International Journal of TESOL Studies (2026)
260802, 1-22 https://doi.org/10.58304/ijts.260802

Article

Differential Effects of Generative AI Editing on the Language Quality of Essays across Proficiency Levels

Tho Doan Vo
University of Economics Ho Chi Minh City, Vietnam

Jessie S. Barrot*
University of Economics Ho Chi Minh City, Vietnam
National University, Manila, Philippines

Received: 1 May 2026 / Received in revised form: 22 July 2026 / Accepted: 24 July 2026 / Available online: 11 August 2026

Abstract
While numerous studies have explored the affordances of generative artificial intelligence (AI) for writing, few have empirically examined its automated language editing capabilities. To address this gap, this paper examines this feature and its potential to level the playing field in academic writing across different proficiency levels. In particular, we investigate the extent to which ChatGPT enhances the linguistic quality of essays and reduces disparities in language quality among students with varying proficiency levels. A total of 200 essays (50 from each proficiency level, namely A2-0, B1-1, B1-2, and B2-0) were selected from ICNALE, ensuring equal representation across 10 Asian countries. The results indicate that ChatGPT significantly improves language quality in six of the seven assessed indicators, with these improvements occurring consistently across all proficiency levels. Furthermore, the findings demonstrate its ability to minimise disparities in language quality among essays written by students with different proficiency levels. The study concludes with a discussion of the implications of these findings for writing instruction, assessment, and research.

Keywords
ChatGPT, artificial intelligence, automated assessment, automated writing evaluation, writing accuracy

1 Introduction

Recent developments in artificial intelligence (AI) have given rise to new technologies that provide automated written corrective feedback (AWCF) and engage users in audio- or text-based human-like conversations (Abudalfa & Barrot, 2026; Ding & Zou, 2024; Shi & Aryadoust, 2024). One such technology is generative AI, which is an AI-powered chatbot that responds to questions and prompts,

*Corresponding author. Email: barrotjessie@ueh.edu.vn

admits mistakes and limitations, and refuses inappropriate requests (Feuerriegel et al., 2024; Ronge et al., 2025; Teng 2024a). Within the context of academic writing, generative AI offers a broad range of affordances such as listing essay topics, creating an outline using a preferred format, and providing quantitative and qualitative feedback on written texts (Alsofyani & Barzanji, 2025; Banihashem et al., 2024; Flodén, 2025; Teng, 2024b). Hence, many scholars and practitioners have explored its affordances for language teaching and learning purposes (e.g., Moorhouse et al., 2025; Shin & Lee, 2025; Su et al., 2023). Among the growing number of generative AI applications, ChatGPT has emerged as one of the most widely adopted tools because it integrates many of these writing-related affordances within a single conversational interface. Its increasing use in educational settings has shifted scholarly attention from examining the general potential of generative AI to evaluating the effectiveness of specific writing functions that it performs.

One recent tool that can provide AWCF is ChatGPT, an AI-powered chatbot developed by OpenAI and launched in November 2022. It allows users to engage in interactive conversations by responding to prompts and questions. Powered by computational techniques and a large language model (LLM), this tool can process complex information and engage in human-like conversations appropriately and contextually (OpenAI, 2022). It remembers previous conversations, allowing users to ask follow-up questions and engage in sustained, coherent interaction. It can also define and explain terms in context, write a dialogue tailored to learners' levels, provide vocabulary notes, and generate comprehension questions for reading activities. These affordances make ChatGPT and other similar chatbots a good source of language input for language learning (Huang et al., 2022; Jeon, 2024).

One important writing affordance of generative AI like ChatGPT is its automated language editing, which checks the cohesion, coherence, syntax, vocabulary, mechanics, and grammatical accuracy of a given text. Although some papers addressed this feature, most are at the conceptual level only (e.g., Barrot, 2023a; Otmar et al., 2025; Whang, 2024). Among those who attempted to evaluate this affordance were Lingard et al. (2023) and Alyasiri et al. (2024). However, their studies yielded mixed results: Lingard et al. (2023) concluded that ChatGPT did not perform well as a language editor because it occasionally altered the author's intended meaning and introduced unnecessarily elaborate constructions, whereas Alyasiri et al. (2024) found ChatGPT-4 to be an effective tool for proofreading, grammatical correction, paraphrasing, and language editing. Furthermore, neither study used quantitative measures of language quality to empirically assess the tool's efficacy as an automated language editing tool. This line of research is critical as it may have significant implications for writing instruction, teaching of form, writing practices, and writing assessment. Although recent studies have documented the affordances of ChatGPT for writing support, relatively little attention has been given to the theoretical mechanisms that explain why improvements may occur in some aspects of language quality but not others (Teng & Huang, 2026).

Thus, this paper examines the differential effects of ChatGPT editing on the language quality of essays across proficiency levels through the lens of Probabilistic Language Modelling (PLM) theory (Jurafsky & Martin, 2025). PLM explains not only how ChatGPT generates language through probabilistic prediction but also why its revisions influence learning differently across language dimensions. Because the system relies on statistical regularities rather than explicit linguistic reasoning, its edits may improve the quality of written products without necessarily promoting language development. Undertaking this study is important for several reasons. First, it contributes to the growing body of research on AI-assisted writing by providing quantitative evidence on how automated language editing influences multiple dimensions of language quality across proficiency levels. This multidimensional analysis extends existing evaluations of AI-assisted writing, which have largely relied on qualitative observations or focused on specific aspects of language editing. Second, it provides empirical evidence that ChatGPT reduces measurable disparities in language quality among writers with varying proficiency levels, contributing to ongoing discussions of educational equity in AI-assisted writing. Finally, the findings offer new pedagogical insights into the specific language dimensions that

benefit most from AI-assisted editing and those that continue to require human guidance. This supports more informed integration of generative AI into writing instruction and assessment.

2 Literature Review

2.1 Automated written corrective feedback (AWCF)

One recent writing technology that has emerged is AWCF, which employs latent semantic analysis, AI, and natural language processing techniques to provide real-time, adaptive feedback on any written text (Barrot, 2023b; Bai & Hu, 2017; Ranalli, 2018; Teng 2024b). It refers to the written corrective feedback provided by automated writing evaluation tools or AI-powered technologies on the quality of a written work. Several recent studies have provided evidence of AWCF's viability in improving writing performance, facilitating revision, enhancing motivation, and assisting instruction (Link et al., 2022; Teng & Shen, 2025; Waer, 2023; Wang et al., 2020; Zhang, 2020). However, AWCF has several limitations. It may overcorrect errors and provides largely standardised feedback that does not adequately accommodate individual learner characteristics, such as differences in L2 proficiency, first language background, writing ability, educational background, and metalinguistic knowledge (Ranalli, 2018). It also lacks a human touch, such as giving encouraging words, acknowledging individual strengths and weaknesses, and showing empathy.

Although studies on the efficacy of AWCF as a language editing tool are emerging, they have yielded mixed results. For instance, Chen et al. (2009) examined grammar feedback from MY Access and Criterion, two of the well-known AWCF tools at the time. Using 269 essays, they found these two systems capable of providing about 30 types of grammar feedback, but with less than 80 percent accuracy. Despite this, MY Access generated several false alarms, while Criterion missed many errors in verb tense, modals, collocations, pronouns, word order, and word choice. From these findings, they posited that these tools lacked the ability to support students' English learning. Since statistical significance was not tested in their study, it is difficult to conclude whether the systems were indeed effective or not.

More recently, Ranalli and Yamashita (2022) examined the error-correction performance of two AWCF tools, namely Grammarly and MS-NLP. Using a corpus of 68 essays written by international students from a large research university in the United States, they found that both tools met the 80 percent benchmark set for precision and correction rate. However, Grammarly performed better than MS-NLP by flagging a more diverse (114 error types) and larger (1,412 items) set of error types. Variations in their performance were attributed to the differences in the algorithms used by these two AWCF tools. While the study provided important insights into the robustness of Grammarly and MS-NLP, the analysis mainly focused on the correctness of language, which is one of the many aspects of writing issues that an AWCF can identify. Information about false alarms or overcorrection was also limited.

