Exploring Whether Generative Artificial Intelligence Can Enhance Chinese EFL Learners' Academic Writings

Liu Hongqiao and Jin Jing*

School of Foreign Languages, Southeast University, China
Email: 1009128806@qq.com (L.H.); 101005058@seu.edu.cn (J.J.)
*Corresponding author

Manuscript October 24, 2024; revised November 12, 2024; accepted December 19, 2024; published April 18, 2025

***Abstract*—Generative Artificial Intelligence (GAI) like Chat Generative Pre-Trained Transformer (ChatGPT) has been widely used by English as a Foreign Language (EFL) students to polish their academic writing, but little quantitative research has explored how the chatbot can revise human writing. Addressing this research gap is crucial, as understanding the potential of GAI in enhancing the quality of EFL writings can inform both pedagogical practices and the development of computer-assisted writing tools. In response to the issue, this study compared five textual features of Chinese EFL learners' academic writings and their revised versions of ChatGPT through a computational analysis tool, Coh-Metrix. The results of the Wilcoxon Signed-Rank Test showed that the GAI could significantly enhance the writing quality of handwritten texts only in narrativity and deep cohesion. By illustrating the performance of ChatGPT as an editor, this study has provided insight into pedagogical instruction on English academic writing and computer-assisted writing.**

***Keywords*—generative artificial intelligence (ChatGPT), academic writing, writing quality**

I. Introduction

Given the pivotal role of writing in the language acquisition process, English as a Foreign Language (EFL) learners are struggling to improve their writing with the help of technology. The ability to produce high-quality academic writing is essential for EFL students in Chinese universities, enabling them to publish articles in international journals. In their pursuit of enhancing writing skills, students often turn to computer-based tools, such as Grammarly (https://grammarly.com/) for automated feedback and Pigai (http://www.pigai.org/) for automated evaluation. The advent of technology has empowered students to leverage Generative Artificial Intelligence (GAI), allowing them to revise their work with technological assistance. Compared to GAI, traditional tools like Grammarly and Pigai focus primarily on surface-level corrections (grammar, spelling) and provide limited personalized feedback, lacking depth in content and timely interaction essential for enhancing the level of EFL students' writing. An exemplary instance of the smarter GAI tools is ChatGPT (Chat Generative Pre-Trained Transformer), a sophisticated language model launched by Open Artificial Intelligence in November 2022 and renowned for its capacity to generate articles, stories and other forms of realistic and coherent written content [1].

The growing popularity of ChatGPT has attracted an increasing number of scholars' attention to its role in writing. These researchers are predominantly in the field of medicine [2–4], focusing on evaluating ChatGPT's ability to generate scientifically accurate content. In addition, researchers in applied linguistics have discussed both the advantages and disadvantages of integrating ChatGPT into academic writing [1, 5–7]. However, to the best of our knowledge, the extent to which the chatbot can surpass human writers across all facets of writing remains unknown, and only a limited number of studies have employed a quantitative method to assess the writing quality of human-written texts compared to their ChatGPT-revised versions.

This study aims to investigate whether and how GAI tools like ChatGPT can enhance Chinese EFL learners' academic writing. Specifically, this study analyzed and compared five textual features (i.e., narrativity, syntactic simplicity, word concreteness, deep cohesion, and referential cohesion) in the two kinds of texts through a computational tool called Coh-Metrix-T.E.R.A. (Text Ease and Readability Assessor). These five features, according to Graesser's team [8], cover the most important five levels in the multilevel theoretical framework: the genre, the situation model, the text base, the syntax, and the words and could reflect the text complexity accurately.

II. Literature Review

A. *ChatGPT and Language Learning*

GAI could integrate machine-learning models to produce new content, including text, audio, video, images, software code and simulations, based on large datasets [9]. One of the most popular GAI tools favored by students recently is ChatGPT, a large language model-based chatbot, processing data and information from the Internet to provide users with automated text.

The potential of ChatGPT on language learning and teaching has aroused the academia's great attention since its launch in 2022. Research in this field can be broadly categorized into two distinct perspectives. The first category of studies has demonstrated the chatbot's impressive language understanding and generation capabilities in language learning while the other type of studies has expressed the academia's concern for its misuse.

Several studies have investigated the chatbot's applications in diverse facets of language acquisition, including reading [10, 11], listening [12, 13], translation [14, 15], and writing [16]. Xiao *et al.* [10] reported that ChatGPT can customize reading materials focused on specific language skills for learners of different language proficiency. Similarly, Wang and Feng's experiment [11] discovered that the ChatGPT-assisted group, which utilized ChatGPT for reading assistance and analysis did better than the control group.

doi: 10.18178/ijiet.2025.15.4.2280

According to Xing [12], ChatGPT has opened new avenues for enhancing our English listening skills by helping students improve language processing and comprehension. Meanwhile, Aryadoust *et al.* [13] found the effectiveness of using ChatGPT to develop materials for listening assessment. ChatGPT has also been reported as a good translator [14]. In Sahari's team's study [14], most teachers and students favored the GPT-translated version over the Google translation. Moreover, Chan and Tang's systematic review [15] found that GPT-generated translations, which excel in handling cultural texts, complex structures, and advanced linguistic features, are comparable to human translations, outperform neutral machine translation outputs, and can also be effectively used for post-editing and translation evaluation, raising new challenges and ethical concerns. Among these English skills, writing is the most favored by scholars focusing on the impact of ChatGPT and relevant studies are reviewed in the next section. In general, these findings acknowledge the positive role of chatbots in enhancing the four language subskills.

