

ORIGINAL ARTICLE OPEN ACCESS

AI Chatbot as a Revision Aid in Second Language Writing: From Error Correction to Lexical Sophistication

Yuah V. Chon¹  | Dongkwang Shin²

¹Department of English Education, Hanyang University, Seoul, South Korea | ²Department of English Education, Gwangju National University of Education, Gwangju, South Korea

Correspondence: Yuah V. Chon (vylee52@hanyang.ac.kr)

Received: 20 November 2023 | **Revised:** 3 March 2026 | **Accepted:** 29 April 2026

Keywords: artificial intelligence | ChatGPT | corrective feedback | error analysis | L2 writing | linguistic complexity

ABSTRACT

Background: The use of recent artificial intelligence (AI) chatbots, such as chat generative pretrained transformer (ChatGPT), in second language (L2) writing may face criticism for potentially promoting plagiarism and raising ethical concerns. However, such tools can effectively provide guided suggestions for improving outlines, content, and organisation. Additionally, EditGPT, an extension that tracks changes made by ChatGPT, provides immediate, direct corrective feedback, thereby enhancing L2 learners' control over their writing process.

Objectives: The aim of this study was to compare the quality of L2 learners' writing (LW), ChatGPT-proofread writing (PW), and learners' ChatGPT-supported revisions (RW) by analysing human raters' evaluations, assessing linguistic complexity, and identifying error types.

Methods: A total of 40 university students majoring in English education, all of whom were English as a foreign language (EFL) learners and pre-service teachers, participated in this study. Based on significant differences in their TOEIC scores, they were divided into two groups. The secondary-school pre-service teachers were classified as skilled learners, while the elementary-school pre-service teachers were identified as less skilled learners. Participants completed a guided writing task supported by ChatGPT for idea generation and essay revision. They interacted directly with ChatGPT using standard prompts for brainstorming and outlining and received automated feedback on editing and proofreading through the EditGPT extension.

Results: When LW and RW were compared, human raters found significant improvements in *language use* but not in *content* and *organisation*. The revisions demonstrated greater linguistic complexity, including increased use of academic words, improved lexical sophistication, and more varied sentence structures. Regarding language errors, 86.17% of the errors in LW were successfully rectified. However, a closer breakdown revealed that certain error types—particularly word choice (38.9% remaining) and sentence structure (35.4% remaining)—remained relatively unresolved. Addressing these lingering errors in both PW and RW often required judgements of contextual appropriateness, especially when attempting to preserve the learners' intended meanings and nuanced expressions. As a result, the overall error resolution rate dropped to 77.02% in RW.

Conclusions: The structured AI-mediated revision environment was associated with short-term improvements in the linguistic quality of L2 writing within a controlled session. These effects reflect product-level changes and cannot be attributed to individual writing process components. While AI-supported revision reduced many surface-level errors, limitations remained for context-sensitive language use. Further longitudinal research is needed to determine whether these changes support sustained writing development.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *Journal of Computer Assisted Learning* published by John Wiley & Sons Ltd.

Lay Description

What is Already Known About This Topic

- Automated writing evaluation (AWE) tools help improve grammar and mechanics but have limitations in handling higher-order writing skills.
- ChatGPT can offer grammar correction, vocabulary suggestions, and structural improvements in L2 writing.
- Concerns exist around ethical implications, academic integrity, and over-reliance on AI-generated content in educational settings.

What This Paper Adds

- Empirical evidence that ChatGPT, especially with EditGPT, significantly enhances linguistic accuracy and complexity in L2 writing.
- Demonstrates that ChatGPT-supported revisions improve vocabulary use and syntactic sophistication, particularly among less skilled learners.
- Reveals persistent challenges with context-sensitive language issues like word choice and sentence structure.
- Emphasises the importance of learner agency in evaluating and accepting AI-generated feedback, rather than relying on passive adoption.

Implications for Practice/or Policy

- Teachers should provide guidance to help learners critically assess AI feedback to avoid misuse or over-reliance.
- AI tools should be used as a supplement to—not a replacement for—human instruction, especially in developing higher-order writing skills.
- Instructional designs should promote metacognitive strategies and reflective practices when integrating AI into writing pedagogy.
- Institutional policies should consider structured AI use in classrooms to maximise benefits while mitigating ethical and pedagogical risks.

enhancing writing style (Ali et al. 2023; Baskara 2023; Kostka and Toncelli 2023; Ray 2023).

The integration of AI in education evokes a mix of excitement and apprehension. While some people embrace its potential benefits, others express reservations, particularly regarding plagiarism, considering technological developments and changes in the pedagogical environment (Farrokhnia et al. 2024; Rudolph et al. 2023; Yan 2023). Additional skepticism surrounds AI's reliability in surpassing human raters (Steiss et al. 2024) and uncertainty regarding whether ChatGPT, as an automated writing evaluation (AWE) tool, can accurately perform word- or sentence-level revisions without significantly altering the content or diverging from the writer's intended meaning (Dikli and Bleyle 2014; Li et al. 2015; Warschauer and Grimes 2008). Nevertheless, it is crucial to recognise and embrace the rapid advancement of technology, as it increasingly influences various aspects of education—especially the teaching and learning of L2 writing. Instead of fearing or rejecting ChatGPT, which is poised to become one of the most innovative types of automated-writing feedback technologies to date, adopting a proactive approach to harnessing this educational resource can help L2 learners improve attitudes towards writing and writing performance (Wilson et al. 2023).

Recent research on ChatGPT and other generative AI tools for L2 writing has expanded rapidly, with studies examining learners' perceptions, writing outcomes, and emerging pedagogical and ethical concerns (Barrot 2026; Crosthwaite and Sun 2025; Ding et al. 2026). At the same time, research on AWE has long been accompanied by scepticism regarding the accuracy, pedagogical value, and instructional implications of automated written corrective feedback (AWCF) among L2 writers and teachers (Link et al. 2022). Despite increasing interest, empirical evidence remains limited regarding how generative AI feedback is actually taken up by learners during revision and how such AI-mediated changes are reflected in the linguistic quality of texts across writing stages. ChatGPT may help learners identify most surface-level errors successfully, but learners may still struggle to detect errors related to deep structure and pragmatics. Conversely, human teachers may spot most of these issues (Al-Garaady and Mahyoob 2023). Although these findings suggest ChatGPT can be a valuable tool in identifying surface-level errors, they also highlight the continued importance of human expertise in evaluating complex aspects of writing. This tension highlights the need for human evaluation of learners' original and AI-supported writing, rather than assuming AI-assisted correction ensures higher writing quality (Steiss et al. 2024).

An unresolved issue concerns how AI tools modify linguistic complexity while preserving the writer's intended meaning, a concern that must be interpreted with caution, as greater linguistic complexity does not necessarily reflect higher writing quality (Zhou et al. 2023). While lexical and syntactic measures capture dimensions of writing beyond surface-level accuracy, ChatGPT's tendency to propose alternative phrasings and restructure sentences may influence complexity differently from traditional teacher feedback. Examining linguistic complexity before and after ChatGPT-supported revision can therefore

1 | Introduction

The introduction of tools and applications based on artificial intelligence (AI) has brought about paradigm shifts in technology-enhanced second language (L2) writing (Steiss et al. 2024; Yan 2023). One such AI tool is chat generative pretrained transformer (ChatGPT) (<https://chat.openai.com/>). Developed by OpenAI in November 2022, ChatGPT is a large-language model (LLM) that autonomously learns from data and produces sophisticated and intelligent writing on a wide range of topics (van Dis et al. 2023). This advanced chatbot, powered by generative AI, marks a new era in which humans and AI coexist in L2 learning and teaching (Mizumoto and Eguchi 2023) and promises significant breakthroughs in language learning (Han et al. 2023). For L2 writing, ChatGPT facilitates deeper learner engagement by supporting essay composition, clarifying grammar points, and

clarify whether AI-assisted changes enrich, simplify, or artificially inflate learners' language use.

A further gap motivating the present study is that existing ChatGPT research has yet to provide a systematic account of how the tool corrects different types of errors when asked to "proofread" learners' drafts. Questions remain regarding how corrections are configured across error types and why some errors are addressed while others are left unresolved. Moreover, learner engagement with AI-generated corrections across proficiency levels remains underexplored, despite growing evidence that lower-proficiency learners may benefit more from AI-supported revision (Tsai et al. 2024).

To address this gap, the present study examined how ChatGPT, supported by EditGPT, can assist learners in revising their writing. EditGPT is an extension that allows users to track changes and receive direct written corrective feedback, providing explicit feedback on linguistic aspects (Ellis 2010) identified by the AI tool. University-level learners of English as a foreign language (EFL) were asked to utilise ChatGPT—prior to incorporating EditGPT—for generating writing ideas, structuring essays, and later to obtain corrective feedback through EditGPT during structured writing activities. Specifically, the study compared three phases of writing: original learner writing (LW), ChatGPT-proofread writing (PW), and learner revisions following engagement with EditGPT feedback (RW). In this study, PW refers to texts proofread using EditGPT integrated into ChatGPT, whereas RW refers to texts produced after learners evaluated and responded to AI-generated feedback.

Because learners reviewed and selectively accepted individual suggestions, RW represents the integration of AI feedback through learner evaluation rather than fully independent writing. Recent research highlights the importance of learner engagement in contexts involving automated or AI-generated feedback, showing that revision reflects learners' selective evaluation and use of suggestions rather than mere incorporation (Ranalli 2018, 2021; Zhang and Hyland 2018). This perspective aligns with the concept of uptake in SLA research (Lyster and Ranta 1997), which conceptualises revision as a learner-mediated response to feedback. Accordingly, in the present study, RW is treated as an AI-mediated learner uptake artefact rather than as simple AI-generated output. These phase-based comparisons extend research on AI-assisted L2 writing by tracing learners' uptake of AI-generated corrective feedback across drafting, proofreading, and revision stages.

Despite the rapid adoption of ChatGPT in educational settings, few empirical studies have employed EditGPT as a methodological tool to trace learners' uptake of AI-provided written corrective feedback across drafting, proofreading, and revision stages. In particular, little is known about how AI-mediated feedback influences human raters' evaluations of learner writing, affects the lexical and syntactic complexity of texts across writing phases, or alters the frequency and distribution of language errors across proficiency levels. By addressing these gaps, the present study provides a more nuanced account of both the effectiveness and limitations of AI-assisted feedback in L2 writing. The following research questions guided our study:

1.1 | Research Questions

1. How are learners' original writing (LW) and ChatGPT-supported revisions (RW) evaluated by human raters?
2. How do the phases of writing (LW, PW, and RW) affect the linguistic complexity (lexical and syntactic indices) of L2 texts?
3. What is the frequency of language errors across the three phases of writing (LW, PW, and RW)?

2 | Background

2.1 | Theoretical Foundation: Role of Automated Writing Evaluation

AWE provides an important theoretical foundation for understanding how AI-based tools such as ChatGPT may influence L2 writing. AWE involves computer-based systems designed to assess written texts by assigning scores, offering diagnostic feedback, or providing suggestions for improvement (Fu et al. 2024). Research on AWE has approached the topic from multiple perspectives, including its influence on writing practices (Li et al. 2015; Ranalli 2018) and its effects on accuracy development across multiple drafts (Bai and Hu 2017; Barrot 2021). Comparative research has explored differences between feedback from AWE systems and human instructors (Dikli and Bleyl 2014; Wang et al. 2013), as well as how instructors use these tools (Li et al. 2015) and how students perceive their effectiveness (Bai and Hu 2017). Comprehensive studies (Fu et al. 2024; Stevenson and Phakiti 2014) have further evaluated AWE's overall contributions to L2 writing. Studies focusing on peer feedback have also shown that AWE can be beneficial in review activities, though it is generally less effective in enhancing text cohesion compared to traditional peer feedback (Chen and Cui 2022; Li and Li 2018). Although the literature on ChatGPT and generative AI in L2 writing is expanding, systematic evidence remains scarce regarding how AI-generated feedback influences linguistic complexity and error management. The following review highlights common themes from AWE research.

