MEASURING WHAT MATTERS: A NOVEL METRIC-BASED EVALUATION OF CHATGPT'S EFFECTIVENESS IN EFL WRITING

Xingxing Jia*

Abstract

Since the introduction of ChatGPT in EFL writing, most studies have evidenced its role primarily through qualitative lens, leaving its instructional effectiveness insufficiently verified in a rigorous, measurable way. Addressing this, the study uses a metric-driven framework known as CAOS—Content, Accuracy, Complexity, and Style. It aims to quantify both the benefits and potential trade-offs to guide balanced classroom implementation. Forty essays were sampled and revised with ChatGPT. Analysis revealed significant improvements in lexical sophistication, syntactic complexity, and grammatical accuracy, as well as enhanced academic tone in the revised texts. However, some content measures showed slight declines in information density, highlighting the need for balanced use of AI tools. These findings provide empirical evidence for the pedagogical value of ChatGPT in EFL writing, while underscoring the importance of strategic, informed implementation.

***Keywords*:** EFL writing; ChatGPT; fine-grained evaluation; writing quality

Introduction

Writing, as an integral component of language learning, plays a critical role in EFL learning and instruction. Proficient and effective essays deliver author's intended meaning, tailored to the specific audience and contextual requirements (McCarthy et al., 2022). Successful writing, as indicated by Okasha & Hamdi (2014), includes vocabulary usage, grammar proficiency, content organization, and mechanical accuracy. While the mastery of grammar and spelling is essential, the organization of content also contributes to high-quality writing. Achieving such writing excellence, however, poses challenges for EFL learners, stemming from factors such as inappropriate writing strategies, negative attitudes towards writing, and lack of writing motivation (Tsai, 2017). All these are often exacerbated by limited feedback from instructors due to time constraints or large class sizes (Kamberi, 2013; Tsai, 2022). One practical way to offset these drawbacks in writing is through the modern automated writing evaluation system (AWES).

As the newly developed force in AWES, ChatGPT has gained substantial attention in EFL writing, providing timely and individualized feedback not only on grammar and spelling but on content and accuracy (Guo & Wang, 2024). To be specific, Chatbot can identify word meaning in a given context, correct linguistic errors, offer examples, definitions, and translation, thus increasing learning efficiency and effectiveness, and further boosting

This work is Licensed under a Creative Commons Attribution-Non-Commercial 4.0 International License.

* Xingxing Jia, Teacher, Shenzhen No.2 Experimental Ming Yuan School, Shenzhen, China

learning motivation (Fauzi et al., 2023). It is the benefits provided by ChatGPT that propel researchers to dig deeper into the use of Chatbots in EFL writing. With the advent of ChatGPT in EFL writing, existing research has largely portrayed its impact through qualitative perspectives (e.g., Guo & Wang, 2024; Uyan & Büyükahıska, 2025), while its instructional value remained to be confirmed with rigorous, measurable evidence.

To systematically quantify both the instructional benefits and potential trade-offs of ChatGPT as a writing tool, this study adopts a fine-grained method to examine the role of ChatGPT in EFL students' writing. Such analysis can provide empirical evidence to elucidate the extent to which the integration of a modern automated writing evaluation system enhances writing proficiency, supplementing the existing literature on the effect of AI-based tools on EFL writing. Additionally, the investigation offers practical insights into the quantitative aspects of ChatGPT implementation in educational settings, thereby benefiting both EFL learners and EFL instructors.

Despite its significance, the existing literature primarily focuses on using qualitative analysis to explore its impact on writing performance, with limited exploration of quantitative aspects, specifically the fine-grained method of ChatGPT on EFL writing. To fill in the gap, the study sets out to apply the fine-grained method to analyze the impact of ChatGPT as an educational tool on the writing output produced by EFL students. Guided by this, we proposed two research questions:

1. How does ChatGPT-assisted revision quantitatively impact CAOS in Chinese EFL students' writing?

2. Does ChatGPT assistance produce different degrees of improvement across various aspects of EFL writing?

Literature review

Technology-assisted EFL writing: Emerging trends in the AI era

Within the realm of computer-assisted language learning, contemporary technologies have assumed an increasingly significant role in the field of EFL pedagogy. The application of such technologies has focused on varied aspects in EFL writing: utilizing wiki-based collaboration (Aydın & Yıldız, 2014), relying on e-feedback for revisions (Shang, 2022), and exploring students' behavior in using computers for writing (Li et al., 2019), to name but a few. The technology-assisted language learning has significant effect size on EFL writing performance, according to the meta-analysis made by Seyyedrezaei and his colleagues (2022).

Research on advanced technologies used in EFL writing can be divided into two lines. One line focuses on the effects of replacing existing technologies with AI-based alternatives. For instance, Chung and Ahn (2022) explored the affordance of Google Translate in assessing syntactic complexity and accuracy of students' writing. In a similar vein, Tsai (2019) highlighted the advantage of Google Translate in reducing grammatical errors and improving overall length of composition. Later on, those AI-based technologies have made their way into providing feedback on writing quality. Dizon and Gayed (2021) reported better performance when using Grammarly in writing tasks, contrasting the

findings obtained from study wherein students have limited improvement in their cognitive engagement after learning with Grammarly (Koltovskaia, 2020).