Despite its remarkable language capabilities, the integration of GAI into language learning and teaching has sparked concerns among researchers. Several studies have also discussed the negative role of ChatGPT in practice, such as issues of plagiarism [17] and educational equity [18]. Early in 2021, Dehouche [17] pointed out that the issue of plagiarism looms large in the use of Artificial Intelligence (AI) systems, as students utilizing the technology may be more prone to submitting assignments containing AI-generated content. Additionally, Cotton *et al.* [18] argued that the uneven accessibility to AI could exacerbate educational resource inequalities, potentially providing unfair advantages to certain students.

Like it or not, the recently released ChatGPT is poised to become more prevalent among language learners. However, the amount of studies on ChatGPT is limited and thus further research on the impact of applying this technology on language teaching and learning holds significant importance.

B. ChatGPT and Writing

Given the popular application of ChatGPT on writing, many articles have explored the potential of ChatGPT-assisted writing. A majority of papers are editorials discussing the benefits and threats of ChatGPT-assisted writing [1, 5–7]. These scholars provide a broad overview of ChatGPT's pros and cons, considering it as a mixed blessing that it greatly improves scientists' writing efficiency and quality while risking plagiarism and fabrication. In addition, several studies have examined the feasibility of blending this technology into writing instruction. Schmohl *et al.* [19] are the first scholars to plan to enhance students' writing skills with the help of the text generator language model GPT-2 from the OpenAI. Similarly, a study in 2023 [16] also discussed the possible applicability of ChatGPT in assisting learners with writing tasks. Furthermore, Yan's study [20] has made the application of GAI in language teaching a reality through a practicum allowing students to use ChatGPT in writing. The studies are based on a hypothesis that ChatGPT could contribute to the writing, which is waiting for testing. In other words, scholars have not reached a consensus on whether GAI tools outdo humans in writing.

To answer the question, researchers have compared AI-generated texts with human writings. For example, Zhou *et al.*'s study [21] evaluated five discourse components of the AI-generated narrative writings and those of undergraduate English majors and concluded that ChatGPT outperformed human writers in narrativity, word concreteness, and referential cohesion, but AI underperformed in syntactic simplicity and deep cohesion. Likewise, many researchers in the field of medicine compared ChatGPT's writing to a human-written essay under the same topic [2–4] with different findings. Ho *et al.* [3] acknowledged AI as "a powerful medical writing tool" by conducting a comparative analysis of case reports written by a first-year medical student and ChatGPT. By contrast, Ariyaratne *et al.*'s study [2] found that AI-generated articles were inaccurate in academic knowledge and cited fictitious references and indicated the unreliability of automated generated texts. These studies could be considered the starting point of the comparison between ChatGPT-written and human-written texts.

The overview reveals two prominent characteristics of contemporary studies: Firstly, the qualitative method has prevailed across the majority of the research, relying on expert assessment of texts rather than employing precise and impartial computational analyses. While such assessments provide valuable insights, they are inherently subjective and may lack consistency [22]. To enhance the objectivity and replicability of results, the quantitative method is instrumental in providing precise and impartial analyses. Secondly, these studies employed ChatGPT as an automated writer to produce articles from scratch, ignoring its role as a reviser or editor to polish human writings.

C. Evaluating Writing by Coh-Metrix

How to evaluate the excellence of writing and improve students' writing ability is a universal challenge confronting scholars and teachers. Since the 1970s, researchers have connected particular linguistic features in writing to the proficiency and advancement of writing skills [23]. To them, the text analysis of the academic writing could reflect the writer's writing ability. However, traditional human scoring methods often suffer from subjectivity, inconsistency, and limited scalability, which can hinder their effectiveness in large-scale assessments. In light of the development of technology, more scholars have realized the importance and convenience of automated tools in assessing learners' writing capabilities [24, 25]. One of the most popular approaches to exam textual data is machine learning-based, which is characterized by Natural Language Processing (NLP). NLP tools are employed to calculate and explore the language discourse of texts, especially linguistic features, one of which is Coh-Metrix [26].

