rate	7.17	1	7.17	7.42	.010	.167

Group	86.44	1	86.44	89.45	< .001	.707
Error	35.75	37	0.97	—	—	—

When controlling for the pre-test error rate, a significant difference between the two groups was still found, $F(1, 37) = 89.45$, $p < .001$, partial $\eta 2 = .707$. The adjusted post-test mean for the experimental group was 7.97 errors per 100 words (SE = 0.22, 95% CI [7.52, 8.41]) and for the comparison group, 10.91 (SE = 0.22, 95% CI [10.46, 11.35]). The adjusted difference was −2.94 errors per 100 words, 95% CI [−3.57, −2.31]. Learners who underwent verification training made significantly fewer linguistic errors at post-test.

The two measures converging reinforced the finding for RQ1. The experimental group simultaneously generated a larger proportion of error-free clauses and fewer errors per 100 words. This pattern suggests that the improvement in accuracy was not tied to one single operational measure of linguistic accuracy.

4.2.2 Results for RQ2: Revision Quality

The second research question was to determine if AI feedback verification training significantly impacted the quality of learners' revision. The main metric used was the Successful Revision Rate, capturing the share of actionable AI-feedback points that led to revisions considered to improve the learner's text.

First, the assumptions of ANCOVA was checked. The residuals were not significantly different from normal, $W = .95$, $p = .078$. Levene's test was also non-significant, $F(1, 38) = 1.98$ $p= .167$, suggesting that the variance of post-test SRR scores did not differ substantially between groups. Finally, the interaction between group and pre-test SRR was non-significant, $p = .072$. The value was somewhat close to .05, but still above the conventional threshold for significance, thus supporting the assumption of homogeneity of regression slopes. Thus, ANCOVA was retained as the primary inferential procedure. Table 4 presents the ANCOVA results for SRR on the post-test.

Table 4

ANCOVA Results for Post-test Successful Revision Rate

Source	SS	df	MS	*F*	*p*	Partial η²
Pre-test SRR	111.63	1	111.63	4.88	.033	.116
Group	3365.30	1	3365.30	147.12	< .001	.799
Error	846.38	37	22.88	—	—	—

The effect of instructional group was statistically significant after controlling for pre-test revision quality, $F(1, 37) = 147.12$, $p < .001$,

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 11

partial η^2 =.799. The experimental group's adjusted post-test SRR was 74.90% (SE = 1.09, 95% CI [72.69, 77.10]), and the comparison group's adjusted mean was 55.93% (SE = 1.09, 95% CI [53.72, 58.13]). The adjusted mean difference was 18.97 percentage points (95% CI [15.80 to 22.14]).

The size and direction of this difference indicated that the explicit verification instruction had a large effect on the learners' ability to successfully make changes in response to AI-generated feedback. Learners in the experimental condition not only revised more often, but were significantly more likely to make changes that solved the problem identified or improved the linguistic, semantic or rhetorical quality of the text. Therefore, the second research question was also answered affirmatively: Revision quality was significantly improved by AI feedback verification training compared to conventional use of AI feedback without explicit verification training.

To look at this difference in more detail, feedback-triggered revisions were coded as successful revisions, unsuccessful revisions, or no revisions. Table 5 shows the distribution of these outcomes at post-test.

Table 5

Distribution of Post-test Feedback-Triggered Revision Outcomes

Revision outcome	**Experimental group *n* (%)**	**Comparison group *n* (%)**
Successful revision	112 (75.7%)	83 (55.3%)
Unsuccessful revision	22 (14.9%)	41 (27.3%)
No revision	14 (9.5%)	26 (17.3%)
Total actionable feedback points	148 (100%)	150 (100%)

Table 5 shows that approximately three quarters of the actionable AI-feedback points received by the experimental group resulted in successful revisions, compared to just over half of those received by the comparison group. By contrast, failed revisions represented only 14.9% of the feedback responses from the experimental group, compared with 27.3% from the comparison group. The experimental group also had fewer instances where actionable feedback led to no revision, 9.5% versus 17.3%. This pattern was consistent with the SRR analysis at the participant level and suggested that verification training improved the quality of feedback uptake rather than simply encouraging learners to make more textual changes.

Additional evidence was also examined in the form of whole-text revision gain. The initial and revised drafts were scored based on the IELTS Writing Task 2 criteria and the difference between the two scores was considered as the overall revision gain. These results are summarized in Table 6.