The second research trend centered on integrating AI-based technologies into EFL writing. Recently, more AI-based plagiarism detectors have been developed and applied in EFL writing as a channel for feedback and quality evaluation. In the context of EFL writing, a prominent illustration can be found in the utilization of Turnitin, wherein it assumes a beneficial function by enhancing students' proficiency in generating precise and comprehensive feedback, thereby influencing the overall caliber of their peers' EFL writing (Li & Li, 2018). Similarly, a newly emerged paraphrasing tool named Quillbot has also been applied to EFL writing with diverse functions, integrating technology and human intelligence (Fitria, 2022). In an examination of Quillbot effects on postgraduates' performance in academic writing, these students exhibit positive perceptions towards the use of this tool in their writing class (Bui & Barrot, 2024; Kurniati & Fithriani, 2022). Overall, the literature mentioned earlier has extensively examined the integration of AI-based technologies into EFL writing pedagogy. However, there has been comparatively limited scholarly attention dedicated to the implementation and exploration of generative AI technologies in this domain.

Revolutionizing Education: The Transformative Potential of ChatGPT

Before the advent of ChatGPT, chatbot-based systems had been used in multiple educational domains, including writing practice, dynamic assessment, and language pedagogy. For example, Nagata et al. (2020) contended that the strength of such system is reflected in word choice and self-correction when evaluating the specific effects of chatbots on students' writing performance. Such positive role of AI-based chatbots is confirmed by Guo et al. (2022) and Uyar & Büyükahıska (2025), who used this tool to scaffold EFL argumentative writing.

The importance of chatbot systems was further reinforced with the introduction of ChatGPT, a large-scale language model released by OpenAI. Unlike its predecessors, ChatGPT has attracted attention from researchers and the public for its ability to respond to requests or questions proposed by its users within seconds. It is a state-of-the-art language model designed to generate text that can be indistinguishable from texts written by humans. Additionally, it can engage in conversation with users in a natural and intuitive way. Within a mere five days after its initial release, its users reached one million (Taecharungroj, 2023). ChatGPT exhibits versatility across diverse domains, with particular emphasis on its notable contributions to the field of EFL learning, signifying a noteworthy advancement in this area. For instance, ChatGPT can act as a stimulus that facilitates language learning and interaction by generating interactional pedagogical materials such as quizzes and flashcards (Dijkstra et al., 2022; Gabajiwala et al., 2022). Specifically, a recent investigation conducted by Dijkstra et al. (2022), who used ChatGPT-3 to generate questions and answers in a context of a reading comprehension task, reported that the use of automatically generated quizzes alleviates the burden associated with manual design. Also, its application involves its deployment as a pedagogical agent aiming at fostering curiosity and learning motivation. In an analysis that automated the process of generation cues as an incentive for asking profound and

numerous questions, Abdelghani et al. (2022) found that language models possess the potential to significantly enhance the implementation of curiosity-stimulating learning, while simultaneously serving as an efficient tool for promoting an amplified curiosity expression (Abdelghani et al., 2022), aligning with results obtained from research conducted by Ali et al. (2023), who investigated the attitudes of both students and educators towards the effect of ChatGPT on developing listening and speaking skills.

Another emerging application of ChatGPT in educational domains is AI-assisted writing. From the students' perspective, the language model can help EFL students complete their research or writing tasks by providing them with information and resources on a given topic. ChatGPT's positive role is also seen in the study conducted by Biswas (2023), where ChatGPT is employed to generate a research article on medical topic with manual revisions afterwards, demonstrating the potential of AI-powered language models to better medical writing. Besides improving overall process, it helps enhance writing accuracy. Such enhancement is more evident at the syntactic level, such as typographical error detection and correction. On the semantic front, ChatGPT can be leveraged to pinpoint textual incongruities and provide appropriate and individualized recommendations for improvement. Zhai (2022), for example, conducted a pilot study wherein ChatGPT was asked to generate an academic paper, leading to the conclusion that ChatGPT contributes to the creation of coherent, accurate, informative, and systematic papers. This finding was later echoed by Yan (2023), who evidenced the affordance and potential applicability of ChatGPT in EFL writing by exploring learners' behaviors and reflections in their exposure to the language model in the writing classroom. Results of it suggest positive user experience with ChatGPT.

In the view of EFL instructors, ChatGPT can be employed as a source of inspiration and reference (Qadir, 2023). They use the tool to generate prompts to facilitate classroom efficiency. Added to that, GPT presents a valuable asset for EFL writing educators. Its automated writing feedback enables efficient assessment of student assignment, reducing the burden of manual correction of language errors and saving the time spent on providing tailored feedback on individual pieces of writing, thus allowing educators to allocate more attention to other essential aspects of teaching, such as lesson planning and individualized instruction (Kasneci et al., 2023).