Coh-Metrix, created by Arthur C. Graesser and Danielle S. McNamara in 2004, is a computational tool extensively utilized in L2 writing research to assess linguistic and discourse features, particularly focusing on writing quality analysis [27, 28] and a comparison of differences between variations of texts [29]. McNamara *et al.*'s study [27] concluded that high-proficiency essays were more likely to contain linguistic features associated with text difficulty and sophisticated language by using Coh-Metrix to assess the linguistic features. Likewise, Maamuujav *et al.*'s study [28] analyzed essays using manual sentence coding and quantitative measures from Coh-Metrix to assess syntactic and lexical features. Moreover, Graesser *et al.* [29] reviewed how the five textual features (narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion) that are major factors of diverse measures in Coh-Metrix explained text variations. In conclusion, Coh-Metrix has been proven by many scholars as a reliable automated scoring tool. Therefore, based on these five textual features as shown in Table 1, this study adopted Coh-Metrix to compare the Chinese EFL learners' academic writings and their ChatGPT-revised versions.

In conclusion, there is a limited amount of quantitative research on the recently released GAI tool, ChatGPT, and the question of how it can enhance the quality of written text remains unknown. Therefore, this study employed the text-analysis tool, Coh-Metrix, to quantitatively assess the human-written academic articles and their ChatGPT-refined versions from five major dimensions. This study aims to contribute to the development and regulation of computer-assisted writing by answering the following questions:

1) How do the students perform in English academic writing in terms of the five textual features?
2) How do ChatGPT-revised versions differ from human writing in terms of the five discourse components?

Table 1. The introduction of five textual features [29]

Textual Features	Definition
Narrativity	Narrative text depicts a story, featuring characters, events, locations, and objects familiar to the reader, closely linked to everyday oral communication.
Syntactic Simplicity	Short sentences with familiar, straightforward syntax are easier to comprehend, while complex sentences entail embedded syntactic structures.
Word Concreteness	Concrete words evoke vivid mental images and hold more significance for readers compared to abstract words.
Referential Cohesion	High-cohesion texts feature words and ideas that span sentences and the entire text, creating interconnected threads that link the explicit text base.
Deep Cohesion	Causal, intentional, and other connectives aid in fostering a more coherent and comprehensive comprehension of the text, particularly at the level of the causal situation model.

III. Methods

This study built two mini-corpora that compromised human-written texts and ChatGPT-revised texts and quantitatively compared the Coh-Metrix scores of five indicators of the two corpora, aiming to explore the feasibility of the new form of computer-assisted writing.

A. *Instrument*

Following the paradigm employed by Zhou *et al.* [21], this study utilized Coh-Metrix for the analysis of texts. Coh-Metrix is a computational tool, exceling in analyzing text across various dimensions related to cohesion and text difficulty [29]. Over 50 published studies have underscored its effectiveness in detecting subtle differences in text and discourse [27]. This study employed the free version of Coh-Metrix, Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.) (https://soletlab.adaptiveliteracy.com:8443/) [30]. It is a tool specifically developed to assess the "ease" and readability of texts. T.E.R.A. evaluates texts based on five key components: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion (see Table 1). These components are each assigned an "ease" score for a given text, indicating how it stands relative to thousands of other texts.

B. *Data Collection*

Two corpora were built in the study, the human-written corpus and the ChatGPT-revised corpus.

The researchers established the human-written corpus by collecting English writings produced by doctoral students in a top-20 university in South China during the final exam of an English academic writing course. These students, totaling 21, were non-English major students specializing in STEM (Science, Technology, Engineering, and Mathematics) subjects. All the students were native Chinese speakers who had successfully passed the College English Test-6, a national examination for non-English majors in China, ensuring they possessed an intermediate level of English proficiency and could write fluently and accurately in English. They all hailed from the class taught by the second author and consented to the use of their writings for research purposes. The exam required students to summarize Kathleen E. Grogan's passage "General Tips for Success in Practicing the Art of Writing," with no specified word limit. Given that the course focused on academic writing and the summarized article in the exam is related to academic writing, this task also falls within the realm of academic writing. During the exam, students were strictly prohibited from accessing dictionaries, reference books, or automatic grammar and spelling checker tools. Upon completion, the researchers transcribed their handwritten writings into electronic versions, thus creating the initial mini-corpus. The corpus comprised twenty-one essays, each averaging 217 words in length, with a total of 4,567 words.

After collecting their articles, researchers used ChatGPT 3.5 to revise their pieces of English writing. A prompt for ChatGPT 3.5 was designed after recognizing the sensitivity of ChatGPT to phrasing according to the suggestions of previous studies, such as reference [31]. In particular, the researchers offered the AI a specific role, a realistic context, and the rules and tone of the writing assignment. The prompt is as follows:

I'm writing a summary for an academic passage named "General Tips for Success in Practicing the Art of Writing". Please rephrase it for clarity, coherence and conciseness, ensuring each paragraph flows into the next. Remove jargon. Use a professional tone.

To avoid the intervention of previous texts, the same prompt template was entered in separately twenty-one chat boxes and twenty-one independent revised versions of texts were generated. This helped to avoid variation in the generated outputs. These polished texts compromised the second corpora with a total of 4,450 words.

