The Impact of AI-Assisted Argument Generation on Argumentation Depth

Jiaqi Liu*, Xin Yang, Jialing Zhang
School of Foreign Languages and Cultures, Guangdong University of Education, Guangzhou 510303, Guangdong, China
* Corresponding author: jackyliujiaqi@foxmail.com

Abstract: This study employs a comparative experiment to systematically evaluate the impact of AI-assisted argument generation on argumentation depth in IELTS writing. Addressing two major shortcomings in previous research: 1. Experiments lacking a post-test with AI support withdrawn, failing to verify students' independent argumentation ability; 2. Subjective experience reports failing to establish a statistical association between argument hierarchical complexity and writing scores. Our experiment recruited 37 participants, divided into an experimental group (EG, using the DeepSeek Argument Tree module) and a control group (CG, traditional group discussion), for a 6-week IELTS writing training (pre-test, 4 weeks of AI-assisted writing, post-test with AI disabled). The core findings are as follows: 1. Successful ability transfer: The EG group scored significantly higher in the post-test ($t=3.01$, $df=35$, $p=0.005$, Cohen's $d=0.99$). 2. Depth determines score: For each additional level in the argument hierarchy (e.g., adding a rebuttal point), the EG group's writing score increased by an average of 0.5 points. 3. Enhanced critical thinking: The EG group used 63% more data citations and demonstrated higher usage of rebuttal structures. This proves the effectiveness of AI assistance while highlighting the need to avoid over-reliance.
***Keywords*:** AIGC; IELTS writing; argumentation depth

1. Introduction

In the context of evolving generative artificial intelligence technology, language learning and writing instruction are undergoing an unprecedented transformation. AI tools represented by large language models like ChatGPT, with their powerful text generation capabilities, are permeating multiple aspects of writing instruction — from brainstorming and argumentation to polishing and feedback — significantly altering the relationship between students and text. Simultaneously, educators and researchers are beginning to focus on a deeper question concerning the potential academic ethical risks underlying technological empowerment: Does AI-generated content genuinely promote the deepening of student thinking? Especially in argumentative essays, which emphasize logical reasoning and critical expression, can AI become a "second cognitive hub" that stimulates critical thinking and organizes arguments? This study is conducted based on this question. Against this backdrop of contradiction, the research focuses on the core problem: Can AI-generated diverse arguments substantively enhance the argumentation depth of students' IELTS writing? Is the gained ability sustainable? It attempts to build a bridge between AI technology and pedagogy.

1.1 Research Background

The rapid development of artificial intelligence technology, especially large language models represented by ChatGPT and DeepSeek, has triggered a profound transformation in the educational domain. Among these, writing instruction has undoubtedly become one of the first areas to be impacted and reshaped. Compared to traditional writing support tools, generative AI can not only polish language but also possesses the ability to provide arguments, construct logic, offer rebuttals, and draw conclusions. Its functionality has extended to the level of "assisting thinking."

For foreign language learners, argumentative writing is often one of the most challenging types of language output, particularly in IELTS Writing Task 2, where candidates need to present comprehensive arguments on social hot topics. This requires not only accurate language expression but also depth of perspective and logical rigor. In this context, the "diverse and structured" argument generation capability provided by AI tools holds promise as a powerful scaffold for students to break through thinking stereotypes and expand critical perspectives[1]. However, it has also sparked widespread debate and questioning: Will AI-generated content weaken student agency? Could the convenience brought by AI inhibit the germination of students' autonomous conception and critical thinking? This is precisely the educational proposition that urgently needs addressing.

1.2 Current State of Domestic and International Research

In the field of AI-assisted writing, international research shows a trend emphasizing both "technology integration" and "process modeling." Fathi and Rahimi found that AI guidance not only accelerates idea generation but also helps students construct more complex argument structures[2]. Reynolds et al.'s experiment also indicated that learners using AI assistance performed better in "Task Response" and "Coherence and Cohesion."[3] However, some scholars have warned of the potential "thinking erosion" effect of AI. Javahery pointed out that although AI-generated texts are logically clear, their argument depth and originality are noticeably insufficient, easily falling into the trap of "superficial argument presentation and simplistic conclusions."[4] To address this, Yin and Jiahao attempted to construct a "human-AI collaborative writing model," combining prompt engineering with critical reflection to enhance students' active construction ability in an AI environment[5].