Methodology

Research material and procedure

The study began with an initial pool of 50 participants (26 males and 24 females), all native Chinese speakers from China aged 15-18 years. 5 participants were excluded from the final analysis due to non-compliance with instructions (4 essays were off-topic, and 1 writing sample was unreadable due to misunderstanding of the task). The study decided to focus on senior high school students, rather than college students. The rationale for choosing senior students is rooted in the fact that the former group has generally not received extensive formal training in academic writing, thus allowing for a more authentic representation of their inherent writing aptitudes. The study had agreements both from the students and their parents.

The English proficiency levels among participants varied. To objectively assess these differences, all students were required to complete an argumentative essay within 30 minutes. The writing prompt was identical for all participants and designed to elicit personal opinions on a specified topic. This approach aligns with previous research (Révész et al., 2017; Yang et al., 2015), which has shown that writing task type and topic can influence linguistic complexity, thus ensuring comparability across samples.

The initial essays were evaluated using the CEER (Cambridge English Exam Results) scale, which ranges from A1 to C2, via the certified online platform **Write and Improve** (https://writeandimprove.com), developed by Cambridge English. Analysis of these essays showed that the participants' English proficiency ranged from A2 to C2. Most notably, the majority of compositions in both groups fell within the B1 and B2 bands. For analytical consistency, only texts classified as B1 and B2 were selected for investigation; thus, 5 essays were discarded, resulting in a final sample of 40 essays for analysis. To facilitate a more precise evaluation, the CEER proficiency levels were converted to a numerical scoring system (1-7), following the approach proposed by Chung and Ahn (2022): A1=1.0, A2=2.5, B1=4.5, B2=5.5, C1=6.5, C1=6.5, C2=7.0. The summary of participants' demographic and proficiency information is presented in Table 1.

Table - 1
Metainformation of the corpus

Number of essays	Word tokens in total	Tokens per essay	SD	Min	Max
40	9914	248	52.46	129	384

After an interval of two weeks, all participants were required to compose a second argumentative essay in a language lab regarding a certain statement: *Some people think that education is a life-long process, while others don't agree*. Each essay was written within a fixed time limit and saved as a Word document. Upon completion, essays were uploaded to ChatGPT. Students received automated feedback on their writing after the prompt input, revised their essays based on the feedback, and then submitted the revised versions as new Word documents. The prompt design was informed by previous research (e.g., Basic, et al., 2023; Kasneci et al., 2023; Yan, 2023) and specifically aimed to revise the CAOS of EFL writing. To better suit the requirements of this study, the prompts were carefully adapted and refined. Both versions of the prompts were entered into ChatGPT for text revision. The comparative results show that no significant differences exist between the outputs generated by the two prompt versions, as evaluated by the CEER scale. This process ensures the validity and consistency of the prompt design used in the study. Prompts used in the study are listed below:

1. Please review the following text and suggest specific ways to improve its content and development. [insert]

2. Please check the following text and correct any grammar, spelling, or punctuation

errors. [insert]

3. Please provide suggestions to enhance the style and coherence of the following text. [insert]

4. Please revise the following text to use appropriate vocabulary and complex sentence structures, while maintaining clarity. [insert]

To ensure objective measurement and comparison of writing quality between the original and revised texts, the study adopted a computational and automated assessment approach. This methodology is deemed both accessible and theoretically robust, as it reduces human subjectivity and potential biases in evaluation (Crossley & McNamara, 2009; Godwin-Jones, 2018). Four online assessment tools were employed: *1Checker* (http://www.1checker.com), *VocabProfiler* (hereafter *VP*, http://www.lextutor.ca/vp/en), L2 Lexical Complexity Analyzer (LCA, https://aihaiyang.com), and L2 Syntactic Complexity Analyzer (LSA, https://aihaiyang.com).

VP and *1Checker* were used to evaluate three main dimensions of writing performance: content, accuracy, and style. *VP* tool facilitated the classification of words into four categories: K1 (the most frequent 1,000 words), K2 (the second frequent 1,000 words), Academic Word List (AWL), and Off List words. It is noteworthy that K2 and AWL words are considered indicators of more advanced vocabulary. Writing style was analyzed by examining the distribution of these word categories across student texts. (Chen et al., 2019). Content development was assessed by comparing *VP* index of "Idea Words" in the original and revised essays. Idea words, encompassing nouns, verbs, adjectives, and adverbs, are regarded as proxies for the informational content of a text, representing the writers' ideas, beliefs, and arguments. A higher count of idea words signifies a greater information richness (Johansson, 2008). The assessment of accuracy involved spelling and grammar error identification using *1Checker*. The software was used to identify surface-level mistakes in spelling and grammar. The probability of writing errors was calculated as the ratio of total spelling and grammar errors to the overall word count, providing a standardized measure of linguistic accuracy.

