ARTICLE

JOURNAL OF RESEARCH STUDIES IN ENGLISH LANGUAGE TEACHING AND LEARNING

To cite this article in APA 7th format:
Huy, P. (2026). Beyond the AI feedback paradox: How metacognitive re-flection transforms AI corrections into L2 learning. *Research Studies in English Language Teaching and Learning*, *4*(4), 823–846. https://doi.org/10.62583/rseltl.v4i4.150

More Citation Formats

For more citation styles, please visit: https://rseltl.pierreonline.uk/

This article is published by Pierre Online Publications Ltd, a UK publishing house.

eISSN: 2977-0394

KEYWORDS
AI-generated feedback, metacognitive reflection, written corrective feedback, L2 writing, grammatical knowledge

Correspondence concerning this article should be addressed to:
sutnym99@gmail.com

Beyond the AI feedback paradox: How metacognitive reflection transforms AI corrections into L2 learning

Phạm G. Huy

Ho Chi Minh City University of Foreign Languages and Information Technology, Vietnam

Abstract

Despite its importance in L2 writing classes, there has been difficulty in providing prompt and personalised feedback in large classes due to teachers' workload. On the other hand, generative artificial intelligence technologies, including ChatGPT, allow for immediate and consistent feedback on students' writing; however, its efficiency might lead to the emergence of the "AI Feedback Paradox," which can undermine the cognitive processes of deep processing and thereby prevent L2 acquisition. The quasi-experimental research aimed at investigating the possible solution of the problem, namely the possibility of overcoming the AI Feedback Paradox by adding structured metacognitive reflection to the AI-generated feedback. Ninety low-intermediate level Vietnamese EFL learners divided into three groups based on the writing classes received either AI-generated corrective feedback, teacher-generated metalinguistic feedback or AI-generated corrective feedback accompanied with metacognitive reflection log during a 12-week writing class. Students' grammatical knowledge was tested through Grammatical Error Identification Test, while the nature of revision was marked as deep or surface revision. The findings revealed significant superiority of the third condition over the first one concerning the increase of grammatical knowledge and the proportion of deep revisions performed by participants; the second condition proved to be between those two. Thus, the usefulness of the AI-generated feedback depends on the level of cognitive demands imposed on learners by the process.

Introduction

Written corrective feedback (WCF) has traditionally been a cornerstone of L2 writing instruction because of its capacity to improve accuracy and promote acquisition. However, the provision of consistent, comprehensive, and prompt WCF has remained a challenging task in the context of higher education environments, where large class sizes and heavy teaching duties have prevented teachers from focusing on individual errors made by students. The emergence of generative AI technology such as ChatGPT has aroused much optimism since the technology provides instant and consistent corrections that seem to decrease the workload of a teacher while assisting the learner in the process of revising. At the same time, this particular aspect of the application creates what is known as the "AI Feedback Paradox," which consists in the fact that the consistency and immediacy that makes the technology appealing can actually hinder the process of deep cognitive processing necessary for language learning. As soon as learners notice and correct the mistakes automatically, they will be able to revise the text correctly without understanding the grammatical rules.

The recent comparative researches conducted on the issue have shown conflicting results, according to which, while some researchers find that feedback from AI can be as beneficial, if not even more beneficial for the short-term accuracy improvement as the feedback from teachers, the other ones point out that AI is not able to perform as flexibly in terms of context and instruction as the teachers themselves. However, little attention is paid to the process of cognitive interaction of the learner with the provided feedback rather than to the ability of AI to give appropriate feedback. From the perspective of the cognitive-interactionist models of L2 acquisition, the feedback will be the most efficient when it triggers the process of noticing the mismatches between one's own language production and the target language and engaging in the reorganising processing of the acquired linguistic knowledge.

Literature Review

Written corrective feedback and the emergence of AI-assisted feedback

As a widely researched instructional approach in L2 writing, written corrective feedback is acknowledged for its capacity to improve learners' language accuracy and aid L2 acquisition. However, despite extensive empirical research that demonstrates the effectiveness of WCF for language learning (Li & Vuono, 2019; Muñoz & Sáez, 2019), there remains controversy among researchers regarding the extent to which different types of corrective feedback help sustain the learning process. Evidently, WCF is influenced by a number of factors that include learner factors, instruction settings, writing activities, and particular types of feedback provided (Li et al., 2022; Nassaji, 2016). Thus, current research has moved beyond the issue of effectiveness of corrective feedback to investigating the contexts in which different forms of feedback prove the most beneficial.

Traditionally, WCF has been defined as a type of written response which is meant to draw learners' attention to errors in their writing process. This can include correction, code, underlining, reformulation or explanation of local features such as grammar or vocabulary as well as textual features like organization and coherence (Crosthwaite et al., 2022; Montgomery & Baker, 2007). Other than dealing with errors, the function of WCF includes two major aspects. First, it is used to help in making improvements in the performance of learners in their writing. Secondly, it is used to help learners develop languages in the long run by encouraging them to attend to forms and adapt their interlanguage systems (Ferris, 2012; Sheen, 2011). From a sociocultural perspective, corrective feedback also constitutes an important interaction between learners and more knowledgeable others (Cen & Zheng, 2024).

It is common practice for scholars to divide written corrective feedback into two broad categories: namely, direct feedback and indirect feedback (Ellis, 2009; Sheen, 2011; Nassaji, 2020). Direct feedback refers to the process of explicitly correcting learners' errors by presenting them with the correct form of the language while indirect feedback refers to the use of techniques like underlining, circling, highlighting, or error codes to alert learners to the presence of errors so that they come up with the right correction (Muñoz et al., 2023).

The different types of feedback rely on different theoretical assumptions with regard to the process of language acquisition. The application of indirect feedback implies the involvement of more complicated processes of cognition because the learners should themselves find and correct any language problems, which helps develop self-regulation and better memorization. On the contrary, the use of direct feedback involves reduced cognitive load since it immediately gives the learner an answer and thus becomes especially helpful in cases when the learner's own grammar is not sufficient for proper correction (Muñoz et al., 2023; Muñoz & Sáez, 2019).

