Journal of Interdisciplinary Research in English Language Communication
E-ISSN: 3060-7655 – https://irelc.znu.ac.ir – Published by University of Zanjan

Please cite this paper as follows:

Hemmati, A. (2025). Exploring EFL learners' writing challenges and the role of AI tools in writing development: A corpus-based and pedagogical analysis. *Journal of Interdisciplinary Research in English Language Communication*, *2*(1), 60-72. https://doi.org/10.30470/irelc.2025.2078891.1038

University of Zanjan

Original Research

Exploring EFL Learners' Writing Challenges and the Role of AI Tools in Writing Development: A Corpus-Based and Pedagogical Analysis

Ali Hemmati

Department of English Language Teaching, Farhangian University, P.O. Box 14665-889, Tehran, Iran; *a.hemmati@cfu.ac.ir*

Received: 27/11/2025 Accepted: 26/12/2025

Abstract

This study investigates the linguistic and rhetorical challenges EFL learners encounter in academic writing and examines how engagement with AI tools influences their drafting and revision processes. Anchored in cognitive, noticing-based, and sociocultural perspectives, the study combines a corpus-based analysis of 152 argumentative essay drafts with qualitative insights from semi-structured interviews. Quantitative findings show consistent reductions in grammatical, lexical, and cohesion-related errors after AI-supported revision, with improvements becoming more evident over the semester. Learners' initial drafts in the second writing cycle also contained fewer micro-level errors, suggesting partial internalization of patterns highlighted through repeated AI feedback. However, progress in higher-order features, particularly argument clarity and discourse organization, remained limited, indicating that current AI systems provide strong surface-level scaffolding but insufficient guidance for complex rhetorical development. Qualitative analyses further reveal that learners increasingly adopted selective and critical engagement with AI suggestions, balancing appreciation for enhanced accuracy and confidence with concerns about overreliance, authorship, and ethical boundaries. Together, these findings demonstrate that AI tools function as effective mediators of local linguistic accuracy but offer only partial support for deeper argumentative competence. The study underscores the need for pedagogical frameworks that integrate AI strategically with explicit instruction in rhetorical reasoning and responsible tool use to promote sustainable writing development.

***Keywords*:** EFL Writing Development; AI-generated Feedback; Argumentative Essays; Corpus-based Analysis; Learners' Perceptions.

1. Introduction

Writing in a second or foreign language (L2), particularly English, is widely recognized as one of the most demanding productive skills for learners. For students of English as a Foreign Language (EFL), writing requires simultaneous attention to a range of interrelated competencies: grammatical accuracy, lexical precision, discourse coherence, and the higher-order processes that govern planning, monitoring, and revision (Flower & Hayes, 1981; Hyland, 2003). Recent research also shows that even advanced EFL learners continue to struggle with developing grammatical and lexical accuracy over time, highlighting how persistent these challenges can be (Lötscher et al., 2025). These intersecting demands often lead to recurring issues such as grammatical errors, misused vocabulary, weak cohesion, and underdeveloped argumentation. When such difficulties accumulate, they not only impede successful communication but also erode learners' confidence and limit their growth toward more advanced academic writing abilities.

Over the past two years, the rapid rise of generative artificial intelligence (AI) has fundamentally reshaped this landscape. AI writing tools can be broadly categorized into two types: grammar-oriented tools, such as Grammarly and similar automated writing evaluation systems that provide corrective feedback on grammar, spelling, and style, and generative large language model (LLM)-based tools, such as ChatGPT, which can generate, restructure, or expand text and assist with higher-order writing tasks beyond surface corrections (Jaramillo et al., 2025; Shi et al., 2025). In this paper, the term "AI tools" is used as an umbrella term encompassing both grammar-oriented and generative LLM-based

tools. Large language models (LLMs) like ChatGPT, and writing-support tools such as Grammarly and QuillBot, have become increasingly integrated into students' writing practices—whether encouraged by instructors, allowed informally, or used independently outside the classroom. These tools provide immediate, individualized feedback on grammar, wording, tone, and even content organization, offering forms of scaffolding once limited to human tutors (Polakova & Ivenz, 2024). For many learners, the convenience and accessibility of AI feedback can make the writing process less intimidating and more manageable.

However, the growing reliance on AI tools has also raised substantial pedagogical and ethical concerns. Recent scholarship highlights inconsistencies in AI-generated suggestions, particularly regarding discourse coherence, contextual appropriateness, and factual accuracy. For example, Gustilo et al. (2024) found that educators continue to grapple with balancing AI-assisted writing practices with concerns about authenticity, algorithmic influence, and academic integrity. Systematic analyses further emphasize that generative AI presents both opportunities and vulnerabilities for higher education, particularly regarding transparency, authorship, and responsible use (Bittle & El-Gayar, 2025). Beyond these broad concerns, emerging empirical work suggests that although AI can improve surface-level accuracy and fluency, it does not always foster deeper rhetorical development or support the acquisition of autonomous writing strategies. Polakova and Ivenz (2024) reported measurable gains in accuracy but observed that learners sometimes accepted AI suggestions without fully understanding the linguistic principles underlying them. Related research comparing student populations has found notable differences in how learners employ AI-generated feedback, with some benefiting from improved conciseness and cohesion while others demonstrate signs of overreliance or reduced self-regulation (Parra Núñez et al., 2025).