Research consistently indicates that AWE tools can enhance writing accuracy, especially in grammatical and mechanical aspects. Bai and Hu (2017) observed that Chinese university students using Pigai attained 98% mechanical accuracy by successfully revising errors related to spelling, capitalization, and punctuation. Likewise, Barrot (2021) found that English as a second language (ESL) learners using Grammarly notably reduced major grammatical errors, although improvements in minor mistakes were less pronounced, indicating a stronger emphasis on grammar than mechanics. Liao (2016) further demonstrated that integrating AWE into a process-writing approach led to sustained reductions in grammatical mistakes, such as sentence fragments and subject-verb agreement issues, among Taiwanese students. Wang et al. (2013) likewise confirmed AWE's positive impact on grammar, sentence structure, and vocabulary usage.

The impact of AWE corrective feedback on writing practices varies based on the specificity and clarity of feedback provided.

Li et al. (2015) noted that using Criterion in ESL classes markedly reduced grammar and mechanics errors and promoted extensive revision. However, the tool often struggled with complex structures and nuanced language use, and its generic feedback sometimes confused students. Criterion feedback predominantly addressed lower-order issues, offering limited guidance for coherence, content development, or argumentation. Similarly, Ranalli (2018) highlighted that Criterion's specific feedback was generally more helpful than generic advice, except in the case of preposition errors, where both feedback types were equally effective. Students also reported greater cognitive effort when interpreting generic feedback, underscoring the importance of clarity for effective revision.

Comparative studies further suggest that AWE feedback is less effective than human or peer feedback in addressing higher-order writing concerns. Dikli and Bleyle (2014) showed that instructors identified considerably more errors than AWE tools, which frequently overlooked common L2 issues such as verb tense and singular/plural agreement. Link et al. (2022) similarly found that while AWE reduced instructors' workloads by addressing lower-level concerns, teacher feedback was more effective in promoting writing complexity and fluency. Chen and Cui (2022), comparing peer feedback and AWE in continuation writing tasks, concluded that peer interaction significantly enhanced cohesion and coherence through effective use of cohesive devices. Conversely, AWE yielded minimal improvement due to vague and insufficient guidance for higher-order issues. Together, these findings suggest that while AWE is effective for improving accuracy, it requires complementary instruction for writing support.

In sum, prior research confirms AWE's strength in reducing linguistic and mechanical errors, promoting drafting processes, and delivering rapid feedback. As a supplementary resource, AWE enables instructors to prioritise higher-order aspects of writing, such as content and organisation. However, studies comparing AWE with human or no feedback report mixed results (Fu et al. 2024; Stevenson and Phakiti 2014). Although AWE systems are effective for mechanical corrections, they continue to show limitations in addressing grammar, collocations, coherence, and content development, with effectiveness varying across error types (Fu et al. 2024; Li et al. 2015).

2.2 | ChatGPT in Second Language Writing: Feedback Quality, Error Patterns, and Linguistic Complexity

With the emergence of ChatGPT, the scope of AI support for L2 writing has expanded beyond the rule-based feedback characteristic of earlier AWE systems. Unlike conventional tools that primarily detect surface-level errors, ChatGPT can generate extended text, reformulate learner language, and engage in interactive dialogue (Cao and Zhong 2023; Kostka and Toncelli 2023; Mindner et al. 2023; Mizumoto et al. 2024). These affordances distinguish ChatGPT from earlier AI-based writing tools and raise new questions about how AI-generated feedback shapes learners' revisions, linguistic complexity, and correction of language errors.

The benefits of ChatGPT as a writing tool in L2 education have been extensively documented, emphasising its potential to enhance learner engagement and linguistic skills. ChatGPT effectively supports tasks such as idea generation (Su et al. 2023), text correction (Harunasari 2023), and the provision of personalised, real-time feedback (Ali et al. 2023; Baskara 2023; Ray 2023). Harunasari (2023) demonstrated how integrating ChatGPT into the EFL writing curriculum aids students across various stages of writing, from planning and drafting to revising their texts, by overcoming writer's block and enhancing coherence. Additionally, Kohnke et al. (2023) highlighted ChatGPT's strength in providing immediate formative assessments and authentic linguistic input, contributing positively to students' confidence and motivation.

Despite these advantages, several challenges associated with ChatGPT as a writing aid have been identified. A key concern is its impact on academic integrity due to potential misuse leading to plagiarism (Barrot 2023; Farrokhnia et al. 2024; Rudolph et al. 2023; Stokel-Walker 2023). Furthermore, its reliance on pre-existing English-language training data introduces risks of cultural bias and the absence of nuanced, culturally-sensitive feedback (Kohnke et al. 2023). Additionally, ChatGPT's capacity to generate undetectable AI-written content complicates the differentiation between human-authored and AI-generated work, raising ethical issues within educational settings (Barrot 2023).

As ChatGPT increasingly assumes roles traditionally held by human feedback providers, it has become essential to examine how AI-supported writing is evaluated through human judgement. Comparative studies evaluating ChatGPT-generated feedback against human teachers and peers have yielded mixed outcomes regarding its effectiveness. Cao and Zhong (2023) observed that human teachers provided superior translation quality compared to feedback from ChatGPT. Conversely, Escalante et al. (2023) reported no significant differences in learning outcomes between groups receiving AI and human tutoring, noting the clarity and specificity of ChatGPT's feedback. Guo and Wang (2024) found that while teachers offered more nuanced, informative feedback, ChatGPT tended to be more directive, yet still beneficial for guiding student revisions. Steiss et al. (2024) further reinforced this, noting that while human raters generally outperformed ChatGPT, the AI's feedback quality closely approximated human feedback, particularly beneficial in initial writing phases. These mixed findings underscore the need to examine how learners' original writing and ChatGPT-supported revisions are evaluated by human raters, rather than assuming that AI-guided improvements translate into higher-quality writing.

Recent studies consistently highlight ChatGPT's strong performance in correcting surface-level errors and enhancing linguistic accuracy in learners' writing. Al-Garaady and Mahyoob (2023) found ChatGPT exceptionally effective in detecting and correcting surface-level linguistic errors, offering insightful suggestions for sentence structure, clarity, and vocabulary enhancement. Schmidt-Fajlik (2023) emphasised ChatGPT's detailed feedback and user-friendly explanations, which were particularly helpful for learners with lower language proficiency. Tsai et al. (2024) reported significant score gains between learners' original essays and their ChatGPT-assisted revisions, with the largest improvements in grammar

and vocabulary rather than higher-order features such as content and organisation. They cautioned, however, that AI-boosted revisions do not necessarily reflect genuine writing competence: disproportionate gains—especially among lower-proficiency learners—complicate accurate assessment of independent writing ability. Their findings highlight the need for studies that analyse ChatGPT-assisted revisions across different proficiency levels to clarify how AI support interacts with learners' underlying writing skills.

In comparison to research focusing on accuracy gains, emerging work has examined how ChatGPT-generated texts differ from human writing in terms of linguistic and syntactic complexity. Zhou et al. (2023) found that although ChatGPT-generated texts excelled in narrativity, word concreteness, and referential cohesion, they lagged behind human-authored texts in syntactic simplicity and deep cohesion. While later versions showed improvements in syntactic simplicity, human writers continued to outperform the system in producing deeply cohesive texts. Liu and Liu (2025) compared 73 GPT-generated argumentative essays with an equal number of native-speaker (NS) essays produced in response to the same prompt. Their analysis showed that ChatGPT tended to produce shorter sentences characterised by heavy coordination and dense noun phrases, whereas NS writers more frequently employed clausal subordination and richer verb-phrase structures to articulate nuanced argumentative relations. Fredrick and Craven (2025) compared 50 ChatGPT-generated essays with 50 student essays on measures of lexical diversity, syntactic complexity, and readability. ChatGPT outperformed students in lexical diversity (e.g., TTR, MTLT) and produced syntactically denser sentences, as evidenced by a higher ratio of dependent clauses per T-unit. However, its texts were less readable, showing higher Flesch–Kincaid and Gunning–Fog scores. While students produced longer T-units, this length often stemmed from developing L2 features—run-ons, weak punctuation, and heavy coordination—whereas ChatGPT's structural complexity emerged from more consistent clause embedding.

Recent research has increasingly examined the role of generative AI, particularly ChatGPT, in supporting L2 writing through automated feedback and evaluation. A scoping review by Crosthwaite and Sun (2025) indicates that although many studies report improvements in writing quality following AI-supported feedback, they call for more fine-grained, text-based research that examines how AI-supported revisions affect measurable linguistic features and writing performance across stages of the writing process. Complementing this perspective, Barrot (2026) raises validity concerns regarding AI-assisted assessment, noting that AI systems tend to align with human judgement primarily on surface-level features such as grammar and form while remaining less reliable for evaluating higher-order aspects of writing, including rhetorical effectiveness and argument development. This suggests that improvements observed in AI-supported writing may reflect form-level changes rather than broader communicative gains and underscores the continued necessity of human evaluation. In a classroom study, Ding et al. (2026) found that ChatGPT feedback led to significant improvements in mechanics, grammar, and overall writing quality but produced little change in organisation and content, whereas combined human–AI feedback resulted in broader

gains. Importantly, the authors note that many studies treat writing quality as a unified construct and highlight the need to examine changes across specific dimensions of writing. Taken together, these studies point to the need for research that evaluates AI-supported writing using multiple indicators, integrating human ratings with detailed analyses of linguistic complexity and accuracy across different phases of writing.

3 | Method

3.1 | Research Design

This study employed a quasi-experimental, within-subject research design in which each participant produced three versions of the same essay: LW, PW, and RW. This design allowed for direct within-learner comparisons of linguistic changes while minimising between-learner variability. Specifically, the design enabled an examination of (a) human raters' evaluations of LW and RW, (b) differences in linguistic complexity across LW, PW, and RW, and (c) changes in error types across the writing phases.

3.2 | Participants

A total of 40 students participated in this study and were divided into two groups. The first group ($N=20$) consisted of pre-service secondary school teachers majoring in English education at a university in Seoul, the largest metropolitan city in South Korea. They were second-year students enrolled in the course *Practice of English Academic Writing*. The second group ($N=20$) comprised pre-service elementary school teachers in their final year at a university in Gwangju, the country's sixth-largest city. These students, also majoring in English education, were taking the course *Multimedia English Education*.

Diagnostic writing tests conducted at the beginning of the semester revealed a significant difference between the two groups ($p<0.01$). To further validate this, students completed the Test of English for International Communication (TOEIC). The secondary-school pre-service teachers scored an average of 835 ($SD=56$), while the elementary-school pre-service teachers scored 725 ($SD=62$). Although both groups demonstrated functional English proficiency for academic and professional contexts, the secondary-school group showed stronger skills, particularly in reading and listening. This difference was statistically significant ($p<0.01$), supporting the classification of the secondary-school pre-service teachers as *skilled learners* and the elementary-school pre-service teachers as *less skilled learners*. TOEIC scores were used as an indicator of overall English proficiency, as they were the only standardised proficiency measure available. These scores were considered alongside students' diagnostic writing performance to approximate differences in writing ability.

The proficiency distinction was necessary because AI-assisted revisions may not equally reflect students' underlying writing competence. AI support may elevate lower-proficiency learners' performance to a degree that blurs their true writing ability. Analysing the two groups separately therefore enabled a clearer assessment of whether ChatGPT-supported revisions benefit

learners similarly across proficiency levels or whether AI assistance amplifies or masks proficiency-based differences (Tsai et al. 2024).

Both groups of students had entered university with strong academic records, as evidenced by their high scores (1–2 stanine levels) on the Korean national college-entrance exam (Suneung). However, their university training diverged in focus. The elementary-school pre-service teachers followed a broad curriculum covering all 12 core subjects required for teaching at the elementary level, which limited the time and coursework dedicated to English language development. In contrast, the secondary-school pre-service teachers pursued a specialised programme in English education, receiving more intensive training in English language proficiency, pedagogy, and subject-specific content knowledge.