To address these limitations, John and Woll (2020) conducted a parallel study that investigated the efficacy of Grammarly, Virtual Writing Tutor, and Microsoft Word in checking grammar. They evaluated the three tools in terms of the number of errors flagged (coverage), appropriateness of proposed corrections, and rates of forms mistakenly flagged as incorrect (overcorrection) in 50 essays and 129 sentences. Findings indicated that none of these three grammar checkers reached an overall coverage of at least 50 percent, rendering them far from entirely replacing human feedback. However, results showed some advantages of using the online AWCF tools (i.e., Grammarly and Virtual Writing Tutor) over Microsoft Word. False alarms were also observed across three tools, with Virtual Writing Tutor registering the highest frequency (30), followed by Microsoft Word (13) and Grammarly (4). As the authors acknowledged, their descriptive study did not attempt to establish statistical significance on the efficacy of AWCF tools as a language editing tool. Similar to Ranalli and Yamashita (2022), John and Woll (2020) limited their analysis of false alarms to overcorrection.

In recent years, we have also witnessed notable improvements in generative AI-based AWCF. For example, Guo (2025) found that ChatGPT provides diverse lexical, grammatical, structural, and content-related feedback that supports foreign language writing development, while Wu et al. (2025) demonstrated that GPT-4 enhanced with chain-of-thought prompting generates more accurate and consistent written corrective feedback than commercial AWCF tools. Nevertheless, these studies also reported persistent challenges, including occasional overcorrection and varying learner perceptions of AI-generated feedback, suggesting that further refinement is still needed.

2.2 ChatGPT and academic writing

As a writing tool, ChatGPT can perform a wide range of writing-related tasks to address the diverse needs of writers, including adaptive and real-time feedback, dependable writing assistance, and a writing platform (Imran & Almusharraf, 2023; Mizumoto et al., 2024; Teng, 2024a). As a writing assistant, it can produce short and long texts across a wide range of subjects, using coherent, grammatically correct language. It can also generate a list of relevant topics and narrow them down based on the user's prompts. This affordance is particularly useful when students experience writer's block. Once a specific topic is selected, the system can be prompted to create an outline in a preferred format (sentence or topic outline; alphanumeric or decimal system). Students can use this outlining feature to further refine their initial outline. As a writing assessment tool, ChatGPT can perform automated essay scoring using a set of criteria and explain how it arrives at each score (Bui & Barrot, 2025a, 2025b; Gjorevski et al., 2025; Latif & Zhai, 2024). It can also provide feedback on the content and organisation of essays, highlighting their strengths and areas for improvement, and offering recommendations to enhance the text. Recent empirical evidence also supports the viability of ChatGPT for L2 writing assessment, demonstrating that LLMs can produce reasonably dependable evaluations of writing accuracy while highlighting important considerations regarding validity and human oversight (Mizumoto et al., 2024). Despite the above features, some scholars and practitioners raised some concerns about the possible adverse impact of the tool on academic integrity, educational equity, and writing practices of writers (Sun & Hoelscher, 2023; Yan, 2023). Some schools and universities even banned the use of this AI tool due to concerns over cheating and negative effects on students' creativity and critical thinking (Johnson, 2023; Roose, 2023). The tool also has some limitations, such as generating inaccurate responses (e.g., citations and referencing) and the inability to detect plagiarised texts (Barrot, 2023a).

Most existing studies have primarily described ChatGPT's writing affordances from a functional perspective, with limited attention to the computational principles that underlie these capabilities (Imran & Almusharraf, 2023; Teng, 2024a). PLM conceptualises language editing as probabilistic prediction based on statistical regularities learned from large language corpora (Jurafsky & Martin, 2025). This perspective suggests that ChatGPT provides corrective feedback through pattern optimisation rather than explicit language instruction. Consequently, improvements in AI-edited texts do not necessarily indicate that learners notice, understand, or internalise the linguistic changes.

In recent years, empirical studies have increasingly examined the effectiveness of ChatGPT in supporting L2 writing, with generally positive but nuanced findings. Evidence indicates that ChatGPT can improve multiple aspects of writing, including idea generation, lexical diversity, grammatical accuracy, coherence, and writing efficiency, while providing immediate and personalised feedback that supports revision and learner engagement (Song & Song, 2023; Xiao et al., 2026). Studies also suggest that learners perceive ChatGPT as a useful writing partner for brainstorming, drafting, paraphrasing, proofreading, and revising texts, particularly because of its ability to provide timely and accessible feedback (Yan, 2023; Xiao et al., 2026). Furthermore, recent investigations have demonstrated the potential of ChatGPT to function as an automated writing assessment tool capable of detecting grammatical, lexical, and stylistic problems and generating feedback that approaches the quality of human evaluation for many surface-level features of writing (Al-Garaady & Mahyoob, 2023; Mizumoto

& Eguchi, 2023). Nevertheless, empirical findings also reveal important limitations. ChatGPT remains less effective in evaluating higher-order aspects of writing, such as argument development, rhetorical appropriateness, creativity, and contextual meaning, and concerns persist regarding students' overreliance on AI-generated suggestions, diminished authorial voice, hallucinated content, and academic integrity (Al-Garaady & Mahyoob, 2023; Barrett & Pack, 2023; Xiao et al., 2026). These findings indicate that although ChatGPT can substantially enhance the linguistic quality of written texts, its contributions are best understood as computational optimisation of language patterns rather than evidence of learners' acquisition of underlying writing competence.

2.3 ChatGPT as a language editing tool

One notable feature of ChatGPT is its automated language editing capability. The system directly corrects errors and misuses in grammar, syntax, mechanics, language style, cohesion, coherence, semantics, and vocabulary using a specific prompt. Its response as an editor can be shaped by the prompt provided by the user (Barrot, 2026), making it more flexible than other automated editing tools that will just select words or strings of text, point out the "error" and suggest a revision (e.g., Grammarly). There have been papers that have explored ChatGPT's editing feature, but most of them are at the conceptual level (e.g., Barrot, 2023a; Castillo-González et al., 2022; Kim, 2023; Otmar et al., 2025; Whang, 2024). These conceptual papers highlighted ChatGPT's ability to detect and correct grammatical errors, improve text coherence and clarity, process large amounts of text, and generate additional content. On one hand, Kim (2023) noted the tool's tendency to generate false information during the editing process. These claims remain to be empirically validated.

Currently, very few studies have empirically examined this feature, leading to an insufficient understanding of the tool's potential, a lack of robust evidence, and an inability to generalise results across different contexts. One of them is the study by Lingard and her colleagues (2023), who evaluated ChatGPT's performance as a language editor and writing coach as a way to level the playing field for non-native English writers in academic publishing. Using experiential evidence on ChatGPT's effectiveness, their findings indicated that the tool did not excel as a language editing tool because it introduced elaborate constructions and altered the writer's intended meaning. The study highlighted the importance of looking critically at ChatGPT's edits rather than accepting them at face value. Alyasiri et al. (2024) also ventured into this underexplored area by evaluating ChatGPT-4's capabilities in detecting and correcting grammatical errors within academic writing. Unlike Lingard et al.'s findings, Alyasiri and his colleagues considered ChatGPT-4 as a proficient language editor. They found ChatGPT-4 as a useful tool in proofreading academic papers (spelling, grammar, punctuation, and syntax), summarising and paraphrasing, content editing, and references formatting and style.

Recent evidence has begun to broaden this limited empirical base. Nadra and Purdy (2026) compared eight generative AI platforms, including multiple versions of ChatGPT, Claude, Gemini, and Copilot, in editing academic writing against a human baseline. Their results showed that, overall, generative AI systems identified substantially fewer sentence-level errors than human editors, with most platforms detecting only about half of the errors identified by humans. Moreover, editing performance varied considerably across language models, indicating that editing quality depends not only on the presence of AI assistance but also on the specific model used. The study further demonstrated that although author expertise, draft stage, and rubric availability did not significantly influence the total number of errors detected, these factors affected the types of errors identified. Across platforms, AI systems tended to overidentify structural and punctuation-related issues while underidentifying context-dependent features, particularly capitalisation errors, quotation integration, and documentation problems.

