Education, Language and Sociology Research
ISSN 2690-3644 (Print) ISSN 2690-3652 (Online)
Vol. 7, No. 1, 2026
www.scholink.org/ojs/index.php/elsr

Original Paper

Generative AI ChatGPT in College English Writing Teaching: An Empirical Study on Effectiveness and Pedagogical Optimization

Wentao Guo[1], Yanhao Guan[2*] & Xinyi Zhang[1]

[1] Zhongkai University of Agriculture and Engineering, Guangzhou, China

[2] China Unicom (Guangdong) Industrial Internet Co., Ltd., Guangzhou, China

[*] Yanhao Guan, E-mail: 1062293994@qq.com

Received: January 4, 2026 Accepted: January 17, 2026 Online Published: January 26, 2026

doi:10.22158/elsr.v7n1p43 URL: http://dx.doi.org/10.22158/elsr.v7n1p43

Abstract

With the rapid advancement of generative artificial intelligence (AI) technology, college English writing teaching has ushered in new development opportunities. This study quantifies teaching effectiveness from two dimensions—external language knowledge mastery and internal writing thinking abilities, and conducts an in-depth exploration of the application value and practical effects of the generative AI technology ChatGPT in college English writing instruction. The research results indicate that ChatGPT can assume partial functions such as tool support and feedback evaluation across various stages of writing, significantly reducing teachers' instructional burden. Additionally, the intelligent feedback system demonstrates remarkable advantages in improving the accuracy of students' language expression and optimizing the logicality of text structure. Meanwhile, the application of ChatGPT is accompanied by inherent limitations, including lack of contextual awareness, potential for misinformation, and plagiarism risks, which necessitate teachers to play an irreplaceable role in evaluation and screening, rule supervision, and competency training. This study provides empirical references and targeted optimization suggestions for the practical path of AI empowering college English writing teaching.

Keywords

ChatGPT, college English writing, language teaching

1. Introduction

Since the launch of ChatGPT in November 2022 and the release of ChatGPT-5 in January 2026, this representative technology in the field of AI, though still in its nascent stage, has garnered extensive

attention from all sectors of society due to its powerful linguistic capabilities and astonishing iteration speed. A substantial body of research has emerged domestically and internationally to discuss the principles, functions, and application methods of this emerging information technology. Scholars have also integrated ChatGPT with the characteristics of various disciplines to explore its application potential and implementation pathways (Dwivedi et al., 2023), including the field of language education. As a deep learning-based natural language processing model, ChatGPT is not specifically designed for language education; however, its natural language understanding capabilities, robust language generation functions, interactive chat-based working mode, massive repository of authentic multilingual resources, and excellent scalability have unlocked new potentials for foreign language education and learning. It can not only serve as a learning tool and partner for students, providing a near-authentic language learning environment and creating favorable conditions for autonomous learning but also act as an intelligent assistant for teachers, supporting teaching and research activities, lesson preparation, real-time teaching evaluation, and the completion of simple repetitive tasks (Qin, 2023).

In fact, the field of second language teaching in China has long attached great importance to the application of modern information technology. The 2020 edition of "*College English Teaching Guidelines*" (China University Foreign Language Teaching Steering Committee, 2020) explicitly proposes, in terms of teaching methods and means, that in the current information and intelligent era, modern information technologies such as multimedia, big data, virtual reality, and AI should become core components of college English education and teaching. This places new requirements on the development of teachers' digital competence in the new era. Against the backdrop of the in-depth integration of modern information technology and education, as a representative of the latest advancements in AI, the college foreign language teaching model integrated with ChatGPT is worthy of in-depth exploration. Particularly from the teacher's perspective—who assumes the crucial responsibility of guiding and supervising students' learning activities (Cai, 2023)—teachers' proficiency in applying emerging information technologies directly influences the advancement of teaching digitalization.

