Comparing human and LLM proofreading in L2 writing: Impact on lexical and syntactic features

Hakyung Sung[1] **Karla Csuros**[2,1*] **Min-Chang Sung**[3,1*]
[1]LCR-ADS Lab, Linguistics, University of Oregon
[2]West University of Timisoara
[3]Gyeongin National University of Education

Abstract

This study examines the lexical and syntactic interventions of human and LLM proofreading aimed at improving overall intelligibility in identical second language writings, and evaluates the consistency of outcomes across three LLMs (ChatGPT-4o, Llama3.1-8b, Deepseek-r1-8b). Findings show that both human and LLM proofreading enhance bigram lexical features, which may contribute to better coherence and contextual connectedness between adjacent words. However, LLM proofreading exhibits a more generative approach, extensively reworking vocabulary and sentence structures, such as employing more diverse and sophisticated vocabulary and incorporating a greater number of adjective modifiers in noun phrases. The proofreading outcomes are highly consistent in major lexical and syntactic features across the three models.

1 Introduction

The use of generative large language models (LLMs) in second language (L2) writing has gained popularity for providing real-time feedback on vocabulary, grammar, and style (e.g., Han et al., 2024; Meyer et al., 2024). These models offer immediate corrective suggestions, enhancing the precision and quality of L2 writing—a role once largely filled by human editors with expertise. As LLMs increasingly replace or supplement human intervention, questions arise about their impact on L2 writings.

While previous studies have concentrated on general error correction through LLM proofreading (e.g., Heintz et al., 2022; Su et al., 2023; Wu et al., 2023; Katinskaia and Yangarber, 2024), recent studies have shown that LLMs do not consistently outperform state-of-the-art supervised grammatical error correction models on minimal-edit benchmarks, often producing more fluency-oriented rewrites instead (Davis et al., 2024). This tendency stems in part from the fact that LLMs, by default, generate transformative fluency corrections rather than minimal edits when processing ungrammatical text (e.g., Coyne et al., 2023; Fang et al., 2023; Loem et al., 2023). However, little research has examined how this generative rewriting behavior affects broader lexical and syntactic characteristics of L2 writing compared to human proofreading, especially when the proofreading goal extends beyond grammatical accuracy to overall intelligibility. Moreover, it remains unclear whether different LLMs yield consistent proofreading outcomes. This study addresses these gaps by posing three guiding questions: (1) What are the similarities and differences in lexical features between human proofreading and LLM proofreading of L2 writings? (2) What are the similarities and differences in syntactic features between human proofreading and LLM proofreading of L2 writings? (3) Do three different LLMs provide consistent proofreading outcomes in terms of lexical and syntactic features in L2 writing?

Our findings show that while both human and LLM proofreading enhance lexical and syntactic features, LLMs are more likely to make more extensive lexical and syntactic edits. By quantifying these changes through a range of lexical and syntactic indices, we reveal that LLMs favor more generative rewrites, which may improve fluency but risk altering nuance or inflating perceived proficiency.

2 Background

2.1 Proofreading in L2 writing

Proofreading is a complex issue in writing research, particularly for L2 writers, as it involves varying scopes of interventions. Traditional definitions of proofreading often restrict it to surface-level error correction that focuses on resolving orthographic and grammatical errors without altering content (Carduner, 2007; Hyatt et al., 2017). However, research shows that professional human proofreaders

*Contributed equally to the study.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–23
July 31 - August 1, 2025

occasionally restructure content to improve the logical flow of ideas and make the writing easier to understand (Salter-Dvorak, 2019). Noting these varying practices in proofreading, Harwood et al. (2009, p. 167) provided a quite general definition of proofreading as "[any] third-party interventions (entailing written alteration) on assessed work in progress."

Previous studies have shown that human proofreading displays variability not just in scope, but also in quality. Harwood (2018) found that 14 proofreaders made between 113 and 472 changes to the same L2 learner essay, with some interventions improving clarity and others introducing new errors, leading to inconsistent quality. Similarly, Shafto (2015) argued that proofreading is a highly attention-dependent task, meaning that symptoms such as tiredness can heavily impact human proofreaders' ability to detect and correct ungrammatical and unnatural expressions.

The debate surrounding the adequacy of L2 proofreading is also characterized by varying perspectives from stakeholders (i.e., students, faculty, researchers). While L2 students often seek proofreading services to improve their grades or enhance their writing skills, some faculty view such assistance as a form of academic dishonesty (Salter-Dvorak, 2019; Turner, 2011). Despite these divergent opinions, there is a general consensus that proofreaders can significantly enhance language accuracy and clarity in L2 writing, provided that the original authorial voice is maintained (Turner, 2024; Warschauer et al., 2023; Zou and Huang, 2024).

2.2 LLMs in L2 writing and proofreading

While automated written corrective feedback has been present in L2 classrooms for over a decade (cf. Wilson et al., 2014), recent research is now exploring how LLM assistants can be incorporated into holistic writing workflows (Zhao, 2024). Researchers examine the integration of the LLM in prewriting (Xiao, 2024) and postwriting stages (Osawa, 2024), as well as its role in fostering metacognitive skills through iterative revisions that include editing and proofreading (Su et al., 2023; Warschauer et al., 2023; Zou and Huang, 2024).