While these studies collectively demonstrate the promise of ChatGPT as an automated language editing tool, the evidence remains limited and inconclusive. Lingard et al. (2023) and Nadra and Purdy (2026) reported that generative AI still falls short of human editing performance and may introduce

unnecessary revisions or fail to detect context-sensitive errors. In contrast, Alyasiri et al. (2024) highlighted ChatGPT-4's effectiveness for proofreading and surface-level language correction. More importantly, existing empirical studies have relied predominantly on users' qualitative evaluations of AI-generated revisions or comparisons with human judgements. Few studies have employed objective linguistic indicators to quantify improvements in language quality following AI-assisted editing. Consequently, there remains insufficient empirical evidence regarding the magnitude of ChatGPT's effects on specific dimensions of language quality and the extent to which these improvements generalise across different writing contexts and learner populations. This limitation underscores the need for quantitative investigations that systematically measure changes in linguistic quality after AI-assisted language editing.

2.4 Theoretical framework

This study is framed within PLM theory, which provides the computational foundation for understanding how generative AI systems such as ChatGPT process, generate, and refine language. The framework extends beyond a description of how ChatGPT functions computationally. It also provides the theoretical rationale for the selection of language quality variables, the formulation of the research questions, and the interpretation of the empirical findings. PLM posits that linguistic behaviour can be modelled through statistical inference, where the probability of a given linguistic unit, such as a word, phrase, or syntactic structure, is conditioned on preceding elements (Jurafsky & Martin, 2025). Rather than relying on explicit grammatical rules or human-like comprehension, PLM-based systems predict the most likely continuations of text sequences using massive language corpora. Through this mechanism, ChatGPT learns to optimise outputs for fluency, coherence, and syntactic correctness, which underpins its performance in language editing and text enhancement tasks. Recent corpus evidence lends empirical support to this probabilistic view of language generation. Comparing more than three million words of human- and ChatGPT-generated academic writing, Zhou et al. (2026) demonstrated that ChatGPT consistently favoured conventional lexical combinations and highly predictable collocational patterns, whereas human writers employed greater lexical flexibility and contextual variation.

From a linguistic standpoint, PLM aligns with corpus-driven and usage-based perspectives on language, which view linguistic competence as emerging from the frequency and distribution of language forms in use (Bybee, 2006). ChatGPT's probabilistic nature enables it to detect grammatical inconsistencies, enhance lexical variation, and improve stylistic fluency, effectively mirroring the processes of statistical learning observed in human cognition. However, this computational strength also introduces epistemological and pedagogical limitations. Because PLM systems rely on surface-level regularities rather than deep semantic or rhetorical understanding, they may overlook contextual nuance, misinterpret pragmatic intent, or produce stylistically "standardised" text that lacks authorial voice (Bryant et al., 2019).

In the context of language editing and academic writing, these characteristics have significant implications. On the one hand, ChatGPT can scaffold linguistic accuracy and promote awareness of form-based conventions; on the other hand, its reliance on probabilistic generalisations may reinforce dominant discourse norms and obscure the creative or disciplinary dimensions of writing. This tension underscores the need to examine how users negotiate between AI-generated linguistic output and their own rhetorical intentions. Consequently, the PLM framework serves as both a computational and pedagogical lens for this study: computationally, it explains how ChatGPT performs linguistic refinement through probability estimation; pedagogically, it frames the implications of such algorithmic mediation for learning, authorship, and the development of writing proficiency.

Guided by PLM theory, the present study conceptualises language quality as a multidimensional construct comprising both statistically predictable language features and higher-order discourse qualities. This perspective predicts stronger improvements in correctness, vocabulary, delivery, and engagement

than in clarity because these dimensions differ in the extent to which they depend on recurrent linguistic regularities that probabilistic language models can estimate reliably. In contrast, clarity depends more heavily on semantic precision, discourse organisation, audience awareness, and communicative intent, which extend beyond statistical optimisation alone. PLM further predicts that the same probabilistic optimisation process will be applied consistently across texts regardless of writer proficiency. Consequently, AI-assisted editing should reduce disparities in language quality across proficiency levels by bringing essays closer to conventional patterns of academic English. These theoretical propositions provide the conceptual basis for the selection of language quality variables, the formulation of the research questions, and the interpretation of the study's findings. Situating ChatGPT's language-editing capacities within PLM theory contributes to broader discussions in applied linguistics and information science about how statistical language models reshape the processes and practices of text production in educational contexts.

Although PLM explains how ChatGPT produces linguistically refined texts, its educational value becomes clearer when interpreted through theories of L2 writing and corrective feedback. Corrective feedback promotes language development when learners notice linguistic problems, understand the rationale for revisions, and subsequently apply those forms in future writing (Ellis, 2009). Within this perspective, ChatGPT functions as an automated source of corrective feedback that provides immediate language support and scaffolds revision, particularly for grammatical accuracy, lexical choice, and sentence formulation. From a sociocultural perspective, this support resembles temporary scaffolding that enables learners to perform writing tasks beyond their current level of independent performance. However, improvements in AI-edited texts should not be interpreted as evidence of language acquisition. Improved writing products do not necessarily indicate that learners have internalised the linguistic knowledge underlying the revisions or can independently transfer those forms to subsequent writing. Long-term language development requires reflection, deliberate practice, and independent application of feedback rather than exposure to corrected text alone (Ellis, 2009). PLM, Sociocultural Theory, and corrective feedback theory explain complementary aspects of AI-assisted writing. PLM accounts for the computational processes that generate probabilistic language revisions, Sociocultural Theory explains how AI functions as a mediating tool that supports learner agency through scaffolded assistance, and corrective feedback theory explains the conditions under which learners notice, interpret, and subsequently use those revisions for language development. Consequently, the present study conceptualises ChatGPT as an AI-mediated language editing and feedback tool that supports higher-quality writing products while recognising that improvements in writing development depend on learners' engagement with AI-generated feedback rather than the feedback itself.

2.5 Rationale for the study

Despite growing interest in the use of generative AI for writing support, empirical evidence regarding its automated language-editing capability remains limited. Existing studies have produced inconsistent findings, with some reporting that ChatGPT may alter authors' intended meaning while others conclude that it is an effective language-editing tool. Furthermore, these studies have relied primarily on qualitative observations or experiential evidence, providing little quantitative evidence of how ChatGPT affects specific dimensions of language quality. Little is also known about whether its language-editing effects differ across learners with varying levels of English proficiency or whether it reduces disparities in language quality among writers of different proficiency levels. To address these gaps, the present study quantitatively examines the effects of ChatGPT's language editing on multiple dimensions of language quality and investigates whether these effects vary across proficiency levels. Thus, we addressed the following research questions:

1. How much does ChatGPT improve the language quality of essays?
2. Is there a significant difference in the language quality of students' essays across proficiency levels after ChatGPT editing?

3 Materials and Methods

3.1 Data source

This study analysed 200 essays, which were randomly selected from the International Corpus Network of Asian Learners of English (ICNALE). ICNALE is a learner corpus of 5,600 essays written by A2-0 (Waystage), B1-1 (Threshold, lower), B1-2 (Threshold, upper), and B2-0 (Vantage) college students from 10 Asian countries, namely Hong Kong, Pakistan, the Philippines, Singapore, Indonesia, Japan, China, South Korea, Thailand, and Taiwan (Ishikawa, 2013). ICNALE essays employed a controlled writing task in a computer laboratory using the following details (Ishikawa, 2013, p. 97): writing prompt (Do you agree or disagree with the following statement? Use reasons and specific details to support your opinion: It is important for college students to have a part-time job), text type (argumentative essay), time allotment for writing (20 to 40 minutes), length (200 to 300 words), and writing platform (word processor). Students were also prohibited from using a dictionary during the writing task to avoid any unnecessary influence on their written output.

A total of 200 essays were selected through stratified random sampling, with proficiency level serving as the primary stratum. Within each proficiency level, 50 essays were randomly selected, with equal representation from the 10 participating countries (five essays per country), thereby ensuring that no single national or linguistic group disproportionately influenced the findings.

These proficiency levels align with the Common European Framework of Reference (CEFR), a widely recognised scale for assessing language proficiency. A2-0 (Waystage) represents learners with basic proficiency, capable of understanding and using simple expressions. B1-1 (Threshold, lower) and B1-2 (Threshold, upper) correspond to intermediate users who can communicate effectively in familiar situations but may still struggle with complexity. B2-0 (Vantage) denotes an upper-intermediate level, where learners demonstrate greater fluency, coherence, and grammatical control in their writing. These classifications provide a structured framework for analysing how ChatGPT's editing performance varies across different proficiency levels.