The evidence regarding differences between direct and indirect WCF is still contradictory. The research suggests that the effectiveness of direct corrective feedback lies in its ability to help learners identify correct forms of language immediately as it is suitable when teaching rule-governed grammatical constructions, such as verbs and articles (Bitchener & Knoch, 2009; Ellis et al., 2008). On the contrary, some researchers argue that indirect corrective feedback is better for developing deep learning as it helps learners to participate in the process actively, think more and develop autonomy in the process of correction (Ferris, 2003; Muñoz & Sáez, 2019). In turn, the use of metalinguistic feedback has often proved to be useful in helping students understand their errors better through the explanations of why those mistakes should be corrected (Bitchener, 2008). However, there are studies that indicate that the increase in the level of explicitness of explanation is not necessarily linked to better learning outcomes, as it was found out that highly detailed metalinguistic feedback had no extra effect on students' knowledge (Shintani & Ellis, 2013; Stefanou & Révész, 2015).

Another significant difference pertains to the nature of written corrective feedback. In focused feedback, correction is provided only for some specific grammatical structures, while comprehensive feedback aims at pointing out most or all of the errors found in a learner's writing (Ellis et al., 2008). It is often observed that focused WCF contributes better to learning as the focus on a limited number of linguistic elements makes the task less cognitively demanding and allows learners to focus their attention on certain grammatical structures (Lee, 2019; Sheen, 2007). Conversely, comprehensive feedback is a closer representation of actual classroom practices and, hence, it is ecologically valid (Colpitts & Howard, 2018; Reynolds & Kao, 2022; Storch, 2010). Most teachers prefer comprehensive feedback in their classes as they consider correction of all apparent mistakes an integral part of good writing instruction (Sheen et al., 2009). Nevertheless, comprehensive feedback requires a lot of time and effort from teachers (Han & Sari, 2024), especially in the context of higher education where the number of students per teacher is very high (Nassaji, 2016, 2020). It is precisely due to such practical difficulties that there have arisen new approaches toward finding technological means to facilitate correctional feedback on written work. The large number of students makes it quite hard for the instructor to give each learner a comprehensive and individualised feedback on his or her writing (Bitchener & Knoch, 2008). Such delayed feedback could reduce the learning experience as it would be too late to make any alterations as the writing has already been done some time ago. Moreover, not enough teaching time would allow for giving more than one cycle of revisions with individualised feedback (Hyland & Hyland, 2006). The presence of many students can also be responsible for inconsistency since an instructor will be unable to give equal consideration to all students' writings (Lee, 2008).

The emergence of AI, or to be more specific, Automated Writing Evaluation (AWE) tools, has significantly expanded the ways of providing corrective written feedback. The benefits of AI-based feedback systems include such aspects as standardised implementation of correction criteria, quick response, and repeated revision cycle (Koltovskaia, 2020; Nassaji, 2024a, 2024b). Cognitive theories of second-language acquisition support using automated feedback due to the opportunity of receiving the necessary repetitions in language forms and error noticing through revision (Kaivanpanah et al., 2020; Nassaji, 2024a). However, the results of comparative studies of AI-generated and teacher-generated feedback are inconclusive. Despite the fact that some researchers still claim the supremacy of human feedback (Kaivanpanah et al., 2020), others state that similar or even better results can be achieved by using AI-generated feedback depending on context (Sistani & Tabatabaei, 2023). Therefore, the effectiveness of AI-assisted corrective feedback seems to depend on a number of factors that need to be examined (Nassaji, 2025).

Among recently developed AI applications, ChatGPT has become exceptionally popular owing to its highly sophisticated language processing abilities. ChatGPT was created by OpenAI (2023) and trained on huge amounts of internet and digitised texts, which allows it to produce coherent, contextually

relevant, and linguistically correct language. Apart from language generation, ChatGPT has proven itself to be a very promising AI in educational contexts, being able to help with writing, summarising, paraphrasing, question answering, and language practice (Baktash & Dawodi, 2023; Escalante et al., 2023). It has also been suggested that ChatGPT may help in designing personalised educational experience, conducting assessments, and addressing instructional challenges (Baidoo-Anu & Owusu, 2023; Baskara & Mukarto, 2023; Deng & Lin, 2023; Farrokhnia et al., 2023; Kasneci et al., 2023; Rudolph et al., 2023; Xiao & Zhi, 2023; Zhang, 2023). In language education, it has also been associated with increased engagement, improvement in language skills and better writing performance owing to planning, ideas generation, revision, and instant feedback (Baskara & Mukarto, 2023; Hong, 2023; Kohnke et al., 2023; Li et al., 2024; Song & Song, 2023; Songsiengchai et al., 2023; Yan, 2023). Nevertheless, despite the mentioned promising trends, the question of whether and under what circumstances ChatGPT or any other AI can be more beneficial than teachers remains open. The mentioned gaps form the background of the present research which aims to expand the knowledge gained during previous studies by comparing written corrective feedback from AI and teachers with a combination of AI feedback and structured metacognitive reflection in a longitudinal L2 writing context. AI-generated and teacher-generated written corrective feedback in L2 writing. With the increased integration of artificial intelligence into language education, the question of whether AI-generated feedback provides learners with similar or even more effective learning opportunities in comparison with those provided by teachers becomes extremely important. Within L2 writing, it goes beyond error correction and includes such issues as learner engagement, cognitive processing, writing development, and the evolving role of teachers in technology-mediated classrooms. Therefore, the recent research tends to make comparisons between ChatGPT-generated feedback and teacher-generated feedback to reveal their differences and common aspects.

One of the areas where comparisons are made quite frequently concerns the quality and effectiveness of ChatGPT-generated written corrective feedback. Previous research showed that ChatGPT is able to generate timely and personalised feedback that can help learners throughout the whole process of writing (Farrokhnia et al., 2023; Han & Li, 2024; Song & Song, 2023). The instantaneous nature of the feedback is one of the important qualities of the effective written corrective feedback because it allows learners to revise their writing when their cognitive representation of the task is still active. Unlike the usual teacher-generated feedback that requires a lot of time to be written owing to teacher workload, AI systems can generate feedback instantly, thus increasing chances for revision and continuous language practice. Additionally, ChatGPT shows good consistency when identifying grammatical errors, reducing teacher workload, and facilitating multiple revision cycles, which can add to the effectiveness of written corrective feedback in big educational environments (Gul et al., 2016; Zou et al., 2023).

Besides, the described advantages allow teachers to spend more time on high-order teaching activities instead of wasting time on the identification of grammatical errors (Kasneci et al., 2023).