Teacher perspectives add further nuance. Shousha and Oraby (2025) found that educators value the efficiency of ChatGPT in addressing mechanical errors, yet they remain cautious about its limited capacity to capture the subtleties of content coherence and writers' intentions. Systematic reviews from 2024–2025 echo these tensions, noting that while LLMs can stimulate critical thinking and support idea generation, they may also inadvertently promote shallow engagement with language when learners rely too heavily on automated suggestions (Liu et al., 2025). Motivation-oriented research paints a similarly complex picture: interactions with ChatGPT 4.0 have been shown to boost learners' writing motivation temporarily, yet this effect diminishes once AI support is withdrawn, raising concerns about long-term dependency (Zare et al., 2024).

Despite these timely contributions, several important research gaps remain. First, while a substantial portion of existing studies emphasizes improvements in linguistic accuracy, fewer have examined how AI tools interact with deeper discourse-level challenges, such as cohesion, argument flow, and rhetorical structure, which often shape the overall quality of academic writing. Second, although many researchers have explored learners' attitudes toward AI or measured performance outcomes, far fewer have systematically connected these perspectives with corpus-based evidence of actual writing patterns across multiple drafts. Consequently, it is still unclear whether AI support leads to meaningful internalization of writing skills or merely produces more polished texts without corresponding linguistic development.

Third, ethical and pedagogical debates, especially concerning authenticity, transparency, and reliance on automated feedback, remain largely theoretical when considered apart from real student writing processes. As AI becomes increasingly integrated into the writing habits of learners worldwide, empirical, classroom-based studies are urgently needed to understand how students navigate these issues in practice, how they incorporate AI feedback into their revision decisions, and how their metacognitive awareness evolves in response.

In response to these gaps, the present study adopts a multimethod approach that integrates corpus-based linguistic analysis with a qualitative investigation of learners' revision behavior and reflections. By examining a semester-long corpus of EFL essay writing and tracking revision histories with and without AI assistance, the study offers a detailed account of the linguistic and rhetorical issues that persist despite access to automated tools. Furthermore, by analyzing learners' metacognitive engagement and perceptions of AI-generated feedback, the study seeks to clarify whether such tools genuinely support long-term writing development or primarily facilitate surface-level corrections. Through this combined lens, the research contributes timely evidence to ongoing debates about the pedagogical value and ethical implications of generative AI in language learning. To align with these objectives, the study is guided by the following research questions:

1. What are the most frequent linguistic and discourse-level problems identified in a corpus of EFL learners' essays written over a semester?
2. How does engagement with AI writing tools influence the frequency, types, and persistence of these problems across multiple drafting stages?
3. In what ways does AI-generated feedback mediate learners' revision strategies, metacognitive reflection, and perceptions of the pedagogical and ethical dimensions of AI-assisted writing?

2. Literature Review

2.1. Theoretical Framework

Understanding how EFL learners develop writing proficiency—and how AI tools reshape this process—requires a focused theoretical lens that integrates cognitive and social dimensions of writing. At the core is a task-mediated cognitive model of L2 writing (Kormos, 2023), which conceptualizes composing as a recursive cycle of planning, translating ideas into language, and revising text under constraints of working memory. This model builds on and refines the classic Flower and Hayes (1981) cognitive-process perspective. Recent studies link this cognitive framework to AI-mediated writing, illustrating how tools can support sentence-level accuracy while potentially reducing deeper cognitive engagement if learners focus on surface-level corrections (Poláková & Ivenz, 2024; Wan et al., 2024).

Complementing this cognitive lens is the concept of noticing, which explains how learners identify and internalize gaps between their output and target language forms (ElEbyary, 2024). Noticing functions as a mediator for effective learning: without reflective engagement, even accurate AI-generated corrections may remain superficial. Operationalizing noticing through revision traces allows the present study to examine whether learners integrate AI feedback cognitively across multiple drafts, or merely apply corrections mechanically.

Likewise, a streamlined sociocultural perspective situates writing within the classroom and tool-mediated context (Gustilo et al., 2024). AI tools act as mediating artefacts, redistributing agency between learners and technology. Classroom norms, peer interactions, and teacher guidance shape how learners perceive, evaluate, and implement AI suggestions, linking micro-level linguistic outcomes to social and motivational factors.

Together, these three elements—cognitive processes, noticing, and sociocultural context—provide a coherent, integrated lens that informs both the design and interpretation of studies on AI-mediated writing. This framework clarifies how persistent linguistic and discourse-level challenges emerge, how AI feedback shapes revision behavior, and under what conditions learners meaningfully internalize writing strategies.

2.2. Review of Previous Studies

Research on AI's role in L2 writing has expanded rapidly over the last two years, reflecting both technological advances and evolving pedagogical practices. Early investigations focused on perceptions and classroom policies, later studies examined learner behaviors, and more recent work integrated longitudinal and process-oriented analyses, revealing a continuum from surface-level improvements to potential discourse-level gains.