Among these LLM integrations, several studies have highlighted the capabilities of LLM proofreading (or more broadly, editing). For instance, Su et al. (2023) found that ChatGPT effectively assessed grammar, clarified meaning, and suggested lexical and syntactic refinements. Similarly, Yan and Zhang (2024) observed that ChatGPT identified and corrected a range of linguistic errors—including lexical (e.g., word choice, idioms), grammatical (e.g., verb tense, articles), structural (e.g., run-on or fragmented sentences), mechanical (e.g., spelling, punctuation), and stylistic (e.g., formality) aspects.

Few studies have compared LLM proofreading directly to human revisions. For instance, Heintz et al. (2022) compared outputs edited by LLMs with those revised by human editors using sentences written by non-native English speakers. They found that while Wordvice AI[1] achieved near-human accuracy (77%) in correcting grammar and spelling errors, it lagged behind human editors in areas like vocabulary refinement and fluency adjustments. Similarly, Jiang et al. (2023) analyzed 2,197 T-units[2] and 1,410 sentences from weekly writing samples of 41 Chinese students in an online high school language program at a U.S. university. They found that ChatGPT-4 achieved high precision (88%) in correcting errors at the T-unit level (in comparison to human judgments), but sometimes overcorrected valid sentences or misinterpreted context-dependent issues, such as ambiguous word order and culturally embedded idioms.

2.3 Summary of findings and research gaps

To briefly summarize, previous research has demonstrated that proofreading in L2 writing is highly variable in both scope and quality, with interventions ranging from surface-level corrections to content restructuring. Recently, LLMs have been shown to offer performance comparable to, or even surpassing, that of human editors in L2 writing proofreading, although they exhibit limitations in context-sensitive judgment and cultural awareness.

Despite these insights, still little is known about the fine-grained linguistic interventions that could be made by LLMs compared to human proofreaders. Additionally, existing research has focused primarily on grammatical error detection and correction, overlooking broader language use. For example, although LLMs may facilitate vocabulary expansion, it remains unclear how their suggestions differ from those of human proofreaders, and detailed syntactic changes remain underexplored.

[1] https://wordvice.ai/proofreading

[2] A T-unit is often defined as the minimal grammatical unit, comprising a single independent clause plus any subordinate clauses or dependent phrases attached to it (Lu, 2010).

Moreover, most studies have examined only one type of LLM, leaving open the question of whether these linguistic changes are specific to one model or generalizable across other LLMs.

3 Methods

3.1 Dataset

This study utilizes the ICNALE Edited Essays dataset, one of the publicly available corpora within the International Corpus Network of Asian Learners of English (ICNALE) project (Ishikawa, 2018, 2021). The dataset comprises 656 essays written by 328 L2 learners and their edited versions produced by professional native English-speaking proofreaders.

The L2 participants were college students learning English in ten regional contexts: Japan (JPN), Korea (KOR), China (CHN), Taiwan (TWN), Indonesia (IDN), Thailand (THA), Hong Kong (HKG), the Philippines (PHL), Pakistan (PAK), and Singapore (SIN). Each participant wrote two argumentative essays in response to the prompts: (1) "It is important for college students to have a part-time job" and (2) "Smoking should be completely banned at all restaurants".

3.1.1 Rationale for dataset selection and representativeness

The ICNALE dataset was chosen for three main reasons. First, it provides paired original and professionally proofread versions, allowing for direct comparison with LLM-generated outputs. Second, it includes explicit L2 proficiency labels, facilitating stratified analyses across proficiency levels. Last, it offers balanced regional coverage across ten Asian countries or regions (see Table 1). However, we acknowledge that broad generalizations to other genres or demographic groups (e.g., narrative writing, younger learners) must be made with caution.

3.1.2 Proficiency band

All participants were classified into four L2 proficiency bands (linked to the Common European Framework of Reference for Languages) based on their recent scores in standardized English tests (e.g., TOEFL, TOEIC) or their performance in a standard receptive vocabulary test[3] (Nation and Beglar, 2007). Table 1 shows the proficiency distribution of each regional learner group.

Region	A2_0	B1_1	B1_2	B2_0	Total
JPN	10	10	10	10	40
KOR	10	10	10	10	40
CHN	10	10	10	10	40
TWN	10	10	10	10	40
IDN	10	10	10	3	33
THA	10	10	10	2	32
HKG	–	10	10	10	30
PHL	–	10	10	10	30
PAK	–	10	10	3	23
SIN	–	–	10	10	20
Total	60	90	100	78	328

Table 1: Distribution of participants by region and proficiency

3.1.3 Proofreading process and proofreader profiles

The ICNALE project recruited five experienced proofreaders with strong academic backgrounds and extensive experience in editing scholarly work. Their profiles are summarized in Table 2.

ID	Age	Sex	Degree	Experience (years)	L1 English
A	28	Female	BA	3	Canadian
B	32	Female	MS	5	Australian
C	27	Female	BS	3	American
D	38	Female	BS	10	British
E	31	Female	PhD	2	Australian

Table 2: Profiles of proofreaders in the ICNALE project

As documented in the ICNALE project, the professional proofreaders were tasked with editing errors and inappropriate wording to ensure that each essay became fully intelligible (Ishikawa, 2021, p. 496). No standardized rubric or adjudication mechanism was imposed at the original corpus compilation stage. All revisions were performed in MS Word using the Track Changes function, which allowed every edit, addition, or deletion to be recorded.