This paper is structured into five chapters. Chapter 1 serves as the introduction, outlining the development history and core functions of generative AI ChatGPT, and elaborating on its application potential and research significance in college English writing instruction. Chapter 2 presents the research background, clarifying the division of writing teaching into four stages—pre-writing, writing, evaluation, and reflection—and analyzing the specific application scenarios and functional values of ChatGPT in each stage, such as material supply, tool support, and evaluation feedback. Chapter 3 details the research design, including the grouping of research subjects, the construction of the ChatGPT feedback system, the selection of double-blind evaluation mechanisms and research tools, as well as the definition of evaluation dimensions and data analysis methods. Chapter 4 analyzes students' English writing abilities, comparing differences in external language knowledge mastery and internal writing thinking abilities (relevance, fluency, and logicality) between the experimental group and the control group based on pre-test and post-test data, and revealing the heterogeneous impacts of ChatGPT feedback on students with

varying proficiency levels. Chapter 5 offers reflections on the teaching experiment, identifying the limitations of ChatGPT (e.g., lack of contextual awareness, misinformation, and plagiarism risks), clarifying the irreplaceable role of teachers in each stage, and proposing targeted pedagogical optimization suggestions.

Based on the application of generative AI ChatGPT in the four stages of college English writing teaching (pre-writing, writing, evaluation, and reflection), this study focuses on addressing the following two research questions: (1) What effects does generative AI ChatGPT exert on improving external English language knowledge mastery and internal writing thinking abilities (logicality, fluency, and relevance) among students with different proficiency levels? (2) What challenges and limitations arise in the practical application of ChatGPT in college English writing instruction?

2. Research Background

Writing teaching often adopts diverse teaching models and flexibly leverages the advantages of various pedagogical approaches based on the formulation of teaching objectives and the needs of learners (Yang, 2012). Regardless of the adopted model, the writing process typically comprises three fundamental stages: pre-writing, writing, and evaluation (Xu, 2011). When discussing the role of AI technologies represented by ChatGPT in second language writing, Liu (2023) added a reflection stage to the writing teaching and learning process. Drawing on these two classification frameworks, this study divides the writing teaching process into four stages—pre-writing, writing, evaluation, and reflection—and analyzes the specific participation of ChatGPT in writing teaching activities at each stage.

In the pre-writing stage, teachers traditionally serve as both designers of classroom teaching activities and key sources of target language and target language culture. These roles can be partially assumed by ChatGPT following its integration into writing teaching. During the writing process, ChatGPT can function as tools such as an online dictionary, translator, and encyclopedia. Liu et al. (2023) proposed that teachers can fully leverage the advantages of large language model chatbots to facilitate learners' brainstorming through dialogue practice and foster transferable language skills. For instance, teachers can provide open-ended questions to guide students in stimulating writing inspiration via interaction with ChatGPT. Meanwhile, as a powerful information retrieval tool, ChatGPT can assist students in accessing background knowledge related to writing topics or explanations of professional terminology, thereby enhancing their understanding and preparation for the writing task. Furthermore, ChatGPT can offer time management strategies, including schedule formulation and timed reminders, to support learners' writing activities (Liu, 2023). Teachers can utilize ChatGPT to guide students in developing writing plans, providing time management advice, and offering suggestions on organizational structure.

In classroom writing teaching, teachers act as organizers of classroom activities, supervisors of target language and target culture input and output, and constructors and maintainers of classroom social relations (Xu & Lei, 2018). ChatGPT can assist non-native English learners in addressing difficulties related to grammar or word usage, with explanations that are sufficiently detailed to supplement

definitions in dictionaries and textbooks. In the initial stage of writing, students may encounter challenges in vocabulary selection, expression methods, sentence structure, and paragraph organization. For example, when students face vocabulary dilemmas, teachers can guide them to use ChatGPT to obtain synonyms, collocation suggestions, and even relevant example sentences, enabling students to accurately express their ideas and viewpoints. Additionally, ChatGPT's conversational agent model can record data based on users' preferences and interests. Research has demonstrated that this conversational agent can adjust the presentation of answers according to students' learning needs, progress, and comprehension levels, providing personalized retrieval feedback (Liu, 2023). Thus, ChatGPT can facilitate targeted improvements in students' writing skills.

