TYPE Original Research
PUBLISHED 29 July 2025
DOI 10.3389/feduc.2025.1614673

AI vs. teacher feedback on EFL argumentative writing: a quantitative study

Areen Alnemrat[1], Hesham Aldamen[2], Mohamad Almashour[2], Mutasim Al-Deaibes[3]* and Rami AlSharefeen[4]

[1]Department of English Language and Literature, Yarmouk University, Irbid, Jordan, [2]Department of English Language and Literature, The University of Jordan, Amman, Jordan, [3]Department of English, American University of Sharjah, Sharjah, United Arab Emirates, [4]Rabdan Academy, Abu Dhabi, United Arab Emirates

Check for updates

OPEN ACCESS

EDITED BY
Siv Gamlem,
Volda University College, Norway

REVIEWED BY
Irina Engeness,
Østfold University College, Norway
Mohammad Mahyoob Albuhairy,
Taibah University, Saudi Arabia

*CORRESPONDENCE
Mutasim Al-Deaibes
✉ deaibesm@gmail.com

RECEIVED 22 April 2025
ACCEPTED 30 June 2025
PUBLISHED 29 July 2025

CITATION
Alnemrat A, Aldamen H, Almashour M, Al-Deaibes M and AlSharefeen R (2025) AI vs. teacher feedback on EFL argumentative writing: a quantitative study.
Front. Educ. 10:1614673.
doi: 10.3389/feduc.2025.1614673

Introduction: This study investigates the effectiveness of AI-generated feedback compared to teacher-generated feedback on the argumentative writing performance of English as a Foreign Language (EFL) learners at different proficiency levels.

Methods: Sixty undergraduate students from a writing-focused EFL course in Jordan participated in a quasi-experimental, pretest-posttest study. Participants were stratified into two ACTFL proficiency levels (Intermediate-Low and Advanced-Low) and assigned to either an AI feedback group or a teacher feedback group. Students completed an argumentative writing task, received feedback based on their group, and revised their essays accordingly. An analytic rubric was used to assess writing performance, and inter-rater reliability was established on a stratified 30% subsample to support the validity of the scoring process, with pre- and post-test scores analyzed for gains.

Results: Results showed significant improvement in writing performance across all groups, regardless of feedback source or proficiency level. Importantly, no statistically significant difference was found between the AI and teacher feedback groups, and the effect size for this comparison was small (Cohen's d = 0.10). A two-way ANOVA revealed a significant main effect for proficiency level but no significant interaction between feedback type and proficiency. Intermediate-Low learners demonstrated the greatest within-group gains, suggesting that both feedback types were particularly impactful for lower-proficiency students.

Discussion: The findings underscore the potential of large language models (LLMs), when carefully scaffolded and ethically deployed, to support writing development in EFL contexts. AI-generated feedback may serve as a scalable complement to teacher feedback in large, mixed-proficiency classrooms, particularly when guided by well-developed prompts and pedagogical oversight.

KEYWORDS

AI-generated feedback, argumentative writing, EFL learners, large language models (LLMs), second language writing, mixed-proficiency classrooms

1 Introduction

Argumentative writing proficiency constitutes a fundamental component of academic literacy acquisition for learners of English as a Foreign Language (Zhu, 2001), involving not only formulating texts characterized with linguistic accuracy but also constructing coherent and persuasive evidence-based arguments, integrating counterarguments, and maintaining a logical flow (Ferretti and Graham, 2019; Su et al., 2023; Su et al., 2021). Effective feedback represents a pivotal mechanism in the developmental trajectory of argumentative writing competence, with instructor-mediated evaluative commentary traditionally serving as the primary intervention modality for facilitating rhetorical enhancement and structural coherence (Banihashem et al., 2022; Zhang and Hyland, 2018). However, providing effective feedback to learners is significantly constrained by structural barriers inherent in contemporary educational settings characterized with high student-to-teacher ratios and heterogeneous proficiency distribution within the same classroom. These systemic limitations transform personalized evaluative feedback into a resource-intensive pedagogical intervention, frequently resulting in delayed feedback and diminished instructional efficacy (Guo et al., 2024; Liu et al., 2024; Ouahidi, 2021; Wisniewski et al., 2020). The advent of Generative Artificial Intelligence (GenAI) in educational contexts has introduced novel avenues for feedback delivery. AI-powered tools, such as ChatGPT, can generate immediate, detailed, and personalized feedback on student writing, potentially alleviating the workload of educators and providing timely assistance to learners (Guo et al., 2024; Lee and Moore, 2024). However, there is limited empirical research on how GenAI tools may assist in scaling feedback, especially in large English as a Foreign Language (EFL) writing classes (Li et al., 2024). For example, Wang and Dang's (2024) systematic review pointed to a significant dearth of empirical investigations examining applications of GenAI in EFL settings. Within the limited scholarly discourse, studies have begun to explore the efficacy of AI-generated feedback in comparison to traditional teacher feedback. For instance, research indicates that GenAI feedback can be as effective as teacher feedback in improving certain aspects of writing, such as coherence and cohesion (Yoon et al., 2023). However, concerns persist regarding the depth, accuracy, and contextual appropriateness of AI-generated feedback, particularly in addressing higher-order writing skills (Chan and Hu, 2023; Yoon et al., 2023). Moreover, the integration of AI into educational feedback mechanisms raises ethical considerations, notably regarding data privacy, the potential for overreliance on technology and plagiarism (Sánchez-Vera et al., 2024). Ensuring that AI tools are used responsibly, and that student data is protected is paramount (Yan et al., 2024). In the Arab world, very few studies looked at the impact of GenAI on higher education settings (Fadlelmula and Qadhi, 2024). This study aims to fill this gap by investigating the comparative effectiveness of AI-generated feedback and teacher-generated feedback on the argumentative writing performance of EFL students in Jordanian settings. Specifically, it seeks to determine whether AI feedback can match or surpass the quality and impact of traditional teacher feedback in enhancing students' abilities to construct and refine arguments, integrate counterarguments, and maintain logical coherence in their writing. The questions this paper aim to answer are as follows:

1. To what extent does the type of feedback (AI-generated vs. teacher-generated) influence EFL students' improvement in argumentative writing performance?
2. To what extent does language proficiency level (Intermediate-Low vs. Advanced-Low) affect writing performance after revision?
3. Is there an interaction between feedback type and proficiency level in determining EFL students' post-revision writing outcomes?