However, researchers tend to argue that ChatGPT should not be considered a substitute for teacher feedback but only a supplementary teaching tool. Previous researches have revealed several limitations of AI-generated feedback, such as wrong answers, cultural bias, inappropriate suggestions, and difficulties interpreting context-dependent language (Baskara & Mukarto, 2023; Borji, 2023; Hacker et al., 2023; Nassaji, 2024a; Sallam, 2023; Owais, 2025). Due to the fact that ChatGPT generates text based on statistical predictions instead of actual linguistic comprehension, it has problems in dealing with culturally sensitive statements, stylistics, or conventions in different fields of knowledge. Besides, the usage of AI requires learners to have a certain degree of AI literacy and prompt engineering skills to get a right response (Strobelt et al., 2023). There are ethical considerations as well, such as plagiarism, lack of academic integrity, and over-reliance on AI during writing (Farrokhnia et al., 2023). Song and Song (2023) stated that although learners appreciated the accessibility and instantaneous nature of the ChatGPT, many of them were careful not to become over-reliant on AI out of the fear of losing their critical thinking, creativity, and independence. Empirical studies that compare AI-generated feedback and teacher-generated feedback have shown encouraging but contradictory results. Escalante et al. (2023) conducted pre-test – post-test study, in which 48 university English language learners were compared in terms of AI-generated and teacher-generated feedback on their academic writing. They had to write 300-word compositions and receive an evaluation according to the analytic rubric which assesses content, coherence, language use, and evidence. The results showed no significant differences in writing performance between the two conditions; however, the learners preferred different feedback sources. While half of them preferred teacher feedback due to personal interaction and opportunities to clarify things, the rest of the participants preferred ChatGPT feedback because it was clearer and more detailed. Based on these results, the researchers concluded that the integration of AI-generated and teacher feedback should be used to exploit benefits of each approach while reducing teachers' feedback workload.

Similar research, but with additional self-correction, was carried out by Cao and Zhong (2023). Their study of Chinese postgraduate translation students showed that ChatGPT was especially effective in improving lexical choice and textual cohesion. Nevertheless, teacher feedback and learner self-correction were more effective in improving translation quality and syntactic development. One of the key limitations of ChatGPT in this regard was related to the fact that it could not correctly evaluate culturally specific translation because of its English language-centric knowledge base. Additionally, there were inconsistencies in the output of different participants that could undermine the reliability of ChatGPT-generated feedback. Finally, the researchers pointed out that their study did not define which

version of ChatGPT had been used, although the rapid progress was observed in successive versions of the system (Zhang, 2023).

Moreover, the complementary role of AI was supported by research of Dai et al. (2023), which characterised ChatGPT-generated feedback based on analysis of postgraduate assignments that were already assessed by the teacher. It was revealed that ChatGPT gave much more comprehensive and coherent feedback than the teachers while giving often process-oriented comments that facilitated deep learning as described in Hattie and Timperley (2007) feedback model. Nevertheless, ChatGPT did not agree with teachers in their evaluation of tasks and often generated overly positive comments as a result of too vague prompts. Although AI and teachers generated task-level feedback, it was concluded that ChatGPT could only supplement teacher judgement as experienced teachers still provided more balanced and context-dependent evaluations. The reliability of ChatGPT was also studied by Mizumoto and Eguchi (2023) who evaluated more than 12,000 TOEFL essays written by learners of eleven first-language backgrounds. The results showed that ChatGPT-based scoring agreed with benchmark TOEFL proficiency levels and greatly improved the prediction of learners' scores. Therefore, AI can be used for the efficient and consistent writing assessment. However, the researchers also noted that ChatGPT did not reach a full agreement with the human expert ratings, which means that AI-assisted assessment should be considered as a supportive tool rather than an alternative to the human experts.

Apart from grammatical correction, the recent research started paying more attention to educational impact of ChatGPT on the writing development. According to Song and Song (2023), in their pre-test – post-test study of fifty Chinese undergraduate EFL learners, it was revealed that ChatGPT learners improved their writing organization, coherence, vocabulary, grammar, and writing motivation in comparison with traditionally instructed learners. They appreciated the instantaneous nature of ChatGPT responses, personalization, and opportunities for independent practice. However, they also had some concerns about the contextual errors and future educational roles of AI. The results of the study match the findings of Cao and Zhong (2023), both researches pointing out the persistent problem of ChatGPT with culturally nuanced language and context-dependent writing.

The collaborative nature of AI and teachers was confirmed by Han and Li (2024) in their investigation of 95 undergraduate students who used ChatGPT-supported feedback for text revision. Indirect coded and holistic rhetorical feedback was studied, the latter being focused on argument development, coherence, and organization (Aljasir, 2021; Carter & Thirakunkovit, 2019). It was revealed that the combination of teacher feedback with AI-generated feedback improved the quality of revision, prevented recurring writing errors, and promoted writing development in general. Additionally, the usage of AI in classroom feedback reduced the teacher workload and provided learners with timely and personalised feedback. Rather than working independently, ChatGPT was effective as a collaborative partner.

In terms of teachers' perceptions, Lin and Crosthwaite (2024) compared the feedback practices of twenty-five experienced EFL and ESL teachers and ChatGPT. It was found that the teachers used balanced combinations of direct and indirect feedback and addressed sentence-level and discourse-level writing issues. In turn, ChatGPT used mostly metalinguistic explanations and reformulations, occasionally over-correcting learners' writing by rewriting unnecessary sentences. It also failed to take into account learners' proficiency level, educational goals, and the types of mistakes when generating feedback. Nevertheless, AI was more accurate in correction of isolated grammatical mistakes.

Similar findings were obtained by Wang et al. (2024) when they compared ChatGPT 3.5 with the evaluation of argumentative essays by teachers. ChatGPT generated the faster and more comprehensive linguistic feedback than teachers, but it lost accuracy when evaluating longer essays and complex argumentation. In contrast to AI, teachers generated personalised comments that reflected students' educational background and needs. ChatGPT also had difficulties in evaluation of the logical development of arguments because of the inability to analyse discourse-level connections. The described limitations were closely related to prompt quality and confirmed Jacobsen and Weber's (2025) findings that high-quality feedback is possible only with carefully prepared prompts. Similar conclusions were also made by Lin and Crosthwaite (2024).