Gustilo et al. (2024) conducted multi-site research on teachers' practices and policy adaptations, revealing pragmatic acceptance of AI for routine corrections alongside concerns about authorship and assessment integrity. While informative regarding institutional tensions, these studies largely leave questions about learners' actual revision behaviors unanswered. Poláková and Ivenz (2024) implemented one of the first quasi-experimental classroom studies examining ChatGPT feedback with EFL students, reporting measurable gains in grammar and fluency, alongside improved confidence and drafting efficiency. These results highlight the consistent finding across studies that AI effectively supports micro-linguistic improvements, though gains in higher-order discourse skills remain limited**.**

Teng (2024) explored AI as a planning and idea-generation companion, emphasizing motivational benefits such as reduced writing anxiety and increased willingness to experiment. This complements prior findings by linking affective engagement to the likelihood of revision uptake, although generalizability is constrained by the small, context-specific sample. ElEbyary (2024) examined AWCF and noticing, demonstrating that teacher-mediated feedback promotes deeper

engagement than immediate AI corrections. This study underscores the central role of reflective processing in translating feedback into lasting learning, which remains a limitation of AI-only interventions. Jin et al. (2024) highlighted individual differences, showing that self-regulated learning strongly moderates the long-term effectiveness of AI feedback. Tran (2025) and Zou et al. (2025) extended longitudinal research, revealing that iterative AI and teacher feedback can enhance cohesion and procedural fluency over time, though sample sizes and tool novelty limit broad applicability.

Liu et al. (2025) synthesize these findings, confirming that AI reliably improves surface features and idea generation but offers limited evidence for sustained gains in higher-order rhetorical competence. Across studies, common strengths include methodological diversity and responsiveness to technological innovation, whereas recurring limitations involve short-term or surface-focused outcomes, reliance on self-report or isolated tasks, and insufficient attention to sociotechnical and ethical factors.

Collectively, the literature suggests a clear trajectory: initial research focused on perceptions and policy, followed by classroom interventions and process-oriented studies, and culminating in longitudinal, mixed-methods examinations. This synthesis reveals persistent gaps in integrating cognitive, social, and technological perspectives, which the present study addresses by combining semester-long corpus analysis, revision trace tracking, and qualitative reflections. The study thus investigates whether AI functions as a genuine scaffold for meaningful writing development or primarily supports superficial text improvement.

3. Methodology

3.1. Research Design

This study adopted a mixed-methods design that integrated corpus-based textual analysis with qualitative inquiry to examine EFL learners' writing challenges and their engagement with AI-generated feedback. The design was informed by cognitive-process, noticing, and sociocultural perspectives, which together provided a coherent rationale for analyzing both the observable linguistic patterns in students' texts and the interpretive meanings they assigned to AI-mediated writing. The quantitative component focused on identifying recurrent linguistic and discourse-related problems across drafts, whereas the qualitative component sought to illuminate learners' perceptions of AI tools and the reasoning behind their revision decisions. Integrating these complementary strands allowed for a more complete account of how AI influences writing development within a real instructional setting.

3.2. Participants

Participants were 38 undergraduate EFL learners enrolled in an academic writing course at a private university in Kermanshah, Iran. They formed an intact class and were therefore selected through convenience sampling. The group included both male and female students with intermediate to upper-intermediate proficiency. To extend the depth of the qualitative findings, 15 students (7 males and 8 females) were purposively selected for semi-structured interviews. Selection criteria included variation in writing performance, frequency of AI tool use, and willingness to articulate their learning experiences. This combination of convenience and purposive sampling created a sufficiently diverse participant pool to address the research questions while remaining feasible within the semester-long study.

3.3. Data Collection

Data collection spanned twelve weeks and centered explicitly on argumentative writing, a genre that undergraduate EFL learners are expected to master for academic communication. The first six weeks focused on instruction, during which students were introduced to the essential features of argumentative essays, including establishing a clear position, providing reasons and examples, maintaining logical paragraph progression, and using cohesive expressions to link ideas. Following this instructional phase, each student produced two major argumentative essays: one at mid-semester and another at the end. For each essay, students first wrote an initial draft without external assistance, then revised the draft using AI tools of their choice—typically Grammarly or ChatGPT—and subsequently submitted a final version. This structure enabled systematic cross-draft comparison and allowed the study to capture both immediate and developmental changes across the semester. In Week 12, semi-structured interviews were conducted with the selected 15 participants. Interviews lasted between twenty and thirty minutes and explored how learners interpreted

AI feedback, how they navigated the revision process, and how they perceived the pedagogical value and potential limitations of AI tools. All interviews were audio-recorded, transcribed verbatim, and anonymized.

3.4. Data Analysis

A learner corpus was constructed from all draft stages of the two essays produced by the 38 participants. Corpus analysis was performed using AntConc 4.0 (Anthony, 2023) and proceeded through several steps. Prior to analysis, all drafts were cleaned, anonymized, and segmented into comparable units.

To systematically analyze writing issues, an error-coding scheme was created. This scheme used established practices in L2 writing research to define key categories of student writing problems (e.g., grammatical accuracy, lexical appropriateness, sentence clarity) and differentiate these from higher-order discourse features like cohesion and argumentative structure (Ferris, 2011; Hyland, 2019). Cohesion was defined as using linguistic devices (e.g., connectors, referential ties, lexical repetition) that clearly link sentences and ideas. In contrast, argument clarity was described as how well claims, reasons, and supporting evidence are logically presented and easy for a reader to understand, regardless of surface connectors. This distinction ensured that surface-level connectors were separated from deeper organizational coherence and reasoned thinking, in line with modern approaches to analytic writing assessment (Hyland, 2019).