The findings of this study hold significant implications for EFL writing instruction. By elucidating the effectiveness of AI-generated feedback, educators can make informed decisions about integrating AI tools into their pedagogical practices. Understanding the comparative advantages and limitations of AI and teacher feedback can aid in designing hybrid feedback models that leverage the strengths of both approaches. Furthermore, addressing the ethical considerations associated with AI in education will contribute to the development of guidelines and policies that ensure responsible and effective use of technology in language learning contexts.

2 Literature review

Argumentative writing is widely recognized as one of the most cognitively and linguistically demanding genres in EFL instruction (Ferretti and Graham, 2019). It requires learners to formulate a clear stance, develop logical reasoning, integrate counterarguments, and adhere to academic discourse conventions (Ferretti and Graham, 2019; Su et al., 2023; Su et al., 2021). These requirements make argumentative writing particularly difficult for students whose linguistic proficiency is still developing (Pelenkahu et al., 2024). In the EFL context, students often struggle not only with surface-level features such as grammar and vocabulary but also with deeper genre-related demands, such as organizing their arguments logically and integrating rebuttals effectively (Mallahi, 2024). This dual challenge underscores the need for scaffolding that supports both linguistic and rhetorical development. Several studies emphasize that EFL learners commonly exhibit a "one-sided" argument structure, with minimal integration of opposing viewpoints (cf. He and Du, 2024; Qin and Karabacak, 2010; Wagner et al., 2017). Moreover, as learners tend to rely heavily on formulaic expressions, they demonstrate limited use of critical thinking strategies during the planning and revision stages of writing (Ferretti and Graham, 2019; Kuhn, 1991). These limitations are often exacerbated in mixed-ability classrooms, where less proficient students may lack the metacognitive strategies necessary to revise content meaningfully, while more proficient learners still require structured support for advanced rhetorical moves such as counterargument and rebuttal (Hazaea, 2023).

Feedback has been identified as one of the most pivotal pedagogical tools in improving argumentative writing skills. According to Hattie and Timperley's (2007, p. 86) model, feedback clarifies goals (*feed up*), points out progress (*feed back*), and closes gaps between current achievement and learning outcomes (*feed* *forward*). Numerous studies confirm that high-quality, formative feedback significantly enhances learners' writing performance, especially when it is timely, specific, and focused on meaning-level aspects (Bitchener and Ferris, 2012; Peltzer et al., 2024; Zhang and Hyland, 2018). In argumentative writing, feedback is particularly crucial for helping learners recognize logical gaps, improve coherence, and refine argumentative strategies (Guo et al., 2024). However, providing individualized feedback in large and mixed-ability classrooms remains a persistent challenge, especially in resource-constrained settings (Liu et al., 2024; Ouahidi, 2021; Wisniewski et al., 2020).

Recent advances in large language models (LLMs) such as ChatGPT have opened new possibilities for AI-supported writing instruction. Unlike traditional automated writing evaluation (AWE) tools, which primarily target grammar and syntax, LLMs can provide nuanced, contextualized, and semi-structured feedback on content-level aspects of writing (Yoon et al., 2023). ChatGPT, in particular, has shown promise in supporting learners with argument structure, organization, and even idea generation, especially when guided by carefully designed prompts (Guo et al., 2024). Studies by Guo et al. (2022), Mahapatra (2024) show that AI-supported scaffolding can improve feedback quality, revision depth, and learner motivation. These tools are especially effective when students are trained to use structured prompts that focus AI output on genre-specific goals, such as argument strength and counterargument inclusion (Mollick and Mollick, 2023). In a study on chatbot-assisted argumentative writing, Guo et al. (2022) found that students who received AI support produced stronger arguments and engaged in deeper revision than those relying solely on peer feedback. These findings are echoed in research by Mahapatra (2024), who found GenAI feedback to have a significant positive impact on student learning coupled with positive student perceptions (p. 13).

Feedback effectiveness in writing instruction is not solely determined by its quality or timing but also by how learners engage with, interpret, and apply the feedback they receive; a process often referred to as feedback uptake (Carless and Boud, 2018). The Feedback Engagement Model proposed by Winstone et al. (2017) emphasizes that feedback is inherently dialogic, requiring active learner agency for it to be pedagogically impactful. This model outlines key stages in feedback engagement, including noticing, sense-making, and implementation each of which can vary significantly depending on whether the feedback originates from a human or an AI source.

When interacting with teacher-generated feedback, learners often benefit from interpersonal trust, contextual knowledge, and nuanced scaffolding that is aligned with classroom dynamics. However, such feedback may be delayed due to time constraints and can sometimes be inconsistent in tone or focus (Zhang and Hyland, 2018). By contrast, AI-generated feedback (e.g., from ChatGPT) offers immediacy and consistency, and can be tailored through prompt engineering to target specific rhetorical or genre-related issues (Guo et al., 2024). Yet, studies show that learners may interact with AI feedback more passively, often accepting suggestions without critical evaluation or reflection (Yoon et al., 2023). This passive uptake raises concerns about surface-level revision, overreliance, and reduced metacognitive engagement. Effective uptake of AI feedback, therefore, depends not only on the technical quality of the output but also on how learners are trained to critically engage with it. Furthermore, scaffolded reflection such as asking students to justify how they revised based on AI feedback can mitigate overreliance and encourage critical thinking (Woo et al., 2024). Research suggests that embedding reflective prompts (e.g., "Which of these AI suggestions will you use, and why?") and requiring justification for revisions can enhance cognitive engagement and support deeper learning (Mollick and Mollick, 2023; Winstone et al., 2017). Ultimately, while both AI and teacher feedback can be effective, their pedagogical value is mediated by how learners perceive their credibility and interact with the feedback in the revision process.

However, the limitations of AI feedback are also widely acknowledged. Yoon et al. (2023) caution that while ChatGPT can generate plausible feedback on coherence and logic, it may also "hallucinate" critiques, offer generic advice, or miss contextually nuanced issues. There is also the risk that students may accept AI suggestions uncritically, leading to overreliance and potential plagiarism (cf. Alshurafat et al., 2024; Esmaeil et al., 2023; Hostetter et al., 2024; Sánchez-Vera et al., 2024). These limitations underscore the importance of prompt design, feedback framing, and teacher mediation in AI-supported writing instruction.