ChatGPT has become an important tool that allows generating the rapid, consistent, and comprehensive written corrective feedback. However, its limitations in the context of contextual interpretation, cultural understanding, discourse-level evaluation, and personalization of the instruction show that AI is now more efficient when combined with the teacher feedback instead of operating independently. The problem is that almost all researches that were conducted so far compared AI-generated feedback with the teacher feedback. Very few researches explored the possibility that cognitive engagement with AI feedback can enhance its efficiency. Thus, the presented study aims to extend the knowledge gained during previous research by adding the third condition where AI feedback will be supplemented by metacognitive reflection. By asking learners to describe the principles underlying the corrections generated by AI before writing revisions, the study will find out whether deeper cognitive processing can enhance AI feedback beyond AI feedback or teacher feedback only. This study is sought to answer the following questions:

Q1: To what extent do AI-generated corrective feedback, teacher-generated metalinguistic feedback, and AI-generated corrective feedback combined with structured metacognitive reflection differ in improving EFL learners' grammatical knowledge?

Q2: How do AI-generated corrective feedback, teacher-generated metalinguistic feedback, and AI-generated corrective feedback combined with structured metacognitive reflection influence EFL learners' revision behaviour, particularly the use of deep and surface revisions?

Method

Research design

A quasi-experimental design, extending for 12 weeks, was used in this study. The independent variable in this intervention was three different conditions for providing corrective feedback to the participants: (a) artificial intelligence (AI)-generated corrective feedback, (b) teacher-generated metalinguistic corrective feedback, and (c) AI-generated corrective feedback and structured metacognitive reflection. The primary goal was to find out whether there was a greater level of improvement in grammatical knowledge from incorporating metacognitive reflection with AI-generated feedback than from using either of these two types of feedback on its own.

A quasi-experimental research design was the best methodological choice in this case because this research took place in a natural educational setting where the participants were already enrolled in the courses. Therefore, it was not possible to assign individual students to different treatments as is typically done in experimental designs. Instead, three intact classes were allocated to each of the feedback conditions. While quasi-experiments are usually considered more vulnerable to threats to internal validity, some methodological precautions were taken to reduce their effect. They include pre-screening of participants by the use of the standardised placement test, equivalent instructional and assessment procedure across groups, identical schedule of teaching, and the same pre-test and post-test instruments being administered in standardised conditions. A between-subjects design was chosen in order to avoid the problem of possible treatment contamination because of exposure to various treatments by the participants. It ensured ecological validity of the experiment.

The independent variable was the type of corrective feedback, consisting of three levels:

- **AI-generated corrective feedback;**
- **teacher-generated metalinguistic corrective feedback; and**
- **AI-generated corrective feedback supplemented by structured metacognitive reflection.**

The main dependent variable was the grammatical knowledge of learners, measured in terms of scores on the Grammatical Error Identification Test (GEIT), both pre- and post-intervention. The secondary dependent variable was the behaviour of revision of learners, measured in terms of the number of deep

revisions made by the learners compared to the number of surface revisions made during the revision process of the essays. A deep revision refers to a revision that involves making substantial change in the grammatical structure of sentences, while surface revisions refer to minor modifications of spelling, punctuation or a grammatical form. The intervention period lasted for 12 weeks; hence the study was able to examine not only the immediate impact of the intervention but also the effects of the prolonged engagement in writing tasks over 12 weeks. During this period, the participants engaged in one argumentative writing task each week, made revisions based on the feedback provided and then submitted the revision. Table 1 gives an overview of the research design.

Table 1.
Research design of the study

Week	Activity	Group A	Group B	Group C
1	Baseline assessment	OPT + GEIT Pre-test	OPT + GEIT Pre-test	OPT + GEIT Pre-test
2–11	Weekly writing tasks	AI feedback	Teacher feedback	AI feedback + metacognitive reflection
12	Final assessment	GEIT Post-test	GEIT Post-test	GEIT Post-test

The longitudinal design allowed to assess the learning outcomes immediately and evaluate the accumulated grammatical progress over the entire period of intervention. Due to repeated writing tasks and multiple rounds of feedback, the study aimed to provide a more holistic assessment of corrective feedback compared to the approaches that use only one treatment or post-treatment measures.

Research context

The study took place in a Vietnamese private higher education institution specialising in foreign languages, information technology, business, and associated subjects. English is a mandatory subject for all undergraduate programmes at this institution and serves as a basis for academic preparation, international communication, and future career development. Adequate command of English is required for students to move to specific major-related subjects. In case of inability to meet the requirements for language competence, the students are placed into the foundation English courses focused on developing the academic reading, writing, listening, and speaking skills.

This study was conducted within the HUFLIT academic writing programme for lower-intermediate English foreign language learners. The programme develops grammatical accuracy, academic vocabulary, paragraphs, essays, and argumentative writing skills via a process-oriented approach to writing instruction. The students should prepare multiple drafts, get feedback on them, revise their writing, and hand over the final version, which makes the programme ideal for testing the efficiency of various types of corrective feedback. All three classes were taught according to the same institutional syllabus, learning outcomes, teaching materials, and assessment criteria. The classes take place weekly during a two-hour session for 12 weeks in total. For the sake of instructional consistency, the same lesson plans, writing tasks, activities in class, home tasks, and methods of assessment were used in all classes except for the type of feedback provided during the intervention phase. No other instructions on grammar were provided to the learners apart from the treatment conditions set by the study.

The application of artificial intelligence in teaching English in higher education is becoming increasingly popular in Vietnamese higher education in light of the wide usage of generative AI, such as ChatGPT. As the institutions begin to implement the AI-assisted learning in order to facilitate language instruction, the need to assess the efficiency of AI-generated feedback becomes evident. This study conducted within the HUFLIT academic writing programme provides an authentic instructional context for assessing the pedagogical efficiency of AI-assisted feedback increasing ecological validity and applicability of the results.

Participants

This study involved 90 undergraduate students taking part in three foundation academic writing classes. Each class consisted of 30 participants making the total number of 90 learners. The students were studying in diverse academic fields like business, engineering, information technology, media studies, and health sciences. The age range of participants varied between 18 and 22 years old, which corresponds to the typical age group of first-year undergraduate students of Vietnam.

For the vast majority of participants, Vietnamese was the first language and English was learned as a foreign language in school before joining university. All participants have been learning English for at least 10 years before entering the higher education system. Nonetheless, their grammatical competence

was still at the lower-intermediate level especially concerning the issues of tenses, subject-verb agreement, articles, prepositions, and sentence structure. In order to guarantee homogeneity of samples, all participants were tested using Oxford Placement Test (OPT) during the first week of the study. Only the students whose English proficiency level according to CEFR was at the B1 level were selected for the study.