A calibration study in which all five proofreaders revised the same eight essays revealed substantial variability in editing behavior (cf. Ishikawa, 2018, p. 122). The number of edited word tokens ranged from 40.00 to 59.63—a difference of 19.63 tokens, or 40.97% of the average. Ishikawa (2021) attributed this variation to the inherent subjectivity of human editing, shaped by individual judgments of intelligibility.

[3]The vocabulary test consists of 50 multiple-choice items designed to measure vocabulary knowledge within the 1,000–5,000 word range. A typical item (from the 4,000-word level) presents a short sentence containing a target word and asks test-takers to select the most appropriate definition.

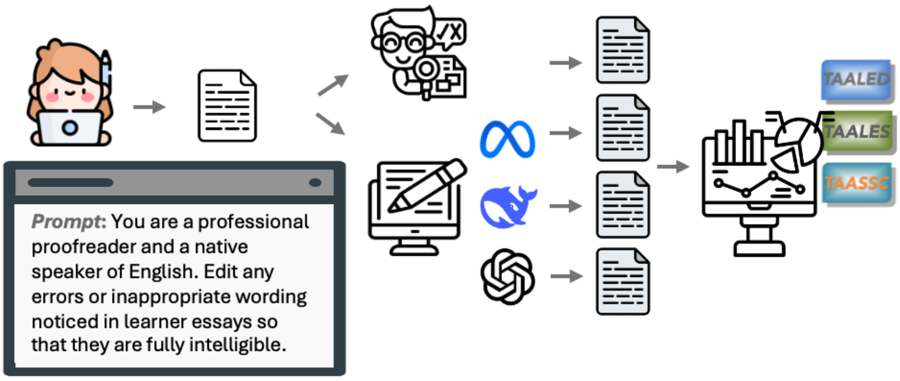

Figure 1: Overview of the experiment

3.2 LLM selection and prompt design

Figure 1 outlines the experiment. First, to compare the human proofreading in the ICNALE project with LLM proofreading, we selected three text-generating LLMs: GPT-4o (used in ChatGPT, accessed via OpenAI's API; Achiam et al., 2023, hence we called them *Chatgpt-4o*), *Llama3.1-8b* (Touvron et al., 2023), and *Deepseek-r1-8b* (Guo et al., 2025). ChatGPT-4o was chosen due to its widespread accessibility, although its underlying parameter count and architecture remain proprietary. In contrast, both Llama3.1-8b and Deepseek-r1-8b are open models with 8 billion parameters that are lightweight enough for local installations, with Deepseek-r1-8b being a distilled version of Llama3.1-8b.

Each model was tasked with reading the original L2 writings and generating a proofread version based solely on a standardized prompt, with no access to additional learner information. The exact prompt used was as follows: "You are a professional proofreader and a native speaker of English. Edit any errors or inappropriate wording noticed in learner essays so that they are fully intelligible. Return only the final edited version of the essay. Do not include any explanations, comments, reasoning, or additional thoughts in your response." This prompt was designed to align with the instructions given to ICNALE proofreaders—"They were asked to edit any error or inappropriate wording noticed in learner essays so that they could be fully intelligible. They were also required not to 'rewrite' the original texts, that is, not to add new content or to alter organization" (Ishikawa, 2021, p. 496)—ensuring consistency with the human proofreading protocol for fair comparison.

3.3 Lexical and syntactic analyses

The proofread-and-generated texts, along with the learner and edited texts in the ICNALE dataset, were processed to extract lexical and syntactic features using the source codes of publicly available NLP tools: TAALED (cf. Kyle et al., 2024), TAALES (cf. Kyle et al., 2018) and TAASSC (cf. Kyle and Crossley, 2018). We measured lexical and syntactic aspects of the learner and proofread essays based on the concept of linguistic complexity, which provides a descriptive-analytic framework for L2 production (Bulté and Housen, 2012; Bulté et al., 2024).

3.3.1 Lexical features

Lexical features were evaluated in terms of two aspects: diversity and sophistication. Lexical diversity indices reflect vocabulary variation and repetition, with higher scores indicating a broader vocabulary range and fewer repetitions. In this study, we employ common measures such as the number of unique words and the moving-average type-token ratio—the latter mitigating the impact of text length on traditional lexical diversity measures (Kyle et al., 2024).

Lexical sophistication indices, on the other hand, focus on measuring the use of advanced words (Laufer and Nation, 1995; Meara and Bell, 2001). They are typically assessed based on relative word frequency, semantic concreteness, and domain or register distinctiveness, with less frequent, less concrete, and more domain-specific words generally considered more sophisticated (Kyle et al., 2018). We also incorporate the concept of ngram sophistication by analyzing associations and dependency relations within bigrams (Kyle and Eguchi, 2021).

3.3.2 Syntactic features

Syntactic features can be examined from multiple perspectives. Traditional approaches, such as measuring the average length of T-units, focus on the overall length of syntactic structures and operate under the assumption that longer units generally indicate greater complexity (Lu, 2010, 2011).