3 Methodology

3.1 Research design

Following the recommendations outlined in Rose et al. (2019), this study employed a quasi-experimental, pretest-posttest, between-subjects design to examine the impact of AI-generated and teacher-generated feedback on EFL students' argumentative writing performance. The two independent variables were Feedback Type (AI-only vs. teacher-only) and Proficiency Level (Intermediate-Low vs. Advanced-Low). The dependent variable was the total score on an analytic rubric evaluating students' performance on a single argumentative writing task. A 2 × 2 factorial design was used to assess both main effects and their interaction. Participants were stratified by proficiency level and then assigned to feedback conditions based on intact class groupings to maintain ecological validity. This design was chosen to reflect the realities of classroom implementation while preserving sufficient experimental control for statistical analysis.

3.2 Participants

The participants were 120 (83 females and 37 males) undergraduate EFL students enrolled in a writing-focused course at a large public university in Jordan. All were native speakers of Arabic and had received a minimum of 4 years of formal English instruction at the university level. Based on prior institutional placement procedures, including Oral Proficiency Interviews aligned with ACTFL guidelines, participants were classified as either Intermediate-Low or Advanced-Low in proficiency. To ensure balanced representation, stratified sampling was used to assign participants to two feedback conditions: AI-generated feedback and teacher-generated feedback, with equal distribution across proficiency levels. This resulted in four subgroups:

AI/Intermediate-Low, AI/Advanced-Low, Teacher/Intermediate-Low, and Teacher/Advanced-Low each containing 30 students. Group assignment was not randomized but followed intact classroom sections to maintain ecological validity. Students were informed about the nature of the study, assured of confidentiality, and provided informed consent. The study received ethical approval from the university's Institutional Review Board.

3.3 Writing task

All students completed the same argumentative writing task, responding to the prompt: "Should university education be free for all students?" This topic was selected for its relevance, accessibility, and capacity to elicit critical reasoning, supporting the integration of claims, counterclaims, and rebuttals. Students were required to produce an essay of 250–300 words within a 45 min time limit. Prior to the task, they received brief instruction on the structural expectations of argumentative writing, including thesis formulation, body development, and conclusion. Students submitted a first draft (Draft 1), received feedback according to group assignment, and then submitted a revised version (Draft 2) within 1 week.

3.4 Feedback conditions

3.4.1 AI feedback group

Participants in the AI group received feedback from ChatGPT (GPT-4) using a structured, piloted prompt designed to elicit genre-specific, rhetorical-level feedback on argumentative writing (see Supplementary Appendix C for representative samples of feedback). The AI prompt guided the model to focus on argument structure, clarity, use of counterarguments, and revision support without rewriting any part of the student's text. Students were trained to input their essays in privacy mode ("chat history off") and instructed to apply the feedback independently. The full version of the tested AI mentor prompt, including the analytic rubric used to guide feedback interpretation, is provided in Supplementary Appendix A. The prompt was developed iteratively and grounded in recent AI pedagogy literature (e.g., Mollick and Mollick, 2023). The same rubric was used to guide feedback delivery in both the AI and teacher conditions to ensure consistency in focus, expectations, and assessment criteria (see Supplementary Appendix B for full rubric). The revised AI prompt was designed to simulate an interactive, step-by-step mentoring session, fostering a collaborative learning environment between the AI and the student. The process begins with the mentor gathering key contextual information about the student's writing goals, proficiency level, and specific areas of concern. By asking one question at a time and pausing for a response, the prompt supports a focused and responsive dialog, ensuring that feedback is tailored to the student's needs and aligned with their current level of development.

A central feature of the prompt is its emphasis on balanced, scaffolded feedback. The AI mentor is instructed to begin by identifying the strengths of the student's work, thereby establishing a supportive tone and recognizing effort. Constructive feedback follows, targeting rhetorical features such as argument clarity, evidence use, organization, and counterarguments. The student is then guided through the revision process, encouraged to apply feedback thoughtfully and to reflect on the changes made promoting metacognitive awareness and deeper engagement with the writing task. Moreover, the prompt supports iterative learning by allowing for follow-up feedback after revision. The AI mentor reviews the revised section, invites further reflection, and provides additional suggestions if needed. Whether the student feels ready to finalize the work or seeks continued support, the session concludes with affirming, forward-looking encouragement. This mentoring model not only reinforces writing development but also fosters learner autonomy, confidence, and a growth-oriented mindset.

3.4.2 Teacher feedback group

Participants in the teacher group received individualized feedback from their course instructor. Comments were handwritten on printed copies of Draft 1, guided by a feedback checklist aligned with the rubric dimensions: argument clarity, supporting evidence, counterargument integration, organization, and language use. Feedback was formative, non-evaluative, and provided within 48 hours of submission. All students were encouraged to reflect on their feedback and revise their drafts accordingly.

3.5 Instruments

3.5.1 Analytic writing rubric

An analytic rubric was developed and validated to assess argumentative writing performance. It included five equally weighted dimensions adapted from established frameworks used in prior EFL writing research (e.g., Ferretti and Graham, 2019; Qin and Karabacak, 2010):

1. Claim Clarity and Relevance
2. Support and Evidence
3. Counterarguments and Rebuttal
4. Organization and Coherence
5. Language Use (Grammar and Vocabulary)

Each category was scored on a five-point scale (1 = very weak; 5 = excellent), with a maximum total score of 25. The rubric (see Supplementary Appendix B for full rubric) was reviewed by two L2 writing experts for content validity.

3.5.2 Data collection procedure

The study protocol comprised six sequential phases designed to ensure methodological rigor and data integrity. Initially, all participants underwent a comprehensive orientation session during which they were informed of the study objectives, feedback mechanisms, and ethical protections governing their participation. Following this briefing, students completed their initial writing task (Draft 1) under standardized classroom conditions to ensure consistency across all participants. The intervention phase involved the systematic delivery of feedback according to predetermined group assignments, with participants receiving either AI-generated feedback through ChatGPT or traditional instructor feedback. Subsequently, students were afforded a 1 week period to independently revise their compositions and submit their final drafts (Draft 2), allowing for adequate reflection and incorporation of the provided feedback. Assessment procedures employed a validated analytic rubric administered by a primary rater who evaluated both initial and revised drafts. To establish inter-rater reliability and minimize scoring bias, a second trained evaluator independently assessed a stratified random sample comprising 36 essays (30% of the total corpus) using identical rubric criteria while remaining blind to group assignment conditions. Finally, all scoring data were systematically recorded in a structured Excel database, with entries organized by participant identification number, experimental group assignment, proficiency level classification, and pre-test (Draft 1) and post-test (Draft 2) performance scores.