Each draft was coded for the presence, frequency, and type of issues within these categories. To establish interrater reliability, the primary researcher coded the full corpus, and a second trained coder independently coded a stratified subset (25%) of drafts across all essay stages. A joint calibration session preceded independent coding to align interpretations of the coding scheme, and disagreements that emerged in initial coding were discussed until consensus was reached. Cohen's kappa and percent agreement were used to quantify agreement, following best-practice recommendations for qualitative and mixed methods research (McHugh, 2012). To check intrarater consistency, the primary researcher recoded 10% of the data after a four-week interval. Coding consistency across rounds was compared quantitatively to assess stability of interpretation (McHugh, 2012). After coding, AntConc's frequency and concordance tools were used to track recurrent linguistic and discourse-level patterns across initial, AI-assisted, and final drafts. Cross-draft comparisons were then conducted to determine the extent to which coded issues decreased, and whether indices of cohesion and argument clarity improved, particularly in relation to suggestions generated by AI tools.

Qualitative data were analyzed using Braun and Clarke's (2022) reflexive thematic analysis. The researcher first engaged in repeated reading of the interview transcripts to achieve deep familiarization and then generated initial codes capturing meaningful patterns related to learners' interpretation and uptake of AI feedback. Codes were subsequently refined, grouped into candidate themes, and iteratively reviewed for internal coherence and distinctiveness. MAXQDA 2022 was used for data organization, systematic retrieval, and audit-trail maintenance. To enhance the trustworthiness of qualitative analysis, a second coder independently reviewed a purposeful subset of transcripts. Divergences were discussed and resolved collaboratively, with recordable justification for final thematic assignments; this approach aligns with methodological guidance emphasizing coder verification to support qualitative rigor (Nowell et al., 2017).

3.5. Validity, Reliability, and Rigor

Multiple strategies were implemented to enhance methodological rigor across both quantitative and qualitative components. The error-coding scheme was defined explicitly with clear distinctions between cohesion and argument clarity to ensure construct clarity and alignment with theoretical definitions in L2 writing research. Interrater reliability checks, including calibration sessions and formal agreement indices (e.g., Cohen's kappa), and intrarater consistency checks contributed to the robustness of the corpus annotations (McHugh, 2012).

The credibility of interpretations was reinforced through triangulation of corpus findings, cross-draft comparisons, and qualitative interview insights (Braun & Clarke, 2022; Creswell & Creswell, 2023). In addition to coder verification in thematic analysis, prolonged engagement with the data and iterative code refinement were employed to promote depth of understanding and reduce interpretive bias (Nowell et al., 2017).

Construct validity was supported by anchoring analytic categories in well-established frameworks in L2 writing research (e.g., categorization of linguistic accuracy, cohesion, and argumentative structure; Ferris, 2011; Hyland, 2019), thereby ensuring that the operational definitions used in coding reflected meaningful constructs rather than ad hoc

interpretations. Ethical principles—including informed consent, confidentiality protection, and explicit guidelines regarding responsible AI use—were upheld consistently.

3.6. Procedure

The study was carried out over twelve weeks during a regular academic semester. At the beginning of the term, the instructor introduced the course objectives and explained the instructional expectations for argumentative writing, after which all students provided informed consent to allow their written work to be analyzed for research purposes. During the first six weeks, instruction focused on essential argumentative writing skills expected at the undergraduate level, including presenting a clear stance on an issue, supporting claims with relevant reasons and examples, organizing paragraphs logically, and using cohesive expressions to link ideas effectively.

In Week 6, students wrote the first major argumentative essay. They produced an initial draft without the use of external tools, revised the draft using AI assistance if they wished, and submitted a final version for course evaluation. The same process was repeated in Week 11 for the second essay, allowing comparison of writing development across two instructional points. All drafts were collected systematically and compiled into a learner corpus to enable cross-draft and cross-essay analysis.

During the final weeks of the semester, semi-structured interviews were conducted with the 15 selected participants. Interviews were scheduled across several days to accommodate students' availability and to prevent interference with final course activities. Each interview explored students' experiences with AI-mediated revision, the aspects of feedback they engaged with most, and their perceptions of how AI tools influenced their writing development. All interviews were audio-recorded, transcribed verbatim, and anonymized for analysis after the semester had concluded.

4. Results

To provide a clear and coherent response to the three research questions, the findings are presented in two complementary parts. The quantitative results appear first and trace how linguistic and discourse-level difficulties evolved across the two rounds of writing, with particular attention to the influence of AI-assisted revision. The qualitative results then deepen these findings by illuminating learners' revision strategies, their metacognitive engagement with AI feedback, and their perceptions of the pedagogical and ethical dimensions of AI-supported writing. Together, these strands offer a comprehensive account of both measurable textual changes and learners' interpretive experiences.

4.1. Quantitative Results

4.1.1. Error Frequencies and Percentages in Round 1 Essays

To address RQ1, the first stage of analysis identified the most frequent linguistic and discourse-level challenges present in the Round 1 initial drafts. Table 1 summarizes these error types and shows how their frequency changed after students revised their drafts using AI tools, thereby also speaking to RQ2, which concerns the effect of AI engagement on error reduction.

Table 1. *Error Frequencies and Percentages in Round 1 Essays (N = 38; 76 drafts)*

Error Type	Initial Drafts	%	Revised Drafts	%	Reduction
Grammar	214	38.9%	126	28.3%	41% decrease
Lexical choice	142	25.8%	101	22.7%	29% decrease
Cohesion	118	21.3%	94	21.1%	20% decrease
Argument clarity	76	13.8%	69	15.0%	9% decrease

The error distribution reveals that grammar and lexical choice were the most persistent challenges in the Round 1 initial drafts. Following AI-assisted revision, all error categories showed measurable declines, with the most substantial gains occurring in grammar accuracy. Learners' reliance on AI tools for sentence-level refinement is evident here. In contrast, argument clarity saw only marginal improvement, suggesting that higher-order rhetorical issues were less effectively supported by the AI feedback typically accessed by students. Consistent with these patterns, paired-samples t-tests confirmed significant reductions in grammar, lexical choice, and cohesion errors, while the change in argument clarity was not statistically significant (see Table 4 above).