In contrast, fine-grained syntactic complexity indices (Kyle and Crossley, 2018) provide a more nuanced analysis by capturing specific structural characteristics rather than relying on surface-level measures like sentence length. These indices are often categorized into clausal-level (e.g., nominal subjects per clause), phrasal-level (e.g., dependents per nominal, including adjectives and prepositions), and morphosyntactic-level features (e.g., use of past tense).

To the best of our knowledge, there is no consensus on which fine-grained indices reliably capture syntactic complexity as perceived by human judges. Nevertheless, L2 writing studies suggest that higher-proficiency learners (identified by human ratings) tend to use more elaborated noun phrases (e.g., Biber et al., 2011).

3.4 Statistical methods

3.4.1 Evaluating linguistic features across groups

Prior to statistical analyses, we confirmed that the five groups of texts (i.e., original [ORIG], human-proofread [EDIT], and the three LLM-proofread versions) were largely comparable in length.[4] This comparability, with the exception of Deepseek-r1-8b, indicates that subsequent improvements in lexical and syntactic domains are not simply due to different text lengths.

We calculated a range of 49 lexical and 143 syntactic indices from every text in the five groups and identified features showing significant between-group variance in two stages. First, we conducted visual inspection of box plots to exclude the indices with a great number of outliers, little individual variance, and/or unnoticeable mean differences. Second, we applied a linear mixed-effects model to each index, using Group (e.g., ORIG, EDIT, ChatGPT-4o) as a categorical fixed effect with ORIG as the baseline. Proficiency was included as a fixed effect that interacted with Group, and Participants were included as a random effect. We retained only those models that converged successfully to ensure reliable estimates. From these convergent models, we focused primarily on the main effect of the proofreading mode, while also examining whether any observed mode effects were moderated by Proficiency. These procedures yielded six lexical and nine syntactic indices. Detailed descriptions of each index are provided in Appendix A.

For each of these indices, we reported the results of four pairwise comparisons, between ORIG and human or LLM proofreading, from the linear mixed-effects models. To avoid a Type I error due to multiple comparisons, we applied a Bonferroni adjustment to the alpha level, reducing it from .05 to .0125.

3.4.2 Evaluating consistency across LLMs

The linear mixed-effects analyses informed us that the cross-model evaluation should exclude five more syntactic features, which showed multicollinearity or overlapping metrics. For the rest ten features,[5] we calculated the standardized z-scores so that each metric contributed equally to a composite measure of overall lexical and syntactic complexity.

Next, we restructured the data so that each row represented an essay and each column contained the composite score derived from the output of a different model, treating these composite scores as "ratings" of the same essay. We then calculated the Pearson correlation coefficients between the ratings for every pair of models' proofread output and computed Cronbach's alpha (Cronbach, 1951) across these scores to assess their overall consistency. All datasets and code used for this analysis are available in the supplementary repository: `https://osf.io/mhtpg/?view_only=13ce0959a80e4d498b6761aba197bc83`.

4 Results

4.1 Lexical features

Table 3 summarizes the analysis of the selected lexical sophistication and diversity features. First, all proofreading modes, including human editing, led to significantly higher bigram mutual information (`raw_bg_MI`) scores. This finding suggests that

[4]The differences in the number of word tokens relative to the original text were: EDIT: –1.02, ChatGPT-4o: +6.13, Llama3.1-8b: –3.38, and Deepseek-r1-8b: –15.11***.

[5]Lexical features: `mattr`, `b_concreteness`, `mcd`, `usf`, `cw_lemma_freq_log`, and `raw_bg_MI`; Syntactic features: `nonfinite_prop`, `amod_dep`, `nominalization`, and `be_mv`.

Index	EDIT	ChatGPT-4o	Llama3.1-8b	Deepseek-r1-8b
`raw_bg_MI`	+0.35 / 1.80***	+0.65 / 3.30***	+0.62 / 3.17***	+0.60 / 3.03***
`usf`	-1.37 / 0.15	-9.21 / 0.99***	-8.48 / 0.91***	-12.09 / 1.30***
`b_concreteness`	+0.00 / 0.02	-0.15 / 0.83***	-0.12 / 0.67***	-0.21 / 1.11***
`cw_lemma_freq_log`	-0.02 / 0.03	-0.30 / 0.54***	-0.26 / 0.47***	-0.37 / 0.67***
`mattr`	+0.01 / 0.18	+0.07 / 2.20***	+0.08 / 2.63***	+0.10 / 3.41***
`ntypes`	+0.63 / 0.05	+19.98 / 1.68***	+16.68 / 1.40***	+16.80 / 1.41***

Table 3: Lexical features compared; For each index, two numbers are shown: the value on the *left* indicates the unstandardized main effect coefficient, while the value on the *right* (following the backslash) represents the standardized coefficient, calculated as the ratio of the coefficient to the residual standard deviation of the dependent variable; Significance vs. ORIG is marked ($*p < 0.0125$, $**p < 0.0025$, $***p < 0.00025$); negative values are red and positive values are blue; interaction effects are omitted.