The results of the Oxford Placement Test confirmed that the three instructional groups were statistically comparable before the intervention. Descriptive statistics indicated highly similar proficiency levels across groups, with mean scores ranging from 40.23 to 40.30. A one-way analysis of variance revealed no statistically significant differences among the groups, $F(2, 87) = 0.04$, $p = .965$, demonstrating baseline equivalence. Furthermore, Levene's test confirmed homogeneity of variance across groups, $F(2, 87) = 0.06$, $p = .940$. These findings indicate that participants commenced the intervention with comparable English language proficiency, thereby strengthening the internal validity of the study and reducing the likelihood that post-intervention differences could be attributed to pre-existing language ability rather than the feedback treatments.

Materials and instruments

A variety of instruments were used to collect quantitative and qualitative data throughout the intervention. They included the Oxford Placement Test (OPT), the researcher-developed Grammatical Error Identification Test (GEIT), ten argumentative writing tasks, AI-generated corrective feedback provided by ChatGPT-4, teacher-generated metalinguistic feedback, metacognitive reflection log, and a revision coding framework.

Oxford Placement Test

The Oxford Placement Test (OPT) was administered at the beginning of Week 1 to assess participants' English language proficiency and determine equivalence among three instructional groups. The test checked grammar, vocabulary, and reading skills of participants. It was carried out in a supervised classroom setting. Participants who scored the same as B1 (lower-intermediate) level of the CEFR were

selected to participate in the study. One-way ANOVA indicated the absence of any statistically significant difference between the three groups at the beginning of the intervention, which means that participants were of comparable proficiency.

Grammatical Error Identification Test

To measure learners' grammatical development, the researcher-developed 40-item Grammatical Error Identification Test (GEIT) was administered as both pre-test and post-test. The test assessed participants' ability to identify and fix grammatical errors in verb tense, subject-verb agreement, articles, prepositions, pronouns, sentence structure, and word order. Every correct answer was worth one point, thus, the maximum score was 40. This instrument was pilot-tested on a comparable group of participants before the study began. Content validity was ensured by assessing the instrument by three university English-language specialists.

Argumentative writing tasks

Participants were asked to write ten argumentative essays of approximately 250 words during Weeks 2-11. Topics concerned problems relevant to university students, such as the role of technology in education, online learning, and social media. To control the comparability of lexical and syntactic complexity, all writing tasks were analysed by Coh-Metrix. Uniform writing instructions, time allocation, and assessment criteria were applied in the course of the intervention.

AI-generated corrective feedback

Participants in the AI Feedback and AI + Metacognitive groups received AI-generated corrective feedback provided by ChatGPT-4 using a standardised prompt. ChatGPT-4 was supposed to identify the grammatical errors, suggest corrections, and explain the corrections in case of necessity, but not rewrite the whole essay. Only grammatical accuracy of texts was corrected by the AI and feedback was delivered electronically within one hour after participants submitted their essays. Students revised and resubmitted their essays on the same day.

Teacher-generated corrective feedback

Participants in the Teacher Feedback group received feedback from an instructor based on the standardised metalinguistic coding system. The errors were identified using correction codes that corresponded to verb tense, subject-verb agreement, articles, prepositions, spelling, punctuation, word order, and other grammatical elements. Feedback was returned within five days; students revised and resubmitted their essays within 48 hours. Instructor followed uniform correction procedures throughout the whole intervention.

Metacognitive reflection log

Participants in the AI + Metacognitive group completed a reflection log after receiving AI-generated feedback. Five corrections were selected by students for each essay and they explained the grammatical rules behind these corrections in their own words before revising their work. Such logs were designed to engage students more with feedback and generate qualitative data about grammatical awareness of learners. The completion rates were recorded for further quantitative analysis.

Revision coding framework

Revisions made by participants were compared to original essays to analyse the revision behaviour. The revision was either classified as surface revision (e.g. small corrections to spelling, punctuation, and grammatical errors), or deep revision (e.g. changes of sentence structure and clauses, etc.). A sample of participants' essays was independently coded by two experienced EFL instructors before analysing the rest of the essays.

Procedure

The study was conducted over 12 weeks. During Week 1, participants completed the Oxford Placement Test and the GEIT pre-test. Between Weeks 2 and 11, students wrote one 250-word argumentative essay each week.

Table 2
Summary of the intervention procedures for each experimental group

Group	Feedback provider	Feedback procedure	Revision procedure
Group A (AI Feedback)	ChatGPT-4	Participants submitted their essays electronically and received AI-generated grammatical feedback within one hour. The feedback identified grammatical errors, suggested appropriate corrections, and provided brief explanations where necessary without rewriting the entire essay.	Participants revised their essays immediately after receiving the feedback and resubmitted the final version on the same day.
Group B (Teacher Feedback)	Course instructor	Participants received teacher-generated metalinguistic feedback using a standardised coding system within five days of essay submission. The feedback highlighted grammatical errors using correction codes rather than direct corrections.	Participants revised their essays within 48 hours of receiving the instructor's feedback and submitted the revised version for assessment.
Group C (AI Feedback + Metacognitive Reflection)	ChatGPT-4 and metacognitive reflection	Participants received the same AI-generated grammatical feedback as Group A. In addition, they completed a structured metacognitive reflection log in which they explained the grammatical rule underlying five AI-generated corrections before beginning the revision process.	After completing the reflection log, participants revised their essays using the AI feedback and resubmitted the final version on the same day.

During Week 12, all participants completed the GEIT post-test under identical testing conditions. The original and revised essays were collected and analysed to determine the proportion of deep and surface revisions completed by each participant as shown in Table 2.

Ethical considerations

Approval of ethical concerns related to the research was acquired from the Research Ethics Committee of the university before collecting any data. Participation in the study was completely voluntary, and all participants were provided with written information about the study, its procedures, and participants' rights. All participants gave written informed consent before the intervention. Participants were assured that their participation or non-participation in the research would have no impact on their grades in courses. All data obtained were anonymised using participant ID numbers without any personal identifying details included in the data set or further publications. Electronic data were stored in

password-protected computers that could be accessed only by the research team. Use of ChatGPT-4 in the research was limited to grammatical correction of participants' writing for educational purposes. Participants were warned against submitting AI-generated texts as their own work. AI technology was used only as a tool of feedback, which ensured that all written texts produced by the participants were entirely their work.

Results

The descriptive statistics presented in Table 3 reveal that the three experimental groups demonstrated highly comparable levels of English language proficiency prior to the intervention. The AI Feedback group obtained a mean Oxford Placement Test (OPT) score of 40.23 (*SD* = 1.10), while both the Teacher Feedback group and the AI + Metacognitive group obtained identical mean scores of 40.30 (*SD* = 1.15 and 1.12, respectively). The small standard deviations observed across all groups indicate limited variability in participants' proficiency levels within each group, suggesting that the participants were relatively homogeneous in terms of their English language ability.