3.5.3 Inter-rater reliability procedures

To ensure robust assessment of rater agreement, both Pearson's correlation (r) and the Intra-Class Correlation Coefficient (ICC) were used, recognizing ICC as the preferred metric for ordinal scoring in educational research (Koo and Li, 2016). Pearson's *r* was calculated for initial comparison, yielding moderate agreement for pre-test scores (r = 0.653; MAD = 2.11) and lower agreement for post-test scores (r = 0.316; MAD = 2.25). However, ICC (2,1) estimates provided a more robust evaluation of inter-rater consistency, with the post-test ICC = 0.61, indicating moderate agreement. These results support the overall reliability of the scoring procedure, though some variability remained particularly at higher score ranges. Rater discrepancies were discussed *post hoc* to calibrate interpretations and ensure consistent rubric application (see Table 1).

As shown in Figure 1, Pre-test scores show moderate alignment between Rater 1 and Rater 2, clustering along the diagonal. Figure 2, on the other hand, show more dispersion, especially at higher score ranges, indicating lower agreement.

3.6 Data analysis

Data analysis was conducted using Python programming language with specialized statistical libraries including pandas for data manipulation, scipy for statistical computations, and statsmodels for advanced statistical modeling. The analytical framework encompassed multiple complementary approaches to comprehensively evaluate the intervention effects. Descriptive analyses were initially performed to characterize the dataset, including computation of means, standard deviations, and score ranges stratified by experimental group and proficiency level classifications. To examine within-group performance changes, paired-sample *t*-tests were employed to assess the statistical significance of improvement from pre-intervention (Draft 1) to post-intervention (Draft 2) scores within each feedback condition. Between-group comparisons utilized independent-sample *t*-tests to evaluate differences in revision gains, calculated as the difference between post-test and pre-test scores, across the two feedback modalities. Additionally, a two-way analysis of variance (ANOVA) was implemented to simultaneously examine the main effects of feedback type and proficiency level, as well as their potential interaction, on post-intervention writing performance. Effect size calculations accompanied all inferential tests to assess practical significance beyond statistical significance, with Cohen's d computed for *t*-test comparisons and partial eta squared (η^2) calculated for ANOVA results. Prior to conducting parametric analyses, fundamental statistical assumptions were rigorously evaluated through Shapiro-Wilk tests for normality of distributions and Levene's tests for homogeneity of variance across groups, ensuring the appropriateness of the selected analytical procedures.

TABLE 1 Inter-rater reliability metrics for writing scores (*n* = 36).

Metric	Pre-test	Post-test
Pearson correlation (r)	0.653	0.316
Mean absolute difference	2.11	2.25

3.7 Ethical considerations

This study complied with institutional and international guidelines for ethical research in education. Informed consent was obtained from all participants. No personally identifiable information was collected or shared with the AI tool. All AI interactions occurred with "chat history" disabled to avoid data retention. To mitigate risks associated with AI-generated feedback (e.g., hallucinations, generic responses), prompts were standardized and tested extensively prior to deployment. Students were explicitly instructed to critically evaluate the feedback they received and revise their work accordingly. Teacher support was available for clarification. All writing samples and scores were anonymized prior to analysis, and data were stored securely with access limited to the research team.

4 Results

4.1 Descriptive statistics

A total of 120 participants were recruited for this study and included in the final analysis. The sample was systematically stratified to ensure balanced representation across two key variables: feedback modality (artificial intelligence-generated vs. teacher-provided feedback) and initial language proficiency level (Intermediate-Low vs. Advanced-Low, as determined by standardized placement assessments). This balanced factorial design resulted in equal cell sizes (*n* = 30 per condition), thereby optimizing statistical power and enabling robust between-group comparisons. Tables 2–4 present comprehensive descriptive statistics including means, standard deviations, and score ranges for pre-test performance, post-test performance, and gain scores (calculated as post-test minus pre-test scores) disaggregated by experimental condition and proficiency level. The descriptive data reveal several noteworthy patterns that warrant detailed examination. Across all experimental conditions and proficiency levels, participants demonstrated measurable improvement in writing performance from the initial draft (pre-test) to the revised draft (post-test). This universal pattern of improvement suggests that both AI-generated and teacher-provided feedback were

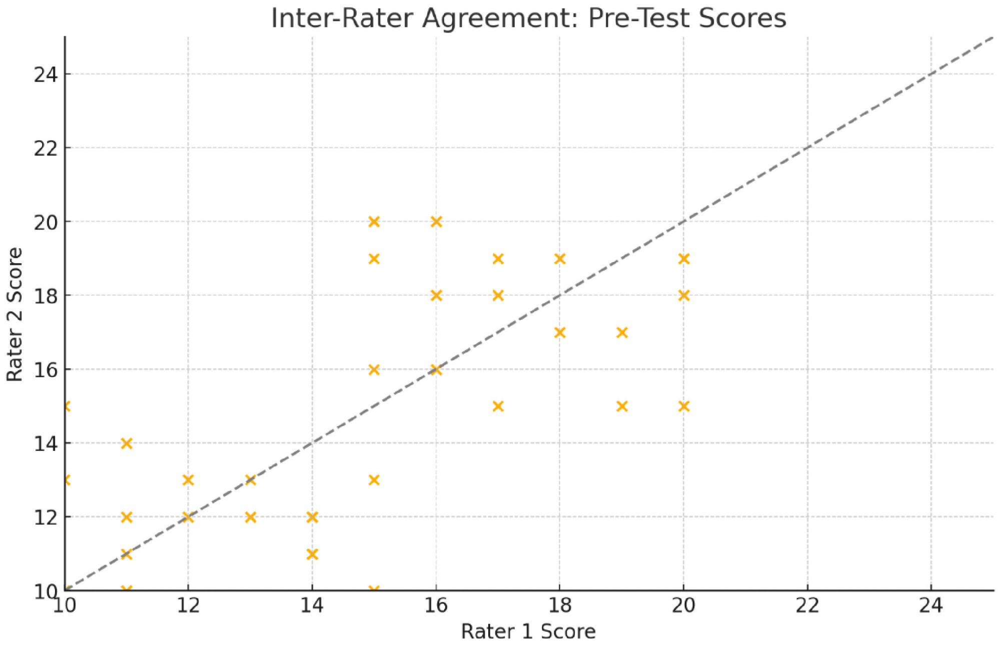

FIGURE 1
Pre-test scores.