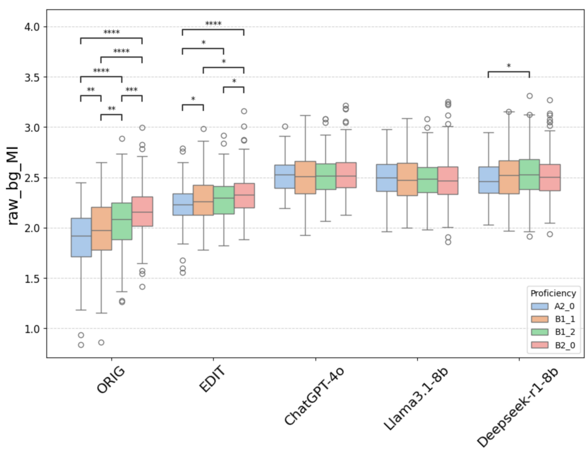

Figure 2: `raw_bg_MI` compared across ORIG, EDIT, and LLM-proofread texts by proficiency

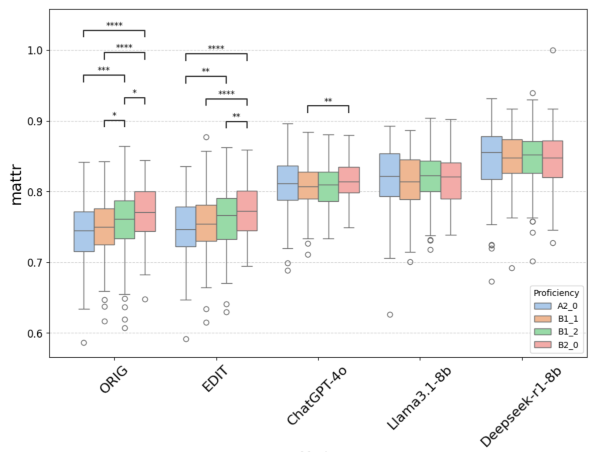

Figure 3: `mattr` compared across ORIG, EDIT, and LLM-proofread texts by proficiency

both human and LLM proofreading improved the lexical sophistication in terms of the coherence or contextual connectedness of adjacent words. However, LLM proofreading substantially increased `raw_bg_MI` to the extent that differences between lower and higher proficiency levels became less distinguishable (Figure 2).

In contrast, only the LLM-proofread texts showed significant changes in additional lexical sophistication measures, including a shift toward more contextually distinctive words (`usf`), less concrete words (`b_concreteness`), and lower-frequency content words (`cw_lemma_freq_log`). Human proofreading, by comparison, did not produce significant differences in these measures.

As for lexical diversity, significant improvements were observed only in the LLM-proofread texts, with increases in metrics such as `mattr` (Figure 3) and `ntypes`, indicating a broader range of vocabulary use.

4.2 Syntactic features

Table 4 summarizes the analysis of the selected syntactic features. Regarding the mean length of T-units (`mltu`), neither human nor LLM proofreading produced a consistent pattern: human proofreading (EDIT) and ChatGPT-4o tended to reduce T-unit length, while Llama3.1-8b and Deepseek-r1-8b tended to increase it, suggesting no uniform effect on the length of minimal grammatical units.

At the clause level, all LLM-proofread texts showed a significant increase in the total number of clauses (`all_clauses`) compared to the original learner essays, with Deepseek-r1-8b exhibiting the largest effect. Moreover, LLM-proofread texts contained a higher proportion of nonfinite clauses (`nonfinite_prop`), whereas human editing resulted in a slight reduction in this index.

At the phrase level, LLM proofreading increased the number of noun phrases (`np`), along with a rise in noun phrase dependencies (`np_deps`). This

Index	EDIT	ChatGPT-4o	Llama3.1-8b	Deepseek-r1-8b
mltu	-115.49 / 0.31	-105.73 / 0.28	+44.26 / 0.12	+118.42 / 0.31
all_clauses	+15.55 / 0.10	+133.76 / 0.84***	+99.12 / 0.62***	+179.00 / 1.12***
nonfinite_prop	-1.33 / 0.29	+2.01 / 0.44***	+2.63 / 0.57***	+5.52 / 1.20***
np	-21.30 / 0.08	+91.96 / 0.36**	+41.27 / 0.16	+194.91 / 0.76***
np_deps	-35.03 / 0.08	+79.21 / 0.17	+91.91 / 0.20	+217.81 / 0.47**
amod_dep	+17.54 / 0.01	+137.65 / 0.75***	+127.44 / 0.70***	+204.54 / 1.12***
nominalization	+58.12 / 0.40**	+152.04 / 1.05***	+102.85 / 0.71***	+213.63 / 1.47***
be_mv	+10.37 / 0.12	-56.53 / 0.63***	-41.60 / 0.47**	-84.02 / 0.94***
past_tense	-15.80 / 0.29	-17.38 / 0.32	-17.77 / 0.32	-19.31 / 0.35**

Table 4: Syntactic features compared; Interpretation of the table follows the same conventions described in Table 3