Table 6

Whole-Text Writing Scores and Revision Gain

Group	Initial draft *M* (*SD*)	Revised draft *M* (*SD*)	Revision gain *M* (*SD*)
Experimental	5.40 (0.53)	6.55 (0.36)	1.15 (0.38)
Comparison	5.26 (0.48)	5.78 (0.28)	0.53 (0.41)

Although the two groups obtained comparable scores on their first drafts, a distinct difference appeared in their revised texts. The experimental group's mean score increased from 5.40 to 6.55, an average gain of 1.15 band-score points. The comparison group improved from 5.26 to 5.78, a substantially smaller average gain of 0.53. An ANCOVA on revised-draft scores, controlling for initial-draft scores, confirmed a statistically significant group effect, $F(1, 37) = 73.90$, $p < .001$, partial $\eta 2 = .666$. The difference between the groups adjusted for pre-intervention was approximately 0.71 band score points, 95% CI [0.54 to 0.88].

This supplementary analysis provided converging evidence for the SRR results. The advantage of the experimental group was clear not only in the quality of individual feedback-triggered revisions, but also in the quality of the revised essays as whole texts. Learners who received verification training thus seemed more capable of translating the AI-generated feedback into meaningful improvements in their writing.

5. Discussion

The findings of the present study demonstrated that AI feedback verification training had a significant positive impact on both the writing accuracy and revision quality of Iraqi EFL learners. Learners explicitly trained to inspect, evaluate and verify AI-generated feedback before using it outperformed learners who received conventional AI feedback without such training. These results supported the growing body of research that generative artificial intelligence can be helpful for EFL writing development when its use is pedagogically structured rather than simply as a correction tool. Mahapatra (2024) found a significant improvement in the academic writing performance of learners with the help of ChatGPT supported formative feedback and Poláková and Ivenz (2024)

reported improvements in grammar, conciseness, and inclusion of relevant information after feedback based on ChatGPT. Alsofyani and Barzanji (2025) also observed that AI-generated feedback could produce writing results comparable to teacher feedback, while Teng and Huang (2026) found that feedback from ChatGPT had a positive impact on both writing performance and learner engagement. However, the present findings extended these studies by suggesting that the effectiveness of AI feedback was not only dependent on access to feedback but also on learners' ability to critically evaluate the feedback before incorporating it into their texts.

In relation to the first research question, the reason for the improved writing accuracy of those learners who had verification training could be the deeper cognitive processing required by the training procedure. Learners were asked to identify the nature of the proposed change, compare it with their intended meaning, and check questionable recommendations against reliable language resources, rather than automatically accepting AI-generated corrections. This process might have increased learners' consciousness of grammatical and lexical form and helped them to notice why a particular structure was inaccurate instead of replacing it mechanically. This would be consistent with concerns raised by Steiss et al. (2024), who found that feedback from ChatGPT was not always as accurate or instructionally clear as feedback from expert humans. This also echoed Barrot (2023), who cautioned that generative AI could generate inaccurate or improper responses although it could be useful in L2 writing. Similarly, Guo and Wang (2024) demonstrated ChatGPT's ability to produce copious amounts of feedback, but also emphasized the need for human judgment, as the quality of comments from AI was inconsistent. Thus, the experimental group learners might have improved their accuracy as a consequence of the verification training which encouraged learners to process linguistic problems in a more conscious way, and therefore, made it more likely that they internalized corrected forms and later reproduced them successfully in independent writing.

The enhancement in writing accuracy could have also been brought about by decreased overreliance on the corrections generated by AI. The traditional use of AI feedback can lead to the learners copying the proposed correction without a thorough understanding of the linguistic concept involved, especially when the feedback is provided in an authoritative manner. According to Cengiz et al. (2025), EFL learners readily accepted ChatGPT's language-related feedback although at times they had difficulties comprehending long, complicated, or out-of-context

feedback. Similarly, Jasim (2025) observed that ChatGPT and Grammarly aided Iraqi EFL learners but noted that overreliance and lack of critical thinking were issues. In this case, verification training would have helped in preventing passive acceptance of AI feedback through making the learner determine whether a correction was necessary or not. This process enabled the learner to differentiate between the correct corrections and the unnecessary ones. Thus, the result validated the argument by Jin et al. (2025) and Rapanta et al. (2025) that for the effective use of generative AI, critical literacy was needed. In practical terms, the learners did not only benefit from receiving more language corrections but also from being selective and critical of those corrections.