Although the sampling procedure balanced country representation across proficiency levels, the study did not explicitly control for first language background, educational context, or country-specific writing conventions. These factors may influence learners' linguistic features and error patterns and, consequently, the types of revisions generated by ChatGPT. Therefore, the findings should be interpreted as reflecting overall trends across a multilingual Asian learner corpus rather than country-specific patterns.

3.2 ChatGPT editing procedure

Each essay was subjected to ChatGPT-4 language editing using the following prompt: You are an academic English editor. Edit the text below for language only. Improve grammar, punctuation, spelling, word choice, and sentence structure; enhance coherence and cohesion where possible without adding new ideas. Preserve the original meaning, stance, and level of detail. Do not introduce new examples, claims, or information. Do not change the length substantially. This prompt was selected to increase procedural transparency and replicability and to ensure that the task aligned with the study's focus on language quality rather than content development.

The same ChatGPT-4 model was used for all essays throughout the study period. To minimise potential carryover effects, each essay was edited in a new chat session with no prior context. The identical prompt was applied to every essay, and no follow-up prompts or iterative revisions were used. Model settings remained constant across all editing sessions. Although ChatGPT does not provide users with direct control over generation parameters such as temperature, the use of a single model version, a standardised prompt, and separate chat sessions enhanced procedural consistency and supported the

reproducibility of the editing process. The edited output was recorded verbatim and paired with the original version for subsequent analysis.

3.3 Language quality measurement

The language quality of the unedited and edited essays was operationalised as surface-level linguistic accuracy and appropriacy (e.g., grammar, mechanics, and local phrasing) plus lexical sophistication (lexical diversity and lexical rarity). Because the dataset included 200 essays and multiple proficiency strata, we used an automated assessment tool to ensure consistent, scalable scoring across the full sample while maintaining standardised evaluation conditions for all texts. We used Grammarly Premium (https://www.grammarly.com/) as an automated tool for detecting language issues and lexical features. Based on the errors it identifies, Grammarly provides an overall text score (OTS) ranging from 1 to 100. This means that a text that received many suggestions and corrections gets a lower score. Conversely, a text that receives fewer suggestions and corrections scores higher overall. Grammarly also checks and counts specific linguistic features of a written text, such as vocabulary (unique words and rare words) and writing issues (clarity, delivery, correctness, and engagement). Under vocabulary, unique words refers to vocabulary diversity measured by the percentage of words used only once in the text. Meanwhile, rare words refers to vocabulary depth measured by the words that do not belong to the 5,000 most common English words. For writing issues, Grammarly counts problems in correctness (presence or absence of errors in spelling, grammar, and punctuation), clarity (keeping the messages concise and readable), delivery (tone, politeness, and formality of language), and engagement (use of language that makes the text more interesting).

Grammarly was selected because it provides (a) stable, replicable output when identical text is submitted under the same settings and (b) fine-grained, count-based indicators that can be analysed quantitatively (Barrot, 2023b; Gain et al., 2019; Thi & Nikolov, 2022). Its indicators broadly align with language constructs commonly examined in L2 writing research, although they operationalise these constructs through proprietary algorithms. Correctness reflects grammatical accuracy, spelling, and punctuation. Clarity represents sentence-level readability and conciseness. Delivery captures tone, formality, and stylistic appropriateness. Engagement reflects language choices that enhance reader interest. Vocabulary is represented through two lexical indicators. Unique words refer to the proportion of words that occur only once in a text and serve as a proxy for lexical diversity. Rare words refer to lexical items outside Grammarly's list of the 5,000 most frequent English words and serve as a proxy for lexical sophistication. These indicators capture lexical variation rather than vocabulary development or proficiency. A higher proportion of rare words should not be interpreted as better writing because lexical quality also depends on contextual appropriateness and semantic precision. Accordingly, these indicators were treated as proxy measures of specific language dimensions rather than direct measures of writing proficiency. Grammarly does not evaluate higher-order dimensions such as argument quality, content development, rhetorical organisation, audience awareness, cohesion, or authorial voice. Therefore, the findings should be interpreted as improvements in measurable language quality rather than comprehensive improvements in writing quality.

To address concerns about automated scoring validity and potential "black-box" bias, we implemented a human validation procedure. An expert in academic writing instruction and second language writing independently reviewed a stratified subset of Grammarly flags across proficiency levels to determine whether each flag represented a legitimate language problem and an appropriate correction. The review focused on language features prone to automated misclassification, including passive voice, technical terminology, acceptable non-native constructions, and stylistic variation. Only validated flags were retained for analysis. The analysis revealed 98% agreement between Grammarly's classifications and the expert review, indicating high reliability. Most discrepancies involved Grammarly's tendency to overcorrect passive voice, technical terminology, and word choice. Because the validation involved only

one expert, inter-rater reliability could not be established. Future studies should include multiple expert raters to strengthen the validity of the measurement procedure.

Finally, because Grammarly's algorithms are proprietary and may privilege standardised norms of English, we explicitly interpret results as improvements in standardised language-form measures aligned with general academic English conventions, rather than as universal improvements in "good writing." This interpretation is consistent with the study's aim: to quantify whether generative AI editing reduces measurable language-form disparities across proficiency levels.

3.4 Statistical analysis

All quantitative data were entered into SPSS version 21 and were subjected to descriptive and inferential statistics. To answer research question 1, we used the mean (M) and standard deviation (SD) to determine the average scores for the following indicators: OTS, unique words, rare words, clarity, delivery, correctness, and engagement. Then, we employed a paired t-test to assess the difference in language quality between unedited and edited essays across proficiency levels. To answer research question 2, we used analyses of variance (ANOVAs) for each component of language quality within each essay type (i.e., unedited and edited). We also ran univariate analyses and pairwise comparisons to determine whether the language quality component differed between each pair of proficiency levels for each essay type. Because of the multiple analyses involved, we applied a Bonferroni correction to control the risk of false-positive conclusions.

4 Results

4.1 Impact of ChatGPT on the language quality of essays

Table 1 presents the impact of ChatGPT editing on the seven language quality areas of 200 essays written by college students. Overall, ChatGPT editing led to substantial improvements in six of the seven language quality areas, as indicated by medium to large effect sizes and statistical significance. *Overall Text Score* (t[199] = 62.024, $p < .001$, $d = 4.39$), *Unique Words* (t[199] = 23.441, $p < .001$, $d = 1.66$), *Rare Words* (t[199] = 30.662, $p < .001$, $d = 2.17$) registered a significant improvement. A significant decrease in writing issues related to *Correctness* (t[199] = -33.540, $p < .001$, $d = 2.37$), *Delivery* (t[199] = -12.627, $p < .001$, $d = 0.89$), and *Engagement* (t[199] = -9.880, $p < .001$, $d = 0.70$) was also observed. Meanwhile, ChatGPT's editing did not significantly impact the clarity of essays (t[199] = 1.038, $p = 0.301$, $d = 0.07$), which is an area that may require more nuanced human intervention.

Table 1
Differences in Language Quality of Unedited and ChatGPT-edited Essays

Areas	Unedited		Edited		*p*-value	Cohen's *d*
	M	*SD*	*M*	*SD*		
OTS	31.8	10.4	84.8	7.3	< .001	4.39
UW	48.5	5.6	57.8	4.9	< .001	1.66
RW	16.5	5.3	29.3	4.0	< .001	2.17
Co	27.4	11.1	1.0	1.1	< .001	2.37
Cl	5.6	2.7	5.8	2.3	.301	0.07
De	11.5	8.9	5.7	6.2	< .001	0.89
En	2.4	1.8	1.1	1.1	< .001	0.70

Notes: OTS (Overall Text Score), UW (Unique Words), RW (Rare Words), Co (Correctness Issues), Cl (Clarity Issues), De (Delivery Issues), En (Engagement Issues); significance: p < .05

Table 2 shows the language quality of unedited and ChatGPT-edited essays produced by students with different proficiency levels. When the language quality of unedited and edited essays between each proficiency level is compared, the data reveal that *Overall Text Score*, *Unique Words*, *Rare Words*, *Correctness*, and *Delivery* significantly improve after ChatGPT editing in all proficiency levels ($p < .001$) with large effect sizes. One notable finding we observed is that the correctness issues were drastically reduced after editing, from high scores in the unedited essays to near zero in the edited essays. In the case of *Engagement*, significant improvement was also observed, with medium effect sizes in A2-0, B1-1, and B1-2 and a large effect size in B2-0 essays. There were no significant differences between the unedited and edited essays in each of the four proficiency levels in terms of *Clarity*, with effect sizes ranging from very small (A2-0 and B1-2) to small (B1-1 and B2-0). The benefit of ChatGPT editing on the language quality appears to be more dramatic for those at A2-0 to B1-2. This might be because writers with higher proficiency already have a better foundation, so the AI's editing mainly refined the B2-0 essays further. Data from Tables 1 and 2 suggest that ChatGPT could be a valuable tool for language editing, especially for users looking to correct grammatical and typographical errors and enhance the sophistication of language use. However, for aspects that require a nuanced understanding of conciseness, readability, context, and intent, its utility may be limited and may require combining AI editing with human judgement.