In the evaluation stage, teachers typically assume the role of essay evaluators. Following the integration of ChatGPT into writing teaching, teachers can guide students to use ChatGPT for first-draft evaluation. ChatGPT can identify linguistic errors (e.g., spelling, punctuation, and grammar), detect logical issues (e.g., coherence, paragraph structure, and argument support), and provide specific suggestions on vocabulary selection, writing style, and context. Compared with teacher evaluation, ChatGPT exhibits distinct advantages in certain aspects and serves as an effective tool for essay evaluation. Guo and Wang (2023) from the University of Hong Kong compared the evaluation results of 50 undergraduate argumentative essays by ChatGPT and five college English teachers, identifying differences in the type, quantity, and style of comments: ChatGPT's comments were evenly distributed across content, structure, and language, whereas teachers' comments focused more on content and language; ChatGPT provided a greater number of comments and completed the evaluation within an extremely short time; in terms of style, ChatGPT directly indicated areas requiring revision, while teachers preferred indirect forms (e.g., questions); ChatGPT offered more praise in content evaluation, provided more holistic suggestions in structure evaluation, and employed more flexible language expressions. Overall, ChatGPT can comprehend the content of students' essays and provide appropriate revision suggestions, and participating teachers held a positive attitude toward ChatGPT's evaluation feedback. Thus, the teacher's role as an evaluator in this stage can be partially delegated to ChatGPT.

In the reflection stage, teachers act as implementers and guides of reflection activities. Teachers themselves need to continuously reflect on and accumulate experience in the process of teaching with emerging AI technologies; after students complete the final draft, teachers should also guide students to reflect on the entire writing process. Firstly, teachers can utilize ChatGPT's question-answering function to facilitate students' personal reflection. Students can ask ChatGPT questions such as "How to overcome challenges encountered in writing?" and "What new writing skills have been acquired?" Through discussions on these questions, students can conduct in-depth analysis and summary of their writing experiences, clearly recognize their progress and shortcomings, and clarify directions and strategies for future writing improvement. In addition to personal reflection, peer evaluation and group discussions are important components of the reflection stage. With the support of ChatGPT's language generation and review functions, students can exchange works and provide feedback and suggestions. This interaction

not only helps students examine their writing from diverse perspectives but also promotes cooperation and communication among students, enhancing their sense of teamwork and critical thinking skills. Furthermore, teachers can guide students to use ChatGPT to develop personal learning plans, such as exploring questions like "How to continuously improve writing ability?" and "How to utilize resources to support learning?" By setting clear learning goals and plans, students can achieve continuous improvement in future writing practice.

3. Research Design

3.1 Research Subjects

This study selected four first-year classes from an undergraduate university in Guangdong Province as research subjects, including two regular classes (Class A and Class B) and two experimental classes (Class A and Class B), with a total of 145 students (73 in regular classes and 72 in experimental classes). To explore the impact of generative AI feedback on students' English writing proficiency, the College English Test Band 4 (CET-4) (Note 1) scores of the four classes were compared, confirming no statistically significant differences in English writing proficiency among the classes ($p > 0.05$). Based on this, Regular Class B and Experimental Class B were designated as the experimental group, receiving generative AI feedback for writing guidance; Regular Class A and Experimental Class A served as the control group, receiving traditional teacher feedback. To ensure balanced group sizes, the 73 students in regular classes were divided into Eg1 ($n = 37$) and Cg1 ($n = 36$), while the 72 students in experimental classes were equally divided into Eg2 ($n = 36$) and Cg2 ($n = 36$). All group sizes were sufficient for independent samples t-test and hierarchical linear model (HLM) analysis.

3.2 Research Material: Generative AI System

This study adopted ChatGPT (GPT-4.0 architecture) developed by OpenAI as the core experimental platform, constructing a three-stage AI feedback system encompassing pre-training fine-tuning, real-time interaction, and quality monitoring. At the technical deployment level, the system was embedded into the university's online learning platform via an API interface, establishing a dedicated writing feedback channel with a maximum request limit of 50 times per second to ensure service stability. At the interaction design level, a customized interface was developed with three modules: error diagnosis, optimization suggestions, and learning path. The error diagnosis module integrated an NLP error classifier, capable of automatically identifying 12 types of common English writing errors; the optimization suggestion module adopted transfer learning technology, loading a fine-tuned dataset containing 500,000 CET-4 model essays; the learning path module was based on reinforcement learning algorithms, dynamically adjusting the difficulty of suggestions according to students' historical performance. At the quality control level, a dual verification mechanism was established: (1) The output format was constrained through preset prompt engineering specifications (see Table 1), with the temperature parameter set to 0.70 to balance creativity and standardization; (2) A review team consisting of three certified teachers was arranged to conduct regular random inspections of 25% of AI feedback content, and the revised Coh-