Table 3.

The descriptive statistics

OPT

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
AI Feedback	30	40.23	1.104	.202	39.82	40.65	39	42
Teacher Feedback	30	40.30	1.149	.210	39.87	40.73	39	42
AI + Metacognitive	30	40.30	1.119	.204	39.88	40.72	39	42
Total	90	40.28	1.112	.117	40.04	40.51	39	42

Furthermore, the narrow and substantially overlapping 95% confidence intervals indicate that the mean scores of the three groups were highly similar. These findings provide preliminary evidence that the participants entered the study with equivalent levels of English language proficiency, thereby reducing the likelihood that subsequent differences in learning outcomes could be attributed to pre-existing proficiency differences rather than the feedback intervention.

Table 4.

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
OPT	Based on Mean	.062	2	87	.940
	Based on Median	.106	2	87	.900
	Based on Median and with adjusted df	.106	2	83.762	.900
	Based on trimmed mean	.064	2	87	.938

Before conducting the one-way ANOVA, the assumption of homogeneity of variances was examined using Levene's test. As shown in Table 4, the test was not statistically significant, $F(2, 87) = 0.06$, $p = .940$. Since the significance value exceeded the conventional alpha level of .05, the assumption of equal variances across the three groups was satisfied. Meeting this assumption indicates that the variability of OPT scores was statistically comparable among the AI Feedback, Teacher Feedback, and AI + Metacognitive groups. Consequently, the use of a one-way ANOVA was considered appropriate because unequal variances were unlikely to bias the comparison of group means. This finding strengthens the validity of the subsequent statistical analysis.

Table 5.

The one-way analysis of variance

OPT

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.089	2	.044	.035	.965
Within Groups	109.967	87	1.264		
Total	110.056	89			

The one-way analysis of variance presented in Table 5 was conducted to determine whether any statistically significant differences existed among the three groups in their English language proficiency before the intervention. The analysis revealed no statistically significant difference in OPT scores, $F(2, 87) = 0.04$, $p = .965$. The very small F statistic indicates that the variation in mean OPT scores between the three groups was negligible when compared with the variation observed within the groups. Moreover, the extremely high p value demonstrates that any observed differences in mean scores were likely due to random sampling variation rather than genuine differences in language proficiency. From these

findings, one can conclude that all three groups are statistically equivalent at the baseline level. This fact is especially important for quasi-experimental design since the participants are assigned to existing classes rather than being randomly assigned to experimental conditions. The confirmation of the equivalence between the groups in terms of pre-existing differences helps to achieve higher validity in this research and provides stronger grounds for attributing the differences obtained during post-test analysis to feedback strategies used in this study rather than initial English proficiency of the groups.

Discussion

This experiment evaluates three types of written corrective feedback administered to L2 students in a 12-week writing course: AI feedback, teacher-provided metalinguistic feedback, and AI feedback coupled with the structured metacognitive reflection activity. While automated writing evaluation tools allow for instant and consistent feedback, their effectiveness in fostering sustained and durable language acquisition remains questionable. This question gives rise to the so-called "AI feedback paradox": the very same properties of AI feedback – immediacy and accuracy of correction along with constant repetitions – may hamper the cognitive process necessary for sustained learning. The results of this study shed light on the phenomenon of "feedback paradox" and offer a way to overcome it.

The first main result was obtained through the analysis of scores on the test on grammatical error identification. It turned out that the gain in explicit knowledge of grammatical rules was uneven among groups. Learners receiving the AI feedback along with the metacognitive reflection log performed much better compared to those receiving just AI feedback. The teacher feedback condition proved to be intermediate, which shows that providing the learners with the necessary information per se is not sufficient to ensure learning. This result is in line with another important principle of the cognitive-interactionist theory of second language acquisition: feedback is especially helpful when it allows the learner to realise that there is a gap between his/her utterance and the target form and engages the learner in deeper processing that results in restructuring of interlanguage (Sheen, 2011; Ferris, 2012). In the AI condition, the provision of both the erroneous form and the correct one along with a brief explanation made the learning process less challenging. While this immediacy allowed making the necessary correction of mistakes instantly, it prevented the learners from the realization of the rules underlying the error, thus limiting cognitive conflict that is vital for acquisition.

In the AI + Metacognitive group, however, learners had to externalise their knowledge by picking up 5 corrections generated by AI per essay and providing explanations of the rules behind these errors in their own words before making a revision. The activity of verbalising the metalinguistic information thus turned the simple reception of corrections into a more complicated sense-making process. From the Vygotskian perspective, the use of the metacognitive reflection log was the tool that facilitated the mediation of learners' transition from other-regulation to self-regulation. It is this transition that presumably accounts for the superior scores on error identification in this group. Previous research has shown that the benefit from explicit feedback is magnified when learners are asked to generate their explanations of the rules rather than simply to receive the pre-prepared ones (Shintani & Ellis, 2013; Stefanou & Révész, 2015). In this way, the current study confirms that the pedagogical value of AI-generated tools depends on the interaction with the learner rather than on the AI itself.

The analysis of the revision behaviour provided the complementing perspective on learners' activity. The surface revisions, such as spell-checking or isolated word substitution, prevailed in the AI-only group. The learners in this condition tended to accept the suggested form without making significant adjustments to the syntactic structure of the text. In the AI + Metacognitive condition, however, learners performed a larger number of deep revisions, i.e. the ones that involved restructuring of clauses and reconfiguration of the grammatical relations within sentences. This result shows that the use of the reflection log not only facilitated explicit knowledge, but also led to the change in the approach to the revision process. After articulating the grammatical rules behind the error, the learners were able to identify similar problems in other places in the text and make necessary revisions. The teacher feedback, which involved the use of the metalinguistic coding system, also involved certain degree of problem-solving, but the five-day delay (which occurred in the current study) probably weakened the connection between detecting the mistake and revision. In contrast, the AI + Metacognitive condition allowed maintaining the immediacy while preserving the cognitive challenge; ChatGPT-4 provided feedback within an hour, thus activating the mental representation of the text, while the metacognitive reflection ensured that immediacy did not turn into intellectual laziness. This result explains the divergence in the findings of previous comparative research of AI feedback and teacher feedback. Some studies showed that AI feedback was equally effective, or even more effective than human feedback in terms

of short-term improvement in accuracy (Escalante et al., 2023; Sistani & Tabatabaei, 2023), while some pointed to the lack of contextual awareness and pedagogical flexibility of artificial intelligence (Lin & Crosthwaite, 2024; Wang et al., 2024). The current findings show that both of these opinions may be true, depending on how "effectiveness" is evaluated and, most importantly, on the nature of learners' interaction with the feedback. If the task is to make learners correct errors immediately, then the AI-alone approach is fine. However, if the task is to help the learners acquire the underlying grammatical knowledge and transfer it to a new task, then AI needs to be used in a more sophisticated learning context. In this way, the paradox of AI feedback can be solved not by discarding the technology but by changing the activity system around it. AI offers the constant noticing trigger, while metacognitive reflection provides the depth of processing needed to turn the noticing into acquisition.