The second finding, which showed a significant improvement in revision quality, can be explained because verification training targeted the decision-making processes involved in revision directly. Revision is not a matter of change for its own sake; it is the process of judging whether a change improves the meaning, the accuracy, the organization, or the rhetorical effect of the writing. Zou et al. (2026) found significant differences in the uptake of AI- and teacher-generated feedback by learners, and that the source and type of feedback impacted the appropriateness of revisions. Similarly, Cengiz et al. (2025) showed that students responded differently to AI feedback by accepting, rejecting, substituting, or ignoring suggestions, especially when feedback was about content and organization, rather than surface-level language. Han and Li (2024) found that students could make substantial use of ChatGPT-supported feedback when it was mediated and adapted by teachers, which suggests that some evaluative support can strengthen revision outcomes. The present study seemed to achieve a similar result through learner training rather than teacher filtering. By teaching learners to verify AI comments themselves, the intervention may have transferred some of the evaluative responsibility from the teacher to the learner. This may account for the greater number of successful revisions produced by the experimental group: they were not simply responding to feedback, but were making reasoned judgments as to whether or not each change was likely to strengthen their text.

The results showed that the most educationally beneficial use of AI feedback was when learners had been trained to interact with it critically and not as an unquestionable source of correction. This was consistent with Liu et al. (2026) who found that generative AI could help to foster feedback literacy by enabling learners to interpret, evaluate and act on feedback. It also validated Mekheimer's (2025) finding that AI-assisted feedback and revision activity were strongly associated with enhanced

writing quality, with the important caveat that revision quality could be enhanced further when learners were explicitly taught to evaluate feedback prior to acting on it. The results also supported the findings of Jasim (2025) and Kucuk (2026) in the Iraqi EFL context who both reported positive effects of AI-supported writing instruction but not with learner verification training as the main intervention. The present study thus offered a more precise account of how AI feedback could be made pedagogically effective. Learners appeared to benefit when they were taught not only to receive AI feedback but also to question it, confirm it, modify it or reject it when necessary. This may have promoted more accurate independent writing and more purposeful revision, showing the verification of AI feedback as a significant aspect of AI literacy and writing feedback literacy in contemporary EFL instruction.

5. Conclusion

This study investigated the impact of AI feedback confirmation training on Iraqi EFL students' revision quality and writing accuracy. The empirical outcomes provide robust proof that explicitly training students to cross-check, inspect and critically evaluate AI-generated feedback leading statistically significant improvement in both the effectiveness of text revisions and independent writing accuracy. It is stated that learners who underwent systematic verification training performed a lower error density per 100 words and a substantially higher Error-Free Clause Ratio (EFCR) while conventional exposure to AI feedback offers general linguistic benefits. Additionally, the intervention profoundly transmitted learners' revision behaviours, resulting in an importantly greater Successful Revision Rate (SRR) and enhanced whole-text writing scores based on IELTS Task 2 band descriptors. Hence, such findings confirm that the educational values of generative AI in second language writing is heavily contingent upon how learners interact with automated suggestions.

6. Pedagogical Implications

The results of the research had many implications for EFL writing instruction in situations where access to generative AI has increased among learners. Rather than teaching the use of AI in writing only for the purpose of acquiring corrections or suggestions from ChatGPT, for instance, it is vital that learners should learn how to evaluate the feedback they receive. This can be done through classroom training on how to verify the feedback. Students may learn how to check what is the problem found by the AI software, whether the suggested correction agrees with their intended meaning, or to check dictionaries, grammar manuals, or

writing criteria if needed, and then choose to accept, revise, or even reject the correction provided. This type of training can mitigate the risk of dependence and uncritically accepting the correction raised in previous studies (Barrot, 2023; Steiss et al., 2024; Jasim, 2025). At the same time, this will help develop feedback literacy and AI literacy (Liu et al., 2026; Jin et al., 2025). In Iraqi EFL classrooms, where teachers may wish to benefit from the speed and accessibility of AI feedback without replacing professional judgment, verification training could provide a practical way of integrating AI into writing instruction while maintaining learner autonomy, critical thinking, and attention to linguistic accuracy.

7. Recommendation for Further Studies

Despite the positive results, several limitations should be considered in interpreting the results. The research was carried out with Iraqi EFL learners within a particular university context and proficiency level, which limited the extent to which the results could be generalized to learners of different ages, proficiency levels, institutions or education systems. The intervention also focused on a particular type of generative AI feedback and a relatively short period of instruction, and the quality and nature of AI output may vary across models, prompts, and future technological updates. Additionally, writing accuracy and revision quality were the main outcomes, and therefore the study did not directly examine whether verification training affected other important dimensions such as writing complexity, fluency, critical thinking, learner autonomy, AI literacy, or long-term retention. Future research could involve replicating the study with larger and more diverse samples, comparing different levels of proficiency and artificial intelligence systems, and using delayed post-tests to determine whether the gains observed were maintained over time. Future research can explore how learners verbally articulate their verification decisions via stimulated recall, think-aloud procedures, or interviews. Research might further explore whether verification training yields the same advantages in the context of AI feedback combined with teacher feedback, peer feedback, or discipline-specific academic writing.