Target measures for lexical and syntactic complexity

LCA generates lexical results pertaining to lexical density, variation, and sophistication. Among its 25 indices, only 15 indices were selected for the investigation, discarding those that transformed from others based on the work of Chung and Ahn (2022). Additionally, we excluded another 2 measures due to the absence of sample values (VS1 and VS2), resulting in a final set of 13 measures for assessment of the lexical density, sophistication, and variation. A comprehensive listing of the selected measures can be found in Table 2.

The syntactic complexity is measured by SCA, involving 14 measures. Of the 14 indices, 7 indices are included for the analysis. The inclusion of the 7 target indices was guided by three principles: validity, redundancy, and distinctiveness. For example, the measure of clauses per sentence (C/S) captures both subordination and coordination, which are conceptually distinct constructs. To ensure a more nuanced understanding of language development, separate measures were chosen to respectively assess subordination

(clauses per T-unit, C/T) and coordination (T-unit per sentence, T/S). Thus, the C/S measure was excluded from the analysis. In addition, research by Lu (2010) indicated that verb phrases per T-unit (VP/T) and complex phrases per T-unit (CP/T) exhibit lower validity in capturing language development. Therefore, these two indices are excluded from the analysis. In cases where multiple indices tapped into similar constructs (e.g., complex nominals per T-unit and clause), only one measure was selected, prioritizing the use of clause as the basic unit of measurement (CN/C). This decision is consistent with the approach taken by Yang et al. (2015).

As a result, the investigation included three length-based (mean length of sentence, T-unit, and clause: MLS, MLT, and MLC); two clausal-level (clauses per T-unit for subordination and T-unit per sentence for coordination: C/T and T/S), and two phrasal-level indices (coordinate phrases per clause and complex nominals per clause: CP/C and CN/C). Detailed information regarding the selected measures can be seen in Table 3. Following the application of the four writing evaluation tools to the selected samples from both groups, the compositions were subsequently modified using ChatGPT (version 4) to enhance language quality, overall organization, and the adequacy of content relevance. These modifications align with the essential requirements of Chinese argumentative writing (Zhang, 2018). To examine the impact of ChatGPT revision, paired simple *t*-test were employed, aiming to assess whether significant differences in CAOS measures existed before and after the ChatGPT use in EFL writing.

This systematic and fine-grained approach allowed for a comprehensive and objective analysis of EFL student writing before and after ChatGPT intervention, ensuring both methodological rigor and theoretical relevance.

Table - 2
Target lexical measures used in the study

Category	**Measure (code)**	**Formula**
Lexical complexity		
Lexical density	Lexical density (LD)	N_{lex}/ N
Lexical sophistication[1]	Lexical sophistication-I (LS1)	N_{slex}/ N_{lex}
	Lexical sophistication-II (LS2)	T_s/T
Lexical variation	Number of different words (NDW)	T
	Type/Token ratio (TTR)	T/N
	Corrected TTR (CTTR)	$T/\sqrt{2N}$
	Lexical word variation (LV)	T_{lex}/ N_{lex}
	Verb variation-I (VV1)	T_{verb}/ N_{verb}
	Verb variation-II (VV2)	T_{verb}/ N_{lex}
	Noun variation (NV)	T_{noun}/ N_{lex}
	Adjective variation (ADJV)	T_{adj}/ N_{lex}
	Adverb variation (ADVV)	T_{adv}/ N_{lex}
	Modifier variation (MODV)	$(T_{adj} + T_{adv})$/ N_{lex}

Notes. N = the number of words; N_{lex}= the number of lexical words; N_{slex}= the number of sophisticated lexical words; N_{verb}= the number of verbs; T = the number of word types; T_{lex}= the number of lexical word types; T_s= the number of sophisticated word types; T_{sverb}= the number of sophisticated verb types; # = number; / = divided by; * = multiply; T-unit: one main clause + any subordinate clause.

[1] According to American National Corpus, words that do not occur in the most frequent 2000 wordlist are sophisticated word

Table - 3
Target syntactic measures used in the study

Category	Measure (code)	Formula
Length of production	Mean length of sentence (MLS)	# of words/# of sentences
	Mean length of T-unit (MLT)	# of words/# of T-units
	Mean length of clause (MLC)	# of words/# of clauses
Subordination	Clauses per T-unit (C/T)	# of clauses/# of T-units
Coordination	T-unit per sentence (T/S)	# of T-units/# of sentences
Phrasal structures	Coordinate phrases per clause (CP/C)	# of coordinate phrases/# of clauses
	Complex nominals per clause (CN/C)	# of complex nominals/# of clauses

Results and Discussion

Comparative Analysis of CAS

Table 4 illustrates the comparative results of content, accuracy, and style (CAS) for student essays before and after revision with the assistance of ChatGPT. The results reveal several noteworthy trends that reflect the influence of AI-assisted revision on writing style and accuracy.