Metrix tool was used to evaluate feedback quality, ensuring a content accuracy rate of over 90%. In the specific implementation process, students in the experimental group logged into the system through unified accounts, and each essay underwent a complete process of "first draft submission → AI diagnosis → independent revision → final draft generation." The system recorded timestamps for each revision and generated behavior logs, ultimately producing a personalized writing analysis report with multiple indicators. (see Table 1 for details)

Table 1. Detailed List of Prompt Process Design

Stage	Prompt
Background Establishment & Learning	Assume you are a college English teacher. Please explain the common genres of English essays, as well as their expression methods and article structures. The following is a writing task on the theme of "…" in a first-year English unit test. Please analyze and summarize the writing theme, task requirements, word count limit, and expression notes. The following is the corresponding model essay… Please identify the total number of words, content that meets the key points/views, spelling errors, improper grammar usage, and other issues in the essay.
Error Type Identification	Please list the common writing errors made by college students… Classify college students' writing errors into five categories: vocabulary, grammar, content, logic, and discourse, and explain the specific meaning of each category respectively…
Essay Evaluation	Identify all errors in the following English essay, list them according to the above-determined error types, and present the type, cause, and revision method of each error in a table…
Essay Optimization	Please optimize this student's English essay, ensuring it meets the word count requirement of the writing task, uses CET-4 vocabulary appropriate for college students as much as possible, and list the parts of speech and meanings of the new words and phrases used.
Provision of Appropriate Learning Suggestions	Provide appropriate and effective writing learning suggestions for the student based on their "weak areas."

3.3 Research Tools

To accurately quantify the effectiveness of AI feedback, this study constructed a double-blind evaluation mechanism. For external competence evaluation, an expert-validated automated scoring system was

adopted, integrating the Coh-Metrix 3.0 text analysis tool and a self-built academic vocabulary analyzer. This system could calculate 85 linguistic feature indicators, including core parameters such as lexical complexity (MTLD ≥ 60) and syntactic complexity (D ≥ 8). For internal competence evaluation, a review team consisting of five college English teachers from the same grade was established, adopting the validated 3D-EVAL scoring system, which included 18 secondary evaluation dimensions. For example, the relevance dimension was subdivided into three sub-items: theme relevance, viewpoint novelty, and example adaptability.

To ensure evaluation consistency, the review team received a two-week calibration training prior to the evaluation, and the Intraclass Correlation Coefficient (ICC) was used to test inter-rater reliability, which ultimately reached an excellent level of 0.89. In the data analysis stage, a HLM was used to process nested data, controlling for random effects such as class and teacher, and a three-level prediction model including time, feedback type, and interaction terms was established. Mplus 8.3 software was employed for path analysis to accurately interpret the direct effects and moderating effects of AI feedback.

1). Internal English Writing Ability: Focusing primarily on students' writing thinking, this dimension is subdivided into three sub-dimensions: essay relevance, fluency, and logicality. Relevance reflects students' understanding of the writing theme, as well as their ability to analyze the topic and develop ideas; fluency reflects students' language expression ability; logicality reflects students' ability to structure and organize essays. This study conducted manual evaluation of students' writing text data based on this classification standard (see Table 2 for details).

2). External English Writing Ability: This dimension mainly evaluates students' mastery of English language knowledge. Referring to the CET-4 writing scoring standards, the study measured the improvement of students' external English writing ability by comparing changes in their essay scores across pre-tests, post-tests, and during the experiment.

Table 2. Rating Scale for College Students' Internal English Writing Ability

Level	Logicality	Fluency	Relevance
A	The essay has clear logic and a distinct structure, with sentences arranged in a reasonable logical order and described appropriately.	Complex sentence cohesion devices can be accurately used, resulting in compact and coherent sentences that are easy to evaluate.	The central idea of the essay is expressed extremely clearly, with extended relevant content closely related to the theme.
B	The essay structure is logically clear, and sentence arrangement and description are basically in line with	Some simple sentence cohesion devices can be used, resulting in relatively coherent	The central idea of the essay is expressed relatively clearly, with appropriate content highly relevant to the

	logic.	expressions that are easy to evaluate.	theme.
C	The essay structure has basic logic, and sentence arrangement is in line with logic, but there are a few cases of inappropriate sentence description.	The cohesion devices between sentences are single, and the content expression is basically coherent, allowing for smooth completion of evaluation.	The central idea of the essay is basically clearly expressed, with a small amount of content not closely related to the theme.
D	The writing content has certain logic, but some sentence arrangements and descriptions are not appropriate.	There is a lack of cohesion devices between sentences, and the content expression is not coherent, making the evaluation process laborious.	The central idea of the essay is vaguely expressed, with some content unrelated to the theme.
E	The writing content has no logic at all, and both sentence arrangement and description are unreasonable.	The cohesion between sentences is chaotic, and the content expression lacks coherence, making it difficult to complete the evaluation.	The central idea of the essay is inaccurately expressed, with content obviously unrelated to the theme.