C. Data Analysis

After constructing the two corpora, comprised of human-written texts and their corresponding GPT-revised versions, each corpus underwent separate analysis using Coh-Metrix-T.E.R.A., focusing on five textual features (refer to Table 1 for details). Each of the five features was scored ranging from 0.01 (1%) to 1 (100%) for a single essay based on a comparison to thousands of other texts within the predefined corpora. The researchers entered each article into the website, recorded the scores of five components in Excel, and then transferred the data to SPSS. This process was repeated 42 times for a total of 42 essays.

Table 2. Shapiro-Wilk tests for normality (1-Original Version; 2-Revised Version)

Features		Statistic	N	Sig.
Narrativity	1	0.90	21	0.04
	2	0.93	21	0.12
Syntactic Simplicity	1	0.95	21	0.41
	2	0.97	21	0.77
Word Concreteness	1	0.98	21	0.85
	2	0.93	21	0.16
Referential Cohesion	1	0.78	21	0.00
	2	0.87	21	0.01
Deep Cohesion	1	0.73	21	0.00
	2	0.86	21	0.01

Subsequently, the non-parametric Wilcoxon Signed-Rank Test was employed to examine the difference between the two versions, as the data violated normality and were not suitable for parametric tests. The results of Shapiro-Wilk tests for normality in Table 2 reveal that the majority of the data adhered to normality, as indicated by significance levels exceeding 0.05. However, data pertaining to deep cohesion and referential cohesion exhibited significance levels considerably below 0.05. Therefore, the researchers replaced the paired sample t-test with the non-parametric Wilcoxon Signed-Rank Test to improve the accuracy of the results.

IV. Results and Discussion

A. Overall Performance Evaluation of Two Corpora

This study aims to conclude the performance in the Coh-Metrix of the two corpora through statistical tests and explore the effectiveness of ChatGPT editing on human writing.

A Wilcoxon Signed-Rank test was conducted to investigate the difference between the ChatGPT-revised version and the original one. Coh-Metrix generated percentile scores ranging from 0.01 (1%) to 1 (100%). Table 3 demonstrates the descriptive results in terms of the five dimensions. As Graesser *et al.* [8] noted, the five dimensions are articulated about comprehension ease. Therefore, text difficulty, conversely defined as the opposite of ease, is the reversal of principal component scores in measures. It means that the lower score in these components means the higher quality of academic writing.

Table 3. Descriptive statistics of five features (1-original version; 2-revised version)

Features	Versions	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
							25th	50th (Median)	75th
Narrativity	1	21	0.38	0.18	0.13	0.66	0.24	0.34	0.58
	2	21	0.20	0.13	0.03	0.48	0.10	0.16	0.30
Syntactic Simplicity	1	21	0.58	0.24	0.08	0.98	0.40	0.61	0.74
	2	21	0.59	0.15	0.26	0.88	0.50	0.58	0.73
Word Concreteness	1	21	0.57	0.20	0.22	0.96	0.42	0.54	0.73
	2	21	0.57	0.23	0.21	0.93	0.30	0.59	0.76
Referential Cohesion	1	21	0.32	0.31	0.04	0.98	0.10	0.17	0.47
	2	21	0.19	0.17	0.01	0.71	0.05	0.19	0.29
Deep Cohesion	1	21	0.88	0.16	0.48	1	0.82	0.95	1.00
	2	21	0.79	0.22	0.22	1	0.66	0.79	0.98

Table 4 shows the results of the non-parametric test. For Narrativity, a significant difference was found between the two versions ($Z = -3.27$, $p = 0.001$), with a large effect size ($r = -0.71$). However, no significant differences were observed for Syntactic Simplicity ($Z = -0.19$, $p = 0.85$, $r = -0.04$) and Word Concreteness ($Z = -0.19$, $p = 0.85$, $r = -0.04$). Referential Cohesion showed a trend towards significance ($Z = -1.72$, $p = 0.09$) with a medium effect size ($r = -0.38$), while Deep Cohesion exhibited a significant difference ($Z = -2.54$, $p = 0.01$) with a relatively large effect size ($r = -0.55$). Accordingly, the ChatGPT-revised version and the original version differ significantly in their performance on narrativity and deep cohesion and a relatively small difference was detected in referential cohesion. Each textual feature is feature is discussed as follows:

Table 4. Overall performance of two corpora (1-original version; 2-revised version)

Pairs of Features (2 – 1)	Negative Ranks			Positive Ranks			Ties	Test Statistics		
	N	Mean Rank	Sum of Rank	N	Mean Rank	Sum of Rank	N	*Z*-Value	Sig. (2-Tailed)	Effect Size (*r*)
Narrativity	17	12.32	209.50	4	5.38	21.50	0	−3.27[a]	0.00	−0.71
Syntactic Simplicity	9	12.22	110.00	12	10.08	121.00	0	−0.19[b]	0.85	−0.04
Word Concreteness	11	10.00	110.00	9	11.11	100.00	1	−0.19[a]	0.85	−0.04
Referential Cohesion	14	11.79	165.00	7	9.43	66.00	0	−1.72[a]	0.09	−0.38
Deep Cohesion	16	9.88	158.00	3	10.67	32.00	2	−2.54[a]	0.01	−0.55

[a] Based on positive ranks; [b] Based on negative ranks.