A significant reduction in total word count was observed after revision, with the mean falling from 248.53 to 218.38 (p = .000), corresponding to a 12.13% decrease. This reduction in length suggests that ChatGPT tends to help EFL learners make their writing more concise and focused. In line with this trend, the mean number of K1 words (the most frequent English words) also dropped significantly by 18.38% (p = .000), indicating a decrease in the reliance on basic vocabulary. On the contrary, the mean number of K2 words (less frequent but still common words) increased by 9.89%. While this increase was not statistically significant, it suggests a subtle shift toward more varied vocabulary. The most notable increase was observed in the use of AWL words. The number of those words doubled with the help of ChatGPT, with the mean number rising from 7.48 to 15.53 (p = .000), marking a 107.62% increase. This demonstrates that ChatGPT not only helps reduce wordiness but also enhances the academic tone of students' essays. Added to that, the use of Off List words increased significantly by 21.43% (p = .041), implying that students' writing became more diverse in vocabulary usage after revision.

Table - 4
Writing Parameters of the EFL Writing Before and After the ChatGPT Assistance

writing measures	before (mean)	after (mean)	improvement rate of after vs. before
Total words	248.53	218.38*** (p = .000)	-12.13%
K1 words	222	181.2*** (p = .000)	-18.38%
K2 words	13.45	14.78	9.89%
AWL	7.48	15.53*** (p = .000)	107.62%
Off List words	5.6	6.8* (p = .041)	21.43%
Different words	124.98	125.05	0.06%
Idea words	121.62	113.58** (p = .005)	-6.61%
Mistakes	9.1	1.6*** (p = .000)	-82.42%
Probability of writing errors	0.0376	0.0076*** (p = .000)	-79.79%

Note, only measures with notable effects are reported for space reasons, *p < .05; **p < .01; ***p < .001.

Despite such changes, some parameters, particularly different words, remained virtually unchanged (124.98 vs. 125.05), suggesting that hat conciseness did not come at the expense of variety in word choice. However, a slight decrease was noted in the number of idea words (from 121.62 to 113.58, p = .005), which may indicate a modest reduction in the explicit presentation of content, possibly due to the more concise and focused structure.

Crucially, the most significant improvements were found in writing accuracy. Both the mean number of mistakes and the probability of writing errors drastically fell with the help of ChatGPT. The number of mistakes dropped by 82.42% (from 9.1 to 1.6, p = .000) and the probability of writing errors decreased by 79.79% (from 0.0376 to 0.0076, p = .000), highlighting the effectiveness of ChatGPT in minimizing grammatical and mechanical errors in student writing.

Overall, the texts with the help of ChatGPT yielded noticeable gains in measures associated with *VP* word categories, K2, AWL, and Off List words in particular, indicating that ChatGPT recommended a wide range of uncommon and low-frequency vocabulary. This shift can be attributed to the limited proficiency of EFL writers. When assigned a writing task, they are expected to generate, plan, and organize their thoughts in line with the given theme. Accomplishing a well-developed piece of English writing

with rich content, accurate vocabulary usage, and complex linguistic expressions in a limited time frame, however, may be challenging for EFL learners since they most often encounter issues related to language proficiency and knowledge, resulting in limited information or content, simplified vocabulary, and syntax (Tsai, 2022). Consequently, they tend to choose familiar vocabulary and simple sentence structures, which explains why EFL students' English texts are typically short, contain more errors, and present a limited use of vocabulary. These issues, however, can be alleviated with the assistance of ChatGPT. With the intervention of ChatGPT, students' writing included a greater variety of higher-order vocabulary. Added to that, ChatGPT-assisted revision leads to shorter, more academically sophisticated, and more accurate student essays. The revisions not only elevate the academic register of writing but also ensure greater linguistic accuracy, reflecting the potential of AI tools to enhance the overall quality of EFL student compositions.

Comparative Analysis of Lexical Complexity

Table 5 presents the comparative results of various lexical complexity measures in student essays before and after the assistance of ChatGPT. The findings illustrate clear patterns regarding lexical density, sophistication, and variation, revealing significant enhancements in lexical quality attributable to AI-assisted revision.

Lexical density presented a statistically significant increase from 0.51 to 0.53 (p = .000), suggesting that the proportion of content words in the essays rose after ChatGPT intervention. This indicates a higher informational content per sentence, reflecting a move towards more content-rich academic writing. As for lexical sophistication, both measured indices (LS1 and LS2) demonstrated highly significant improvements. LS1 increased from 0.12 to 0.22 (p = .000), while LS2 rose from 0.12 to 0.2 (p = .000). Such findings highlight that ChatGPT encouraged students to use more advanced and less frequent vocabulary, thus enhancing the lexical sophistication. Regarding lexical variation, several metrics displayed significant upward trends. Such increase was statically significant in TTR from 0.48 to 0.54 (p = .000) and CTTR from 5.25 to 5.64 (p = .000). These improvements can also be observed on LV from 0.71 to 0.77 (p = .000), further illuminating that lexical diversity of EFL writing was markedly enhanced after ChatGPT's assistance.