The teacher-provided metalinguistic feedback condition gave results that need to be carefully interpreted. The fact that this group did not consistently outperform the AI + Metacognitive group in terms of all criteria does not mean that the role of the teacher is not crucial. Rather, the current study reinforces the notion that teaching writing in the era of AI means that teachers should become the designers of the learning experiences rather than the providers of corrective feedback. The teachers' expertise is necessary in order to determine the target structures, adjust the task complexity, model the process of reflection, and provide the type of dialogue-based feedback that current AI is unable to provide (Han & Li, 2024). The five-day delay, which was ecologically valid for the context of higher education in the current study, may be one of the factors that weakened the effect of teacher's feedback compared to instantaneous AI conditions. Future research may investigate whether the effect of the teacher feedback is preserved if it is provided synchronously or the next day, for example, through learning management system.

Some of the pedagogical implications can be derived from the results of this study. First, when implementing AI into writing curricula, institutions should not position ChatGPT-like tools as autonomous tutors. Instead, the teachers should be trained to provide the students with metacognitive prompts to accompany AI feedback, for example, in the form of the structured error journal, peer discussion of the corrections, or the tasks like the ones used in the current study. Second, the fact that surface-level revisions prevail in the AI-only conditions should result in rubric adjustments in the way that the rubrics

start rewarding restructuring of texts and rule articulation in addition to error-free final versions. Otherwise, AI may contribute to the development of culture of shallow editing that masks learners' persistent grammatical weaknesses. Third, given the heavy dependence of the quality of AI-generated feedback on the prompts provided to the AI (Jacobsen & Weber, 2025), teacher-guided prompt design becomes the new literacy to develop for language educators. A well-prepared prompt can limit the AI to provide feedback that is selective, rule-focused, and correctly calibrated according to the learners' proficiency level, avoiding excessive corrections and overly generic metalinguistic explanations reported in previous studies (Lin & Crosthwaite, 2024).

Study limitation

Limitations of this study include the limited scope of investigation to lower-intermediate EFL learners in one Vietnamese university; the generalization of results to other proficiency levels, educational cultures, and L1 backgrounds should be further researched. Moreover, while the test on grammatical error identification allowed measuring the explicit knowledge, the further research should also involve tests of spontaneous written production and post-tests in order to measure the durability of learning. The condition with teacher feedback included a temporal discrepancy between it and AI conditions (five-day delay versus one-hour); in order to separate these two variables, disentangle the effects of delay from the effects of human response, the designs comparing feedback latency should be used in the future. Also, only one version of ChatGPT was used in the current study; given the rapid development of the system, the results might not be applicable to the future versions that possess advanced capabilities for contextual reasoning. Finally, the qualitative analysis of the metacognitive logs could shed light on the process of articulating rules.

Conclusion

This study compared the effects of AI-generated written corrective feedback, metalinguistic teacher feedback, and AI-generated feedback plus structured metacognitive reflection on grammatical knowledge acquisition and revision strategies used by lower intermediate Vietnamese students learning English as a foreign language. The results suggest that even though AI-generated feedback can provide quick, consistent and easily available grammar corrections, the educational value of this type of feedback is greatly increased when metacognitive reflection is done by learners. The participants who were

using both AI-generated feedback and reflection had better progress in their grammatical knowledge and conducted more deep revisions than the other two groups, which suggests that cognitive processing is a necessary element to translate corrective feedback into language learning. Moreover, metalinguistic teacher feedback is still a good method of teaching; however, the delayed delivery of this kind of feedback can limit the chance for quick revision and cognitive processing. Instead of thinking about artificial intelligence as a replacement for teachers, it can be seen from the results of this study that the most efficient way to use it is as a complement to pedagogical techniques that would help students think about rules and learn independently. The AI feedback paradox becomes clear from this analysis, as it can be concluded that technology works best when it is combined with pedagogy to facilitate learning process.

The study adds to the body of literature on the topic of AI assisted learning by providing an insight into the fact that in order to succeed with AI-generated written corrective feedback one needs to consider not only the quality of AI but also the ways the technology is utilised. Using metacognitive activities in AI-assisted writing instruction proves to be an effective way to improve grammatical development, deep revisions, and learner independence.

AI acknowledgements

The researcher used AI tools to support language editing and improve clarity during manuscript preparation, in accordance with the recommendations of the Committee on Publication Ethics (COPE). All research design, analysis, interpretation, and final content remained the sole responsibility of the researcher.

Acknowledgements

The author would like to express sincere gratitude to the participating students for their support and cooperation during the data collection process. Their participation made this research possible.

Conflict of interest

The researcher confirms that there is no conflict of interest associated with this study.

Financial support

The researcher confirms that this study did not receive any form of financial support.

References

Aljasir, N. (2021). *Matches or mismatches? Exploring shifts in individuals' beliefs about written corrective feedback as students and teachers-to-be*. *Journal of Teaching and Teacher Education, 9*(1), 1–10. https://doi.org/10.12785/jtte/090101

Baidoo-Anu, B., & Owusu, M. (2023). *Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning*. SSRN. https://doi.org/10.2139/ssrn.4337484

Baktash, J., & Dawodi, M. (2023). *GPT-4: A review on advancements and opportunities in natural language processing*. arXiv. https://arxiv.org/abs/2305.03195

Baskara, F., & Mukarto, F. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics, 7*(2), 343–358.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*, 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research, 12*(3), 409–431. https://doi.org/10.1177/1362168808089924

Bitchener, J., & Knoch, U. (2009). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics, 31*(2), 193–214. https://doi.org/10.1093/applin/amp016

Borji, A. (2023). *A categorical archive of ChatGPT failures*. arXiv. https://doi.org/10.48550/arXiv.2302.03494

Cao, S., & Zhong, L. (2023). *Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation*. arXiv. https://doi.org/10.48550/arXiv.2309.01645

Carter, T., & Thirakunkovit, S. (2019). A comparison of L1 and ESL written feedback preferences: Pedagogical applications and theoretical implications. *Journal of Response to Writing, 5*(2), 139–174.