B. Comparative Analysis of Two Corpora Across Various Components

1) Narrativity

The Wilcoxon Signed-Rank test implied a significant difference between the two versions in narrativity, hinting that ChatGPT lowered the scores of narrativity. It means that ChatGPT could decrease the narrativity of human writings and increase the abstractness.

As illustrated in the website of Coh-Metrix-T.E.R.A., texts high in narrativity are characterized by a high proportion of verbs, common words, and pronouns, and a low narrativity score indicates a higher occurrence of uncommon words and potentially an increased abundance of information and ideas. Academic papers tend to use abstract nouns or terminologies to convey more information while narratives employ more common verbs or daily expressions [32]. Because the scaled texts were academic summaries of an exposition, the reduced narrativity observed in the ChatGPT-revised version highlights the chatbot's adeptness in refining academic texts.

The data shows EFL learners preferred to use the first-and second-person pronouns. The researchers searched 70 results of second-person pronouns and 16 results of first-person pronouns in human writing texts. By contrast, only four results of "your" and four results of "we" were found in the ChatGPT-revised version. This discovery is identified with the conclusion of Gao's research [33]. Gao found that the plural first-person pronoun "we" was used more often by Chinese EFL writers than English natives. MacIntyre [34] also pointed out that using personal pronouns, such as "I" and "we", is often seen as contradicting the need for objectivity and formality in academic writing. Jiang and Hyland [35] illustrated the use of first personal pronouns from Hyland's metadiscourse framework. They found more explicit authorial self-mention bundles in British students' essays than in GPT-generated texts, indicating students' desire to gain recognition for an individual voice from readers.

Another reason for the lower narrativity score in the refined texts is that ChatGPT deleted the common daily language in the original texts. The chatbot prefers to use formal or technical words. For instance, "initiate" in the revised version was used to replace "start" in the original text and "keep writing" was refined into "perseverance in writing". This assumption is also identified with the study by Zhou *et al.* [21] that concluded ChatGPT outperformed the Chinese Intermediate English learners in narrativity.

2) Syntactic simplicity

Syntactic simplicity is determined by several indices, including the average number of clauses per sentence, the word count per sentence, and the number of words preceding the main verb of the principal clause. Texts with fewer clauses, shorter sentences, and fewer words preceding the main verb tend to score higher for syntactic simplicity. However, the high *p*-value in the test revealed no obvious effectiveness of ChatGPT in improving the syntactic simplicity of learners' writings.

This conclusion is against the findings of Zhou and the team's study [21], which not only identified a significant difference between learners' writing and the initial version of ChatGPT-generated writing but also found no distinct difference between learners' writing and the revised version of ChatGPT-generated writing. On the one hand, the insignificant effect in this study may be attributed to the smaller sample size in the current study. On the other hand, as Zhou *et al.* [21] wrote, syntactic simplicity is not always a reliable sign reflecting the quality of writing. Different educators and scholars have different definitions of a good essay. For instance, Broadhead *et al.* [36] offered two indicators of good academic writing: shorter base clauses and a higher percentage of words in free modifiers, particularly when placed after a base clause. However, teachers would not pay excessive attention to the sentence structure.

A possible reason for the large *p* value is that ChatGPT rarely changes the sentence structure unless a clear demand is given while preferring to refine the words. For example, the sentence "it is significant for scientists to improve writing practices, and to become prolific and confident writers" is polished by ChatGPT into "it is essential for scientists to enhance their writing practices and become proficient and confident writers". AI only changed the words "significant" and "improve" without altering its sentence structure.

3) Word concreteness

Concrete words are contrasted with abstract words. In academic genres, writers are more likely to use abstract nouns or concepts [32]. Therefore, the lower percentile of word concreteness in writing suggests more academic rigor. Pitifully, a significant difference in terms of word concreteness was not found in the current study, which, again, may be related to the small sample size.

In terms of word concreteness, Zhou *et al.* [21] found that ChatGPT outperformed the Chinese intermediate English learners in this aspect when writing narratives. From their perspective, ChatGPT's more use of concrete words reflected the chatbot's ability of logic-based narration, such as recounting events in spatial or chronological order, but lacked depth in addressing event-related contexts. By contrast, Jiang and Hyland [35] found Bundles specifying abstract qualities are a common feature in the essays, appearing twice as often in the ChatGPT texts as in the students' essays after comparing the 3-word bundles in argumentative essays written by native English speakers with those generated by ChatGPT. Their contradictory results just demonstrate the uncertainty and flexibility of Generative Artificial Intelligence tools in writing different genres. Future studies could enlarge the sample size and clarify whether the AI revisor prefers to use abstract or concrete words in diverse genres.