A closer examination of other metrics of lexical variation also evidenced improvement. For instance, VV1 increased significantly (p = .026), and VV2 also showed a significant rise (p = .001). Similarly, NV (p = .000) and ADJV (p = .001) exhibited statistically significant gains. The measure for MODV likewise increased significantly from 0.22 to 0.24 (p = .000). These results collectively indicate a broader and more varied use of different parts of speech in the revised essays. However, ADVV did not show a significant difference, suggesting that ChatGPT may not have notably impacted the diversity of adverb usage.

Table - 5
Comparative Results of Lexical Complexity Before and After the Assistance of ChatGPT

	measures	**before**	**after**	***t***	***df***	***p***
Lexical density	LD	0.51±0.03	0.53±0.02	-5.30	39	.000***
Lexical sophistication	LS1	0.12±0.06	0.22±0.06	-10.44	39	.000***
	LS2	0.12±0.05	0.2±0.05	-10.50	39	.000***
Lexical variation	NDW	116.85±20.03	117.95±16.7	-0.53	39	.600
	TTR	0.48±0.05	0.54±0.04	-8.81	39	.000***
	CTTR	5.25±0.46	5.64±0.4	-5.80	39	.000***
	LV	0.71±0.11	0.77±0.09	-4.45	39	.000***
	VV1	17.58±5.58	19.46±4.29	-2.31	39	.026*
	VV2	0.2±0.03	0.22±0.03	-3.78	39	.001**
	NV	0.61±0.07	0.7±0.06	-6.69	39	.000**
	ADJV	0.12±0.03	0.14±0.03	-3.56	39	.001**
	ADVV	0.1±0.03	0.1±0.03	-0.86	39	.393
	MODV	0.22±0.04	0.24±0.04	-4.47	39	.000***

Note, *$p < .05$; **$p < .01$; ***$p < .001$.

In a nutshell, the date displayed in Table 5 reveals that the use of ChatGPT as a writing assistance tool leads to significant improvements in multiple dimensions of lexical complexity in EFL writing, most notably marked enhancements in lexical density, sophistication, and variation, with students employing more advanced and varied vocabulary after ChatGPT' s help. Such findings underscore the potential of ChatGPT to elevate the lexical quality of EFL student compositions, promoting not only richer vocabulary use but greater lexical diversity and sophistication.

Comparative Analysis of Syntactic Complexity

Table 6 displays a comparative analysis of various measures of syntactic complexity in EFL student writing before and after the assistance of ChatGPT. The results reveal several statistically significant improvements in syntactic features, implying that AI-assisted revision can effectively enhance the structural sophistication of EFL student compositions.

Table - 6
Comparative results of syntactic complexity before and after the assistance of ChatGPT

	measures	**before**	**after**	**t**	**df**	***p***
Length of production	MLS	15.15±2.82	16.72±2.3	-3.55	39	.001**
	MLT	13.17±2.13	14.19±2.38	-2.70	39	.010*
	MLC	8.11±1.3	9.17±1.25	-5.83	39	.000***
Subordination	C/T	1.64±0.27	1.56±0.23	1.77	39	.084
Coordination	T/S	1.15±0.14	1.19±0.17	-1.17	39	.250
Phrasal structures	CP/C	0.16±0.12	0.33±0.15	-7.09	39	.000***
	CN/C	0.75±0.23	0.88±0.25	-4.08	39	.000***

Note, *$p < .05$; **$p < .01$; ***$p < .001$.

The most prominent findings relate to the increased length of production units. The mean length of sentence increased significantly from 15.15 to 16.72 ($p = .001$), and the mean length of MLT rose from 13.17 to 14.19 ($p = .01$). Such increase also occurred in MLC from 8.11 to 9.17 ($p = .000$). These findings indicate that students tended to produce long and more elaborated sentences, T-units and clauses. This trend illuminates an enhancement in the ability to construct more complex and informative syntactic structures, which is often considered a marker of advanced language proficiency (Norris & Ortega, 2009). A coherent pattern can be observed in the use of phrasal structures, manifested by the marked rise in the frequency of CP/C ($p = .000$) and CN/C ($p = .000$).

Both sets of measures share a capacity to encode more information within a single syntactic unit, thus contributing to more advanced, academic-like writing.

The underlying cause for the similar pattern likely lies in the linguistic models that underpin ChatGPT. Trained on vast corpora of formal and academic texts, ChatGPT is predisposed to generate and suggest language that mirrors native-like academic discourse (Lee, 20222). As a result, students who revise their writing with the help of ChatGPT receive implicit exposure to advanced forms, especially complex noun phrases and information-dense structures that they might not otherwise encounter or produce independently in a typical classroom setting.

The combined increases of the two dimensions display a broader shift in writing style of EFL students: from reliance on shorter, simpler toward a more condensed and information-rich mode of expression. This trend is particularly salient in academic contexts, where efficiency and precision are higher valued. As Biber et al. (2011) highlight academic English is distinguished by its high degree of syntactic density, often achieved through the use of complex nominals and extended noun phrases rather than increased subordination. The present findings, therefore, suggest that ChatGPT not only helped lengthen syntactic units but promoted phrasal complexity that features expert academic writing.