Although 40 students initially participated in the study, only 39 students' essays were included in the final analyses. One student did not adhere to the task instructions provided by the researcher and submitted an essay that appeared to have been fully generated by ChatGPT. As the submission contained no identifiable learner-generated content, linguistic errors, or evidence of revision, it was excluded to preserve the validity and integrity of the analysis.

3.3 | Instruments

3.3.1 | ChatGPT and EditGPT

In this study, ChatGPT serves two primary functions: (1) providing guided writing support in the form of ideas and outlines, and (2) offering automated feedback on student essays. For generating ideas and outlines, students interact directly with ChatGPT through its standard user interface. However, to deliver automated feedback, students use the EditGPT Chrome extension.

Participants installed EditGPT from the Chrome Web Store, integrating it into the ChatGPT interface, where it initially appears as an "Editing Disabled" button above the dialogue box. Clicking this button changes the status to "Editing Enabled," thus activating the proofreading and editing functionality. EditGPT displays suggested edits in a format analogous to the "track changes" feature in Microsoft Word, clearly indicating each modification along with comprehensive editing statistics (e.g., original vs. revised word counts and percentage of revised content). Students can review each suggested change individually, choosing to accept or reject edits, which ensures active engagement with AI-generated feedback rather than passive acceptance. Figure 1 illustrates the EditGPT platform integrated with ChatGPT.

Compared to traditional grammar and proofreading tools such as *Grammarly*, *Ginger*, *LanguageTool*, *ProWritingAid*, and *After the Deadline*, which primarily focus on grammar and syntax corrections with varied levels of transparency and accuracy (Sahu et al. 2020), EditGPT uniquely provides explicit, traceable revisions through direct corrective feedback. Its use may also influence how learners engage with feedback. From a sociocognitive perspective (e.g., Ellis 2010), the tool acts as a form of mediated

assistance, scaffolding the editing process. At the same time however, it may encourage an overemphasis on surface-level accuracy rather than deeper content or structural development. Although EditGPT offers greater transparency than previous grammar checkers, it may still introduce biases in what students attend to during revision, particularly if not paired with explicit instruction or reflective practice. Consequently, the degree to which learners critically evaluate AI-generated suggestions remains an open question.

3.3.2 | Guided Writing Task

The learners engaged in a guided writing task by utilising a set of prompts provided by the researchers to interact with ChatGPT. These prompts helped learners generate possible essay topics, select an outline, and revise their drafts. Although learners were free to choose their own topics, the topics were generated through Prompt 1 and refined through Prompt 2; thus, each learner's essay topic was the one they selected from the ChatGPT-generated list of ideas. This procedure ensures that the essay topics are clearly identifiable and consistently derived from the same prompt structure. Learners were strictly instructed to use only these prompts and were not allowed to use any additional prompts or engage in other activities. This restriction ensured control over learner actions and supported the validity of the writing data.

The prompts used in the guided writing task were as follows:

Prompt 1: "I'd like to write an essay. Could you give me any ideas or outlines?"

Prompt 2: "Could you give me more details about Topic O? (O = topic chosen)."

Prompt 3: "Can you proofread the following essay? [Insert your essay]."

At each stage of the guided writing process, the learners were directed to copy and paste ChatGPT's output. After using Prompt 3, they submitted both the EditGPT-proofread version of their essays and information on the percentage of text rewritten (as indicated by EditGPT), along with the word count. Commonly selected topics included issues such as online education, environmental protection, and the role of technology in learning.

The decision to restrict learners to a fixed set of prompts was a deliberate methodological choice aimed at standardising the mode of interaction with ChatGPT. This controlled approach ensured consistency across participants' experiences, enabling the researchers to isolate the effects of specific AI support on writing outcomes. By limiting variability in how learners engaged with the tool, the study minimised potential confounding factors and facilitated a more reliable comparison between the groups.

3.4 | Procedure

All 40 participants completed the same sequence of writing tasks in a single 80-min class session. The goal was to examine

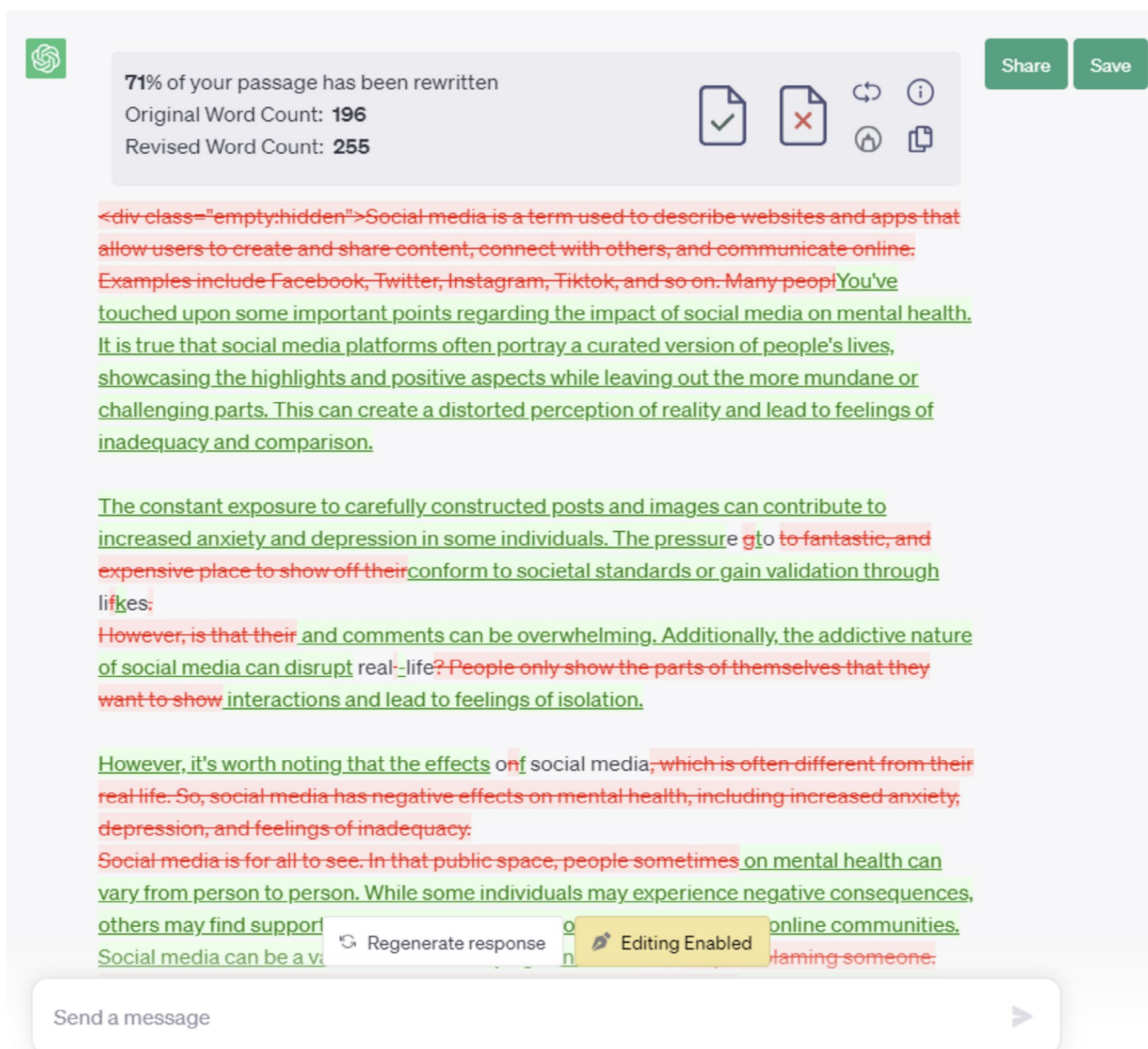


FIGURE 1 | EditGPT integrated with ChatGPT.

changes in L2 writing across three stages: LW, PW, and RW. Each learner produced all three versions of the same text, allowing for within-participant analysis of changes across writing phases.

During the first 20 min, learners generated ideas and outlined their essays using Prompts 1 and 2. In the next 60 min, they drafted an initial essay (LW), received feedback from ChatGPT via Prompt 3 (producing PW), and then critically evaluated and selectively incorporated that feedback to create a revised version (RW). All three writing phases occurred consecutively within the session.

To prevent direct adoption of AI-generated text, learners were restricted to predefined prompts and were not permitted to generate alternative drafts. During the proofreading stage, EditGPT presented feedback in a tracked-changes format

requiring learners to review and decide on each suggested edit individually. Learners also reported the percentage of text revised based on EditGPT's output. One student was later excluded from analysis due to non-compliance with task instructions, resulting in a final sample of 39. No follow-up tasks or delayed post-tests were administered, as the aim was to assess immediate linguistic changes resulting from AI-supported proofreading and revision.

3.5 | Data Analyses

3.5.1 | Assessment of Writing Products

To answer Research Question 1, LW and RW were the subject of analyses. The two raters selected for this study were native speakers of English who taught L2 writing at the university

where one of the researchers worked. Both had more than 8 years' experience in rating a large pool of essays written by learners of English. For assessing writing, we chose Jacobs et al. (1981) analytic scale, ESL Composition Profile (Weigle 2002), which rates five aspects of writing: content, organisation, vocabulary, language use, and mechanics. The raters underwent a day of training to ensure that they reached an agreement on the descriptors of the analytic scale. During this session, the raters jointly reviewed the rubric criteria for each sub-construct (content, organisation, vocabulary, language use, and mechanics) and discussed sample essays representing a range of proficiency levels. Discrepancies in initial ratings were examined in detail, and consensus was reached through discussion until both raters demonstrated consistent application of the scale across benchmark essays.

The raters assessed LW and RW, resulting in 80 essays (2 × 40 learners) for each rater to mark. The raters independently graded the essays while unaware of LW and RW. The essays were assigned scores on a scale of 100, and the scores were calculated for inter-rater reliability. For essays with a difference of more than 10 points, the leading researcher of the study advised the raters to reconsider their scores. The raters were asked to negotiate until an agreement could be reached. Finally, inter-rater reliability for the essays, assessed using Pearson's correlation, was high and statistically significant ($p < 0.01$) for both LW (0.933) and RW (0.973). High levels of agreement were also observed across the analytic sub-constructs, including content (0.982), organisation (0.848), vocabulary (0.963), language use (0.834), and mechanics (0.911). For subsequent analyses, the mean of the two raters' scores was used to compare performance across LW and RW.

3.5.2 | Analysis of Written Texts

To address Research Question 2, the written materials were examined for linguistic complexity, which includes lexical and syntactic complexity. The analyses involved the use of computer-assisted tools to derive measures of complexity across the three phases of writing: LW, PW, and RW. Prior to computational analysis, minor preprocessing was performed manually by the researcher using a standardised protocol and applied uniformly to LW, PW, and RW texts. Corrections were limited to mechanical issues that interfered with automated processing (e.g., obvious typographical errors, missing sentence-final punctuation, and spacing errors affecting word segmentation). No lexical, grammatical, structural, or content changes were made. The preprocessed texts were used only for computational analysis, while all ratings and coding were conducted on the original texts.

Jarvis (2013) proposed that lexical diversity could be evaluated based on volume (text length), rarity (frequency of words in the language), and variability (type-token ratio (TTR) adjusted for text length). Lexical diversity was measured using the classic version of the Compleat Web Vocab Profile (VP) (Cobb, n.d.), which analyses the lexical profile of a given text by using indices such as 1000- (K1) and 2000-word families (K2), *academic words*, *off-token words*, and *total token*. Coh-Metrix 3.0 was used for calculations of TTRs and the *measure of textual lexical*

diversity (MTLD; McCarthy and Jarvis 2010). MTLD was utilised to complement the TTR, as the latter is usually affected by the length of a text sample. However, MTLD can overcome the potential confounding of text length by using sampling and estimation methods (Jarvis 2013).