4.1.2. Error Frequencies and Percentages in Round 2 Essays

To evaluate developmental change across the semester (RQ2), Table 2 presents error patterns from the Round 2 drafts. The comparison with Round 1 allows assessment of whether repeated engagement with AI feedback contributed to greater control over linguistic and discourse features.

Table 2. *Error Frequencies and Percentages in Round 2 Essays (N = 38; 76 drafts)*

Error Type	Initial Drafts	%	Revised Drafts	%	Reduction
Grammar	162	36.1%	82	23.5%	49% decrease
Lexical choice	113	25.2%	74	21.2%	35% decrease
Cohesion	97	21.7%	72	20.5%	26% decrease
Argument clarity	74	16.9%	60	17.2%	19% decrease

Compared with Round 1, reductions in error frequency during Round 2 were more pronounced, especially in grammar and lexical choice. Importantly, the initial drafts of Round 2 also contained fewer errors, reflecting some internalization of linguistic patterns reinforced during earlier AI-mediated revisions. Nevertheless, improvements remained concentrated at the surface level. Argument clarity, although slightly improved, continued to pose difficulties, highlighting the limited reach of AI tools in addressing higher-order reasoning and discourse organization. Inferential tests showed that all Round 2 error-type reductions reached statistical significance, although the effect size for argument clarity remained comparatively small (Table 4).

4.1.3. Development Across Rounds

Table 3 isolates learners' independent writing performance by comparing only the initial drafts from both rounds. This comparison clarifies whether gains observed across the semester reflect genuine development rather than only successful AI-mediated revision.

Table 3. *Development Across Round 1 and Round 2 Essays 1 and 2 (Initial Draft Comparison)*

Error Type	Round 1Essays Initial Drafts	Round 2 Essays Initial Drafts	Improvement
Grammar	214	162	24%
Lexical choice	142	113	20%
Cohesion	118	97	18%
Argument clarity	76	74	3%

The comparison shows meaningful improvements in grammar, lexical choice, and cohesion even before AI intervention, suggesting that repeated cycles of drafting, instruction, and AI-supported revision contributed to the internalization of certain linguistic skills. However, the near-stagnant improvement in argument clarity (3%) demonstrates that learners continued to struggle with constructing logically developed arguments. These findings reinforce the observation that AI tools, especially those emphasizing surface-level correction, do not sufficiently promote growth in discourse-level competencies without complementary pedagogical support. As shown in Table 4, paired-samples comparisons confirmed that the longitudinal improvements in grammar, lexical choice, and cohesion reached statistical significance, whereas the very small change in argument clarity did not.

Table 4. *Paired-Samples t-Test Results for Error Reductions Across Rounds*

Error Type	Comparison	$t(37)$	p	Cohen's d
Grammar	Round 1 Initial → Revised	6.84	< .001	1.11
Lexical choice	Round 1 Initial → Revised	5.02	< .001	0.81
Cohesion	Round 1 Initial → Revised	3.21	.003	0.52
Argument clarity	Round 1 Initial → Revised	1.12	.27	0.18
Grammar	Round 2 Initial → Revised	8.15	< .001	1.32
Lexical choice	Round 2 Initial → Revised	5.67	< .001	0.92
Cohesion	Round 2 Initial → Revised	3.74	.001	0.61
Argument clarity	Round 2 Initial → Revised	2.14	.039	0.35
Grammar	Round 1 Initial → Round 2 Initial	4.92	< .001	0.80

Error Type	Comparison	*t*(37)	*p*	Cohen's *d*
Lexical choice	Round 1 Initial → Round 2 Initial	4.01	< .001	0.65
Cohesion	Round 1 Initial → Round 2 Initial	3.58	.001	0.58
Argument clarity	Round 1 Initial → Round 2 Initial	0.89	.38	0.14

As evident in Table 4, the inferential statistics reinforce the patterns visible in the descriptive summaries. Grammar, lexical choice, and cohesion all showed clear and statistically reliable reductions after revision in both rounds, with effect sizes reflecting moderate to substantial improvement (p-values reported as < .001 or exact values where appropriate). These results align with the relatively large percentage drops reported in the earlier tables. In contrast, argument clarity changed very little in Round 1 and only modestly in Round 2, which explains why the inferential statistics provide limited evidence for improvement in this area. When comparing the initial drafts across rounds, a similar pattern emerges: students made measurable progress in managing sentence-level concerns, but their ability to articulate a clearer argumentative line did not shift significantly over the course of the semester.

4.2. Qualitative Results

To explore RQ3, thematic analysis was conducted to understand how AI feedback shaped learners' revision behaviors, their metacognitive engagement, and their perceptions of AI's pedagogical and ethical roles. Table 5 presents the main themes, sub-themes, and illustrative participant excerpts.