No statistically significant differences were found in C/T (p = .084) and T/S (p = .250), revealing that the use of subordination and coordinating structures remained relatively stable despite the fact that students produced longer syntactic units. Such illumination is in line with the results of previous research wherein it shows that increasing sentence length does not necessarily correlate with higher use of subordination and coordination (Lu, 2011), and that EFL learners often find it challenging to diversify their use of these complex structures without explicit instruction.

Both the significant and unmarked findings point to the potential of AI-assisted tools not just to correct surface-level errors, but to scaffold the acquisition of deeper structural features of academic English. Importantly, these changes occurred without a corresponding rise in subordination or coordination, indicating a selective but substantial effect on particular aspects of syntactic complexity.

Conclusion

This study sets out to explore the effectiveness of using ChatGPT as a writing assistant tool for EFL students in senior high school. Their writing performance before and after the help of ChatGPT has been evaluated with fine-grained indices, specifically content, accuracy, complexity, and style. Given the substantial advantages demonstrated by ChatGPT in this investigation, it is crucial to leverage this tool effectively throughout the writing process to enhance writing outcomes. Empirical evidence from the present research shows that EFL students not only used ChatGPT as a resource for word selection, informational enrichment, and syntactic modeling but treated it as a "second audience" that provides valuable feedback during revision. Specifically, the significant improvements in content, style, accuracy, and complexity indicate that ChatGPT can assist learners in acquiring new lexical items, enriching their writing content, mastering

more advanced sentence structure, and gaining exposure to idiomatic expressions typical of official academic English.

Furthermore, by offering immediate error correction and suggestions for lexical and syntactic refinement, ChatGPT supports EFL learners in producing more accurate and varied language. This process-oriented use of AI encourages EFL students to view it as an iterative activity rather than a one-time product, thereby fostering greater learning autonomy and engagement. This perspective aligns with results from Lee (2020) and Niño (2008), which underscore the importance of process writing and learner autonomy for successful writing development. Overall, the integration of ChatGPT into EFL writing classrooms holds promise not only for improving linguistic performance but for cultivating positive learner attitudes towards writing and self-regulated learning.

While ChatGPT provides efficiency, convenience, and accuracy in EFL students' writing, there are still several potential challenges users should be aware of, such as innovation, accuracy, and transparency. Given the fact that ChatGPT is based on prior data fed to it, it will result in repetitive text generation and a lack of creativity and originality that human writers bring to their writing. Additionally, easily automated text generation in assignments or coursework may disengage learners in school education (Kasneci et al., 2023). Another concern about using ChatGPT in generating text is that the content of those texts may be misleading, thus necessitating users to develop competencies and literacy necessary to critical thinking and strategies for fact checking (Kasneci et al., 2023). Regarding transparency, it is acknowledged that using ChatGPT in the writing process or inclusion of text generated by machine should be made clear (Biswas, 2023).

Despite such challenges, we admit the powerfulness of those AI-powered language models, such as ChatGPT in assisting writers in writing tasks. Nonetheless, careful consideration of potential risks associated with the utilization of AI in EFL writing is imperative if EFL students move forward with adopting the language model. Moreover, it is crucial for EFL instructors to have a comprehensive understanding of ChatGPT use in the context of EL. This understanding will help develop suitable instructional models for integrating ChatGPT into diverse educational settings rather than discouraging its usage as suggested by McMurtrie (2022). Analysis of his study illuminated that tools like ChatGPT will inevitably become an integral part of everyday writing, similar to how calculators and computers have become indispensable tools in the field of mathematics and science.

While ChatGPT offers notable advantages in terms of efficiency, convenience, and linguistic accuracy for EFL student writing, it is important to recognize and address several potential challenges associated with its use. One of the main concerns is the risk of reduced creativity and originality, as ChatGPT generates content based on large datasets of existing texts. This reliance on prior data can lead to repetitive phrasing and a lack of innovative thought - qualities that are best fostered through human agency in the writing process. The tendency towards formulaic or generic text generation may ultimately undermine the development of students' unique voices and creative expression. Additionally, the ease of automated text production may diminish students' engagement and investment in their own learning. As noted by Kasneci et al. (2023), excessive

dependence on AI-generated writing in academic settings can result in passive participation, potentially hampering the development of essential writing and critical thinking skills. Another critical issue is the potential for ChatGPT to produce misleading or factually inaccurate content, given its probabilistic language modeling. This challenge underscores the importance of cultivating critical digital literacy among learners, equipping them with the ability to evaluate and fact-check AI-generated output (Kasneci et al., 2023). Transparency is another key consideration. It is increasingly recognized that the use of AI tools like ChatGPT in the writing process, including the incorporation of machine-generated text, should be disclosed to ensure academic integrity and maintain trust in educational settings (Biswas, 2023).