3.4 Analysis of Students' English Writing Ability

Based on the research design and tools described above, this section analyzes students' English writing abilities before and after the experiment, focusing on group differences and intervention effects.

3.4.1 College Students' English Writing Ability Before the Experiment

In terms of English writing ability, this study compared the writing outputs of the experimental and control groups across all classes during the first writing lesson. SPSS analysis results exhibited that the external writing ability of students in the experimental classes was significantly higher than that of students in the regular classes, with a score difference of approximately 6 points. Specifically, students in the experimental classes achieved an upper-middle level of language knowledge mastery, while those in the regular classes were close to the lower-middle level. In terms of internal writing ability (e.g., topic analysis, idea development, and structure organization), first-year students overall performed at an upper-middle level, with students in the experimental classes slightly outperforming those in the regular classes. However, significance analysis indicated no statistically significant differences in external or internal

writing ability between the two types of classes ($p > 0.05$), confirming that the writing abilities of students in different classes were comparable prior to the introduction of generative AI feedback (see Table 3 for details).

All descriptive statistics (Mean, SD, SEM) were calculated using SPSS 26.0. The Standard Error of Mean (SEM) was computed as $SD/\sqrt{n}$ (n = group sample size) to ensure accuracy. Anomalous values in the preliminary analysis (e.g., SD = 0.00) were verified against original scoring records and corrected to reflect the actual distribution of students' abilities. The inter-rater reliability for all subjective evaluations (e.g., logicality, fluency) reached ICC = 0.89, confirming the consistency of the scoring process.

Table 3. Independent Samples T-Test of Pre-Test Data on College Students' Writing Ability

Writing Ability	Dimension	Group	Mean	Standard Deviation	Standard Error of Mean	p-value
Internal	Logicality	Eg1	3.73	0.53	0.12	0.902
		Cg1	3.74	1.35	0.24	
		Eg2	3.96	0.22	0.03	0.791
		Cg2	3.95	0.48	0.08	
	Fluency	Eg1	2.99	0.37	0.07	0.413
		Cg1	2.71	0.81	0.14	
		Eg2	3.00	0.22	0.03	1.003
		Cg2	3.01	0.38	0.06	
	Relevance	Eg1	3.76	0.98	0.17	0.531
		Cg1	3.65	1.23	0.23	
		Eg2	4.02	0.20	0.04	0.562
		Cg2	3.96	0.48	0.07	
External	CET-4 Essay Score	Eg1	6.24	2.61	0.48	0.773
		Cg1	6.35	2.68	0.55	
		Eg2	9.82	2.34	0.34	0.258
		Cg2	10.34	1.87	0.29	

Note. Eg1 = Regular Class B (Experimental Group 1); Cg1 = Regular Class A (Control Group 1); Eg2 = Experimental Class B (Experimental Group 2); Cg2 = Experimental Class A (Control Group 2). CET-4 essay scores were rated based on official criteria (ICC=0.91); internal ability scores were rated via 3D-EVAL system (ICC=0.89).

3.4.2 College Students' English Writing Ability After the Experiment

To explore the impact of generative AI on students' English writing ability, the study analyzed data from the final experiment. Results demonstrated that students in the experimental group outperformed those in the control group in both internal and external writing abilities. Specifically, there was a statistically significant difference in external writing scores between the experimental group and the control group in the regular classes ($p = 0.035$), with an overall improvement of approximately 1.78 points. The difference between the experimental group and the control group in the experimental classes was more pronounced ($p = 0.001$), with an overall improvement of approximately 2.18 points (see Table 4 for details). In terms of internal writing ability, Eg1 exhibited statistically significant differences in relevance ($p < 0.001$), fluency ($p = 0.006$), and logicality ($p = 0.002$), whereas the experimental group in the experimental classes only showed a significant difference in relevance ($p = 0.002$), with no significant differences in fluency or logicality (see Table 4 for details).