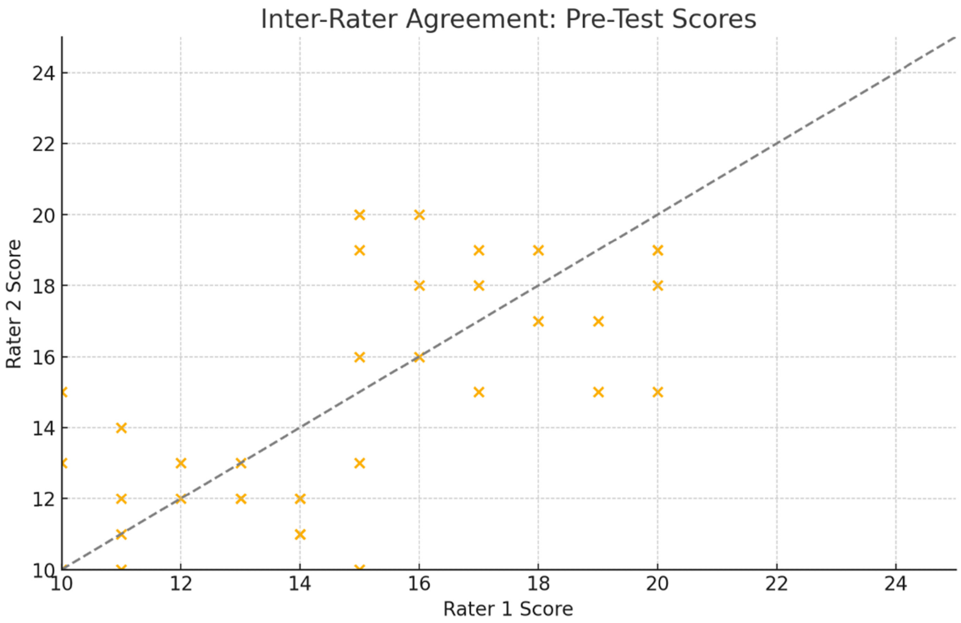

FIGURE 2
Post-test scores.

effective in facilitating writing enhancement, regardless of students' initial proficiency levels. The consistency of this finding across all subgroups provides preliminary evidence for the general efficacy of corrective feedback in second language writing contexts. The data also reveal a compelling interaction between initial proficiency level and learning outcomes. Intermediate-Low proficiency students, while demonstrating lower absolute scores on both pre-test and post-test measures compared to their Advanced-Low counterparts, exhibited notably higher average gain scores across both feedback conditions. This pattern suggests that students with lower initial proficiency may derive greater benefit from corrective feedback interventions, potentially due to greater room for improvement or increased sensitivity to explicit error correction at earlier stages of language development. Conversely, Advanced-Low proficiency students, despite achieving higher absolute performance scores, showed more modest gains from pre-test to post-test. This ceiling

TABLE 2 Pre- and post-test means and standard deviations.

Group	Proficiency	Pre M (SD)	Post M (SD)	Gain M (SD)
AI	Advanced-low	17.23 (1.76)	20.4 (1.65)	3.17 (2.09)
AI	Intermediate-low	12.5 (1.72)	18.1 (1.54)	5.6 (2.18)
Teacher	Advanced-low	17.2 (1.71)	20.43 (1.65)	3.23 (2.33)
Teacher	Intermediate-low	12.4 (1.67)	17.43 (1.87)	5.03 (2.53)

TABLE 3 Score ranges.

Group	Proficiency	Pre min–max	Post min–max	Gain min–max
AI	Advanced-low	15–20	18–23	−1 to 7
AI	Intermediate-low	10–15	15–20	1–10
Teacher	Advanced-low	15–20	18–22	−1 to 7
Teacher	Intermediate-low	10–15	15–20	0–10

effect phenomenon is consistent with previous research in second language acquisition, which suggests that learners at higher proficiency levels may require more sophisticated or targeted interventions to achieve measurable improvement (cf. Biber et al., 2011).

4.2 Within-group comparisons (paired samples *t*-tests)

To evaluate whether the feedback conditions led to statistically significant improvement in writing, paired samples *t*-tests were conducted within each of the four subgroups. Cohen's d values were calculated to assess the magnitude of the change, as shown in Table 5. Table 5 reveals that all groups demonstrated statistically significant improvement in writing scores from pre- to post-feedback, with large or very large effect sizes observed across conditions.

4.3 Between-Group comparison (independent samples *t*-test)

To assess whether AI or teacher feedback led to greater gains, an independent samples *t*-test was conducted on the gain scores across feedback groups: t(118) = 0.55, *p* = 0.586, Cohen's d = 0.10. The difference in gain scores between the AI and teacher feedback groups was not statistically significant, and the effect size was very small, suggesting practical equivalence.

4.4 Two-way ANOVA

A two-way ANOVA was conducted to test the main effects of Feedback Type and Proficiency Level, and their interaction on Post-Test Total Score, as shown in Table 6. The two-way analysis of variance yielded a statistically significant main effect for proficiency level. Post-revision performance scores were significantly higher among Advanced-Low participants compared to Intermediate-Low participants, indicating that language proficiency level was a

TABLE 4 Descriptive statistics by group and proficiency.

Group	Proficiency	Pre_total_mean	Pre_total_std	Pre_total_min	Pre_total_max	Pre_total_count	Post_total_mean	Post_total_std	Post_total_min	Post_total_max	Post_total_count	Gain_mean	Gain_std	Gain_min	Gain_max	Gain_count
AI	Advanced-low	17.23	1.76	15.0	20.0	30.0	20.4	1.65	18.0	23.0	30.0	3.17	2.09	−1.0	7.0	30.0
AI	Intermediate-low	12.5	1.72	10.0	15.0	30.0	18.1	1.54	15.0	20.0	30.0	5.6	2.18	1.0	10.0	30.0
Teacher	Advanced-low	17.2	1.71	15.0	20.0	30.0	20.43	1.65	18.0	22.0	30.0	3.23	2.33	−1.0	7.0	30.0
Teacher	Intermediate-low	12.4	1.67	10.0	15.0	30.0	17.43	1.87	15.0	20.0	30.0	5.03	2.53	0.0	10.0	30.0

TABLE 5 Paired *t*-test results and effect sizes by group and proficiency.