Following Norris and Ortega's (2009) framework, the syntactic indices were calculated for sentence structure (*numbers of sentences, clauses, and t-units*) and syntactic complexity (*mean lengths of sentences, t-units, and clauses*) using Coh-Metrix 3.0. To measure phrasal complexity, the mean number of *modifiers per noun phrase* was calculated. To evaluate the level of syntactic sophistication, the *syntactic structure similarity* index was calculated. This measure estimates the extent to which syntactic structures are consistent in a text. That is, a lower syntactic structure similarity index indicates a larger selection of structures.

3.5.3 | Error Analysis of Written Texts

To identify errors, the researchers identified syntactic and lexical components of L2 writing that hindered the precise expression of meaning. Once the errors were identified, they were categorised according to their types. Using Ferris (2011) analytical framework of error taxonomy (p. 102), we could derive a classification of errors, which comprised (1) article, (2) mechanics, (3) noun ending, (4) preposition, (5) pronoun, (6) sentence structure, (7) subject-verb agreement, (8) verb form, (9) verb tense, (10) word choice, and (11) word form. Ferris (2011) categorises run-on and fragment separately; however, because both are related to structuring sentences, the two types were combined in our study. In our taxonomy, a category for "mechanics" was added to the taxonomy to include punctuation and abbreviation errors. Owing to crosslinguistic influence, "preposition" errors were also added due to their occurrence in the essay.

Two raters were chosen to identify the types of errors that occurred. One of the English native speakers who had rated the writing products was asked to corroborate the appropriateness of the error taxonomy. The other was the author of this study. To mark the errors, the two raters independently engaged in a reiterative process of reading the texts. When inter-rater reliability was initially calculated for 10% of the error instances, the agreement rate was 86.1%. After further calibration and clarification of error categories, agreement rate increased to 96.3% based on additional 10% of the data. Category-level agreement rates were as follows: article (97.5%), mechanics (100%), noun ending (96.8%), preposition (96.7%), pronoun (98.9%), sentence structure (88.5%), subject-verb agreement (98.2%), verb form (96.3%), verb tense (99.7%), word choice (92.3%), and word form (94.6%). Disagreements were primarily noted in cases of ambiguous sentence structures or word choices requiring interpretation of learners' intended meanings. Such discrepancies were resolved through discussion between raters, with detailed examination of the context of each sentence, until a consensus was reached.

For the remaining 80% of the dataset, one rater completed the coding, and all cases that were ambiguous or potentially

TABLE 1 | Classification of writing errors with examples.

Error type (Code)	Description
1. Article (Art)	Cases in which an article is missing or an article is used needlessly
2. Mechanics (MC)	Spelling, punctuation, capitalization, and abbreviation errors
3. Noun Endings (N)	Refers to cases in which plural nouns are in incorrect forms or nouns are incorrectly written in the plural form
4. Preposition (Prep)	Cases in which prepositions are misused or omitted
5. Pronoun (Pro)	Involves the incorrect use of personal or relative pronouns
6. Sentence Structure (SS)	Includes cases of (i) run-on sentences (including the comma splice) and fragments, (ii) verbal phrase errors, and (iii) word or phrase order errors.
7. Subject-Verb Agreement (SV)	Cases in which the predicate does not agree with the number of the subject
8. Verb Form (VF)	Cases of wrong voice (active/passive) or confusion between infinitives and gerunds
9. Verb Tense (VT)	Cases of wrong verb tense (e.g., not using the perfect tense when necessary) or inconsistency in verb tense use
10. Word Choice (WC)	Includes lexical errors for (i) incorrect word choice or (ii) unclear messages or awkward expressions
11. Word Form (WF)	Instances of (i) incorrect part of speech or (ii) ill-formed word

inconsistent with the established coding scheme were flagged and jointly reviewed by both raters. This procedure ensured that the coding for the full dataset adhered to the guidelines established during calibration. Discrepancies were resolved through discussion and close examination of the sentence context until consensus was reached. Table 1 shows the complete version of the error taxonomy and examples.

3.5.4 | Statistical Analysis

Holistic writing scores, indices for the linguistic complexity of L2 texts, and frequencies of the distinct types of errors were calculated to derive descriptive statistics (mean and

frequency). Paired-samples *t*-tests were conducted to compare raters' assessment of LW and RW. A repeated-measures, one-way analysis of variance (ANOVA)¹ was conducted on indices of linguistic complexity and frequency of errors to calculate differences among LW, PW, and RW. Prior to analysis, the assumptions of normality and sphericity were examined. For paired-samples *t*-tests, the normality of the difference scores was assessed using Shapiro–Wilk tests and inspection of Q–Q plots. For the repeated-measures ANOVAs, normality of residuals was similarly evaluated. The assumption of sphericity was tested using Mauchly's test. When the assumption of sphericity was violated, Greenhouse–Geisser corrections were applied to adjust the degrees of freedom. Effect sizes for repeated-measures ANOVAs were reported as partial eta squared (η^2) and interpreted using conventional benchmarks, with 0.01 indicating a small effect, 0.06 a medium effect, and 0.14 a large effect.

4 | Results

4.1 | Raters' Evaluation of Learners' Writing and ChatGPT-Supported Revisions

The raters' evaluations of the two types of writing indicated a mean difference of 6.41 between LW ($M = 68.85$, $SD = 7.67$) and RW ($M = 75.26$, $SD = 7.07$), which was statistically significant ($t = -7.738$, $df = 38$, $p < 0.001$). When analysed by writing sub-constructs as presented in Table 2, significant differences between LW and RW occurred for *vocabulary* and *language use* at $p < 0.01$ (Bonferroni adjustment). In contrast, improvements in *content* ($t = -1.865$, $p > 0.01$) and *organisation* ($t = -2.961$, $p > 0.01$) did not reach statistical significance. As shown in Table 2, when analysed further by proficiency groups, the *less skilled learners* differed in their performance between LW and RW in *vocabulary* and *language use*. In comparison, the *skilled learners* differed in their performance only in *language use*. Across proficiency levels, learner-mediated revision with AI support was associated with the greatest reductions in errors related to language use. These improvements were observed in sentence construction as well as in grammatical features such as agreement, tense, articles, and prepositions. In contrast, changes in content and organisation did not reach statistical significance for either less skilled or skilled learners, indicating that learners' AI-supported revisions were more likely to affect linguistic-level features than higher-order aspects of writing.

4.2 | Effects of Three Writing Phases on Linguistic Complexity in L2 Writing

Table 3 presents the descriptive statistics and results of the repeated-measures one-way ANOVA for lexical and syntactic indices across LW, PW, and RW with the alpha level set at $p < 0.003$ with Bonferroni correction (0.05/16). The percentage of academic words differed significantly across phases ($F = 19.181$, $p < 0.001$), with a large effect size ($\eta^2 = 0.330$), indicating a substantial increase from LW to RW ($p < 0.001$). However, there was no difference in the use of academic words between PW and RW, indicating learners' integration

TABLE 2 | Raters' assessment of less skilled and skilled learners' writing.

		Learners' writing (LW)		ChatGPT-supported revisions (RW)		<i>t</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Less skilled (<i>N</i> =19)	Content	20.90	3.39	21.29	3.60	−1.481
	Organisation	14.63	2.30	15.21	2.34	−2.480
	Vocabulary	14.11	1.99	17.05	2.01	−6.214**
	Language Use	16.24	3.30	20.68	2.46	−5.663**
	Mechanics	4.58	0.84	4.95	0.23	−2.111
Skilled (<i>N</i> =20)	Content	20.85	1.89	21.03	2.17	−1.129
	Organisation	14.38	1.05	14.55	1.04	−1.926
	Vocabulary	17.75	0.98	18.23	0.77	−2.033
	Language Use	18.70	2.96	22.43	2.30	−6.794**
	Mechanics	4.75	0.64	5.00	0.00	−1.751
TOTAL	Content	20.87	2.69	21.15	2.92	−1.865
	Organisation	14.50	1.75	14.87	1.80	−2.961
	Vocabulary	15.97	2.40	17.65	1.60	−5.156***
	Language Use	17.50	3.33	21.58	2.51	−8.636***
	Mechanics	4.67	0.74	4.97	0.16	−2.768

Note: Due to non-compliance with task instructions, one student's data was excluded from analysis (*N*=39). ****p*<0.001, ***p*<0.01 (Bonferroni correction).

of the words in their revisions. As shown in Table 4, post hoc analyses of the percentage of academic words indicated that less skilled learners incorporated significantly more academic vocabulary in the RW phase than in LW ($F=15.761$, $p<0.003$; LW: $M=11.41$, $SD=5.93$; RW: $M=13.31$, $SD=5.03$). In contrast, no significant change was observed for skilled learners ($F=5.210$, $p=0.025$).

As shown in Table 3, *TTR* for all words differed significantly across writing phases ($F=12.008$, $p<0.003$), whereas *TTR* for content words was not significant. *MTLD*, another measure of lexical diversity, also showed a significant phase effect ($F=14.010$, $p<0.003$). Both measures indicated large effects (*TTR*: $\eta^2=0.235$; *MTLD*: $\eta^2=0.264$), with post hoc analyses showing greater lexical diversity in PW and RW than in LW ($p<0.017$). However, no significant differences were observed within either proficiency group. In contrast, the overall percentage of off-token words did not differ significantly across the three writing phases. Taken together, these findings suggest that the observed increase in lexical diversity from LW to RW was primarily associated with a greater use of academic vocabulary.

Examination of mean word counts (total tokens) revealed minimal differences across LW ($M=271.90$), PW ($M=271.50$), and RW ($M=267.48$), with no significant differences overall or within proficiency groups. These findings indicate that essay length remained stable across writing phases, suggesting that observed differences in linguistic complexity reflect qualitative changes in language use rather than variation in text length.

Regarding syntactic structures and complexity, significant phase effects were observed for the number of sentences ($F=22.055$, $p<0.003$), clauses ($F=25.506$, $p<0.001$), *t*-units ($F=28.558$, $p<0.003$), as well as for mean length of sentences (MLS) ($F=17.718$, $p<0.003$), mean length of *t*-units (MLT) ($F=19.187$, $p<0.003$), and mean length of clauses (MLC) ($F=36.758$, $p<0.003$). Effect sizes were large across measures (e.g., *t*-units: $\eta^2=0.423$; *MLC*: $\eta^2=0.485$), indicating substantial structural changes across writing phases.

Post hoc comparisons showed that the numbers of sentences, clauses, and *t*-units decreased in both PW ($p<0.017$) and RW ($p<0.017$), whereas MLS, MLT, and MLC increased in both PW ($p<0.017$) and RW ($p<0.017$) relative to LW. This pattern suggests that texts produced during AI-supported revision contained fewer but longer and more syntactically dense units. The same trend was observed for both proficiency groups; however, among skilled learners, reductions in the number of sentences and increases in MLS and MLT did not reach statistical significance.

Additional syntactic complexity indices indicated that PW and RW exhibited higher syntactic complexity than LW in terms of *Modifiers per Noun Phrase* (NP) (phrasal complexity) ($F=7.541$, $p<0.003$) and *Structural Similarity* (sentence syntax similarity) ($F=9.016$, $p<0.003$). The number of words or phrases used to describe or modify nouns increased in PW and RW compared to LW ($p<0.017$). The index for Modifiers per NP suggests that texts produced during AI-supported revision contained denser noun phrases with more modifiers. As indicated in Table 4, this pattern was primarily driven by increases among less skilled learners from

TABLE 3 | Descriptive statistics and inferential statistics for linguistic complexity.