Cen, Y., & Zheng, Y. (2024). The motivational aspect of feedback: A meta-analysis on the effect of different feedback practices on L2 learners' writing motivation. *Assessing Writing, 59*, Article 100802. https://doi.org/10.1016/j.asw.2023.100802

Colpitts, B., & Howard, L. (2018). A comparison of focused and unfocused corrective feedback in Japanese EFL writing classes. *Lingua Posnaniensis, 60*(1), 7–16. https://doi.org/10.2478/linpo-2018-0001

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of Scopus-indexed research on L2 WCF. *Journal of Second Language Writing, 58*, Article 100934. https://doi.org/10.1016/j.jslw.2022.100934

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y., Gasevic, D., & Chen, G. (2023). *Can large language models provide feedback to students? A case study on ChatGPT*. EdArXiv. https://doi.org/10.35542/osf.io/hcgzj

Deng, J., & Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems, 2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal, 62*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353–371. https://doi.org/10.1016/j.system.2008.02.001

Escalante, J., Pack, A., & Barret, 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

Farrokhnia, M., Banihashem, S., Norooz, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Ferris, D. (2003). Responding to writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 119–140). Cambridge University Press.

Ferris, D. (2012). Written corrective feedback in second language acquisition and writing studies. *Language Teaching, 45*(4), 446–459. https://doi.org/10.1017/S0261444812000250

Gul, R., Tharani, A., Lakhani, A., Rizvi, N., & Ali, S. (2016). Teachers' perceptions and practices of written feedback in higher education. *World Journal of Education, 6*(3), 10–20. https://doi.org/10.5430/wje.v6n3p10

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models*. arXiv. https://doi.org/10.48550/arXiv.2302.02337

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

Han, T., & Sari, E. (2024). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning, 37*(4), 961–985. https://doi.org/10.1080/09588221.2022.2067179

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hong, W. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation, 5*(1), 37–45. https://doi.org/10.61414/jeti.v5i1.103

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jacobsen, L., & Weber, K. (2025). The promises and pitfalls of LLMs as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI, 6*(2), 1–17. https://doi.org/10.3390/ai6020035

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

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

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback. *Assessing Writing, 44*, Article 100450. https://doi.org/10.1016/j.asw.2020.100450

Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(3), 144–164. https://doi.org/10.1016/j.jslw.2007.12.001

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching, 52*(4), 524–536. https://doi.org/10.1017/S0261444819000247

Li, B., Lowell, V., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education. *Computers and Education: Artificial Intelligence, 7*, Article 100266. https://doi.org/10.1016/j.caeai.2024.100266

Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in *System*. *System, 84*, 93–109. https://doi.org/10.1016/j.system.2019.05.006

Li, S., Hiver, P., & Papi, M. (2022). Individual differences in second language acquisition: Theory, research, and practice. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of SLA and individual differences* (pp. 3–34). Routledge.

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

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

Montgomery, J., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing, 16*(2), 82–99. https://doi.org/10.1016/j.jslw.2007.04.002

Muñoz, B., & Sáez, K. (2019). Indirect written corrective feedback in the treatment of subject–verb agreement in third person singular among students of English as a foreign language. *Alpha, 49*, 275–290.

Muñoz, B., Ortiz, M., & Sáez, K. (2023). Preferencias y opiniones de estudiantes de un Programa de Pedagogía en Inglés con distinto nivel de competencia lingüística acerca del tratamiento de los errores en la escritura en LE. *Literatura y Lingüística, 47*, 279–306. https://doi.org/10.29344/0717621X.47.2736

Nassaji, H. (2016). Interactional feedback in second language teaching and learning: A synthesis and analysis of current research. *Language Teaching Research, 20*(4), 535–562.

Nassaji, H. (2020). Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges. *Language Teaching, 53*(1), 3–28.

Nassaji, H. (2024a). *Corrective feedback in the age of AI: Discoveries, methodological challenges, and future directions*. Keynote speech.

Nassaji, H. (2024b). *The AI paradox: Enhancing L2 writing or facilitating L2 learning?* Keynote speech.
Nassaji, H. (2025). *AI and L2 acquisition: A tool to enhance writing or facilitate learning?* Keynote speech.
OpenAI. (2023). *GPT-4 system card*.
Owais, A. K. (2025). AI in English teaching with ChatGPT: Does it really work? In *Using AI tools in text analysis, simplification, classification, and synthesis* (Chap. 7). IGI Global. https://doi.org/10.4018/979-8-3693-9511-0.ch007
Reynolds, B., & Kao, C. (2022). A research synthesis of unfocused feedback studies in the L2 writing classroom. *Journal of Language and Education, 8*(4), 5–13.
Sallam, M. (2023). The utility of ChatGPT as an example of large language models in healthcare education, research and practice. *medRxiv*.
Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*(2), 255–283.
Sheen, Y. (2011). *Corrective feedback, individual differences and second language learning*. Springer.
Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556–569.
Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation. *Journal of Second Language Writing, 22*(3), 286–306.
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*. https://doi.org/10.3389/fpsyg.2023.1260843
Songsiengchai, N., Sereerat, B., & Watananimitgul, W. (2023). Leveraging artificial intelligence (AI): ChatGPT for effective English language learning among Thai students. *English Language Teaching, 16*(11), 68–79.
Stefanou, C., & Révész, A. (2015). Direct written corrective feedback, learner differences, and the acquisition of second language article use. *The Modern Language Journal, 99*(2), 263–282.
Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies, 10*(2), 29–46.
Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H., & Rush, A. (2023). Interactive and visual prompt engineering for ad-hoc task adaption with large language models. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 1146–1156.
Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, Article 101440.
Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3).
Yan, X. (2023). Impact of ChatGPT on learners in an L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*, 13943–13967.
Zhang, B. (2023). *Preparing educators and students for ChatGPT and AI technology in higher education*.
Zou, D., Xie, H., & Wang, F. (2023). Effects of technology-enhanced peer, teacher and self-feedback on students' collaborative writing, critical thinking tendency and engagement in learning. *Journal of Computing in Higher Education, 35*(1), 166–185.

Copyrights