Table 5. *Main Themes, Sub-Themes, and Sample Narratives*

Main Theme	Sub-Themes	Sample Narrative
AI as Surface-Level Linguistic Support	Grammar correction	"It helps me catch mistakes I didn't realize I made, like wrong verb tenses." (P3)
	Vocabulary refinement	"Sometimes it suggests better academic words, and I use them to sound more formal." (P8)
	Sentence smoothing	"When my sentences feel awkward, it rewrites them in a smoother way." (P11)
	Confidence building	"I feel safer submitting my essay after checking it with AI." (P5)
Selective and Critical Engagement With AI	Rejecting inaccurate suggestions	"Some suggestions change my meaning, so I don't accept everything it gives." (P9)
	Evaluating suggestions against intended meaning	"I compare what it suggests with what I want to say before deciding." (P2)
	Learning personal error patterns	"It shows things I repeat, so I try not to do them next time." (P12)
	Detecting unnatural tone	"Sometimes it sounds too formal or robotic, so I fix the tone myself." (P1)
Ambivalence About Dependence and Authenticity	Fear of overreliance	"I'm worried I might depend on it too much and stop learning." (P4)
	Authorship concerns	"If I let AI change too much, I don't know if it's still my writing." (P7)
	Ethical uncertainty	"I'm not sure how much use is allowed by teachers." (P10)
	Concern about shallow learning	"It improves my writing, but that doesn't mean my skills are really improving." (P6)

P=Participant

4.2.1. Theme 1: AI as Surface-Level Linguistic Support

Participants consistently described AI as most effective for grammar, vocabulary, and sentence fluency. These patterns suggest that learners primarily view AI as a corrective device rather than a source of conceptual or rhetorical guidance. While this support increased writing confidence, it also indicates that AI is functioning as a *performance enhancer* rather than a *learning catalyst*. The heavy reliance on surface-level corrections mirrors the quantitative finding that micro-linguistic areas experienced the greatest improvement.

4.2.2. Theme 2: Selective and Critical Engagement With AI

Students demonstrated a growing ability to appraise AI feedback critically, particularly as the semester progressed. This selectivity indicates a developing sense of linguistic ownership and an emerging metacognitive orientation toward revision. Such engagement is consistent with the theoretical framework emphasizing noticing: learners became more aware of which suggestions were beneficial and which threatened accuracy or intended meaning. This theme provides qualitative support for quantitative trends showing improved initial drafts in Essay 2.

4.2.3. Theme 3: Ambivalence About Dependence and Authenticity

Despite appreciating AI's utility, learners expressed concerns about overreliance, authorship, and ethical boundaries. This ambivalence aligns with broader concerns in the literature about transparency and academic integrity. Importantly, participants worried that AI-mediated improvements might not reflect genuine skill development—a concern echoed in the limited advancement in argument clarity observed quantitatively. Their reflections reinforce the need for pedagogical frameworks that clarify appropriate AI use while promoting learner agency.

5. Discussion

The current study examined how undergraduate EFL learners engaged with AI-generated feedback while producing and revising argumentative essays over a semester, integrating corpus-based measures of error frequency with interview-derived insights into revision behavior. Anchored in a task-mediated cognitive model of L2 writing (Kormos, 2023), the noticing construct (ElEbyary, 2024), and a sociocultural view of tool mediation (Gustilo et al., 2024), the findings illuminate how AI tools shape both the micro-level mechanics of revision and learners' emergent evaluative practices.

A striking outcome is the clear and repeated reduction in micro-linguistic errors—grammar and lexical choice—after AI-supported revision. Statistical analyses indicated that these reductions were reliable and substantial, showing moderate to large effect sizes that provide strong evidence that learners consistently benefited at the sentence level. This micro-level improvement is consistent with recent empirical work showing that AI tools and modern AWCF systems are particularly effective at flagging and remediating sentence-level issues (ElEbyary et al., 2024; Poláková & Ivenz, 2024; Yan & Zhang, 2024). Interpreted through Kormos's cognitive lens, the immediacy of AI feedback likely reduces the processing burden during revision cycles, making form-level corrections cognitively cheaper and therefore more frequent. From the noticing perspective, repeated, salient cues to the same error types appear to have heightened learners' sensitivity to those forms: learners described becoming aware of and attending to recurrent problems (e.g., tense or collocation errors), which in turn contributed to cleaner initial drafts in the second round.

Yet the quantitative and qualitative data converge on an important asymmetry: macro-rhetorical competence—especially argument clarity—showed only marginal gains. By contrast, improvements in argument-level writing were minimal and often failed to reach conventional thresholds of statistical reliability, highlighting that higher-order discourse skills were less influenced by AI-assisted revision alone. While surface accuracy improved, learners rarely invoked AI feedback as a prompt to reorganize argument structure or to deepen claim-evidence reasoning. This result mirrors patterns observed in other recent studies and reviews: AI and AWCF often increase fluency and correctness but do not reliably foster discourse-level revision unless pedagogically scaffolded (Wan et al., 2024; Teng, 2024). The CoCo Matrix perspective is instructive here: many AI tools operate in regions of the matrix that optimize low-entropy, local transformations (Wan et al., 2024), which explains why they efficiently smooth phrasing yet often leave the argumentative arc unchanged.

The cross-round comparison in this study—where Round 2 initial drafts already showed fewer micro-errors—suggests partial internalization of correction patterns. Longitudinal analyses confirmed that learners maintained significant gains in grammar, lexical choice, and cohesion across the semester, whereas argument clarity remained largely unchanged, indicating that surface-level improvements can be internalized more readily than discourse-level skills. This trajectory aligns with results from longitudinal interventions that combine instruction with iterative AI use (Tran, 2025) and with modeling work indicating that self-regulation determines how learners appropriate AI-mediated practices (Jin et al., 2024). In other words, when learners actively process and adapt AI suggestions, repeated exposure can shift some corrective strategies from externally prompted to internally applied. Nevertheless, the persistence of weak gains in

argument clarity underscores a boundary condition: internalization of form does not automatically extend to internalization of rhetorical problem-solving. That boundary is apparent in both the corpus traces and the interview narratives.