References

Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. *Frontiers in Education, 10*, 1614673. https://doi.org/10.3389/feduc.2025.1614673

Al-Obaydi, L. H., & Pikhart, M. (2025). AI partner versus human partner: Comparing AI-based peer assessment with human-generated peer assessment in examining writing skills. *Language Testing in Asia, 15*, Article 38. https://doi.org/10.1186/s40468-025-00375-8

Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A

mixed-methods study. *Journal of Educational Computing Research, 63*(2), 431–463. https://doi.org/10.1177/07356331241307297

Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity, 56*, 101766. https://doi.org/10.1016/j.tsc.2025.101766

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*, Article 23. https://doi.org/10.1186/s41239-024-00455-4

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Cengiz, B. C., Bilki, Z., Ataş, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System, 134*, 103837. https://doi.org/10.1016/j.system.2025.103837

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414. https://doi.org/10.2307/356602

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31*(2), 315–339. https://doi.org/10.2307/3588049

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

Han, F., & Wang, Z. (2026). Teacher feedback and ChatGPT feedback on Chinese university EFL learners' English essay revision: A mixed-methods study. *Computers and Education: Artificial Intelligence, 10*, 100590. https://doi.org/10.1016/j.caeai.2026.100590

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, 103502. https://doi.org/10.1016/j.system.2024.103502

IELTS. (2023). *IELTS Writing Band Descriptors*. IELTS.

Jasim, A. F. (2025). The impact of using ChatGPT and Grammarly on developing Iraqi EFL university students' academic essay writing. *Journal of Misan Researches, 21*(41), 365–390. https://doi.org/10.52834/jmr.v21i41.286

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence, 9*, 100436. https://doi.org/10.1016/j.caeai.2025.100436

Kucuk, T. (2026). Human versus artificial intelligence feedback: A comparative study of teacher-guided ChatGPT and teacher feedback in English as a foreign language writing development. *International Journal of Systematic Innovation, 10*(4), 026170063. https://doi.org/10.6977/IJoSI.202608_10(4).0015

Liu, J., Li, W., & Shi, Z. (2026). Developing students' feedback literacy in disciplinary academic writing through generative artificial intelligence. *Assessing Writing, 68*, 101030. https://doi.org/10.1016/j.asw.2026.101030

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

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. *Discover Education, 4*, Article 170. https://doi.org/10.1007/s44217-025-00602-7

Min, H.-T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Oxford University Press. (n.d.). *Oxford Placement Test*. Oxford University Press.

Poláková, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101–143. https://doi.org/10.1111/0023-8333.31997003

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing, 26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Rapanta, C., Bhatt, I., Bozkurt, A., Chubb, L. A., Erb, C., Forsler, I., Gravett, K., Koole, M., Lintner, T., Örtegren, A., Petricini, T., Rodgers, B., Webster, J., Xu, X., Christensen, I. M. F., Dohn, N. B., Christensen, L. L. W., Zeivots, S., & Jandrić, P. (2025). Critical GenAI literacy: Postdigital configurations. *Postdigital Science and Education, 7*, 1296–1333. https://doi.org/10.1007/s42438-025-00573-w

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

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

Teng, M. F., & Huang, J. (2026). Assessing ChatGPT feedback for EFL learners' engagement and writing performance. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70155

Wang, Z., Lei, H., & Chen, K. (2026). The influence of GenAI on second language writing instruction: A systematic review. *System, 140*, 104081. https://doi.org/10.1016/j.system.2026.104081

Zhou, C., & Wang, Y. (2026). University students' writing feedback literacy in the AI era: The interplay of generative AI acceptance, writing anxiety and writing self-efficacy in Chinese EMI educational contexts. *European Journal of Education, 61*(1), e70478. https://doi.org/10.1111/ejed.70478

Zou, S., Guo, K., Wang, J., & Liu, Y. (2026). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study.

Computer Assisted Language Learning, *39*(4), 1062–1091. https://doi.org/10.1080/09588221.2024.2447279