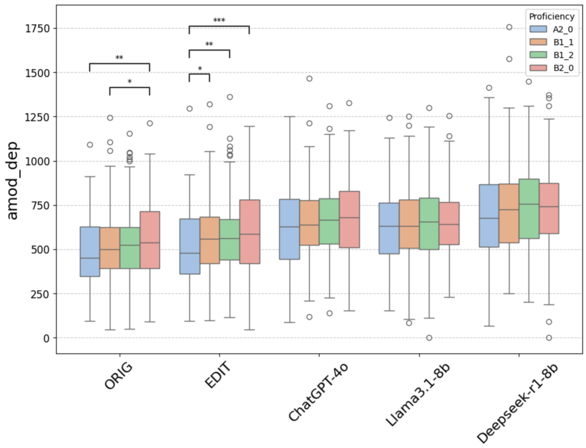

Figure 4: amod_dep compared across ORIG, EDIT, and LLM-proofread texts by proficiency

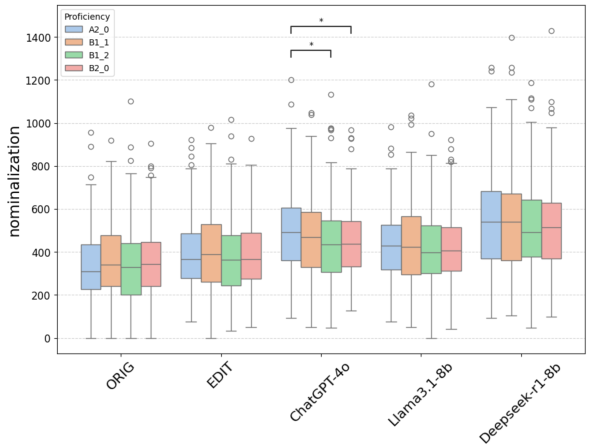

Figure 5: nominalization compared across ORIG, EDIT, and LLM-proofread texts by proficiency

suggests that LLM proofreading not only added more noun phrases but also enriched their internal structure. In particular, the marked increase in adjective modifier dependencies (amod_dep; e.g., "various jobs") suggests that LLM outputs favor more descriptive noun phrases (Figure 4).

At the morphological-syntactic level, both human and LLM proofreading showed significant increases in nominalization, but the increases were more pronounced in the LLM outputs (Figure 5). In contrast, the non-auxiliary use of the main verb "be" declined significantly under LLM proofreading, while human proofreading showed only a slight increase (be_mv). Additionally, all proofreading modes consistently reduced the use of past tense (past_tense).

4.3 Cross-model consistency

Based on the features that demonstrated meaningful group differences—and after removing indices with multicollinearity and conceptual overlap—we selected ten lexical or syntactic features. The composite lexical and syntactic scores exhibit strong internal consistency across the LLMs, with Cronbach's alpha values of 0.83 and 0.81, respectively.

Pair	Lexical	Syntax
ChatGPT-4o – Llama3.1-8b	0.70	0.62
ChatGPT-4o – Deepseek-r1-8b	0.60	0.53
Llama3.1-8b – Deepseek-r1-8b	0.56	0.65

Table 5: Pairwise Pearson correlations for lexical and syntactic features across LLMs

Table 5 presents the pairwise Pearson correlations among the three LLM proofreading models. For lexical features, ChatGPT-4o and Llama3.1-8b correlate at 0.70, while Deepseek-r1-8b correlates at 0.60 with ChatGPT-4o and 0.56 with Llama3.1-8b. For syntactic features, the corresponding correlations are 0.62, 0.53, and 0.65. These findings suggest that, despite minor variations, particularly with Deepseek-r1-8b, the LLMs tended to modify vocabulary and syntactic structures in a relatively consistent manner when proofreading L2 writings, as measured by our selected indices.

5 Discussions

We compared the lexical and syntactic features of original L2 writings with those of texts that were proofread by human and LLMs. We also evaluated the consistency of LLM proofreading across different models.

Lexical features We found significant increases in bigram association strength, a ngram-level index of lexical sophistication, across all the proofreading modes. However, only LLM-proofread texts demonstrated notable changes in both word-level sophistication and diversity. Together, these results suggest that while both human and LLM proofreading improved the natural sequence of vocabulary–thus, enhancing the intelligibility of L2 writings–LLM proofreading provided an additional boost in lexical diversity and sophistication. In fact, this boost sometimes reduced or even eliminated typical differences between proficiency levels. Given that lexical sophistication and diversity are important constructs when evaluating L2 writing proficiency (Kyle et al., 2018, 2021), texts produced using LLM proofreading may obscure learners' true writing abilities and artificially inflate their advanced language skills, ultimately undermining accurate assessment and long-term development.

We also observed that LLMs often replaced repeated words with alternative expressions—even when such changes are unwarranted—calling for caution. For example, "I often can smell" became "I often catch a whiff", altering the intended meaning. Consequently, L2 writers using LLM proofreading should be mindful of unintended shifts in meaning or style and double-check suggested edits.

Syntactic features Compared with the marked lexical shifts, syntactic edits were subtler but still distinct pattern of edits. First, both human and LLM proofreading consistently reduced past-tense verbs, favoring present or neutral tense—a pattern often associated with factual, persuasive prose (Burrough-Boenisch, 2003; Fang and Maglio, 2024).

However, LLMs made more extensive structural modifications, including a higher proportion of non-finite clauses (e.g., "Because the company that need worker will ask the job experiences" $\rightarrow$ "Companies looking to hire often require prior work experience") and a marked increase in adjective modifier dependencies (e.g., "become the social problem" $\rightarrow$ "become a significant social problem"). They also introduced more nominalizations (e.g., "we should..." $\rightarrow$ "(our) primary responsibility") and reduced the non-auxiliary use of the main verb "be" (e.g., "is not the first" $\rightarrow$ "should not take precedence").