Table 2
Language Quality of Unedited and ChatGPT-edited Essays across Proficiency Levels

Areas	Unedited Essays								Edited Essays							
	A2-0		*B1-1*		*B1-2*		*B2-0*		*A2-0*		*B1-1*		*B1-2*		*B2-0*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
OTS	28.7	9.4	28.9	10.2	29.3	9.0	40.2	8.3	86.0	7.0	84.5	8.2	83.9	7.4	84.7	6.6
UW	48.7	6.2	46.8	6.1	49.2	4.9	49.5	4.9	58.9	5.4	56.9	5.0	58.0	4.4	57.4	4.6
RW	15.9	5.0	15.0	5.1	15.8	5.5	19.2	4.8	29.1	4.3	28.9	4.1	29.3	3.7	29.7	4.1
Co	31.7	12.9	30.1	10.2	28.7	10.6	19.3	4.7	0.6	0.9	1.0	1.0	1.1	1.2	1.0	1.1
Cl	5.3	3.2	5.2	2.4	5.1	2.3	6.8	2.6	5.8	2.4	6.0	2.5	5.6	2.0	6.0	2.3
De	9.8	7.1	12.0	10.3	12.7	9.6	11.4	8.2	4.4	4.5	5.8	7.3	6.2	7.2	6.1	5.3
En	2.3	1.6	2.6	1.9	2.1	1.8	2.5	1.8	1.3	1.2	1.1	1.2	1.1	1.0	0.9	0.9

Note: OTS (Overall Text Score), UW (Unique Words), RW (Rare Words), Co (Correctness Issues), Cl (Clarity Issues), De (Delivery Issues), En (Engagement Issues)

4.2 ChatGPT's efficacy in reducing the gap in the language quality of essays produced by students across proficiency levels

The data from Table 3 provides a quantitative look at ChatGPT's efficacy in editing essays from students with varying levels of language proficiency, specifically in terms of reducing the disparities in language quality. Before editing, there were significant differences in the *Overall Text Score* ($t[196] = 18.551$, $p < .001$, $d = 0.53$), *Rare Words* ($t[196] = 6.790$, $p < .001$, $d = 0.32$), *Correctness* ($t[196] = 15.171$, $p < .001$, $d = 0.48$), and *Clarity* ($t[196] = 4.501$, $p < .001$, $d = 0.26$) between proficiency levels, with small to medium effect sizes. Pairwise comparisons further indicate that differences were significant between A2-0 and B2-0, B1-1 and B2-0, and B1-2 and B2-0. After ChatGPT's intervention, these differences are no longer statistically significant, with *p*-values well above the .05 threshold and Cohen's *d* values dropping to indicate small effect sizes. This change suggests that ChatGPT's editing had a homogenising effect on the language quality of essays across different proficiency levels. The lack of significant *p*-values in the edited essays implies that the differences in language quality between proficiency levels become negligible after ChatGPT editing. This is particularly notable in the *Overall Text Score* category, where the initial medium effect size of 0.53 drops to a small 0.10 after editing. In categories such as *Unique*

Words, *Delivery*, and *Engagement*, ChatGPT's homogenising effects remain tentative primarily because these areas were homogenous already among unedited essays, and that same homogeneity was observed after subjecting the essays to ChatGPT editing.

Table 3
Differences in Language Quality across Proficiency Levels

Areas	Unedited		Edited	
	p-value	Cohen's *d*	*p*-value	Cohen's *d*
OTS	< .001	0.53	.562	0.10
UW	.079	0.19	.209	0.15
RW	< .001	0.32	.747	0.08
Co	< .001	0.48	.118	0.17
Cl	< .001	0.26	.831	0.07
De	.405	0.12	.461	0.12
En	.598	0.10	.264	0.14

Notes: OTS (Overall Text Score), UW (Unique Words), RW (Rare Words), Co (Correctness Issues), Cl (Clarity Issues), De (Delivery Issues), En (Engagement Issues); significance: $p < .05$

5 Discussion

5.1 The extent to which ChatGPT improves the language quality of essays

The first research question examined the extent to which ChatGPT improves the language quality of students' essays. Findings revealed that ChatGPT editing significantly enhanced Overall Text Score, lexical diversity (Unique Words), lexical sophistication (Rare Words), Correctness, Delivery, and Engagement across all proficiency levels, whereas Clarity did not improve significantly. Among the specific language-issue indicators, the largest improvement occurred in Correctness, indicating that ChatGPT was particularly effective in correcting grammatical, spelling, punctuation, and sentence-level language problems. Although improvements appeared across all proficiency levels, the magnitude of change was more pronounced among students at the lower proficiency levels, suggesting that learners with greater linguistic needs benefited more from AI-assisted language editing.

These findings reinforce and extend previous research on AWCF and generative AI-assisted writing. Earlier studies on traditional AWCF systems primarily evaluated error detection, correction accuracy, coverage, and false alarms in systems such as Grammarly, Criterion, MY Access, and Microsoft Word (Chen et al., 2009; John & Woll, 2020; Ranalli & Yamashita, 2022). Although these systems demonstrated value for grammatical revision, they focused largely on isolated language errors rather than comprehensive language refinement. More recent research on generative AI indicates that ChatGPT provides broader linguistic support through grammatical correction, lexical enhancement, stylistic refinement, and personalised feedback (Guo, 2025; Teng & Huang, 2026; Wu et al., 2025). The present findings provide quantitative evidence supporting these claims. ChatGPT improved multiple measurable dimensions of language quality simultaneously rather than grammatical accuracy alone, suggesting that LLMs function as comprehensive language editors instead of conventional grammar checkers.

The results also corroborate empirical studies that evaluated ChatGPT as a language editing tool. Alyasiri et al. (2024) concluded that ChatGPT-4 effectively supports proofreading, grammatical correction, paraphrasing, and language editing, whereas Lingard et al. (2023) cautioned that AI editing occasionally alters authorial intent and introduces unnecessarily elaborate constructions. Nadra and Purdy (2026) likewise reported that although generative AI performs reasonably well in sentence-level editing, it still detects substantially fewer language problems than human editors and struggles with context-

dependent features. Our findings provide a more nuanced perspective. ChatGPT substantially improved measurable language quality, particularly correctness and vocabulary, while showing limited influence on clarity. This pattern indicates that ChatGPT functions more effectively as a language refinement system than as a comprehensive writing improvement tool.

PLM theory provides a compelling explanation for these findings (Jurafsky & Martin, 2025). The theory predicts that LLMs perform most successfully on language features governed by recurrent statistical regularities because these patterns are strongly represented in their training data. Grammar, spelling, punctuation, lexical choice, and conventional sentence structures exhibit relatively stable probabilistic distributions that language models can estimate with considerable accuracy. Consequently, ChatGPT can reliably identify and revise subject-verb agreement errors, tense inconsistencies, article misuse, preposition errors, punctuation problems, awkward phrasing, and lexical imprecision. Thus, improvements in Overall Text Score, Correctness, Unique Words, and Rare Words reflect probabilistic optimisation of highly predictable linguistic patterns rather than explicit linguistic reasoning.