	Learners' Writing (LW)		ChatGPT-proofread Writing (PW)		ChatGPT-supported Revisions (RW)				
(N= 39)	M	SD	M	SD	M	SD	F	ηp^2	Post hoc
Lexical Indices									
■ Lexical Diversity									
K1 token (%)	77.915	5.749	75.803	5.074	93.295	110.028	0.910	0.023	N/A
K2 token (%)	5.945	2.421	6.366	2.725	6.335	2.808	3.445	0.081	N/A
Academic words token (%)	10.625	4.720	12.212	4.192	12.060	4.276	19.181*	0.330	1 < 2* 1 < 3* 2 = 3
Off-token words (%)	5.162	3.339	5.077	3.369	5.129	3.546	0.076	0.002	N/A
Total token	271.900	82.638	271.500	70.150	267.475	75.786	0.412	0.010	N/A
Type-token ratio (content word lemmas)	0.663	0.074	0.684	0.069	0.682	0.070	8.670	0.182	N/A
Type-token ratio (all words)	0.521	0.063	0.541	0.059	0.540	0.061	12.008*	0.235	1 < 2* 1 < 3* 2 = 3
MTLD (all words)	71.643	16.737	79.936	18.144	79.417	18.953	14.010*	0.264	1 < 2* 1 < 3* 2 = 3
Syntactic Indices									
■ Syntactic Structures									
No. of sentences	17.775	5.461	15.200	4.508	15.550	4.971	22.055*	0.361	1 > 2* 1 > 3* 2 = 3
No. of clauses	25.150	7.741	22.125	6.153	22.425	6.752	25.506*	0.395	1 > 2* 1 > 3* 2 = 3
No. of t-units	19.150	5.985	16.175	4.651	16.500	5.079	28.558*	0.423	1 > 2* 1 > 3* 2 = 3
■ Syntactic Complexity									
Mean length of sentences (MLS)	15.714	3.357	17.806	3.516	17.596	3.319	17.718*	0.312	1 < 2* 1 < 3* 2 = 3
Mean length of t-units (MLT)	14.551	3.033	16.677	3.416	16.492	3.109	19.187*	0.330	1 < 2* 1 < 3* 2 = 3
Mean length of clauses (MLC)	11.050	2.440	12.146	2.080	12.114	2.099	36.758*	0.485	1 < 2* 1 < 3* 2 = 3
Number of noun phrases per sentence, mean	0.804	0.171	0.848	0.172	0.851	0.186	7.541*	0.162	1 < 2* 1 < 3* 2 = 3

(Continues)

TABLE 3 | (Continued)

(N=39)	Learners' Writing (LW)		ChatGPT-proofread Writing (PW)		ChatGPT-supported Revisions (RW)		F	ηp^2	Post hoc
	M	SD	M	SD	M	SD			
Sentence syntax similarity, all combination, across paragraphs	0.128	0.030	0.116	0.023	0.118	0.023	9.016*	0.188	1 > 2* 1 > 3* 2 = 3

Note: 1 = Learners' writing, 2 = ChatGPT-proofread writing, 3 = ChatGPT-supported revised writing. Due to non-compliance with task instructions, one student's data was excluded from analysis (N = 39).

Main Effects: * $p < 0.003$ (Bonferroni correction).

Post Hoc Tests: * $p < 0.017$ (Bonferroni correction, post hoc test).

LW ($M = 0.815$, $SD = 0.174$) to PW ($M = 0.896$, $SD = 0.174$, $p < 0.017$) and RW ($M = 0.908$, $SD = 0.190$, $p < 0.017$), whereas no significant change was observed for skilled learners ($F = 0.093$, $p > 0.003$). Similarly, the lower structure similarity index observed for both PW and RW relative to LW ($p < 0.017$) suggests a broader range of syntactic structures. However, after Bonferroni correction, these differences were not statistically significant within either proficiency group. This pattern suggests that changes in syntactic variability associated with AI-supported revision were subtle and broadly distributed rather than pronounced within a single proficiency group. Overall, most significant phase effects were associated with medium to large effect sizes, suggesting that the observed differences reflect substantive changes in text features rather than trivial statistical variation.

4.3 | Error Types Across Writing Phases

Analyses of the frequency of errors were conducted to (1) assess the AI tool's capability in rectifying writing errors present in LW and (2) examine how learners responded to the AI tool's corrections by comparing PW with RW. To control for variation in essay length, error frequency was calculated per sentence. As shown in Table 5, descriptive statistics for the total mean frequency of errors indicated that LW errors occurred 0.064 times per sentence. This figure decreased substantially to 0.012 in PW and increased slightly to 0.016 in RW. A repeated-measures one-way ANOVA indicated a significant difference across the three writing phases ($F = 88.430$, $p < 0.001$). The reduction in overall error frequency showed a very large effect size ($\eta p^2 = 0.695$), indicating substantial change across writing phases. Bonferroni-adjusted post hoc comparisons indicated no significant difference between PW and RW ($p = 0.153$), suggesting that most corrections identified during the AI-supported proofreading stage were retained in the final revision. In contrast, the sharp reduction from LW to PW ($p < 0.001$) indicates that substantial decreases in errors occurred during the proofreading phase. Given the overall reduction in errors, a more fine-grained analysis by error type was conducted, especially to examine how the learners responded to correction in PW, and ascertain how this occurred by proficiency groups.

In LW, 11 error types were identified: *articles*, *mechanics*, *noun ending*, *prepositions*, *pronouns*, *sentence structure*, *subject-verb agreement*, *verb form*, *verb tense*, *word choice*, and *word form*.

To trace changes across LW, PW, and RW, a repeated-measures one-way ANOVAs were conducted for each error type with the alpha level adjusted to $p < 0.005$ using a Bonferroni correction (0.05/11). The statistical analyses indicated significant differences for all types of errors. Post hoc tests generally indicated significant reductions in PW relative to LW ($p < 0.001$), with several rule-governed error types (*verb form*, *verb tense*, and *word form*) reaching zero occurrences at the proofreading stage. As illustrated in Figure 2, the per-sentence error frequencies across phases further demonstrate the strong error reduction during the AI-supported proofreading stage. Nevertheless, some error types—particularly word choice and sentence structure—remained less consistently resolved, suggesting limits to the tool's ability to address context-sensitive language use. Across all learners, the total number of errors declined from 470 in LW to 65 in PW (86.17% of errors corrected), but increased to 108 in RW, with the overall error resolution declining to 77.02%. When normalised by the number of participants ($N = 39$), this corresponds to a reduction from 11.75 errors per learner in LW to 1.63 in PW, followed by a slight increase to 2.70 in RW. These patterns indicate that AI-supported proofreading provided generally reliable error correction, although some errors were reintroduced during learner-mediated revision.

To further examine whether these patterns differed by proficiency level as seen in Table 6, separate analyses were conducted for less skilled and skilled learners. The results revealed both shared and proficiency-specific trajectories. Both groups showed significant improvements ($p < 0.005$) in *articles*, *sentence structure*, *subject-verb agreement*, and *word choice* across writing phases, with significant differences observed between LW and PW ($p < 0.017$), as well as between LW and RW ($p < 0.017$). This indicated that ChatGPT-supported proofreading facilitated improvements in both rule-based and meaning-dependent error types regardless of proficiency. In contrast, neither proficiency group demonstrated significant improvements in *mechanics*, *noun ending*, *pronouns*, *verb form*, *verb tense*, or *word form* ($p > 0.005$), suggesting that either these errors were already infrequent or that learners were less responsive to AI-generated corrections in these categories during revision. Notably, improvement in *preposition* use was observed only among skilled learners ($p < 0.005$), suggesting that judgements involving subtle aspects of collocational knowledge—including preposition use—were more readily taken up by higher-proficiency learners during revision.

TABLE 4 | Measures of linguistic complexity by proficiency group.

(N= 39)		Less skilled learners					Skilled learners				
		M	SD	F	ηp^2	Post Hoc	M	SD	F	ηp^2	Post Hoc
<i>Lexical Diversity</i>											
K1 token (%)	LW	76.915	6.693	16.381*	0.463	1 > 2*	78.915	4.576	0.961	0.048	N/A
	PW	73.656	4.716			1 > 3*	77.950	4.571			
	RW	74.195	5.190			2 = 3	112.395	155.095			
K2 token (%)	LW	6.115	2.651	3.944	0.172	N/A	5.775	2.223	0.462	0.024	N/A
	PW	6.908	3.085				5.825	2.260			
	RW	6.715	3.183				5.955	2.397			
Academic words (%)	LW	11.410	5.933	15.761*	0.453	1 < 2*	9.840	3.037	5.210	0.215	N/A
	PW	13.699	4.935			1 < 3*	10.725	2.658			
	RW	13.305	5.025			2 = 3	10.815	3.002			
Off-token words (%)	LW	5.255	3.031	0.126	0.007	N/A	5.070	3.699	0.004	0.000	N/A
	PW	5.110	2.666				5.044	4.024			
	RW	5.189	2.693				5.069	4.307			
Total token	LW	282.450	79.887	0.691	0.035	N/A	261.350	86.031	1.421	0.070	N/A
	PW	277.350	68.203				265.650	73.330			
	RW	283.900	73.913				251.050	75.883			
Type-token ratio (content word)	LW	0.604	0.037	1.790	0.086	N/A	0.592	0.037	4.880	0.204	N/A
	PW	0.613	0.038				0.604	0.039			
	RW	0.610	0.039				0.607	0.042			
Type-token ratio (all words)	LW	0.508	0.058	5.164	0.214	N/A	0.534	0.069	3.020	0.137	N/A
	PW	0.531	0.059				0.542	0.053			
	RW	0.528	0.061				0.550	0.059			
Measure of Textual Lexical Diversity (MTLD)	LW	69.355	14.678	7.975	0.296	N/A	73.932	18.670	6.935	0.267	N/A
	PW	80.172	15.581				79.700	20.806			
	RW	79.510	15.691				79.325	22.162			
<i>Syntactic Structures</i>											
No. of Sentences	LW	19.300	4.330	14.320*	0.430	1 > 2*	16.250	6.129	8.510	0.309	N/A
	PW	15.900	4.254			1 > 3*	14.500	4.752			
	RW	16.600	4.773			2 = 3	14.500	5.063			
No. of Clauses	LW	25.600	7.789	10.105*	0.347	1 > 2*	24.700	7.868	15.618*	0.451	1 > 2*
	PW	22.700	6.441			1 > 3*	21.550	5.960			1 > 3*
	RW	23.250	7.247			2 = 3	21.600	6.295			2 = 3
No. of t-units	LW	20.500	4.883	19.343*	0.504	1 > 2*	17.800	6.771	10.951*	0.366	1 > 2*
	PW	16.500	4.419			1 > 3*	15.850	4.966			1 > 3*
	RW	17.150	4.902			2 = 3	15.850	5.294			2 = 3

(Continues)

TABLE 4 | (Continued)

		Less skilled learners					Skilled learners				
		<i>M</i>	<i>SD</i>	<i>F</i>	η^2	<i>Post Hoc</i>	<i>M</i>	<i>SD</i>	<i>F</i>	η^2	<i>Post Hoc</i>
<i>Syntactic Complexity</i>											
MLS	LW	14.886	3.576	13.118*	0.408	1 < 2*	16.542	2.983	6.000	0.240	N/A
	PW	17.940	4.546			1 < 3*	17.672	2.160			
	RW	17.563	4.189			2 = 3	17.630	2.250			
MLT	LW	14.028	3.265	15.410*	0.448	1 < 2*	15.073	2.764	7.344	0.279	N/A
	PW	17.321	4.539			1 < 3*	16.032	1.577			
	RW	17.007	4.074			2 = 3	15.978	1.637			
MLC	LW	11.415	2.834	13.488*	0.415	1 < 2**	10.686	1.977	29.296*	0.607	1 < 2*
	PW	12.523	2.401			1 < 3**	11.770	1.678			1 < 3*
	RW	12.544	2.372			2 = 3	11.684	1.741			2 = 3
Number of noun phrases per sentence	LW	0.815	0.174	14.227*	0.428	1 < 2*	0.794	0.171	0.093	0.005	N/A
	PW	0.896	0.174			1 < 3*	0.800	0.159			
	RW	0.908	0.190			2 = 3	0.794	0.168			
Syntactic structure similarity across all sentences	LW	0.132	0.034	3.845	0.168	N/A	0.123	0.024	5.874	0.236	N/A
	PW	0.120	0.023				0.113	0.024			
	RW	0.122	0.024				0.114	0.022			

Note: 1 = Learners' writing (LW), 2 = ChatGPT-proofread writing (PW), 3 = ChatGPT-supported revised writing (RW). Main Effects: * $p < 0.003$ (Bonferroni correction). Post Hoc Tests: * $p < 0.017$ (Bonferroni correction, post hoc test).