Table 4. **Independent Samples T-Test of Post-Test Data on College Students' Writing Ability**

Writing Ability	Dimension	Group	Mean	Standard Deviation	Standard Error of Mean	p-value
Internal	Logicality	Eg1	4.43	0.79	0.15	0.002
		Cg1	3.54	1.09	0.20	
		Eg2	4.02	0.26	0.04	0.265
		Cg2	3.96	0.29	0.05	
	Fluency	Eg1	3.21	0.57	0.11	0.006
		Cg1	2.71	0.71	0.14	
		Eg2	3.20	0.40	0.06	0.233
		Cg2	3.11	0.31	0.05	
	Relevance	Eg1	4.54	0.58	0.11	<0.001
		Cg1	3.65	1.13	0.21	
		Eg2	4.22	0.40	0.06	0.002
		Cg2	3.96	0.29	0.04	
External	CET-4 Essay Score	Eg1	9.45	2.47	0.47	0.035
		Cg1	7.57	3.54	0.67	
		Eg2	12.40	1.48	0.22	0.001
		Cg2	10.12	2.11	0.31	

Note. Eg1 = Regular Class B (Experimental Group 1); Cg1 = Regular Class A (Control Group 1); Eg2 = Experimental Class B (Experimental Group 2); Cg2 = Experimental Class A (Control Group 2). CET-4

essay scores were rated based on official criteria (ICC=0.91); internal ability scores were rated via 3D-EVAL system (ICC=0.89).

3.4.3 Analysis of College Students' English Writing Ability Before and After the Experiment

This study conducted a comprehensive analysis of writing outputs during the experiment and changes in abilities before and after the experiment, focusing on comparing scores after each writing lesson as well as scores from the first and last writing lessons to explore the specific effects of generative AI feedback. The analysis covered two dimensions: external writing ability (reflected by changes in writing scores) and internal writing ability (reflected by essay relevance, fluency, and logicality).

1) Analysis of Changes in English Writing Scores

After each writing lesson, the magnitude of score changes among students in the experimental group was generally greater than that in the control group, with scores in the experimental group showing a continuous upward trend. The experimental group in the regular classes exhibited a statistically significant score change after the first writing lesson ($p = 0.001$), with an increase of 2.6 points. The experimental group in the experimental classes also showed a significant score change after the second writing lesson ($p = 0.004$), with an increase of 1.5 points. In contrast, the control groups in both experimental and regular classes showed no significant score changes after each writing lesson ($p > 0.05$). Overall pre-test and post-test results indicated that the writing score changes of the experimental group in the experimental classes were statistically significant ($p < 0.001$), with an increase of 2.58 points, and the score changes of the experimental group in the regular classes were also significant ($p = 0.000$), with an increase of 3.21 points. In contrast, the score changes of the control group in the experimental classes ($p > 0.05$) and the control group in the regular classes ($p > 0.05$) were not significant. These results demonstrate that generative AI writing feedback significantly promoted students' mastery of English language knowledge and basic norms, effectively improving their writing scores. This feedback can accurately identify errors in students' writing and provide detailed explanations and revision suggestions, helping students understand and consolidate relevant knowledge, thereby achieving significant and durable improvements in writing scores in a short period.

2) Analysis of Changes in Essay Logicality

The magnitude of logicality changes among the experimental and control groups in the regular classes was generally greater than that in the experimental classes, with the experimental group in the regular classes exhibiting the largest change in the third round. Although the magnitude of change in the experimental group in the experimental classes was small, it showed a stable and positive trend overall. Significance analysis revealed that the logicality changes of the experimental group in the regular classes were statistically significant in the third round and across the overall pre-test and post-test ($p = 0.006$; $p = 0.001$), while the logicality changes of other groups were not significant ($p > 0.05$).

Comprehensive analysis indicates that the initial logicality of students in the experimental group of the regular classes was relatively weak, and their logicality was significantly improved under the intervention

of generative AI feedback. This suggests that the promotional effect of this feedback on college students' structural organization ability is predominantly evident among students with weaker foundations (pre-test CET-4 essay score < 7 points). Although the feedback can improve writing content and enhance the accuracy of sentence arrangement and narration, its effects are mainly reflected in students with underdeveloped logical thinking. For students in the experimental classes with relatively mature logical thinking, the role of generative AI feedback in improving essay logicality is relatively limited.