The substantial reduction in correctness issues deserves particular attention because it represents the largest effect observed in the study. Grammarly operationalises correctness as grammatical accuracy, spelling, punctuation, and sentence mechanics. Consequently, the dramatic reduction in correctness issues suggests that ChatGPT effectively resolved language features explicitly targeted in the editing prompt. Academic style also improved through more conventional lexical choices and smoother sentence construction, although the prompt intentionally restricted ChatGPT from introducing new ideas or modifying the rhetorical structure of the essays. These findings align with recent systematic reviews showing that generative AI performs particularly well on surface-level language refinement while exhibiting more variable performance on higher-order writing dimensions (Abudalfa & Barrot, 2026; Xiao et al., 2026).

Interpretation of these findings should also acknowledge the characteristics of the measurement instrument. Grammarly operationalises language quality through proprietary algorithms whose computational procedures remain inaccessible. Consequently, the observed improvements cannot be attributed with certainty to specific linguistic processes because the exact scoring mechanisms remain undisclosed. Nevertheless, Grammarly's indicators broadly correspond to language constructs commonly examined in second language writing research. Correctness approximates grammatical accuracy, vocabulary provides proxy measures of lexical diversity and lexical sophistication, clarity represents sentence-level readability and conciseness, whereas delivery and engagement capture dimensions of stylistic appropriateness. These indicators do not directly assess syntactic complexity, discourse organisation, argument quality, rhetorical effectiveness, genre knowledge, or authorial voice. Therefore, the present findings should be interpreted as evidence of improvements in measurable language-form features rather than comprehensive improvements in writing quality or writing proficiency.

One notable finding concerns the absence of significant improvement in clarity despite substantial gains across the other language dimensions. Grammarly conceptualises clarity primarily through readability, conciseness, sentence structure, and ease of comprehension rather than grammatical correctness. The present findings suggest that ChatGPT successfully corrected surface-level language problems but frequently left discourse-level issues unresolved. Weak argumentation, limited explanation, insufficient elaboration, and imprecise communication often remained after editing. For instance, the sentence "I think that a part time job is important to many people while it not important to many people also" became "For some people, having a part-time job holds great significance, while for others, it may not be as important." Although grammatical accuracy and readability improved, the revision still failed to explain why opinions differed. Likewise, "Sometimes their experiences they received worth enough for their futures" became "These experiences acquired during part-time employment can hold significant value for their future careers and personal growth." The revised version demonstrates greater fluency but still lacks sufficient explanatory precision regarding the nature of those experiences. These examples illustrate that ChatGPT refined linguistic expression more effectively than rhetorical development.

This finding also aligns with PLM theory. Semantic precision, discourse organisation, rhetorical purpose, and audience awareness depend less on recurrent statistical regularities and more on contextual interpretation and communicative intention (Jurafsky & Martin, 2025). Current LLMs optimise linguistic probability rather than communicative effectiveness. Consequently, ChatGPT can substantially improve grammatical well-formedness while producing comparatively modest gains in clarity. Similar observations appear in previous research, which reported that ChatGPT performs less consistently on higher-order writing qualities than on sentence-level language features (Al-Garaady & Mahyoob, 2023; Barrett & Pack, 2023; Xiao et al., 2026).

Another noteworthy result concerns improvements in Delivery and Engagement. Although these constructs depend more heavily on stylistic and contextual appropriateness, ChatGPT nevertheless produced significant improvements in both areas. Delivery reflects tone, formality, politeness, and stylistic consistency, whereas Engagement captures lexical variety, sentence flow, and reader interest. ChatGPT appears to enhance these dimensions through probabilistic selection of conventional academic language patterns. For example, conversational expressions were consistently replaced with more formal academic constructions, repetitive wording was reduced, and lexical variety increased while preserving the original meaning. These findings further support PLM theory because conventional stylistic patterns also exhibit stable statistical regularities that LLMs can model effectively.

The findings indicate that ChatGPT functions most effectively as an automated language editor that optimises surface-level linguistic quality rather than as a comprehensive writing tutor. The improvements observed in this study should not be interpreted as evidence that learners acquired greater linguistic competence or internalised the revisions. Corrective Feedback Theory emphasises that language development depends on learners noticing, understanding, and subsequently applying corrective feedback independently (Ellis, 2009). Sociocultural Theory likewise conceptualises AI as a mediating tool that provides temporary scaffolding rather than permanent competence. Hence, the improvements documented in the present study represent enhanced quality of the edited text rather than direct evidence of language acquisition.

5.2 The extent to which ChatGPT reduces differences in language quality across proficiency levels

The second research question examined whether ChatGPT editing reduces differences in language quality among students with different English proficiency levels. As reported, significant differences across proficiency levels existed in Overall Text Score, Rare Words, Correctness, and Clarity before ChatGPT editing. After editing, these differences were no longer statistically significant, indicating that ChatGPT substantially reduced measurable disparities in language quality across proficiency levels. Pairwise comparisons further demonstrated that the largest differences before editing occurred between the B2-0 group and the lower proficiency groups, whereas these differences largely disappeared following AI-assisted language editing. These findings indicate that ChatGPT consistently applied language refinements across essays regardless of writers' initial proficiency, producing a convergence in measurable language quality.

This finding extends current knowledge on AI-assisted writing support. Previous studies suggested that generative AI could assist second language writers through proofreading, grammatical correction, and revision support (Alyasiri et al., 2024; Lingard et al., 2023; Mizumoto et al., 2024), whereas recent reviews concluded that ChatGPT provides immediate and personalised language assistance that benefits learners across different writing tasks (Teng & Huang, 2026; Xiao et al., 2026). However, existing empirical evidence has focused primarily on improvements in individual writing performance or users' perceptions of AI-generated feedback. Few studies have quantitatively examined whether AI-assisted language editing narrows measurable differences among writers with different proficiency levels. Thus, the present study extends the literature by demonstrating that ChatGPT not only improves language

quality but also substantially reduces measurable disparities in language-form features across proficiency levels.

PLM theory offers a coherent explanation for this homogenizing effect (Jurafsky & Martin, 2025). LLMs apply the same probabilistic optimisation processes regardless of the writer's proficiency because every text is revised toward statistically conventional patterns learned from extensive language corpora. Consequently, essays that initially differed in grammatical accuracy, vocabulary, and stylistic quality become more similar after editing. This convergence does not occur because lower proficiency writers acquire linguistic competence during the editing process. Instead, it reflects the model's tendency to normalise language according to dominant statistical patterns of academic English.

Recent corpus evidence provides additional support for this interpretation. Zhou et al. (2026) demonstrated that ChatGPT-generated academic writing relies more heavily on conventional lexical combinations, narrower semantic variation, and more predictable collocational patterns than human-written texts. This tendency naturally reduces linguistic variability because revisions consistently favor highly probable language forms. The convergence observed in the present study therefore appears to represent standardisation of language expression rather than convergence in underlying writing ability.

This interpretation also explains why improvements occurred consistently across proficiency levels despite substantial differences in the quality of the original essays. Lower proficiency writers benefited more because their texts contained greater numbers of statistically predictable language errors, which ChatGPT could correct efficiently. Higher proficiency writers also benefited, although improvements were comparatively smaller because their essays already conformed more closely to conventional academic language patterns before editing. This means ChatGPT functioned primarily as a language equaliser at the textual level rather than a mechanism for equalizing linguistic competence.

Although the reduction in proficiency differences may promote greater equity in language support, this finding also raises important educational and assessment considerations. Standardised language refinement can reduce linguistic disadvantages experienced by less proficient writers and multilingual learners. This expands access to high-quality academic language. This outcome aligns with the broader educational potential of generative AI to provide individualised language assistance at scale (Abudalfa & Barrot, 2026; Guo, 2025). However, greater textual similarity may also reduce stylistic diversity, diminish authorial voice, and encourage formulaic academic writing. Excessive linguistic standardisation could therefore narrow the range of acceptable rhetorical expression within academic discourse.

The findings likewise have important implications for writing assessment. Because ChatGPT substantially reduced measurable differences in grammatical accuracy, vocabulary, and sentence construction, AI-edited texts may no longer accurately represent learners' independent writing ability. Teachers who evaluate only the final submitted product may find it increasingly difficult to distinguish learners' actual language competence from AI-assisted language refinement. This means improvements observed in AI-edited texts should not be interpreted as evidence that students independently mastered the underlying linguistic knowledge or revision strategies. Assessment practices may require greater emphasis on writing processes, revision histories, reflective commentaries, and in-class writing tasks that document learners' independent performance.