Group	Proficiency	t-value	*P*-value	Cohen's d
AI	Advanced-low	8.32	0.0	1.518
AI	Intermediate-low	14.1	0.0	2.575
Teacher	Advanced-low	7.6	0.0	1.388
Teacher	Intermediate-low	10.92	0.0	1.993

TABLE 6 Two-way analysis of variance (ANOVA) results.

Source	SS	df	F	*P*
C (group)	3.008333333333277	1.0	1.0618724008519935	0.3049323061304629
C (proficiency)	210.67500000000322	1.0	74.36342428238272	3.890819445332791e-14
C (group):C (proficiency)	3.6750000000084	1.0	1.2971903844203565	0.25707366633367534
Residual	328.6333333333333	116.0	Nan	Nan

meaningful predictor of writing quality following revision. With respect to feedback modality, the main effect of feedback type (AI-mediated versus instructor-provided) did not reach statistical significance. This finding suggests comparable efficacy between artificial intelligence and human instructor feedback on learners' writing performance. The analysis further revealed no significant interaction effect between proficiency level and feedback type. The absence of a significant interaction indicates that the relative effectiveness of AI-mediated versus instructor-provided feedback remained consistent across proficiency levels.

4.5 Summary of findings

Both AI-generated and teacher-generated feedback led to statistically significant improvements in students' argumentative writing scores. However, no significant difference was observed in the gain scores between the two feedback groups, indicating comparable effectiveness. Proficiency level exerted a strong main effect, with Advanced-Low learners outperforming Intermediate-Low learners on the post-test. No significant interaction was found between feedback type and proficiency level. Effect size calculations underscored the practical relevance of these findings, particularly for Intermediate-Low learners who demonstrated the largest within-group gains. Although all essays were scored by the primary researcher, inter-rater reliability was established on a stratified 30% subsample. The results revealed moderate to acceptable agreement, especially for pre-test score supporting the reliability and validity of the scoring process.

5 Discussion

5.1 Summary of main findings

This study set out to examine the impact of AI-generated versus teacher-generated feedback on the argumentative writing performance of EFL learners with different proficiency levels. Results revealed that both feedback types led to statistically significant improvement in writing scores across all participants, regardless of proficiency. However, no significant differences were found between the AI and teacher feedback groups in terms of overall writing gains. This finding disagrees with those drawn by Li et al. (2024), who found that students who relied on AI feedback (experiment group) demonstrated more improvement than that relied on teacher feedback (control group). In contrast, proficiency level had a significant main effect, with Advanced-Low learners outperforming Intermediate-Low learners in the post-test. This finding comports with the systematic review conducted by Biber et al. (2011), which demonstrated an inverse relationship between initial proficiency level and magnitude of performance gains, whereby students with lower baseline competencies exhibited proportionally greater improvement than their more proficient counterparts. Effect sizes were large across all groups, indicating substantial practical improvement, especially among Intermediate-Low learners who benefited most from the revision process. These findings contribute to a growing body of evidence suggesting that AI feedback, when structured and scaffolded effectively, can match the effectiveness of traditional teacher feedback in supporting meaning-level revisions in argumentative writing (e.g., Guo et al., 2024).

5.2 AI Feedback and writing development

The finding that AI-generated feedback performed comparably to teacher-generated feedback supports prior research on the instructional potential of large language models (LLMs) in EFL writing (Guo et al., 2022). Importantly, students in the AI group used tested prompts that directed the model to provide focused, rhetorical-level commentary that is targeting argument clarity, evidence, counterarguments, and coherence. This aligns with recent work emphasizing the critical role of prompt engineering in shaping the relevance and usefulness of AI feedback (Mollick and Mollick, 2023). The non-significant difference in gains between AI and Teacher groups, coupled with the small effect size (Cohen's d = 0.10), is particularly relevant for scalability in writing instruction. In contexts where teacher feedback is constrained by class size, workload, or time limitations, structured AI feedback can serve as a viable supplement or alternative, especially if embedded within a pedagogically sound writing process.

5.3 Proficiency level as a moderator

The significant effect of proficiency level highlights the importance of learner characteristics in shaping writing outcomes. While all groups improved significantly, Intermediate-Low learners exhibited the largest effect sizes in both feedback conditions. This suggests that when given access to clear, actionable feedback (whether from a human or an AI assistant) lower-proficiency learners can make dramatic improvements in their writing performance. These findings support research that emphasizes the role of feedback scaffolding in mixed-proficiency classrooms (Knoch et al., 2015). Interestingly, the absence of an interaction effect between feedback type and proficiency level suggests that AI feedback did not disadvantage less proficient learners, a concern raised in earlier research (Yoon et al., 2023). This outcome may be attributed to the training and guidance students received on how to interpret and apply AI feedback, as well as the structured prompts used to limit off-topic or generic responses.

6 Conclusion, implications, limitations, and future research

This study explored the comparative effectiveness of AI-generated and teacher-generated feedback on EFL students' argumentative writing performance across two ACTFL proficiency levels. Findings revealed that both feedback types led to statistically significant gains in writing scores, with no significant difference between the AI and teacher groups. This result, coupled with a small effect size for the between-group comparison, suggests that well-structured AI feedback (delivered through prompt engineering) can serve as a scalable and pedagogically meaningful alternative to traditional teacher feedback. Notably, proficiency level emerged as a significant predictor of post-revision performance, with Advanced-Low learners outperforming Intermediate-Low learners. However, the largest within-group effect sizes were observed among Intermediate-Low students, indicating their high responsiveness to structured revision support, regardless of the feedback source. The results underscore the growing potential of large language models (LLMs) in second language writing instruction, especially in contexts where teacher feedback is limited by class size or time constraints. At the same time, the findings reinforce the importance of embedding AI use within guided, ethical, and pedagogically sound frameworks. This study is not without limitations. Essays were scored by a single rater, and only immediate revision gains were assessed. Future studies should include multiple raters, examine long-term effects of AI-assisted revision, and incorporate learner reflections to better understand feedback uptake. As educational technologies evolve, it is imperative to continue evaluating how AI can complement rather than replace the human element in language teaching and learning. Despite the single-rater design, the inclusion of inter-rater reliability metrics for a representative subsample provides reasonable confidence in scoring validity and transparency.