A closer inspection of the proportion of errors remaining in RW further clarifies these patterns. While several rule-governed errors were almost entirely eliminated, *word choice* and *sentence structure* errors remained relatively persistent, with 38.9% and 35.4% of the original errors unresolved, respectively. In fact, among skilled learners, the frequency of *word choice* errors did not differ significantly between LW and RW ($p > 0.017$), indicating that errors related to expressing intended meaning were not resolved during the AI-supported revision process. Other error categories, including mechanics (23.8%), noun ending (25.0%), and prepositions (18.8%), showed moderate reductions. Taken together, these findings suggest that AI-supported revisions are particularly effective for addressing surface-level and structurally constrained errors, while improvements in context-sensitive and lexically nuanced language use are more limited and appear to depend partly on learner proficiency.

5 | Discussion

This study makes several contributions to research on AI-assisted L2 writing. First, it provides empirical evidence that learners' AI-supported revisions were associated with improvements in surface-level language features, while also revealing limitations in handling context-sensitive language, such as word choice and sentence structure. Second, by incorporating EditGPT to trace learners' choice of AI feedback, the study offers new insights into how learners respond to automated

revisions. Importantly, the use of EditGPT may have reduced direct adoption of ChatGPT's revisions by requiring learners to evaluate and selectively incorporate suggested changes. This design ensured that RW reflected learner-mediated uptake rather than automated correction alone and addressed ethical concerns by distinguishing AI-supported feedback from direct AI-generated text production. Third, the study provides pedagogical implications for integrating AI tools to foster reflective engagement with feedback and more deliberate revision decisions.

The findings are consistent with prior research suggesting that ChatGPT, when used as an AWE tool, can improve surface-level language features, including grammatical and mechanical accuracy, sentence structure, clarity, and vocabulary use (Al-Garaady and Mahyoub 2023; Tsai et al. 2024). The large effect sizes observed for several lexical, syntactic, and error-related measures further indicate that these phase-based improvements were practically meaningful within the AI-supported revision context. However, ChatGPT-supported revision was less effective in addressing context-sensitive language, such as appropriate word choice and discourse-level sentence structuring, a limitation also reported in earlier AWE research (Barrot 2026; Li et al. 2015; Ranalli 2018). In addition, although ChatGPT can support idea generation and outlining at earlier stages of writing, the present study did not find statistically significant improvements in content or organisation. This pattern reflects a well-documented limitation of AI-based proofreading tools (Chen and Cui 2022; Dikli

TABLE 5 | Frequency of errors per sentence and sum of errors.

(N=39)	Learners' Writing (LW)		ChatGPT-proofread Writing (PW)		ChatGPT-supported Revisions (RW)		<i>F</i>	<i>ηp2</i>	<i>Post hoc</i>		
	<i>M</i> (SD)	<i>Sum</i>	<i>M</i> (SD)	<i>Sum</i>	<i>M</i> (SD)	<i>Sum</i>					
Articles	0.067 (0.08)	47	0.010 (0.032)	4	0.011 (0.029)	6	19.625*	0.335	1 > 2*	1 > 3*	2 = 3
Mechanics	0.123 (0.137)	84	0.030 (0.072)	15	0.036 (0.07)	20	19.625*	0.335	1 > 2*	1 > 3*	2 = 3
Noun Ending	0.033 (0.05)	24	0.003 (0.016)	2	0.010 (0.023)	6	9.834*	0.201	1 > 2*	1 > 3*	2 = 3
Preposition	0.050 (0.067)	32	0.001 (0.007)	1	0.009 (0.025)	6	16.905*	0.302	1 > 2*	1 > 3*	2 = 3
Pronoun	0.023 (0.034)	16	0.003 (0.015)	2	0.004 (0.019)	2	17.116*	0.305	1 > 2*	1 > 3*	2 = 3
Sentence Structure	0.101 (0.081)	65	0.027 (0.071)	15	0.035 (0.052)	23	20.263*	0.342	1 > 2*	1 > 3*	2 = 3
Subject-Verb Agreement	0.049 (0.049)	34	0.001 (0.008)	1	0.002 (0.01)	2	37.506*	0.49	1 > 2*	1 > 3*	2 = 3
Verb Form	0.037 (0.07)	27	0	0	(0)	0	11.045*	0.221	1 > 2*	1 > 3*	N/A
Verb Tense	0.033 (0.071)	19	0	0	0.002 (0.014)	1	7.603*	0.163	1 > 2*	1 > 3*	2 = 3
Word Choice	0.169 (0.144)	108	0.045 (0.08)	25	0.068 (0.087)	42	24.079*	0.382	1 > 2*	1 > 3*	2 = 3
Word Form	0.020 (0.048)	14	0	0	0	0	7.165*	0.155	1 = 2	1 = 3	N/A
TOTAL	0.064 (0.034)	470	0.011 (0.017)	65	0.016 (0.015)	108	88.916*	0.695	1 > 2*	1 > 3*	2 = 3

Note: 1 = Learners' writing, 2 = ChatGPT-proofread writing, 3 = ChatGPT-supported revised writing. Due to non-compliance with task instructions, one student's data was excluded from analysis (N=39). Main Effects: * $p < 0.005$ (Bonferroni correction). Post Hoc Tests: * $p < 0.017$ (Bonferroni correction, post hoc test).

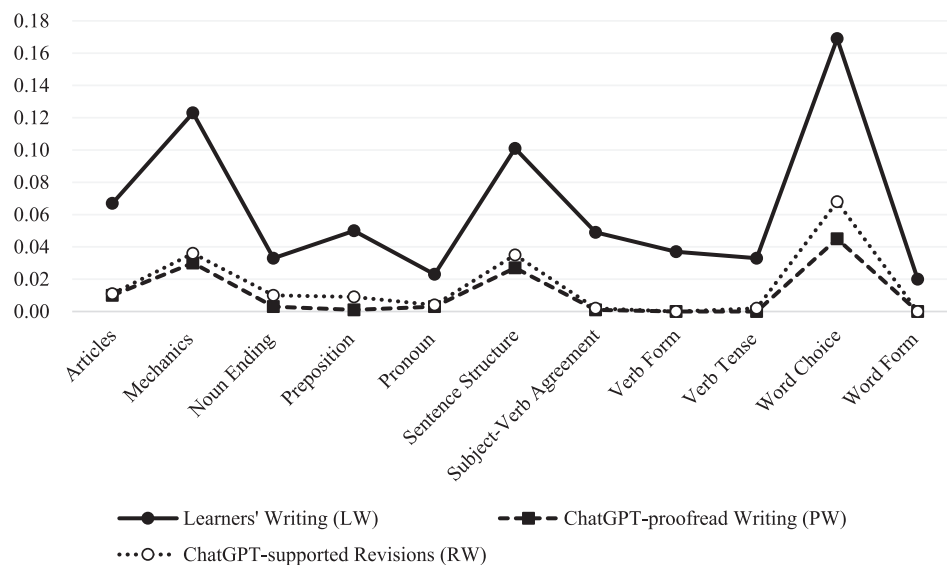
**FIGURE 2** | Frequency of errors per sentence.

TABLE 6 | Trajectories of errors by proficiency group.

(N= 39)		Less skilled learners					Skilled learners				
		<i>M</i>	<i>SD</i>	<i>F</i>	ηp^2	<i>Post Hoc</i>	<i>M</i>	<i>SD</i>	<i>F</i>	ηp^2	<i>Post Hoc</i>
Article	LW	0.057	0.071	10.046*	0.346	1 > 2*	0.078	0.089	9.383*	0.331	1 > 2*
	PW	0.004	0.019			1 > 3*	0.016	0.040			1 > 3*
	RW	0.006	0.019			2 = 3	0.017	0.037			2 = 3
Mechanics	LW	0.117	0.127	9.232	0.327	N/A	0.128	0.149	7.427	0.281	N/A
	PW	0.027	0.076				0.034	0.068			
	RW	0.028	0.071				0.043	0.070			
Noun Ending	LW	0.028	0.046	3.865	0.169	N/A	0.038	0.055	6.196	0.246	N/A
	PW	0.007	0.022				0.000	0.000			
	RW	0.007	0.021				0.012	0.026			
Preposition	LW	0.034	0.048	7.759	0.290	N/A	0.065	0.081	10.250*	0.350	1 > 2*
	PW	0.002	0.010				0.000	0.000			1 > 3*
	RW	0.007	0.016				0.012	0.031			2 = 3
Pronoun	LW	0.018	0.027	7.125	0.273	N/A	0.029	0.040	9.879	0.342	N/A
	PW	0.002	0.010				0.004	0.019			
	RW	0.002	0.010				0.006	0.025			
Sentence Structure	LW	0.068	0.038	32.105*	0.628	1 > 2*	0.134	0.098	8.526*	0.310	1 > 2*
	PW	0.014	0.026			1 > 3*	0.039	0.097			1 > 3*
	RW	0.011	0.023			2 = 3	0.060	0.062			2 = 3
Subject-Verb Agreement	LW	0.042	0.051	13.164*	0.409	1 > 2*	0.057	0.047	25.547*	0.573	1 > 2*
	PW	0.000	0.000			1 > 3*	0.003	0.011			1 > 3*
	RW	0.000	0.000			2 = 3	0.005	0.014			2 = 3
Verb Form	LW	0.042	0.075	6.374	0.251	N/A	0.031	0.066	4.464	0.190	N/A
	PW	0.000	0.000				0.000	0.000			
	RW	0.000	0.000				0.000	0.000			
Verb Tense	LW	0.010	0.020	4.726	0.199	N/A	0.055	0.094	6.047	0.241	N/A
	PW	0.000	0.000				0.000	0.000			
	RW	0.000	0.000				0.005	0.020			
Word Choice	LW	0.124	0.076	24.348*	0.562	1 > 2*	0.214	0.180	9.362*	0.330	1 > 2*
	PW	0.020	0.043			1 > 3*	0.069	0.100			1 = 3
	RW	0.024	0.044			2 = 3	0.112	0.098			2 = 3
Word Form	LW	0.015	0.024	3.705	0.163	N/A	0.026	0.064	3.199	0.144	N/A
	PW	0.003	0.012				0.000	0.000			
	RW	0.003	0.012				0.000	0.000			

Note: 1 = Learners' writing, 2 = ChatGPT-proofread writing, 3 = ChatGPT-supported revised writing. Main Effects: * $p < 0.005$ (Bonferroni correction). Post Hoc Tests: * $p < 0.017$ (Bonferroni correction, post hoc test).

and Bleyle 2014; Ding et al. 2026; Link et al. 2022) and underscores the continued importance of human instruction for higher-order writing development.

Despite these contributions, the findings should be interpreted cautiously. Because the writing was produced within an AI-supported environment, RW should be understood as AI-mediated

performance rather than fully independent learner production. The final texts represent hybrid outcomes shaped by learners' evaluation, selection, and integration of AI-generated suggestions. In this sense, RW reflects learners' capacity to engage with and operationalize external feedback rather than their independent linguistic competence. Moreover, the findings represent short-term text improvement and learner response to feedback rather than evidence of durable learning or writing development. The study did not measure learning gains, transfer, or retention, and such outcomes require longitudinal investigation.