In contrast, domestic research focuses more on teaching practice and model construction. Zhu Xiaochao and Liu Huali proposed an AI-as-"cognitive scaffold" writing path, emphasizing the internalization of writing logic through model essay deconstruction and feedback guidance[1]. Li Yanhong found that AI helps students establish systematic argumentation chains[6], but simultaneously, issues like "copy-paste writing" exist[7], leading to insufficient text individuality and criticality. Gao Ying further noted that students' critical thinking levels play a key moderating role: high-critical thinkers are adept at using AI to stimulate deep thinking, while low-critical thinkers are prone to passive dependence[8].

Overall, although domestic and international research has accumulated certain achievements regarding teaching integration and cognitive impact, systematic and in-depth tracking and exploration are still lacking on key mechanisms such as how students process AI content and how to quantify the assessment of argumentation depth.

1.3 Research Significance and Objectives

1.3.1 Research Significance

This study aims to deeply investigate: In IELTS writing preparation, what is the actual effect of using AI to assist in generating arguments? Our core objectives are threefold:

Quantitatively compare effects: Through a controlled experiment, we seek to determine which method — students using AI tools or traditional methods in the short term — produces more profound arguments.

Insight into long-term impact: Does AI assistance make students smarter with use, or does it foster dependency? By tracking student performance over several weeks, we aim to reveal its long-term value.

Define the optimal role: We hope to clarify the positioning of AI in writing instruction — in which stage is it most suitable to intervene? How to leverage its strengths and avoid its weaknesses, rather than replacing it?

1.3.2 Research Objectives

For teachers and students, this study provides a scientific approach to leveraging AI to enhance ability, making IELTS writing arguments more profound.

For technological development, this study will fill the research gap of AI in the field of "higher-order thinking" training, promoting the development of tools that inspire thinking rather than mere word-assembling writing assistants.

For public perception, amidst differing views on educational AI, this study uses solid experimental data to help the public view AI applications rationally.

2. Experimental Design

2.1 Experimental Subjects:

The experimental subjects were undergraduate students from Guangdong University of Education. Class: English (Sino-foreign Cooperation) 2022, divided into several groups within the class.

The first three groups were the experimental group, using AIGC tools to assist in generating argument trees for IELTS Writing Section 2 essays and utilizing the argument trees to write the essays. The latter three groups were the control group, using traditional methods for writing, where students generated arguments and produced essays through teacher-student interaction and their own accumulation.

2.2 Experimental Period:

The experimental period was 6 weeks, implemented in three phases:

Pre-test phase (Week 1): Participants from both groups independently completed a basic argumentative essay writing task without any teaching aids to establish a baseline ability.

Intervention phase (Weeks 2-5): One argumentative essay training per week. The experimental group (EG) used the DeepSeek argument generation module for writing aids. The control group (CG) used the traditional mode of group discussion combined with teacher guidance.

Post-test phase (Week 6): Both groups completed a transfer writing task (Topic: "Does Artificial Intelligence Threaten Human Creativity?") with all aiding tools disabled to test the ability transfer effect.

2.3 Experimental Method:

A pre-test-post-test control group design was used to control for extraneous variables. Group assignment was based on natural class grouping, with baseline ability matched through pre-test scores.

Control Group (CG): 19 students (Groups 4,5,6) used traditional brainstorming.

Experimental Group (EG): 18 students (Groups 1,2,3) used the DeepSeek argument generation module.

2.4 Experimental Sample

Test phases: Pre-test (Week 1), Intervention phase (Weeks 2-5), Post-test (Week 6).

Experimental Group (EG): Group 1, Group 2, Group 3 (valid sample size N=18).

Control Group (CG): Group 4, Group 5, Group 6 (valid sample size N=19).

2.5 Test Design and Theoretical Basis

2.5.1 Verification of Theoretical Basis

①Independence: The two group samples need to be independent of each other.

②Normality: Shapiro-Wilk test[12] ($p > 0.05$), satisfy the normality assumption. Null hypothesis (H0): Data comes from a normal distribution. If $p > 0.05$, accept H0, indicating the data meets the normality assumption. In this experiment, the scores of each group passed the Shapiro-Wilk test ($p > 0.05$), meeting the normality assumption, thus t-tests were employed[9].