The findings have significant implications for EFL writing instruction across diverse educational contexts. Large language models can serve as effective feedback partners, delivering genre-specific guidance that supports student revision while freeing instructors for higher-order pedagogical activities. Standardized prompt design focusing on genre conventions can mitigate AI inaccuracies and misdirection, while the scalable nature of AI feedback enables differentiated instruction in large, mixed-proficiency classrooms. However, AI should complement rather than replace teacher feedback, functioning as a preliminary revision tool that prepares students for subsequent instructor guidance. This tiered approach preserves essential human pedagogical expertise while leveraging technological capabilities to enhance instructional effectiveness and accessibility across varied resource contexts.

While this study offers valuable insights into the comparative effects of AI and teacher feedback, several limitations must be acknowledged. First, although most essays were scored by a single trained rater, inter-rater reliability was established through a dual-rating process on a stratified subsample of 36 essays. The inclusion of a second rater, use of a validated analytic rubric, and high transparency in scoring procedures help mitigate concerns of bias and subjectivity. Pre-test scores showed moderate agreement (r = 0.653), while post-test scores revealed more variability (r = 0.316), a common pattern in subjective assessment of revised writing. Several factors may have contributed to this issue. First, the nature of the post-test responses, produced after exposure to individualized feedback and revision, may have led to more diverse writing structures and strategies, increasing subjectivity in rating. Second, although raters used a shared rubric, differences in interpretation may have emerged when evaluating revisions. Discrepancies were resolved through discussion to ensure shared understanding of rubric dimensions. Nonetheless, future studies should consider full-scale dual scoring or use intra-class correlation (ICC) to assess agreement more robustly. To address this in future research, we recommend the implementation of full dual scoring for all writing samples rather than a subset, alongside rigorous rater calibration sessions prior to and during scoring. These practices can help improve consistency and minimize subjectivity in evaluating student writing, especially in post-intervention contexts where performance tends to be more heterogeneous.

Second, the study examined short-term effects of feedback on a single revision cycle. Longitudinal data would be needed to evaluate the durability of learning gains and the impact of sustained feedback engagement over time. Third, another limitation of this study is the absence of direct measurement of student feedback uptake. While our approach was guided by theoretical models of feedback engagement (e.g., Winstone et al., 2017), we did not collect empirical data on how students interpreted or applied the formative feedback provided. As a result, the study cannot account for individual differences in feedback engagement or clarify the specific ways in which feedback influenced revisions. Future research should consider incorporating qualitative and process-oriented methods, such as think-aloud protocols, student interviews, or digital revision tracking, to capture how learners interact with feedback and make use of it during revision. Lastly, a further limitation of this study lies in its use of a single argumentative writing prompt as the basis for data collection. While argumentative writing is an important academic genre, relying on a single task restricts the extent to which findings can be generalized to other types of writing, such as narrative, expository, or reflective genres. Moreover, genre-specific features may influence how students interpret and apply feedback, meaning that the observed effects of the intervention may not transfer uniformly across contexts. Future research should replicate this study using a broader range of writing genres and disciplinary tasks to assess whether the effectiveness of rubric-aligned, AI-generated feedback varies by genre or subject area.

Although post-test scores showed lower inter-rater correlation, the inclusion of a second rater on a stratified subsample mitigates the risk of single-rater bias and supports the reliability of scoring. While ICC estimates indicated moderate agreement (ICC = 0.61), full-scale dual scoring across all essays would further enhance the reliability and generalizability of the findings, particularly in the context of post-revision assessment where scoring tends to be more variable.

Future research endeavors should investigate several critical dimensions to advance understanding of AI-assisted writing instruction. Longitudinal studies examining the cumulative effects of iterative AI-supported writing cycles would provide valuable insights into sustained performance trajectories and skill development patterns over extended periods. Additionally, comparative analyses of differentiated AI prompt strategies, including structure-oriented, error-correction focused, and tone-modification approaches, would elucidate the relative efficacy of targeted feedback modalities in facilitating specific aspects of revision behavior. Investigations into student attitudes and confidence levels regarding AI-generated versus instructor-provided feedback represent another essential research direction, particularly given the implications for pedagogical acceptance and implementation. Furthermore, the incorporation of qualitative methodologies would substantially enhance the depth of understanding regarding student engagement with AI feedback systems. Specifically, stimulated recall interviews and reflective journaling protocols could illuminate the cognitive processes underlying students' interpretation, evaluation, and application of AI-generated suggestions, a domain that remains significantly underexplored in current literature. Such mixed-methods approaches would provide crucial insights into the mechanisms through which AI feedback influences writing development and inform evidence-based best practices for educational implementation.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by University of Jordan Board of Ethics. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AA: Writing – review and editing, Writing – original draft. HA: Writing – review and editing, Writing – original draft. MA: Writing – review and editing, Writing – original draft. MA-D: Writing – original draft, Writing – review and editing. RA: Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1614673/full#supplementary-material

References

Alshurafat, H., Al Shbail, M. O., Hamdan, A., Al-Dmour, A., and Ensour, W. (2024). Factors affecting accounting students' misuse of chatgpt: An application of the fraud triangle theory. *J. Financial Report. Account.* 22, 274–288. doi: 10.1108/JFRA-04-2023-0182

Banihashem, S. K., Noroozi, O., Van Ginkel, S., Macfadyen, L. P., and Biemans, H. J. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educ. Res. Rev.* 37:100489. doi: 10.1016/j.edurev.2022.100489

Biber, D., Nekrasova, T., and Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Res. Rep. Ser.* 2011, i–99. doi: 10.1002/j.2333-8504.2011.tb02241.x

Bitchener, J., and Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing.* London: Routledge.