Meanwhile, although the increase in overall noun complexity following LLM proofreading was not statistically robust (`dp_deps`), the gains were primarily driven by the insertion of adjective modifiers rather than by broader grammatical restructuring. For example, the structural complexity of noun phrases involving prepositional phrases (e.g., "disadvantages of works") or coordination (e.g., "advantages and disadvantages") remained largely unchanged.

Cross-model consistency We found that the three LLMs exhibit generally consistent proofreading performance in terms of the major lexical and syntactic features. We speculate that this consistency arises from fundamental similarities in how they are trained and optimized for language generation tasks. Consequently, while different LLMs may produce distinct outputs, their overall patterns of lexical enhancement and syntactic restructuring remain comparable.

6 Conclusions

Our study shows that while both human and LLM proofreading improve lexical and syntactic features in L2 writing, LLMs typically implement more generative edits, reworking vocabulary and sentence structures to a greater extent. Although these changes may enhance clarity and style, they risk overshadowing the original meaning or authorial voice and potentially inflate apparent language proficiency.

This finding has important implications for L2 writing practice. Acknowledging the great similarities in proofreading outcomes across different LLMs, more attention should be given to the question of "how to use LLM-proofreading effectively" rather than "what LLM to use for proofreading." This key question can be addressed in reference to the observations that we have reported above, such as non-mandatory lexical substitution and excessive syntactic restructuring. Being aware of these tendencies in LLM-proofreading, L2 writers can better maintain control over their writing process while strategically making use of LLMs for linguistic improvements.

Limitations

This study has several limitations. First, the same proofreading directive may be interpreted differently by human and LLM proofreaders, potentially affecting the nature and extent of the modifications.

Second, the analysis lacks qualitative comparisons between original and edited texts, which could reveal subtler aspects of the revisions. As one reviewer noted, LLM-proofread essays may appear more sophisticated but sometimes sacrifice coherence or introduce unintended nuances, making them harder to read. A more systematic qualitative analysis (ideally supported by human perception data comparing human- and LLM-proofread texts) would clarify whether LLM edits genuinely improve writing quality or simply enhance surface-level features.

Third, the task effects and proficiency-level constraints limit generalizability: our analysis focused solely on argumentative writing by Asian university-level students who already possess a certain level of L2 English proficiency. Consequently, these findings may not extend to other types of writing or to L2 groups with different backgrounds.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *Tesol Quarterly*, 45(1):5–35.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2024. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*.

Bram Bulté and Alex Housen. 2012. Defining and operationalising l2 complexity. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, pages 21–46. John Benjamins Publishing Company, Amsterdam.

Joy Burrough-Boenisch. 2003. Examining present tense conventions in scientific writing in the light of reader reactions to three dutch-authored discussions. *English for Specific Purposes*, 22(1):5–24.

Jessie Carduner. 2007. Teaching proofreading skills as a means of reducing composition errors. *Language Learning Journal*, 35(2):283–295.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Christopher Davis, Andrew Caines, O Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of english learner text. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11952–11967.

David Fang and Sam J Maglio. 2024. Time perspective and helpfulness: Are communicators more persuasive in the past, present, or future tense? *Journal of Experimental Social Psychology*, 110:104544.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA. Association for Computational Linguistics.

Nigel Harwood. 2018. What do proofreaders of student writing do to a master's essay? differing interventions, worrying findings. *Written Communication*, 35(4):474–530.

Nigel Harwood, Liz Austin, and Rowena Macaulay. 2009. Proofreading in a uk university: Proofreaders' beliefs, practices, and experiences. *Journal of Second Language Writing*, 18(3):166–190.

Kevin Heintz, Younghoon Roh, and Jonghwan Lee. 2022. Comparing the accuracy and effectiveness of wordvice ai proofreader to two automated editing tools and human editors. *Science Editing*, 9(1):37–45.

Jon-Philippe K Hyatt, Elisa Jayne Bienenstock, and Jason U Tilan. 2017. A student guide to proofreading and writing in science. *Advances in Physiology Education*, 41(3):324–331.

Shin'ichiro Ishikawa. 2018. The icnale edited essays: A dataset for analysis of l2 english learner essays based on a new integrative viewpoint. *English Corpus Studies*, 25:117–130.

Shin'ichiro Ishikawa. 2021. Asian learners' knowledge and use of l2 english words and phrases: A corpus-based study on learners in china, japan, korea, and taiwan. In J. Szerszunowicz, editor, *Intercontinental Dialogue on Phraseology 4: Reproducible Language Units from an Interdisciplinary Perspective*, pages 493–510. University of Bialystok Publishing House.

Zilu Jiang, Zexin Xu, Zilong Pan, Jingwen He, and Kui Xie. 2023. Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, 8(4):247.

Anisia Katinskaia and Roman Yangarber. 2024. Gpt-3.5 for grammatical error correction.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): Version 2.0. *Behavior Research Methods*, 50:1030–1046.