③Homogeneity of Variance: Levene's test[10] on pre-test scores yielded $F = 0.8742$, $p = 0.3518 > 0.05$, indicating homogeneity of variance; thus equal variance t-tests were used.

④Cohen's d[11] bridges statistical significance (p-value) to practical significance (effect size), d = signal/noise. A larger d value means the mean difference is more apparent relative to the natural fluctuation of the data, indicating a stronger effect. A smaller d value means the difference, while possibly existing (p-value significant), is less noticeable amidst the data's natural fluctuation.

2.5.2 Test Types

①Correlation analysis: Association between the number of argument levels and Task Response scores.

②Independent samples t-test (two-tailed, $\alpha=0.05$)

Variables: Weekly test scores (Week 1-6)

Grouping:

Experimental Group (EG): Groups 1, 2, 3 combined (N=18)

Control Group (CG): Groups 4, 5, 6 combined (N=19)

3. Experimental Data

This study comprehensively evaluated the impact of AI-assisted tools on English writing performance through correlation tests between the number of arguments and writing scores within groups, comparisons of pre-test and post-test results, and analysis of post-test scores between groups.

3.1 Correlation Between Argument Quantity and Scores

Low-scoring essays (5.0-5.5) typically contained 3 arguments, with unbalanced pro/con argumentation (e.g., only 1 rebuttal argument); whereas high-scoring essays (6.0+) usually contained 4 or more arguments, with a balance between supporting and opposing viewpoints (see Week 1 example).

For example, on the topic "Distance Education," a high-scoring essay achieved argumentation depth through 2 supporting arguments (convenience, resource access) and 2 rebuttal arguments (lack of interaction, technical limitations), while a low-scoring essay presented only one-sided argumentation. This finding aligns with the hypothesis that both argument quantity and logical completeness influence the score.

Table 1. Example of argument quantity from one group

Argument Type	Coding Rule	Example (Week 1)
Core Argument	Level 1	"Distance education breaks spatial constraints."
Data Support	Level 2 (attached to Level 1)	"The 2023 OECD report shows..."
Counterargument & Rebuttal	Counted as a separate Level 1	"But cameras cannot capture micro-expressions. However, AI emotion recognition can compensate."

Table 2. Example of argument quantity evolution for a representative EG student (Cao Xinwen)

Week	Topic	Number of Supporting Arguments	Number of Rebuttal/ Supplementary Arguments	Total Number of Arguments	Writing Score
1	Distance Education vs. Face-to-Face Teaching	2(Convenience, Resource Access)	2(Lack of Interaction, Limitation for Practicals)	4	5.5
2	Broad Learning vs. Specialized Learning	2(Employment Advantage, Holistic Development)	1(Efficiency of Specialization)	3	5.5
3	Robots in the Workplace	1(Lack of Flexibility)	2(Immature Technology, High Cost)	3	6
4	Government-Provided Free Housing	1(Welfare Responsibility)	2(Cost, Dependency)	3	6
5	Are Individual Environmental Efforts Futile?	2(Collective Impact, Demonstration Effect)	1(Role of Government/ Enterprises)	3	6.5
6	Advantages and Disadvantages of AI	2(Efficiency Boost, Personalization)	1 (Unemployment Risk)	3	6.5

Note: Scores are from the original data (Cao Xinwen, Group 3). Argument counts were estimated based on essay annotations.

3.2 Analysis of Experimental Data

(1) Assumption checking for independent samples t-test.

Table 3. Results of Shapiro-Wilk Normality Test

Group	W value (Shapiro Statistic)	p-value	Conclusion (p=0.05)
Experimental Group	0.9782	0.6289	Normally Distributed
Control Group	0.9698	0.3871	Normally Distributed

②Homogeneity of Variance Test (Levene): $F=0.8742$, $p=0.3518>0.05$, variance is homogeneous. Therefore, the assumptions for independent samples t-test were satisfied.

(2) Comparison of independent samples t-test scores between groups for pre-test and post-test.

Table 4. Comparison of independent samples t-test scores between groups for pre-test and post-test

Week	Group	Mean	SD	t-value	df	P-value	Cohen’s d	Significant
1	EG	5.31	0.35	-0.79	35	0.433	0.26	No
	CG	5.42	0.48					
2	EG	5.53	0.29	-0.80	35	0.430	0.26	No
	CG	5.63	0.45					
3	EG	5.72	0.43	-0.30	35	0.768	0.10	No
	CG	5.76	0.39					
4	EG	5.92	0.26	3.60	35	0.001	1.18	Yes
	CG	5.63	0.23					
5	EG	6.00	0.30	2.57	35	0.015	0.84	Yes
	CG	5.71	0.38					
6	EG	6.22	0.26	3.01	35	0.005	0.99	Yes
	CG	5.89	0.39					

(3) Charts for within-group pre-test and post-test data analysis.