Carless, D., and Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assess. Eval. High. Educ.* 43, 1315–1325. doi: 10.1080/02602938.2018.1463354

Chan, C. K. Y., and Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *Int. J. Educ. Technol. High. Educ.* 20, 43–18. doi: 10.1186/s41239-023-00411-8

Esmaeil, A.-A.-A., Dzulkifli, D., Maakip, I., Matanluk, O., and Marshall, S. (2023). Understanding student perception regarding the use of ChatGPT in their argumentative writing: A qualitative inquiry. *J. Komunikasi Malays. J. Commun.* 39, 150–165. doi: 10.17576/JKMJC-2023-3904-08

Fadlelmula, F., and Qadhi, S. (2024). A systematic review of research on artificial intelligence in higher education: Practice, gaps, and future directions in the GCC. *J. Univ. Teach. Learn. Pract.* 21, 146–173. doi: 10.53761/pswgbw82

Ferretti, R. P., and Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Read. Writ.* 32, 1345–1357. doi: 10.1007/s11145-019-09950-x

Guo, K., Pan, M., Li, Y., and Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *Int. High. Educ.* 63:100962. doi: 10.1016/j.iheduc.2024.100962

Guo, K., Wang, J., and Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assess. Writ.* 54:100666. doi: 10.1016/j.asw.2022.100666

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hazaea, A. N. (2023). Process-Genre approach in mixed-ability classes: Correlational study between EFL academic paragraph reading and writing. *Novitas-Royal (Research on Youth and Language)* 17, 1–12. doi: 10.5281/zenodo.10015742

He, H., and Du, Y. (2024). "The effectiveness of dialogical argumentation in supporting low-level EAP learners' evidence-based writing: A longitudinal study," in *English for academic purposes in the EMI context in Asia: XJTLU impact*, eds B. Zou and T. Mahy (Switzerland: Springer), 45–75.

Hostetter, A. B., Call, N., Frazier, G., James, T., Linnertz, C., Nestle, E., et al. (2024). Student and faculty perceptions of generative artificial intelligence in student writing. *Teach. Psychol.* 52, 319–329. doi: 10.1177/00986283241279401

Knoch, U., Rouhshad, A., Oon, S. P., and Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *J. Sec. Lang. Writ.* 28, 39–52. doi: 10.1016/j.jslw.2015.02.005

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropract. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Lee, S. S., and Moore, R. L. (2024). Harnessing Generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learn.* 28, 82–106. doi: 10.24059/olj.v28i3.4593

Li, H., Wang, Y., Luo, S., and Huang, C. (2024). The influence of GenAI on the effectiveness of argumentative writing in higher education: Evidence from a quasi-experimental study in China. *J. Asian Public Pol.* 18, 405–430. doi: 10.1080/17516234.2024.2363128

Liu, Y., Xiong, W., Xiong, Y., and Wu, Y.-F. B. (2024). Generating timely individualized feedback to support student learning of conceptual knowledge in Writing-To-Learn activities. *J. Comp. Educ.* 11, 367–399. doi: 10.1007/s40692-023-00261-3

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

Mallahi, O. (2024). Exploring the status of argumentative essay writing strategies and problems of Iranian EFL learners. *Asian-Pacific J. Sec. For. Lang. Educ.* 9:19. doi: 10.1186/s40862-023-00241-1

Mollick, E., and Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv [Preprint]* doi: 10.48550/arXiv.2306.10052

Ouahidi, L. M. (2021). Teaching writing to tertiary EFL large classes: Challenges and prospects. *Int. J. Linguist. Literat. Trans.* 4, 28–35. doi: 10.32996/ijllt.2021.4.6.5

Pelenkahu, N., Ali, M. I., Tatipang, D. P., Wuntu, C. N., and Rorintulus, O. A. (2024). Metacognitive strategies and critical thinking in elevating EFL argumentative writing proficiency: Practical insights. *Stud. Eng. Lang. Educ.* 11, 873–892. doi: 10.24815/siele.v11i2.35832

Peltzer, K., Lorca, A. L., Krause, U.-M., and Busse, V. (2024). Effects of formative feedback on argumentative writing in English and cross-linguistic transfer to German. *Learn. Instruct.* 92:101935. doi: 10.1016/j.learninstruc.2024.101935

Qin, J., and Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System (Linköping)* 38, 444–456. doi: 10.1016/j.system.2010.06.012

Rose, H., McKinley, J., and Baffoe-Djan, J. B. (2019). *Data collection research methods in applied linguistics*. London: Bloomsbury Academic.

Sá,nchez-Vera, F., Reyes, I. P., and Cedeo, B. E. (2024). "Impact of artificial intelligence on academic integrity: Perspectives of faculty members in Spain," in *Artificial intelligence and education: Enhancing human capabilities, protecting rights, and fostering effective collaboration between humans and machines in life, learning, and work*, ed. M. D. Díaz-Noguera (Spain: Octaedro).

Su, Y., Lin, Y., and Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assess. Writ.* 57:100752. doi: 10.1016/j.asw.2023.100752

Su, Y., Liu, K., Lai, C., and Jin, T. (2021). The progression of collaborative argumentation among English learners: A qualitative study. *System* 98:102471. doi: 10.1016/j.system.2021.102471

Wagner, C. J., Parra, M. O., and Proctor, C. P. (2017). The interplay between student-led discussions and argumentative writing. *TESOL Quar.* 51, 438–449. doi: 10.1002/tesq.340

Wang, H., and Dang, A. (2024). Enhancing L2 writing with generative ai: A systematic review of pedagogical integration and outcomes. *Preprint* 2, 1–30. doi: 10.13140/RG.2.2.19572.16005

Winstone, N. E., Nash, R. A., Parker, M., and Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educ. Psychol.* 52, 17–37. doi: 10.1080/00461520.2016.1207538

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3389/fpsyg.2019.03087

Woo, D. J., Wang, D., Guo, K., and Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Educ. Inf. Technol.* 29, 24963–24990. doi: 10.1007/s10639-024-12819-4

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.* 55, 90–112. doi: 10.1111/bjet.13370

Yoon, S.-Y., Miszoglad, E., and Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv [Preprint]* doi: 10.48550/arXiv.2310.06505

Zhang, Z. V., and Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assess. Writ.* 36, 90–102. doi: 10.1016/j.asw.2018.02.004

Zhu, W. (2001). Performing argumentative writing in English: Difficulties, processes, and strategies. *TESL Can. J.* 19, 34–50. doi: 10.18806/tesl.v19i1.918