Corrective Feedback Theory and Sociocultural Theory further reinforce this interpretation. Corrective feedback facilitates language development only when learners notice revisions, understand the linguistic principles underlying those revisions, and subsequently transfer that knowledge to future writing (Ellis, 2009). Likewise, Sociocultural Theory conceptualises technological support as temporary scaffolding that assists learners during task completion without guaranteeing long-term internalisation. As such, the homogenizing effect observed in this study should be interpreted as evidence of standardised language refinement rather than evidence that learners reached comparable levels of writing proficiency.

The above findings suggest that ChatGPT should be viewed as a tool that promotes greater equity in access to language refinement rather than one that equalises writing competence. Language quality

and writing ability should remain conceptually distinct. ChatGPT can substantially improve the quality of written products, particularly for learners with greater linguistic needs, but independent writing competence still depends on deliberate practice, metalinguistic reflection, and opportunities to apply corrective feedback without AI assistance. Educational practice should encourage students to critically evaluate AI-generated revisions, justify accepted and rejected edits, and gradually develop independent self-editing abilities. Such an approach allows AI-assisted language editing to support learning while preserving assessment validity and learner agency.

6 Conclusion

Our study systematically quantified the impact of ChatGPT's language editing on students' academic essays across varying levels of English proficiency. The results demonstrated ChatGPT's capability to elevate language quality, substantially enhancing correctness, vocabulary richness, and overall textual engagement. These improvements significantly reduced the gap in the language quality of essays produced by students with varying proficiency. While the technology displayed the ability to standardise the technical aspects of language, such as grammar and vocabulary use, it did not significantly enhance the clarity of essays, indicating a need for nuanced human judgement in this domain. These findings should be interpreted as evidence of ChatGPT's immediate editing effects on standardised language-form measures rather than evidence of long-term writing development or language acquisition. The study demonstrates improvements in the quality of edited texts, but it does not establish that learners internalised these revisions or transferred them to subsequent independent writing. Accordingly, the primary contribution of this study lies in providing empirical evidence of ChatGPT's effectiveness as an automated language editing tool that improves measurable language quality across proficiency levels under standardised editing conditions. However, the study also highlights the limitations of AI in replicating the more nuanced, context-dependent aspects of language quality, suggesting that while AI can significantly contribute to language quality, it cannot fully replace the discerning capabilities of a human editor.

This study's findings carry several implications for the field of academic writing, TESOL curriculum development, and AI application in education. The differential effects of ChatGPT across language quality dimensions underscore the need for careful pedagogical integration. Within TESOL writing curricula, ChatGPT should function as a scaffold for revision. Its strongest contribution lies in supporting language refinement during the revision stage of process writing, where students examine successive drafts, evaluate revisions, and strengthen the linguistic quality of their texts before final submission. This integration aligns with process-oriented writing instruction because revision remains a central stage in which learners refine both language and meaning through iterative evaluation. Although ChatGPT substantially improved correctness, vocabulary, delivery, and engagement, its limited effect on clarity suggests that teachers should require students to explain, evaluate, and justify AI edits rather than accept them uncritically. Such activities can strengthen metalinguistic awareness, learner autonomy, and feedback literacy, which enables learners to interpret and use feedback effectively to improve future writing. While ChatGPT effectively corrects grammatical errors and enhances lexical diversity, teachers should position it primarily as a tool for improving surface-level language features rather than higher-order aspects of writing. Students should revise clarity, argument development, cohesion, and rhetorical purpose under teacher guidance because these dimensions require contextual judgement and disciplinary awareness beyond AI-generated language refinement. Genre-based pedagogy also provides an appropriate instructional context because teachers can require students to examine whether AI-generated revisions preserve the communicative purpose and rhetorical conventions of a target genre. Classroom activities can encourage students to compare their original and AI-edited drafts, evaluate the rationale behind each revision, and decide which suggestions should be accepted, modified, or rejected. Teachers can further strengthen these activities through structured revision logs or reflective

commentaries that require students to justify every accepted or rejected AI suggestion with reference to grammatical rules, rhetorical purpose, audience awareness, or disciplinary conventions. This approach also supports authorship, assessment validity, and academic integrity because students who rely solely on AI corrections may produce stronger texts without fully understanding or independently applying the underlying language forms. Overdependence on AI-driven feedback may hinder learners' ability to internalise linguistic structures and self-edit. This concern is particularly relevant for students in the initial stages of English learning, where the development of grammatical awareness is crucial for long-term proficiency. Lower proficiency L2 writers and multilingual learners may particularly benefit from structured AI-assisted revision because immediate language support can reduce cognitive demands associated with grammar and vocabulary, allowing greater attention to idea development and organisation. Academic writing students, in contrast, should engage in more sophisticated evaluation tasks that require them to critique AI revisions for disciplinary appropriateness, authorial voice, precision of meaning, and rhetorical effectiveness. An excessive reliance on AI for surface-level corrections may result in a superficial grasp of language rather than deeper mastery. Teachers should therefore combine AI-assisted editing with explicit instruction on grammar, revision strategies, and rhetorical writing so that learners progressively develop independent writing competence. Additionally, while ChatGPT refines linguistic accuracy, it does not directly enhance rhetorical competence, argumentation, or discourse-level cohesion. Learners who depend heavily on AI-generated corrections may achieve grammatically sound texts but might still struggle with logical organisation, coherence, and persuasive writing. Hence, TESOL curricula should incorporate AI-assisted editing as one component of a broader instructional framework that integrates process writing, genre knowledge, explicit feedback instruction, and reflective revision practices.

Academic integrity is another critical concern. The present findings indicate that ChatGPT standardises language quality across proficiency levels, which may blur the boundary between acceptable language support and substantive textual intervention. Although the underlying ideas remain the learners' own, AI-assisted editing can substantially reshape the linguistic expression of those ideas. The ethical use of ChatGPT hinges on whether students engage with the AI tool as a writing assistant or merely accept its suggestions without critical reflection. If learners uncritically adopt ChatGPT's edits without analysing their rationale, the tool may inadvertently encourage passive learning rather than active skill development. The homogenising effect observed in this study also has implications for assessment validity. Although ChatGPT reduced disparities in measurable language quality, the edited text may no longer accurately reflect learners' independent linguistic competence. At the same time, this standardisation may reduce surface-level disadvantages for lower proficiency L2 writers and multilingual learners. Therefore, institutions should develop clear guidelines on the responsible use of AI in academic writing, ensuring that students benefit from its affordances while maintaining academic integrity. Teachers should also distinguish formative activities, where AI-assisted editing can support learning, from summative assessments that aim to evaluate independent writing ability. Given these considerations, teachers should implement AI tools like ChatGPT within a structured learning framework that encourages active engagement rather than passive dependence. Combining AI-generated feedback with explicit instruction on grammar, self-editing strategies, and rhetorical development could mitigate potential drawbacks while maximising the pedagogical benefits of AI-assisted writing support.

Unlike traditional forms of automated corrective feedback, which rely on fixed grammatical rules, ChatGPT's editing is dynamic and data-driven (Alyasiri et al., 2024; Lingard et al., 2023). This study contributes to ongoing discussions about the role of AI-assisted feedback in writing pedagogy, particularly in terms of balancing AI-generated linguistic corrections with explicit instruction on self-editing and rhetorical coherence. Additionally, by demonstrating how ChatGPT reduces proficiency-level disparities, our findings offer insights into how AI tools can serve as a scaffold for learners at different stages of writing development. These results align with broader concerns in AI-assisted language

learning, where the effectiveness of AI feedback must be assessed not only in terms of accuracy but also in its ability to promote long-term writing competence.

Recognising the limitations of this study is crucial for understanding its scope and framing the context of its findings. First, our study focused on a specific version of ChatGPT and a particular set of essays from the ICNALE corpus. Thus, the results may not be generalisable to other genres of writing, forms of discourse, or other AI platforms. For future research, it would be advantageous to expand the scope of investigation by diversifying AI platforms, broadening the text corpus, and testing updated ChatGPT versions. While providing valuable quantitative data, the study did not extend to a qualitative analysis, leaving the potential improvements in argumentative structure or rhetorical effectiveness unexplored. Future research could incorporate qualitative analyses to examine these dimensions. Finally, this study provides a snapshot of immediate editing outcomes but does not assess the long-term impact of AI editing on writing proficiency. Subsequent research could explore longitudinal effects to better understand the role of AI in writing over time.