The present study further demonstrates that the impact of AI-assisted revision is not uniform across learners. While previous research has cautioned that AI-supported revision may complicate the accurate assessment of lower-proficiency learners (Schmidt-Fajlik 2023; Tsai et al. 2024), the current findings show that the AI tool exerted differential effects across distinct aspects of writing. In particular, less skilled learners benefited not only from reductions in grammatical errors but also from an expanded range of vocabulary use. The analysis further shows that less skilled learners increased their use of academic vocabulary, resulting in greater lexical diversity, whereas comparable improvements were not observed among skilled learners. This pattern suggests a potential proficiency-specific lexical benefit of ChatGPT-assisted revision that has received little attention in existing research.

Between LW and RW, the AI tool proved helpful in facilitating more complex sentence structures, particularly by enabling less skilled learners to produce fewer but more complex sentence structures. This finding is consistent with Fredrick and Craven (2025), who report that ChatGPT can help students perform beyond their current proficiency in lexical diversity and syntactic complexity, as reflected in a higher ratio of dependent clauses per T-unit. Nonetheless, Fredrick and Craven (2025) caution that improvements in lexical diversity and syntactic complexity do not necessarily translate into improved readability or communicative effectiveness.

A qualitative comparison of LW and RW revealed improvements in deeper syntactic complexity. For example, RW demonstrated broader use of embedded clauses, as seen in the sentence, 'One common type of animal experiment is drug and chemical testing, where animals suffer...', which integrates subordinate information more smoothly than the simpler clause structure in LW: '...one type of common animal experiment is drug and chemical testing.' Similarly, the use of a relative clause in 'animals have to undergo gavage procedures, which involve inserting a long, thin tube...' adds descriptive complexity while maintaining clarity, contrasting with 'This is called gavage procedure, and it happens multiple times a day.' These findings confirm Fredrick and Craven's (2025) study that ChatGPT can help L2 learners produce more syntactically denser sentences.

Nonetheless, this structural shift should be interpreted cautiously. Because multiple components of the AI-supported writing process were implemented simultaneously, the observed increases in syntactic density likely reflect the combined influence of the structured revision environment—including repeated drafting, increased time-on-task, and prompt scaffolding—rather than the independent effect of the AI system alone.

In addition to clausal embedding, RW also exhibited greater noun phrase elaboration, particularly among less skilled learners; a pattern comparable to findings reported by Liu and Liu (2025). The absence of statistically significant changes in sentence counts or length measures for skilled learners suggests that higher-proficiency writers were less reliant on external support for syntactic restructuring.

At the discourse level, texts in the RW phase showed clearer transitions, improved cohesion, and a more formal tone, contrasting with earlier AWE tools that offered limited guidance for coherence (Li et al. 2015). Logical connectors (e.g., *Firstly, Secondly, For instance*) and concessive clauses (e.g., *While all animals...*) were used more effectively to signal relationships between ideas, resulting in more coherent arguments. These changes suggest that, even when EditGPT primarily provided direct corrective feedback, ChatGPT—when paired with learner evaluation followed by acceptance or rejection of feedback—can support discourse-level coherence. This finding contrasts with earlier AWE studies reporting limited effects on higher-order writing features (Chen and Cui 2022; Link et al. 2022), suggesting that discourse-level benefits may emerge under specific conditions, such as when learners actively evaluate and integrate AI-generated feedback.

Although the findings indicate clear benefits of AI-assisted writing, ethical and pedagogical concerns remain regarding over-reliance on automated tools and the potential reductions in critical thinking and learner autonomy (O'Connor 2023; Stokel-Walker 2023). To address these concerns, EditGPT was used as a mediation interface to encourage active evaluation and reduce uncritical adoption of AI-generated feedback.

Error patterns provide further insight into both the benefits and limitations of this learner-mediated process. While errors decreased substantially from LW to PW, a modest increase was observed in RW. This increase likely reflects the risks inherent in learner-mediated revision rather than limitations of the proof-reading stage alone. When evaluating AI suggestions, learners may modify or reject revisions to preserve their intended meaning, resulting in partial corrections or the introduction of new errors. In addition, efforts to restructure sentences or introduce alternative lexical choices may create new opportunities for error, particularly in areas requiring contextual or pragmatic judgement. Thus, the rise in errors in RW can be understood as evidence of active learner intervention during uptake rather than a simple failure of AI correction.

Learners' acceptance and rejection behaviours further revealed how they evaluated AI-generated feedback. To examine these processes, the researchers categorised ChatGPT's sentence-level suggestions into four types: (1) instances in which students rejected an appropriate correction, (2) instances in which students accepted a correction, (3) cases where ChatGPT failed to identify an error, and (4) cases where ChatGPT's correction required further refinement.

Category 1, in which a valid correction was rejected, may reflect learners' uncertainty or lack of confidence in evaluating language appropriateness. For example, in the sentence, "As time goes by, the number of people who are using social media is increasing and it became* main platform for everyone," ChatGPT

correctly revised the verb tense, yet the student chose to retain the original, grammatically flawed version. This kind of decision may indicate a lack of trust in AI feedback or difficulty in recognising subtle grammatical errors—highlighting the need for instructional guidance in reviewing and evaluating AI-generated suggestions.

In Category 3, where ChatGPT failed to identify errors, the tool overlooked issues such as sentence fragments (“However, there are also negative effects. First, human alienation.”) and subject-verb agreement (“...the amount and kinds of information... is innumerable*”). These examples underscore the limitations of ChatGPT’s contextual understanding and highlight the necessity of human oversight. As stated by Liao (2016), this suggests that the effectiveness of AWE tools can be significantly enhanced when combined with instructional scaffolding and iterative revision processes. Similarly, Category 4 illustrates how learners might struggle to interpret vague or ambiguous AI suggestions. For instance, when ChatGPT failed to revise the sentence, “Finally, teachers can use strategies for balancing technology use in education,” the awkward phrasing remained untouched—likely because the tool could not infer the writer’s intended meaning. These patterns suggest that decisions were shaped by learner confidence, language awareness, and contextual interpretation rather than uncritical reliance.

Taken together, the findings extend prior research on AI feedback Kim and Chon (2025) by showing that the effectiveness of ChatGPT-based tools depends on learners’ capacity to evaluate meaning and appropriateness. Instructional support is therefore needed to help learners critically evaluate AI suggestions, particularly for context-sensitive language decisions. Instructors can foster learners’ evaluative skills through reflective practices, such as comparing original sentences with AI-generated suggestions and justifying acceptance or rejection decisions, which can sharpen learners’ “sentence adequacy judgements” (Moorkens 2018). Classroom discussions or peer review activities focusing on word choice, tone, and contextual appropriateness can further help learners approach AI output not as definitive solutions but as revision options to be weighed against communicative goals and intended meaning.

6 | Conclusion

The findings indicate that a structured ChatGPT-mediated revision environment can enhance the linguistic quality of L2 writing, particularly in grammatical accuracy, lexical sophistication, and syntactic density. However, improvements in content and organisation were limited, and persistent errors in word choice and sentence structure suggest that context-sensitive language use still requires human judgement and instructional support. These results underscore the complementary role of teachers in supporting higher-order writing while AI tools address lower-level linguistic concerns.

Observed improvements reflect product-level gains rather than deeper learning, as generative AI may encourage “metacognitive laziness” when learners accept suggestions uncritically (Fan et al. 2025). Addressing this risk requires integrating AI use with learners’ feedback literacy so that

it promotes active feedback engagement (Zhan et al. 2025). Accordingly, AI-supported writing should be embedded in structured instruction that develops learners’ ability to evaluate, question, and selectively apply feedback to support ownership and self-regulation.

These instructional implications must be considered alongside several methodological limitations that also inform future research and practice. First, the small sample size ($N=39$) and the focus on Korean EFL university students limit statistical power and generalizability. Second, the use of single-turn, researcher-provided prompts constrained authentic interaction with ChatGPT, reducing ecological validity. Third, the absence of a non-AI comparison condition makes it difficult to isolate AI-specific effects, as improvements may partly reflect general revision processes such as rereading, increased attention, or additional time-on-task. Fourth, the proficiency groups were drawn from different institutional contexts, and TOEIC-based grouping may not fully capture writing-specific ability. Fifth, human ratings were conducted only for LW and RW, limiting the ability to distinguish the relative contributions of AI proofreading and learner mediation. Finally, the single-session design captured short-term text improvement rather than sustained development.

Taken together, the findings suggest that the value of generative AI in L2 writing depends less on text improvement than on its pedagogical integration. When embedded in structured, reflective, feedback-literacy-oriented instruction, AI can support learner-mediated uptake and more deliberate revision. Future longitudinal and ecologically valid research should examine whether such use fosters independent writing and sustained learning.

Author Contributions

Yuah V. Chon: writing – Original Draft Preparation, Writing – Review and Editing, Data Curation, Formal Analysis, Investigation, Project Administration, Software, Validation, Visualization, Funding Acquisition. **Dongkwang Shin:** writing – Review and Editing, Conceptualization, Data Curation, Software, Investigation, Project Administration, Methodology.

Funding

This work was supported by Hanyang University, 202400000003783.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Peer Review

For transparency, the peer review documents associated with this article are available at <https://doi.org/10.1002/jcal.70259>.

Endnotes

¹ Because the dependent variables represented separate constructs rather than a single multivariate outcome, MANOVA was not applied. Instead,

each construct was analysed independently using paired-samples *t*-tests with Bonferroni adjustment and repeated-measures ANOVA.