Qualitative data add nuance to these patterns by showing that students did not uniformly accept AI output; rather, many developed selective engagement strategies. Participants reported rejecting suggestions that distorted meaning, sounded excessively formal, or misaligned with their rhetorical intent—an emergent evaluative stance that reflects growing critical digital literacy. This selective behavior resonates with ElEbyary's (2024) emphasis on noticing and with Gustilo et al.'s (2024) sociocultural account: AI tools function as mediating artefacts only insofar as learners and classroom norms shape how suggestions are interpreted and used. Moreover, recent mixed-method case studies of L2 writer engagement with ChatGPT show similar tripartite engagement (behavioral, cognitive, affective) and underscore the role of digital competence in shaping how learners process AWCF (Yan & Zhang, 2024).

At the same time, learners' comments that AI suggestions sometimes felt "robotic" or risked changing intended meaning corroborate concerns raised by several classroom studies: automated systems may inadvertently prompt surface edits at the expense of rhetorical nuance (Poláková & Ivenz, 2024; Teng, 2024). The CoCo Matrix and related taxonomies (Wan et al., 2024) help explain why: many AI affordances are optimized for local entropy reduction (editing for fluency and grammar) rather than for guiding complex argumentative planning, which requires different kinds of information gain and interactive scaffolding.

Finally, the study's mixed-methods approach reveals an important transactional dynamic: AI feedback and learner cognition interact. When learners used AI suggestions as prompts for reflection (not simple replacement), they demonstrated richer metacognitive processing and better selective uptake—patterns that echo findings in recent AWCF and activity-theory studies (Yan & Zhang, 2024). This interactional dynamic suggests that AI's pedagogical value is conditional: it enables more efficient correction and supports noticing, but its contribution to higher-order development depends on learners' interpretive work and the instructional practices that frame its use.

In sum, the evidence points to a nuanced conclusion. AI tools reliably enhance micro-linguistic accuracy and can foster emergent evaluative skills; however, they do not, by themselves, drive deeper advances in argumentative clarity. Overall, the inferential statistics provide evidence that AI reliably improves sentence-level accuracy but has a limited impact on discourse-level rhetorical skills, reinforcing the boundary between surface-level gains and higher-order argumentative development. Framed by cognitive, noticing, and sociocultural theories, the study shows that AI is a potent mediating artefact for surface accuracy and for prompting selective reflection—but it remains limited as a direct driver of discourse-level rhetorical learning in the absence of complementary human mediation.

6. Conclusion

This study explored how undergraduate EFL learners engaged with AI-generated feedback over a semester of argumentative writing, offering a comprehensive view of how technological support interacts with developing linguistic and rhetorical competence. Quantitative analyses revealed substantial and statistically reliable reductions in grammatical, lexical, and cohesion-related errors across drafts and writing rounds, with only limited improvements observed in argument-level writing, and qualitative accounts showed increased learner confidence and heightened awareness of recurring linguistic issues. These convergent findings reinforce the view that AI tools function as effective scaffolds for surface-level accuracy, providing immediate and individualized input that supports noticing and facilitates more precise local revisions.

However, the minimal improvement observed in argument clarity signals that AI assistance alone does not readily translate into gains in higher-order reasoning or discourse-level organization. Learners' reflections confirmed this pattern, noting that AI suggestions typically offered limited support for refining claims, structuring arguments, or developing coherent lines of reasoning. This gap highlights the enduring importance of instruction in cultivating argumentative logic, genre knowledge, and critical thinking—dimensions of writing that remain largely outside the operational strengths of current AI feedback systems.

The findings also underscore the pedagogical need to integrate AI tools intentionally rather than prescriptively. Instructors can leverage AI's demonstrated efficacy in micro-linguistic refinement while dedicating targeted instructional

time to rhetorical development and strategic revision. Specifically, educators might design revision activities that pair AI-generated feedback with guided prompts asking learners to analyze argument structure, evidence use, and overall coherence. This approach encourages learners to actively evaluate and adapt AI suggestions rather than accepting them passively. Embedding short reflection exercises, such as journals or discussion boards where learners justify their decisions regarding AI feedback, can further strengthen metacognitive awareness and support autonomous learning. Moreover, learners' increasing selectivity in accepting or rejecting AI suggestions suggests that explicit guidance in critical evaluation can further enhance the developmental value of AI-mediated writing. Providing annotated exemplar essays that illustrate both surface-level corrections and greater rhetorical improvements can help learners differentiate between local linguistic adjustments and higher-order argumentative strategies, reinforcing meaningful skill development. Clear institutional guidelines regarding ethical and responsible use are likewise essential, given students' concerns about overreliance, authorship, and the boundaries of acceptable support.

Several limitations shape the interpretation of these outcomes. The study involved a relatively small cohort from a single private university, constraining generalizability. Its focus on argumentative writing narrows applicability to other genres, and the use of a single AI platform does not represent the full diversity of emerging tools. These boundaries nevertheless point toward productive avenues for future research. Longitudinal work could examine whether the improvements observed here endure beyond the instructional context, while comparative studies across genres, proficiency levels, and AI systems could clarify how different learners interact with different forms of automated feedback. Additionally, research integrating AI-generated feedback with explicit instruction in argumentation would illuminate how technological and pedagogical scaffolds jointly support writing development.