Similarly, there is no consensus in academia on the relationship between abstract words and writing quality. Some scholars suggested that more proficient writers tend to use fewer abstract items. McNamara *et al.* [27] observed that higher-quality essays featured a greater abundance of specific and easily visualized words, whereas Crossley *et al.* [37] discovered that essays written at a higher-grade level contained more concrete words with fewer possible interpretations. According to Crossley [23], this conceptually aligns with the idea that advanced writers are likely to provide more detailed evidence to substantiate their claims, which may be reflected lexically by the increased usage of easily visualized, concrete, and precise words.

4) *Referential cohesion*

Referential cohesion involves the link or overlap among words, word stems, or concepts spanning across sentences or paragraphs within a text. It is established through the repetition of words or phrases and the consistent use of terminology or concepts throughout the text [38]. Ultimately, referential cohesion enhances coherence and comprehension by maintaining continuity and reinforcing important concepts across the text.

The test indicates a trend towards significance ($Z = -1.72$, $p = 0.09$) with a medium effect size ($r = -0.38$), slightly over the set significance level (alpha = 0.05). Specifically, certain writings in the original version exhibited a higher percentile of referential cohesion markers compared to those in the revised version. On the one hand, the barely significant t-value could be influenced by the non-normality of the data. The larger sample size could lead to a more normal distribution of data and thus a more reliable result. On the other hand, this discrepancy could be attributed to learners' tendency to use the same words to refer to the same thing due to limited vocabulary, whereas AI, endowed with robust language abilities, can pursue a greater diversity of expressions.

However, research on the utilization of referential cohesive markers among college-level writers yields mixed results. Early research by Witte and Faigley [39] initially noted a greater density of cohesive ties in higher-quality essays crafted by college students. However, recent studies present conflicting results, with some indicating a notable negative correlation [40], while others propose a positive link [41] between referential cohesion across sentences and text quality. Overall, the relationship between referential cohesion and text quality remains subject to debate and further investigation.

5) *Deep cohesion*

Deep cohesion measures the extent to which events, ideas, and information within a text are interconnected. T.E.R.A. accomplishes this by assessing various types of words that establish connections between different sections of the text, known as connectives. These connectives include causal, additive, logical, and adversative connectives. By employing these connectives, the text's coherence is enhanced, facilitating comprehension for the reader. A higher occurrence of connectives, particularly when contextual demands necessitate their use, corresponds to greater ease in achieving deep cohesion within the text [42].

A significant difference in deep cohesion between the two corpora was identified, supported by the p-value ($p = 0.01 < 0.05$). With the large effect size ($r = -0.55$), it is safe to claim the effectiveness of ChatGPT as a reviser on deep cohesion in a study with a larger sample size.

Upon closer examination of the texts, researchers noted a diverse range of connectives employed in both versions, albeit with fewer instances observed in the ChatGPT-revised versions. This finding is understandable given that the selected article for summarization outlined five suggestions for academic writing, allowing students to logically sequence these suggestions using connectives such as "firstly" and "secondly". All texts in the original version predominantly utilized logical connectives. However, quantity does not necessarily equate to quality. Students often utilized repetitive and mechanical sequencing connectives, whereas the GAI accentuated internal cohesion.

Other studies generated different findings. Zhou *et al*. [21] and Zhao *et al*. [43] asserted limitations of coherence in ChatGPT. For instance, according to Zhou and his team's statistics, ChatGPT's revised version used fewer causal connectives (e.g., "therefore," "hence")—averaging 20.29 per sentence—but tended to overuse "and," repeating it more than 15 times. These researchers argue that despite recent advances in NLP, improving coherence in generated texts remains a noteworthy challenge due to AI's dependence on statistical patterns rather than a comprehensive grasp of context and meaning. However, the contrasting results between the current study and others may stem from differences in research design. In this study, ChatGPT was used as a revisor, which allowed the human writers' logic to remain intact in the essays.

V. Conclusion

This study examined the textual quality of writings by EFL learners and those revised by ChatGPT using Coh-Metrix-T.E.R.A., focusing on five textual features to assess ChatGPT's potential as an EFL writing revisor. The results suggested that while ChatGPT was effective in enhancing narrativity, referential cohesion, and deep cohesion, but failed in syntactic simplicity and word concreteness.

The study contributes to the ongoing development of ChatGPT and similar AI technologies, informing efforts to refine their capabilities in assisting language learners and improving overall writing quality. In addition, this study underscores the value of leveraging ChatGPT in EFL writing instruction. It not only provides students with instant feedback and revision suggestions but also offers opportunities for educators to explore innovative approaches to writing pedagogy. The use of AI tools like ChatGPT could foster a more personalized and efficient learning experience, enabling students to receive tailored suggestions for improvement. This is particularly valuable in large classrooms where individualized feedback is often limited. Therefore, EFL teachers could use these AI tools as teaching assistants to grade student essays, save time, and provide quantitative data which could assess student progress. Students can also use AI to edit and restructure sentences or paragraphs, learning various ways to express ideas and organize their work. However, more importantly, educators should warn students to use AI as a learning partner instead of a learning surrogate and teach them to critically evaluate the AI-generated content.