3) Analysis of Changes in Essay Fluency

Analysis results demonstrated that the magnitude of fluency changes among the experimental and control groups in both experimental and regular classes was small after each writing lesson. Only the experimental group in the experimental classes showed a positive increase of 0.2 in the third round, while the fluency of the experimental group in the regular classes exhibited a steady upward trend. Significance analysis indicated that the fluency changes of the control groups in both experimental and regular classes were not significant in each round or across the overall pre-test and post-test ($p > 0.05$). In contrast, the fluency changes of the experimental group in the experimental classes were statistically significant in the second round ($p = 0.046$), third round ($p = 0.008$), and overall pre-test and post-test ($p = 0.006$), while the experimental group in the regular classes only showed a significant difference in the overall pre-test and post-test ($p = 0.032$).

Pre-test and post-test analysis of essay fluency among students in experimental and regular classes showed that the fluency of the experimental group in the experimental classes was significantly improved, but there was no significant difference in fluency levels between the two groups. This may be attributed to the relatively high initial fluency of the control group; although the experimental group improved with the support of generative AI feedback, the control group also achieved a certain degree of improvement with the assistance of traditional teacher feedback. These findings suggest that generative AI feedback has a significant impact on the writing expression ability of students with weaker English foundations. The feedback can identify vocabulary and grammar errors in sentence cohesion, provide improvement suggestions, and demonstrate more fluent expression methods. However, the understanding and application of certain English knowledge points still require a solid foundation. Thus, long-term use of generative AI feedback is necessary for students with weaker foundations to achieve sustained improvements in essay fluency.

4) Analysis of Changes in Essay Relevance

Analysis of relevance changes among the experimental and control groups in each class after each writing lesson indicated that the fluctuation range of relevance in the experimental and control groups of the regular classes was generally greater than that in the experimental classes. The change range of each group in the experimental classes was approximately 0.1, while the fluctuation range in the regular classes could reach up to 0.4. The relevance of the experimental group showed a stable upward trend, whereas the control group exhibited unstable changes. Significance analysis revealed that the relevance change of the experimental group in the regular classes was statistically significant in the third round (p

= 0.048) and across the overall pre-test and post-test (p = 0.001). Although the experimental group in the experimental classes did not show significant changes in each round, the overall pre-test and post-test relevance change was still significant (p = 0.004). In contrast, the relevance changes of the control groups in both experimental and regular classes were not statistically significant in each round or across the overall pre-test and post-test (p > 0.05).

Pre-test results and English writing score analysis of students in experimental and regular classes showed that generative AI feedback significantly improved college students' ability to analyze topics and develop ideas, particularly among students with medium academic performance (pre-test CET-4 essay score 7-10 points). The system provides support for students through four components: error analysis, essay optimization, vocabulary expansion, and learning suggestions. For students with weaker foundations, teachers' explanations of error examples and students' self-correction enhance their understanding of the theme and improve their ability to analyze topics and develop ideas. For students with upper-middle levels, the essay optimization and vocabulary expansion functions can enhance their innovation ability, enabling them to appropriately expand relevant content on the basis of clarifying the theme and ideas, thereby effectively improving the relevance of the essay.

The above results indicate that ChatGPT feedback has heterogeneous effects on students with different proficiency levels—students with weaker foundations (Eg1) show more significant improvements in internal writing abilities (logicality, fluency, relevance), while students with stronger foundations (Eg2) only achieve significant progress in relevance, which addresses the first research question.

4. Reflection on the Teaching Experiment

Based on a one-semester teaching experiment and combined with the experimental process and results, this study proposes the following suggestions for the three stages of writing teaching: pre-writing, writing, and evaluation. Notably, the limited improvement in logicality (p = 0.265) and fluency (p = 0.233) among high-proficiency students (Eg2) may be attributed to ChatGPT's inadequate capacity to deliver advanced logical guidance, which aligns with its inherent limitation of lacking contextual depth.