Overall, the findings suggest that AI tools offer meaningful but partial support for EFL writing. Their core value lies in complementing—rather than replacing—the instruction and guided practice that foster deep, transferable competence in academic argumentation. By translating these insights into structured classroom activities, guided scaffolding, and reflective learner practices, educators can maximize the pedagogical benefits of AI tools while reducing risks of superficial engagement or dependency.

Acknowledgments

The author gratefully acknowledges the students who participated in this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Anthony, L. (2023). *AntConc* (Version 4.0) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information, 16*(4), Article 296. https://doi.org/10.3390/info16040296

Braun, V., & Clarke, V. (2022). Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health, 24*(2), 1–6. https://doi.org/10.1080/26895269.2022.2129597

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE. https://doi.org/10.5539/ijel.v13n3p1

ElEbyary, K., Shabara, R., & Boraie, D. (2024). The differential role of AI-operated WCF in L2 students' noticing of errors and its impact on writing scores. *Language Testing in Asia, 14*(1), 1-24. https://doi.org/10.1186/s40468-024-00312-1

Ferris, D. R. (2011). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press. http://103.203.175.90:81/fdScript/RootOfEBooks/E%20Book%20collection%20%202024%20%20G/ENGLISH/ferris_dana_r_treatment_of_error_in_second_language_student.pdf

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Gustilo, L., Ong, E., & Lapinid, M. R. (2024). Algorithmically-driven writing and academic integrity: Exploring educators' practices, perceptions, and policies in the AI era. *International Journal for Educational Integrity, 20*, Article 3. https://doi.org/10.1007/s40979-024-00153-8

Hyland, K. (2003). *Second language writing*. Cambridge University Press. https://catdir.loc.gov/catdir/samples/cam041/2003041957.pdf

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press. https://www.amazon.com/Second-Language-Writing-Ken-Hyland/dp/1108456413

Jaramillo, J. J., Chiappe, A., & Sáez-Delgado, F. (2025). From struggle to mastery: AI-powered writing skills in ESL learners. *Applied Sciences*, *15*(14), Article 8079. https://doi.org/10.339/app15148079

Jin, Y., Lin, T. J., & Lai, C. (2024). Modeling AI-assisted writing: How self-regulated learning influences writing outcomes. *Computers in Human Behavior, 165*, 108538. https://doi.org/10.1016/j.chb.2024.108538

Kormos, J. (2023). The role of cognitive factors in second language writing and writing to learn a second language. *Studies in Second Language Acquisition, 45*(3), 622–646. https://doi.org/10.1017/S0272263122000481

Liu, J., Sihes, A. J. B., & Lu, Y. (2025). How do generative artificial intelligence (AI) tools and large language models (LLMs) influence language learners' critical thinking in EFL education? A systematic review. *Smart Learning Environments, 12*(1), 48. https://doi.org/10.1186/s40561-025-00406-0

Lötscher, F., Trüb, R., Lohmann, J., Möller, J., Jansen, T., & Keller, S. D. (2025). Development of grammatical and lexical skills in argumentative EFL writing at upper secondary level in Germany and Switzerland. *Frontiers in Education, 10*, 1605658. https://doi.org/10.3389/feduc.2025.1605658

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1–13. https://doi.org/10.1177/1609406917733847

Parra Núñez, R., Castrillo de Larreta-Azelain, M. D. (2025). The impact of using ChatGPT for feedback in EFL writing: a systematic review. *Revista Internacional De Lenguas Extranjeras /International Journal of Foreign Languages, 23*, 95-119. https://doi.org/10.17345/rile23.4221

Polakova, P., & Ivenz, T. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of CHATGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning,* 1–28. https://doi.org/10.1080/09588221.2025.2454541

Shousha, A., & Oraby, A. (2025). EFL teachers' perceptions of ChatGPT's role in teaching English writing. *English Language Teaching, 18*(7), 82–103. https://doi.org/10.5539/elt.v18n7p82

Teng, M. F. (2024). ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*(4), 100270. https://doi.org/10.1016/j.caeai.2024.100270

Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences, 15*(2), Article 232. https://doi.org/10.3390/educsci15020232

Wan, R., Gebreegziabher, S., Li, T. J., & Badillo-Urquiola, K. (2024). CoCo matrix: Taxonomy of cognitive contributions in co-writing with intelligent agents. In *Proceedings of the 16th Creativity & Cognition Conference* (C&C'24'24). Chicago, IL, USA.

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities & Social Sciences Communications*, *11*(1), 1-14. https://doi.org/10.1057/s41599-024-03543-y

Zare, J., Al-Issa, A., & Madiseh, F. R. (2025). Interacting with ChatGPT in essay writing: A study of L2 learners' task motivation. *ReCALL*, *37*(3), 1–18. https://doi.org/10.1017/S0958344025000035

Zhang, Y. (2024). Incorporating ChatGPT as an automated written corrective feedback tool into L2 writing class. *Journal of Language Teaching, 4*(4), 22–34. https://doi.org/10.54475/jlt.2024.024

Zou, B., Wang, C., He, H., Li, C., Purwanto, E., & Wang, P. (2025). Enhancing EFL writing with visualised GenAI feedback: A cognitive-affective theory of learning perspective on revision quality, emotional response, and human–computer interaction. *Learning and Motivation, 91*, Article 102158. https://doi.org/10.1016/j.lmot.2025.102158