However, this study is not without limitations. Firstly, the evaluation solely relied on the Coh-Metrix-T.E.R.A. tool, which may not comprehensively capture all aspects of textual quality and could overlook crucial elements such as contextual understanding and logical coherence, particularly in AI-generated text. Additionally, the study chose the free version, ChatGPT 3.5, instead of the charged latest version as the revisor. Although the free version ensures the tool's accessibility to nearly everyone, this choice might have impacted the results, as newer iterations of the model could provide improved capabilities for text revision. Moreover, the sample size of the EFL learners involved may not be representative of the broader population, limiting the generalizability of the findings. Lastly, the research did not delve into the underlying mechanisms or processes through which ChatGPT contributes to writing revision, leaving room for further exploration and inquiry. Looking ahead, future research could focus on other benefits of Generative Artificial Intelligence tools bring to language education, like exploring how ChatGPT can be used to cultivate creativity and critical thinking in EFL learners, or investigating its potential impact on different writing genres.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

LH conducted the research, analyzed the data, and wrote the manuscript; JJ edited the manuscript, supervised the project, and acquired funding; all authors had approved the final version.

References

[1] S. A. Bin-Nashwan, M. Sadallah, and M. Bouteraa, "Use of ChatGPT in academia: Academic integrity hangs in the balance," *Technology in Society*, vol. 75, p. 102370, 2023.

[2] S. Ariyaratne, K. P. Iyengar, N. Nischal, N. C. Babu, and R. Botchu, "A comparison of ChatGPT-generated articles with human-written articles," *Skeletal Radiology*, vol. 52, no. 9, pp. 1755-1758, 2023.

[3] W. L. J. Ho, B. Koussayer, and J. Sujka, "ChatGPT: Friend or foe in medical writing? An example of how ChatGPT can be utilized in writing case reports," *Surgery in Practice and Science*, vol. 14, p. 100185, 2023.

[4] S. O'Connor, "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse," *Nurse Education in Practice*, vol. 66, p. 103537, 2023.

[5] J. S. Barrot, "Using ChatGPT for second language writing: Pitfalls and potentials," *Assessing Writing*, vol. 57, p. 100745, 2023.

[6] S. W. Beck and S. R. Levine, "Backtalk: ChatGPT: A powerful technology tool for writing instruction," *Phi Delta Kappan*, vol. 105, no. 1, pp. 66–67, 2023.

[7] J. Huang and M. Tan, "The role of ChatGPT in scientific communication: Writing better scientific review articles," *American Journal of Cancer Research*, vol. 13, no. 4, pp. 1148, 2023.

[8] A. C. Graesser, D. S. McNamara, Z. Cai, M. Conley, H. Li, and J. Pennebaker, "Coh-Metrix measures text characteristics at multiple levels of language and discourse," *Elementary School Journal*, vol. 115, no. 2, pp. 211–229, 2014.

[9] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. J. Bamber, J. R. Beltran, P. Boselie, F. Lee Cooke, S. Decker, A. D. DeNisi, P. K. Dey, D. Guest, A. J. Knoblich, A. Malik, J. Paauwe, S. Papagiannidis, C. Patel, V. Pereira, S. Ren, and A. Varma, "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT," *Human Resource Management Journal*, vol. 33, no. 3, pp. 606–659, 2023.

[10] C. Xiao, S. X. Xu, K. Zhang, Y. Wang, and L Xia, "Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications," in *Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Canada, 2023, pp. 610–625.

[11] X. Wang and Y. Feng, "An experimental study of ChatGPT-Assisted improvement of Chinese college students' English reading skills: A case study of dear life," in *Proc. the 15th International Conference on Education Technology and Computers*, pp. 21–26, 2023.

[12] R. Xing, "Advancements in English listening education: ChatGPT and convolutional neural network integration," *Journal of Pedagogical Research*, vol. 7, no. 5, pp. 280–290, 2023.

[13] V. Aryadoust, A. Zakaria, and Y. Jia, "Investigating the affordances of OpenAI's large language model in developing listening assessments," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100204, 2024.

[14] Y. Sahari, A. M. T. Al-Kadi, and J. K. M. Ali, "A cross sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties," *Journal of Psycholinguistic Research*, vol. 52, pp. 2937–2954, 2023.

[15] V.Chan and W. Tang, "GPT and translation: A systematic review," in *Proc. 2024 International Symposium on Educational Technology (ISET)*, pp. 59–63, 2024.

[16] Y. Su, Y. Lin, and C. Lai, "Collaborating with ChatGPT in argumentative writing classrooms," *Assessing Writing*, vol. 57, p. 100752, 2023.