Kristopher Kyle and Scott A Crossley. 2018. Measuring syntactic complexity in l2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Kristopher Kyle and Masaki Eguchi. 2021. Automatically assessing lexical sophistication using word, bigram, and dependency indices. In Frances Blanchette and Constantine Lukyanenko, editors, *Perspectives on the L2 Phrasicon: The View from Learner Corpora*, pages 126–151. De Gruyter Brill, Berlin.

Kristopher Kyle, Hakyung Sung, Masaki Eguchi, and Fred Zenker. 2024. Evaluating evidence for the reliability and validity of lexical diversity indices in l2 oral task responses. *Studies in Second Language Acquisition*, 46(1):278–299.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, 16(3):307–322.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL Quarterly*, 45(1):36–62.

Paul Meara and Huw Bell. 2001. P-lex: A simple and effective way of describing the lexical characteristics of short l2 tests. *Prospect*, 16(3):5–19.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Paul Nation and David Beglar. 2007. A vocabulary size test. *The Language Teacher*, 31(7):9–13.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 1998. The university of south florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, Computers*, 36(3):402–407.

Koji Osawa. 2024. Integrating automated written corrective feedback into e-portfolios for second language writing: Notion and notion ai. *RELC Journal*, 55(3):881–887.

Hania Salter-Dvorak. 2019. Proofreading: How de facto language policies create social inequality for l2 master's students in uk universities. *Journal of English for Academic Purposes*, 39:119–131.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Meredith A Shafto. 2015. Proofreading in young and older adults: The effect of error category and comprehension difficulty. *International Journal of Environmental Research and Public Health*, 12(11):14445–14460.

Yanfang Su, Yun Lin, and Chun Lai. 2023. Collaborating with chatgpt in argumentative writing classrooms. *Assessing Writing*, 57:100752.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Joan Turner. 2011. Rewriting writing in higher education: The contested spaces of proofreading. *Studies in Higher Education*, 36(4):427–440.

Joan Turner. 2024. Afterword: Revisiting the boundaries of editing and proofreading. In Nigel Harwood, editor, *Proofreading and Editing in Student and Research Publication Contexts: International Perspectives*, pages 221–233. Routledge, London.

Mark Warschauer, Waverly Tseng, Soobin Yim, Thomas Webster, Sharin Jacob, Qian Du, and Tamara Tate. 2023. The affordances and contradictions of ai-generated text for writers of english as a second or foreign language. *Journal of Second Language Writing*, 62:101071.

Joshua Wilson, Natalie G Olinghouse, and Gilbert N Andrada. 2014. Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal*, 12(1):93–118.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Qimin Xiao. 2024. Chatgpt as an artificial intelligence (ai) writing assistant for efl learners: An exploratory study of its effects on english writing proficiency. In *Proceedings of the 2024 9th International Conference on Information and Education Innovations*, pages 51–56.

Da Yan and Shuxian Zhang. 2024. L2 writer engagement with automated written corrective feedback provided by chatgpt: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, 11(1):1–14.

Dan Zhao. 2024. The impact of ai-enhanced natural language processing tools on writing proficiency: An analysis of language precision, content summarization, and creative writing facilitation. *Education and Information Technologies*, 30(6):8055–8086.

Min Zou and Liang Huang. 2024. The impact of chatgpt on l2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies*, 29(11):13201–13219.

A Descriptions of the selected indices

Index	Description
Lexical indices	
ntypes	Counts the number of unique words, taking into account their part-of-speech.
mattr	Computes the type-token ratio over a 50-word sliding window.
b_concreteness	Uses psycholinguistic norms to assess word concreteness across categories based on large-scale ratings, indicating how tangible or abstract a word is perceived to be (Brysbaert et al., 2014).
usf	Measures the number of distinct stimuli that elicit a target word in a word association experiment; lower USF scores suggest the use of words that are more contextually distinct (Nelson et al., 1998).
cw_lemma_freq_log	Represents the logarithm of lemma frequencies for content words, computed with reference to an English web corpus (Schäfer and Bildhauer, 2012).
raw_bg_MI	Calculates raw bigram mutual Information to quantify the strength of association between consecutive words, with higher values indicating a stronger collocational relationship; this is measured against an English web corpus.
Syntactic indices	
mltu	Measures the average length of T-units, where a T-unit is defined as a main clause plus any subordinate clause(s) attached to it.
all_clauses	Counts the total number of clauses in the text (normed by 10,000 words).
nonfinite_prop	Computes the proportion of nonfinite clauses (e.g., gerunds, infinitives) relative to the total number of clauses.
np	Counts the total number of noun phrases, highlighting the nominal complexity within sentence structures (normed by 10,000 words).
np_deps	Counts the number of internal dependencies within noun phrases (e.g., adjectives, prepositions, coordinations) (normed by 10,000 words).
amod_dep	Measures the frequency of adjective modifier dependencies (normed by 10,000 words).
nominalization	Counts the frequency of nominalizations (i.e., words that convert verbs or adjectives into noun forms) identified by tokens containing predefined suffixes such as *-al*, *-ness*, among others (normed by 10,000 words).
be_mv	Measures the frequency of the verb “be” when used as a main verb (excluding its auxiliary function) (normed by 10,000 words).
past_tense	Measures the frequency of past tense verbs (normed by 10,000 words).