Rather than discouraging the use of ChatGPT, it is more productive to promote responsible and informed engagement. As McMurtrie (2022) argues, AI-based writing tools are likely to become as ubiquitous in language learning as calculators and computers are in mathematics and science. Therefore, the focus should be on guiding students to use ChatGPT and similar tools as effective supports for language learning, while also fostering the skills necessary for critical evaluation and creative authorship.

References

Abdelghani, R., Wang, Y. H., Yuan, X., Wang, T., Sauzéon, H., & Oudeyer, P. Y. (2022). GPT-3-driven pedagogical agents for training children's curious question-asking skills. *arXiv preprint arXiv:2211.14228*.

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. (2023). Impact of ChatGPT on learning motivation: Teachers and students' voices. *Journal of English Studies in Arabia Felix*, *2*(1), 41-49.

Aydın, Z., & Yıldız, S. (2014). Using wikis to promote collaborative EFL writing. *Language Learning & Technology*, *18*(1), 160-180.

Biswas, S. (2023). ChatGPT and the future of medical writing. *Radiology*, *307*(2), 223312.

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*, 2041-2058.

Chen, F., Tsai, S., & Tsou, W. (2019). The application of translingual practice in a college ESP writing. *English Teaching & Learning*, *43*(1), 65-83.

Chung, E. S., & Ahn, S. (2022). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, *35*(9), 2239-2264.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135.

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). *Reading Comprehension Quiz Generation using Generative Pre-trained Transformers*. Retrieved from https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf

Dizon, G., & Gayed, J. M. (2021). Examining the impact of Grammarly on the quality of mobile L2 writing. *JALT CALL Journal, 17*(2), 74-92.

Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A. M. A., & Hatta, H. R. (2023). Analysing the role of ChatGPT in improving student productivity in higher education. *Journal on Education, 5*(4), 14886-14891.

Fitria, T. N. (2022). Avoiding plagiarism of students' scientific writing by using the QuillBot Paraphraser. *Elsya: Journal of English Language Studies, 4*(3), 252-262.

Gabajiwala, E., Mehta, P., Singh, R., & Koshy, R. (2022). *Quiz Maker: Automatic Quiz Generation from Text Using NLP*. Paper presented at the Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021.

Godwin-Jones, R. (2018). Second language writing online: An update. *Language Learning & Technology, 22*(1), 1-15.

Guo, K., Wang, J. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in RFL writing. *Education and Information Technologies, 29*, 8435-8463.

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing, 54*, 100666.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics, 53*, 61–79.

Kamberi, L. (2013). The significance of teacher feedback in EFL writing for tertiary level foreign language learners. *Procedia-Social and Behavioral Sciences, 70*, 1686-1690.

Kasneci, E., SeBler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450.

Kurniati, E. Y., & Fithriani, R. (2022). Post-graduate students' perceptions of Quillbot utilization in English academic writing class. *Journal of English Language Teaching and Linguistics, 7*(3), 437-451.

Li, J., & Li, M. (2018). Turnitin and peer review in ESL academic writing classrooms. *Language Learning & Technology, 22*(1), 27-41.

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning, 32*(7), 784-804.

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

McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing, 52*, 100608.

McMurtrie, B. (2022). AI and the future of undergraduate writing. *The Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/newsletter/teaching/2023-01-05

Nagata, R., Hashiguchi, T., & Sadoun, D. (2020). *Is the Simplest Chatbot Effective in English Writing Learning Assistance?* Paper presented at the Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics.

Okasha, M. A., & Hamdi, S. A. (2014). Using strategic writing techniques for promoting EFL writing skills and attitudes. *Journal of Language Teaching & Research, 5*(3), 674-681.

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. Paper presented at IEEE Global Engineering Education Conference (EDUCON).

Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning, 67*(1), 208-241.

Seyyedrezaei, M. S., Amiryousefi, M., Gimeno-Sanz, A., & Tavakoli, M. (2022). A meta-analysis of the relative effectiveness of technology-enhanced language learning on ESL/EFL writing performance: retrospect and prospect. *Computer Assisted Language Learning*, 1-34.

Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4-16.

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing, 7*(1), 35.

Tsai, S. C. (2017). Effectiveness of ESL students' performance by computational assessment and role of reading strategies in courseware-implemented business translation tasks. *Computer Assisted Language Learning, 30*(6), 474-487.

Tsai, S. C. (2019). Using Google Translate in EFL drafts: a preliminary investigation. *Computer Assisted Language Learning, 32*(5-6), 510-526.

Tsai, S. C. (2022). Chinese students' perceptions of using Google Translate as a translingual CALL tool in EFL writing. *Computer Assisted Language Learning, 35*(5-6), 1250-1272.

Uyan, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20-32.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 1-25.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53-67.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.4312418

Zhang, Y. (2018). An investigation into the development of structure and evidence use in argumentative writing. *Theory and Practice in Language Studies, 8*(11), 1441-1448.