Acknowledgement

This study was funded by the University of Economics Ho Chi Minh City, Vietnam.

References

Abudalfa, S. I., & Barrot, J. (2026). Generative artificial intelligence for automated writing evaluation: A systematic review of trends, efficacy, and challenges. *Assessing Writing*, *68*, 101041. https://doi.org/10.1016/j.asw.2026.101041

Al-Garaady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal*, *Special Issue on CALL*, *9*, 3–17. https://dx.doi.org/10.24093/awej/call9.1

Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research*, *63*(2), 431–463. Advance online publication. https://doi.org/10.1177/07356331241307297

Alyasiri, O. M., Salman, A. M., & Salisu, S. (2024). ChatGPT revisited: Using ChatGPT-4 for finding references and editing language in medical scientific articles. *Journal of Stomatology, Oral and Maxillofacial Surgery*, 101842. https://doi.org/10.1016/j.jormas.2024.101842

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback?. *International Journal of Educational Technology in Higher Education*, *21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Barrett, A., & Pack, A. (2023). Not quite eye to AI: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, *20*, 59. https://doi.org/10.1186/s41239-023-00427-0

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

Barrot, J. (2023b). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. (2026). The CLARIFIES+ framework: A contextualized prompting approach for generative artificial intelligence. *International Journal of TESOL Studies*, *260215*, 1–22. https://doi.org/10.58304/ijts.260215

Bryant, J., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52–75. https://doi.org/10.18653/v1/W19-4406

Bui, N. M., & Barrot, J. (2025a). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Bui, N. M., & Barrot, J. (2025b). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2521003

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Castillo-González, W., Lepez, C. O., & Bonardi, M. C. (2022). Chat GPT: A promising tool for academic editing. *Data and Metadata*, *1*, 23. https://doi.org/10.56294/dm202223

Chen, H. H. J., Chiu, S. T. L., & Liao, P. (2009). Analyzing the Grammar feedback of two automated writing evaluation systems: MyAccess and Criterion. *English Teaching & Learning*, *33*(2), 1–43. https://doi.org/10.6330/ETL.2009.33.2.01

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*(11), 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*, 111–126. https://doi.org/10.1007/s12599-023-00834-7

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201–224. https://doi.org/10.1002/berj.4069

Gain, A., Rao, M., & Bhat, S. K. (2019). Usage of Grammarly–online grammar and spelling checker tool at the Health Sciences Library, Manipal Academy of Higher Education, Manipal: A study. *Library Philosophy and Practice*, 1–13.

Gjorevski, A., Li, M., & Cox, T. L. (2025). Exploring the potential of ChatGPT for evaluating English essays in a criterion-based assessment. *TESOL Quarterly*, *59*, S251–S279. https://doi.org/10.1002/tesq.70011

Guo, X. (2025). Facilitator or thinking inhibitor: Understanding the role of ChatGPT-generated written corrective feedback in language learning. *Interactive Learning Environments*, *33*(5), 3526–3544. https://doi.org/10.1080/10494820.2024.2445177

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, *15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, *37*(1–2), 1–26. https://doi.org/10.1080/09588221.2021.2021241

John, P., & Woll, N. (2020). Using grammar checkers in an ESL context: An investigation of automatic corrective feedback. *CALICO Journal*, *37*(2), 169–192. https://doi.org/10.1558/cj.36523

Johnson, A. (2023). *ChatGPT in schools: Here's where it's banned—and how it could potentially help students*. Forbes. https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=6c92bf4c6e2c

Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3

Kim, S. G. (2023). Using ChatGPT for language editing in scientific articles. *Maxillofacial Plastic and Reconstructive Surgery, 45*(1), 13. https://doi.org/10.1186/s40902-023-00381-x

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lingard, L., Chandritilake, M., de Heer, M., Klasen, J., Maulina, F., Olmos-Vega, F., & St-Onge, C. (2023). Will ChatGPT's free language editing service level the playing field in science communication?: Insights from a collaborative project with non-native English scholars. *Perspectives on Medical Education*, *12*(1), 565–574. https://doi.org/10.5334/pme.1246

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Moorhouse, B. L., Wan, Y., Wu, C., Wu, M., & Ho, T. Y. (2025). Generative AI tools and empowerment in L2 academic writing. *System*, *133*, 103779. https://doi.org/10.1016/j.system.2025.103779

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

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics*, *3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Nadra, S., & Purdy, J. P. (2026). Is genAI a good editor of academic writing? *Computers and Composition*, *80*, 102993. https://doi.org/10.1016/j.compcom.2026.102993

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. OpenAI. https://openai.com/blog/chatgpt/

Otmar, R., Michael, R., Mullins, S., & Day, K. (2025). Ethics and the use of generative AI in professional editing. *AI and Ethics*, *5*, 1719–1731. https://doi.org/10.1007/s43681-024-00521-7

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it?. *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, *26*(1), 1–25. http://hdl.handle.net/10125/73465

Ronge, R., Maier, M., & Rathgeber, B. (2025). Towards a definition of generative artificial intelligence. *Philosophy & Technology*, *38*, 31. https://doi.org/10.1007/s13347-025-00863-y

Roose, K. (2023). *Don't ban ChatGPT in schools. Teach with it*. The New York Times. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Shin, D., & Lee, J. H. (2025). Leveraging LLM-based chatbots for interactional grammar feedback in L2 writing: Opportunities and challenges. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2549037

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

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Sun, G. H., & Hoelscher, S. H. (2023). The ChatGPT storm and what faculty can do. *Nurse Educator*, *48*(3), 119–124. https://doi.org/10.1097/NNE.0000000000001390

Teng, M. F. (2024a). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies*, *6*(3), 36–57. https://doi.org/10.58304/ijts.20240304

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

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

Teng, M. F., & Shen, X. (2025). Metacognitive awareness and EFL learners' writing performance in ChatGPT-assisted writing context. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70049

Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher*, *31*(6), 767–779. https://doi.org/10.1007/s40299-021-00625-2

Waer, H. (2023). The effect of integrating automated writing evaluation on EFL writing apprehension and grammatical knowledge. *Innovation in Language Learning and Teaching*, *17*(1), 47–71. https://doi.org/10.1080/17501229.2021.1914062

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., ... & Quintana, R. (2020). eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449. https://doi.org/10.1016/j.asw.2020.100449

Whang, Y. (2024). ChatGPT for editors: Enhancing efficiency and effectiveness. *Science Editing*, *11*(1), 84–90. https://doi.org/10.6087/kcse.332

Wu, J., Li, J., Ge, Z., Xu, M., Lin, L., & Zhang, R. (2025). Effectiveness of generative AI in automated written corrective feedback with prompting. *Journal of Educational Computing Research*, *63*(6), 1493–1527. https://doi.org/10.1177/07356331251359430

Xiao, F., Zhu, S., & Xin, W. (2026). Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a Foreign Language education: A scoping review. *ECNU Review of Education*, *9*(1), 1–19. https://doi.org/10.1177/20965311241310881

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*, 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhou, S., Chen, C., & Wu, Q. (2026). Flexibility versus formulaicity: Comparing phrasal verb use between human-written and AI-generated academic essays. *International Journal of TESOL Studies*, *8*(3), 158–182. https://doi.org/10.58304/ijts.260510

Tho D. Vo is an English lecturer at the University of Economics Ho Chi Minh City, Vietnam. He has obtained his Ph.D in Education at Victoria University of Wellington, New Zealand where he worked on the teachers' and students' use of digital technologies in the English-medium context of Vietnamese higher education. His research interests are in the areas of technology in language education and English-medium education.

https://orcid.org/0000-0002-6493-6946

Jessie S. Barrot is a Research Fellow at University of Economics Ho Chi Minh City, Vietnam and a Professor and Vice President for Research and Development at the National University, Philippines. He has published papers on curriculum and instruction, computer-assisted language learning, language teaching, and L2 writing.

http://orcid.org/0000-0001-8517-4058