[17] N. Dehouche, "Plagiarism in the age of massive generative pre-trained transformers (GPT-3)," *Ethics in Science and Environmental Politics*, vol. 2, pp. 17–23, 2021.

[18] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innovations in Education and Teaching International*, vol. 61, no.2, pp. 228–239, 2023.

[19] T. Schmohl, A. Watanabe, N. Fröhlich, and D. Herzberg, "How artificial intelligence can improve the academic writing of students," presented at the Conf. Future of Education, 2020.

[20] D. Yan, "Impact of ChatGPT on learners in an L2 writing practicum: An exploratory investigation," *Education and Information Technologies*, vol. 28, no. 11, pp. 13943–13967, 2023.

[21] T. Zhou, S. Cao, S. Zhou, Y. Zhang, A. He, "Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing," *System*, vol. 118, p. 103141, 2023.

[22] A. Mehrad and M. Zangeneh, "Comparison between qualitative and quantitative research approaches: Social sciences," *International Journal for Research in Educational Studies*, vol. 5, no. 7, pp. 1–7, 2019.

[23] S. A. Crossley, "Linguistic features in writing quality and development: An overview," *Journal of Writing Research*, vol. 11, no. 3, pp. 415-443, 2020.

[24] S. S. Buckingham, Á. Sándor, R. Goldsmith, R. Bass, M. McWilliams, "Towards reflective writing analytics: Rationale, methodology and preliminary results," *Journal of Learning Analytics*, no. 1, pp. 58–84, 2017.

[25] T. D. Ullmann, "Automated analysis of reflection in writing: Validating machine learning approaches," *International Journal of Artificial Intelligence in Education*, vol. 29, pp. 217–257, 2019.

[26] A. Petchprasert, "Utilizing an automated tool analysis to evaluate EFL students' writing performances," *Asian-Pacific Journal of Second and Foreign Language Education*, vol. 6, no. 1, pp. 1–10, 2021.

[27] U. Maamuujav, C. B. Olson, and H. Chung, "Syntactic and lexical features of adolescent L2 students' academic writing," *Journal of Second Language Writing*, vol. 53, p. 100822, 2021.

[28] D. S. McNamara, S. A. Crossley, and P. M. McCarthy, "Linguistic features of writing quality," *Written Communication*, vol. 27, no. 1, pp. 57–86, 2010.

[29] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-Metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 193–202, 2004.

[30] D. S. McNamara, A. Graesser, Z. Cai *et al*. Coh-Metrix Common Core T.E.R.A. version 1.0. [Online]. Available: http://coh-metrix.commoncoretera.com

[31] D. Gruda, "Three ways ChatGPT helps me in my academic writing," *Nature*, 10, 2024.

[32] W. Nagy and D. Townsend, "Words as tools: Learning academic vocabulary as language acquisition," *Reading Research Quarterly*, vol. 47, no. 1, pp. 91–108, 2012.

[33] X. Gao, "A cross-disciplinary corpus-based study on English and Chinese native speakers' use of first-person pronouns in academic English writing," *Text & Talk*, vol. 38, no. 1, pp. 93-113, 2017.

[34] R. MacIntyre, "The use of personal pronouns in the writing of argumentative essays by EFL writers," *RELC Journal*, vol. 50, no. 1, pp. 6–19, 2019.

[35] F. Jiang and K. Hyland, "Does ChatGPT argue like students? Bundles in argumentative essays," *Applied Linguistics*, 2024.

[36] G. Broadhead, J. Berlin, and M. Broadhead, "Sentence structure in academic prose and its implications for College writing teachers," *Research in the Teaching of English*, vol. 16, no. 3, pp. 225–240, 1982.

[37] S. A. Crossley, J. Weston, S. T. M. Sullivan, D. McNamara, "The development of writing proficiency as a function of grade level: A linguistic analysis," *Written Communication*, vol. 28, no. 3, pp. 282–311, 2011.

[38] D. J. Follmer and R. A. Sperling, "Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text," *Learning and Individual Differences*, vol. 67, pp. 177–187, 2018.

[39] S. Witte and L. Faigley, "Coherence, cohesion, and writing quality," *College Composition and Communication*, vol. 32, pp. 189–204, 1981.

[40] D. Perin and M. Lauterbach, "Assessing text-based writing of low-skilled college students," *International Journal of Artificial Intelligence in Education*, vol. 28, pp. 56–78, 2016.

[41] C. A. MacArthur, A. Jennings, and Z. A. Philippakos, "Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?" *Reading and Writing*, vol. 32, no. 6, pp. 1553–1574, 2019.

[42] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-Metrix: Providing multilevel analyses of text characteristics," *Educational Researcher*, vol. 40, no. 5, pp. 223–234, 2011.

[43] W. Zhao, M. Strube, and S. Eger, "Discoscore: Evaluating text generation with BERT and discourse coherence," arXiv preprint arXiv:2201.11176, 2022, 2023.