References

- Al-Garaady, J., and M. Mahyoob. 2023. "Chatgpt's Capabilities in Spotting and Analyzing Writing Errors Experienced by EFL Learners." *Arab World. English Journal* 9: 3–17. <https://doi.org/10.24093/awej/call9.1>.
- Ali, J. K. M., M. A. A. Shamsan, T. A. Hezam, and A. A. Q. Mohammed. 2023. "Impact of ChatGPT on Learning Motivation: Teachers and Students' Voices." *Journal of English Studies in Arabia Felix* 2, no. 1: 41–49. <https://doi.org/10.56540/jesaf.v2i1.51>.
- Bai, L., and G. Hu. 2017. "In the Face of Fallible AWE Feedback: How Do Students Respond?" *Educational Psychology* 37, no. 1: 67–81. <https://doi.org/10.1080/01443410.2016.1223275>.
- Barrot, J. S. 2021. "Using Automated Written Corrective Feedback in the Writing Classrooms: Effects on L2 Writing Accuracy." *Computer Assisted Language Learning* 36, no. 4: 584–607. <https://doi.org/10.1080/09588221.2021.1936071>.
- Barrot, J. S. 2023. "Using ChatGPT for Second Language Writing: Pitfalls and Potentials." *Assessing Writing* 57: 100745. <https://doi.org/10.1016/j.asw.2023.100745>.
- Barrot, J. S. 2026. "Generative Artificial Intelligence for Automated Essay Scoring: Exploring Teacher Agency Through an Ecological Perspective." *Assessing Writing* 67: 100990.
- Baskara, F. R. 2023. "Integrating ChatGPT Into EFL Writing Instruction: Benefits and Challenges." *International Journal of Education and Learning* 5, no. 1: 44–55. <https://doi.org/10.31763/ijelev5i1.858>.
- Cao, S., and L. Zhong. 2023. "Exploring the Effectiveness of ChatGPT-Based Feedback Compared With Teacher Feedback and Self-Feedback: Evidence From Chinese to English Translation." *arXiv*: 2309.01645. <https://doi.org/10.48550/arXiv.2309.01645>.
- Chen, M., and Y. Cui. 2022. "The Effects of AWE and Peer Feedback on Cohesion and Coherence in Continuation Writing." *Journal of Second Language Writing* 57: 100915. <https://doi.org/10.1016/j.jslw.2022.100915>.
- Crosthwaite, P., and S. Sun. 2025. "Generative AI and L2 Written Feedback Studies: A Scoping Review." *RELC Journal* 57, no. 1: 207–219. <https://doi.org/10.1177/00336882251386530>.
- Dikli, S., and S. Bleyle. 2014. "Automated Essay Scoring Feedback for Second Language Writers: How Does It Compare to Instructor Feedback?" *Assessing Writing* 22: 1–17. <https://doi.org/10.1016/j.asw.2014.03.006>.
- Ding, L., D. Zou, and L. Kohnke. 2026. "Enhancing L2 Writing With AI: Effects of ChatGPT Feedback and the Role of Student Engagement." *Artificial Intelligence in Language Education* 2: 103356–103356. <https://doi.org/10.29140/aile.2026.103356>.
- Ellis, R. 2010. "Cognitive, Social, and Psychological Dimensions of Corrective Feedback." In *Sociocognitive Perspectives on Language Use and Language Learning*, edited by R. Batstone, 151–165. Oxford University Press.
- Escalante, J., A. Pack, and A. Barrett. 2023. "AI-Generated Feedback on Writing: Insights Into Efficacy and ENL Student Preference." *International Journal of Educational Technology in Higher Education* 20, no. 1: 57. <https://doi.org/10.1016/j.asw.2014.03.006>.
- Fan, Y., L. Tang, H. Le, et al. 2025. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology* 56, no. 2: 489–530. <https://doi.org/10.1111/bjet.13544>.
- Farrokhnia, M., S. K. Banihashem, O. Noroozi, and A. Wals. 2024. "A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research." *Innovations in Education and Teaching International* 61, no. 3: 460–474. <https://doi.org/10.1080/14703297.2023.2195846>.
- Ferris, D. 2011. *Treatment of Error in Second Language Student Writing*. University of Michigan Press.
- Fredrick, D., and L. Craven. 2025. "Lexical Diversity, Syntactic Complexity, and Readability: A Corpus-Based Analysis of ChatGPT and L2 Student Essays." *Frontiers in Education* 10: 1616935. <https://doi.org/10.3389/educ.2025.1616935>.
- Fu, Q. K., D. Zou, H. Xie, and G. Cheng. 2024. "A Review of AWE Feedback: Types, Learning Outcomes, and Implications." *Computer Assisted Language Learning* 37, no. 1–2: 179–221. <https://doi.org/10.1080/09588221.2022.2033787>.
- Guo, K., and D. Wang. 2024. "To Resist It or to Embrace It? Examining Chatgpt's Potential to Support Teacher Feedback in EFL Writing." *Education and Information Technologies* 29, no. 7: 8435–8463. <https://doi.org/10.1007/s10639-023-12146-0>.
- Han, J., H. Yoo, Y. Kim, et al. 2023. "RECIPE: How to Integrate ChatGPT into EFL Writing Education." In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 416–420. Association for Computing Machinery. <https://doi.org/10.1145/3573051.3596200>.
- Harunasari, S. Y. 2023. "Examining the Effectiveness of AI-Integrated Approach in EFL Writing: A Case of ChatGPT." *International Journal of Progressive Sciences and Technologies* 39, no. 2: 357–368. <https://doi.org/10.52155/ijpsat.v39.2.5516>.
- Jacobs, H. L., S. A. Zingraf, D. R. Wormuth, V. F. Hartfiel, and J. B. Hughey. 1981. *Testing ESL Composition*. Newbury House Publishers.
- Jarvis, S. 2013. "Defining and Measuring Lexical Diversity." In *Vocabulary Knowledge: Human Ratings and Automated Measures*, edited by S. Jarvis and M. Daller, 13–45. John Benjamins.
- Kim, M., and Y. V. Chon. 2025. "Impact of Self-Revision, Machine Translation, and ChatGPT on L2 Writing: Raters' Assessments, Linguistic Complexity, and Error Correction." *Assessing Writing* 65: 100950. <https://doi.org/10.1016/j.asw.2025.100950>.
- Kohnke, L., B. L. Moorhouse, and D. Zou. 2023. "ChatGPT for Language Teaching and Learning." *RELC Journal* 54, no. 2: 537–550. <https://doi.org/10.1177/00336882231162868>.
- Kostka, I., and R. Toncelli. 2023. "Exploring Applications of ChatGPT to English Language Teaching: Opportunities, Challenges, and Recommendations." *Teaching English as a Second or Foreign Language-TESL-EJ* 27, no. 3: 1–19. <https://doi.org/10.55593/ej.27107int>.
- Li, J., and M. Li. 2018. "Turnitin and Peer Review in ESL Academic Writing Classrooms." *Language, Learning and Technology* 22, no. 1: 27–41.
- Li, J., S. Link, and V. Hegelheimer. 2015. "Rethinking the Role of Automated Writing Evaluation (AWE) Feedback in ESL Writing Instruction." *Journal of Second Language Writing* 27: 1–18. <https://doi.org/10.1016/j.jslw.2014.10.004>.
- Liao, H. C. 2016. "Using Automated Writing Evaluation to Reduce Grammar Errors in Writing." *ELT Journal* 70, no. 3: 308–319. <https://doi.org/10.1093/elt/ccv058>.
- Link, S., M. Mehrzad, and M. Rahimi. 2022. "Impact of Automated Writing Evaluation (AWE) on Teacher Feedback, Student Revision, and Writing Improvement." *Computer Assisted Language Learning* 35, no. 4: 605–634. <https://doi.org/10.1007/s10639-022-11458-x>.
- Liu, W., and X. Liu. 2025. "A Comparative Analysis of Syntactic Complexity in Argumentative Essays From a Rhetorical Perspective: ChatGPT vs. English Native Speakers." *PLoS One* 20, no. 8: e0329410. <https://doi.org/10.1371/journal.pone.0329410>.
- Lyster, R., and L. Ranta. 1997. "Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms." *Studies*

- in *Second Language Acquisition* 19, no. 1: 37–66. <https://doi.org/10.1017/S0272263197001034>.
- McCarthy, P. M., and S. Jarvis. 2010. “MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment.” *Behavior Research Methods* 42: 381–392. <https://doi.org/10.3758/BRM.42.2.381>.
- Mindner, L., T. Schlippe, and K. Schaaff. 2023. “Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT.” In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, edited by T. Schlippe, E. C. K. Cheng, and T. Wang, 152–170. Springer Nature. https://doi.org/10.1007/978-981-99-7947-9_12.
- Mizumoto, A., and M. Eguchi. 2023. “Exploring the Potential of Using an AI Language Model for Automated Essay Scoring.” *Research Methods in Applied Linguistics* 2, no. 2: 100050. <https://doi.org/10.1016/j.rmal.2023.100050>.
- Mizumoto, A., N. Shintani, M. Sasaki, and M. F. Teng. 2024. “Testing the Viability of ChatGPT as a Companion in L2 Writing Accuracy Assessment.” *Research Methods in Applied Linguistics* 3, no. 2: 100116. <https://doi.org/10.1016/j.rmal.2024.100116>.
- Moorkens, J. 2018. “What to Expect From Neural Machine Translation: A Practical In-Class Translation Evaluation Exercise.” *Interpreter and Translator Trainer* 12, no. 4: 375–387. <https://doi.org/10.1080/1750399X.2018.1501639>.
- Norris, J. M., and L. Ortega. 2009. “Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity.” *Applied Linguistics* 30, no. 4: 555–578. <https://doi.org/10.1093/applin/amp044>.
- O'Connor, S. 2023. “Open Artificial Intelligence Platforms in Nursing Education: Tools for Academic Progress or Abuse?” *Nurse Education in Practice* 66: 103537. <https://doi.org/10.1016/j.nepr.2022.103537>.
- Ranalli, J. 2018. “Automated Written Corrective Feedback: How Well Can Students Make Use of It?” *Computer Assisted Language Learning* 31, no. 7: 653–674. <https://doi.org/10.1080/09588221.2018.1428994>.
- Ranalli, J. 2021. “L2 Student Engagement With Automated Feedback on Writing: Potential for Learning and Issues of Trust.” *Journal of Second Language Writing* 52: 100816. <https://doi.org/10.1016/j.jslw.2021.100816>.
- Ray, P. P. 2023. “ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope.” *Internet of Things and Cyber-Physical Systems* 3: 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Rudolph, J., S. Tan, and S. Tan. 2023. “ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?” *Journal of Applied Learning and Teaching* 6, no. 1: 342–363. <https://search.informit.com/doi/10.3316/informit.T2025102700003092061569891>.
- Sahu, S., Y. K. Vishwakarma, J. Kori, and J. S. Thakur. 2020. “Evaluating Performance of Different Grammar Checking Tools.” *International Journal* 9, no. 2: 2227–2233.
- Schmidt-Fajlik, R. 2023. “ChatGPT as a Grammar Checker for Japanese English Language Learners: A Comparison With Grammarly and ProWritingAid.” *AsiaCALL Online Journal* 14, no. 1: 105–119. <https://doi.org/10.54855/acoj.231417>.
- Steiss, J., T. Tate, S. Graham, et al. 2024. “Comparing the Quality of Human and ChatGPT Feedback of Students’ Writing.” *Learning and Instruction* 91: 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>.
- Stevenson, M., and A. Phakiti. 2014. “The Effects of Computer-Generated Feedback on the Quality of Writing.” *Assessing Writing* 19: 51–65. <https://doi.org/10.1016/j.asw.2013.11.007>.
- Stokel-Walker, C. 2023. “ChatGPT Listed as Author on Research Papers: Many Scientists Disapprove.” *Nature* 613, no. 7945: 620–621. <https://doi.org/10.1038/d41586-023-00107-z>.
- Su, Y., Y. Lin, and C. Lai. 2023. “Collaborating With ChatGPT in Argumentative Writing Classrooms.” *Assessing Writing* 57: 100752. <https://doi.org/10.1016/j.asw.2023.100752>.
- Tsai, C. Y., Y. T. Lin, and I. K. Brown. 2024. “Impacts of ChatGPT-Assisted Writing for EFL English Majors: Feasibility and Challenges.” *Education and Information Technologies* 29, no. 17: 22427–22445. <https://doi.org/10.1007/s10639-024-12722-y>.
- van Dis, E. A., J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting. 2023. “ChatGPT: Five Priorities for Research.” *Nature* 614, no. 7947: 224–226. <https://doi.org/10.1038/d41586-023-00288-7>.
- Wang, Y. J., H. F. Shang, and P. Briody. 2013. “Exploring the Impact of Using Automated Writing Evaluation in English as a Foreign Language University Students’ Writing.” *Computer Assisted Language Learning* 26, no. 3: 234–257. <https://doi.org/10.1080/09588221.2012.655300>.
- Warschauer, M., and D. Grimes. 2008. “Automated Writing Assessment in the Classroom.” *Pedagogies: An International Journal* 3, no. 1: 22–36. <https://doi.org/10.1080/15544800701771580>.
- Weigle, S. C. 2002. *Assessing Writing*. Cambridge University Press.
- Wilson, J., A. Potter, T. C. Cordero, and M. C. Myers. 2023. “Integrating Goal-Setting and Automated Feedback to Improve Writing Outcomes: A Pilot Study.” *Innovation in Language Learning and Teaching* 17, no. 3: 518–534. <https://doi.org/10.1080/17501229.2022.2077348>.
- Yan, D. 2023. “Impact of ChatGPT on Learners in an L2 Writing Practicum: An Exploratory Investigation.” *Education and Information Technologies* 28: 13943–13967. <https://doi.org/10.1007/s10639-023-11742-4>.
- Zhan, Y., D. Boud, P. Dawson, and Z. Yan. 2025. “Generative Artificial Intelligence as an Enabler of Student Feedback Engagement: A Framework.” *Higher Education Research and Development* 44, no. 5: 1289–1304. <https://doi.org/10.1080/07294360.2025.2476513>.
- Zhang, Z., and K. Hyland. 2018. “Student Engagement With Teacher and Automated Feedback on L2 Writing.” *Assessing Writing* 36: 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>.
- Zhou, T., S. Cao, S. Zhou, Y. Zhang, and A. He. 2023. “Chinese Intermediate English Learners Outdid ChatGPT in Deep Cohesion: Evidence From English Narrative Writing.” *System* 118: 103141. <https://doi.org/10.1016/j.system.2023.103141>.