ChatGPT exhibits certain limitations when providing input corpora in the pre-writing stage, which requires teachers to act as evaluators and screeners of the plans and texts generated by ChatGPT. On the one hand, ChatGPT lacks real contextual information, so teachers need to adjust the difficulty of corpora according to teaching needs and students' proficiency levels. Teachers' accumulated teaching experience, grasp of teaching objectives and progress, and understanding of students serve as important reference bases. On the other hand, ChatGPT's training texts are derived from diverse sources, which may contain misinformation or biased content (Yang & Wang, 2023) or generate language expression styles inappropriate for classroom teaching. Thus, teachers need to remain highly vigilant in this regard.

The use of ChatGPT in the writing process has also raised concerns, with college English teachers' core worry being the potential for students to plagiarize AI-generated texts. This study found that students have a potential risk of cheating by using ChatGPT when investigating their attitudes toward the

technology. Consequently, students are increasingly concerned about whether teachers can timely detect and fairly address academic integrity issues potentially caused by new technologies. At this juncture, teachers' roles need to transform into "defenders" who restrict the use of ChatGPT within the scope of academic rules. Teachers also need to assist and guide students in distinguishing the correctness and appropriateness of texts generated by ChatGPT, and further verify and confirm questionable content. Although the application of emerging intelligent technologies can often stimulate students' interest in writing and enhance their learning motivation, long-term interaction solely with machines may lead to a gradual decline in students' learning enthusiasm (Fryer et al., 2017). Therefore, the role of teachers in writing evaluation remains irreplaceable.

A major advantage of ChatGPT evaluation is its ability to generate revised texts based on students' original essays. In contrast, teacher evaluation, peer evaluation, and other types of intelligent evaluation typically only provide targeted suggestions based on students' essays, without directly generating revised versions. When explaining essay evaluation and revision, teachers generally only provide 1-2 model essays as reference answers for designated writing topics. However, even after correcting all errors, each student's writing ideas and language style remain unique. The reference texts generated by ChatGPT based on students' own essays can achieve "personalization." Students can provide feedback on unsatisfactory aspects of the AI-generated texts and form the final draft through interactive negotiation. Nevertheless, this convenience also brings the risk of plagiarism. Teachers need to guide students to deepen their understanding and reflection on writing objectives, master genre structures, enrich writing content, and improve language expression levels during this process. Teachers and ChatGPT provide scaffolding for students' revisions, helping them continuously expand their Zone of Proximal Development in the learning process. Such targeted guidance can enhance students' confidence in completing tasks, avoid academic misconduct such as plagiarism, and thereby enable teachers to better fulfill their role as "defenders."

Acknowledgments

This paper is supported by: Regular Project of the 2025 Annual "14th Five-Year Plan" for the Development of Philosophy and Social Sciences in Guangzhou [Grant Number: 2025GZGJ385]; Department of Education of Guangdong Province under Grant the Educational Reform and Development Project in Tertiary Education (Grant Number: YueJiaoGaoHan [2024] No.9).

References

Dwivedi, Y. K., Kshetri, N., Hughes, L. et al. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 1-63. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and human task partners. *Computers in Human Behavior*, *75*, 461-468. https://doi.org/10.1016/j.chb.2017.05.045

China University Foreign Language Teaching Steering Committee. (2020). *College English Teaching Guidelines*. Beijing: Higher Education Press.

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing [J/OL]. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12146-0

Liu, L. (2023). Opportunities and Challenges Brought by ChatGPT to International Chinese Language Education—A Compilation of Expert Views from the Joint Forum of Beijing Language and Culture University and Chinese Language Teachers Association USA. *Journal of Chinese Linguistics*, *51*(3), 291-315.

Qin, Y. (2023). Exploration of Foreign Language Teaching Methods in the Human-Machine Symbiosis Scenario—Taking ChatGPT as an Example. *Computer-Assisted Foreign Language Education*.

Yang, M., & W. Yafei. (2023). ChatGPT's "Understanding" and "Meaning": On the Form, Function and Standpoint Behind Its Generated Language. *China Foreign Language*, *3*, 24-32.

Yang, Y. L. (2012). Writing Teaching, Ten Perspectives—From Classical Rhetoric to Digital Writing. *Contemporary Foreign Language Studies*, *3*, 131-136.

Xu, F. (2011). A Multi-Perspective Theoretical Review and Reflection on English Writing Teaching Methods. *Foreign Language World*, *2*, 57-64.

Note

The College English Test Band 4 (CET-4), administered and conducted by the National Education Examinations Authority under the Ministry of Education of China, is a nationwide standardized exam designed to serve educational purposes.