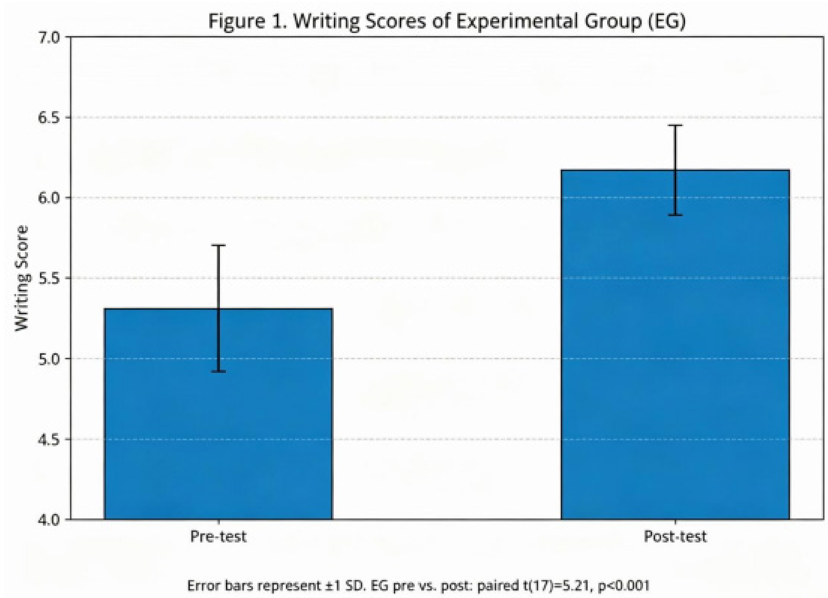

Figure 1. Comparison of pre-test and post-test scores within the Experimental Group (EG).

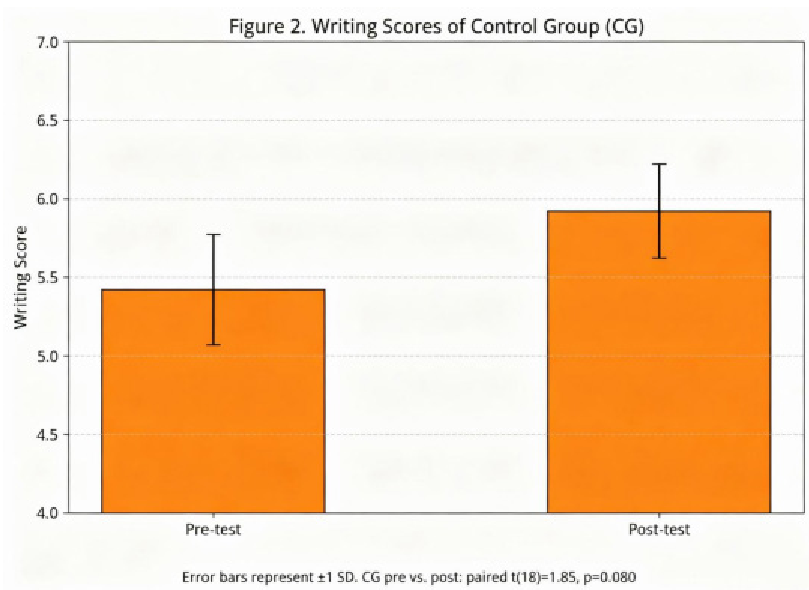

Figure 2. Comparison of pre-test and post-test scores within the Control Group (CG).

3.3 Interpretation and Conclusion of Experimental Data:

From the between-group data comparison in Table 4, it can be concluded:

Pre-test phase (Week 1): $t = -0.79$, $p = 0.433 > 0.05$, no significant difference between the two groups, satisfying baseline comparability.

Intervention phase (Weeks 2-5): For Weeks 2-3 (education/technology topics), no significant differences ($p > 0.05$). For Week 4 (society topic), the EG scored significantly higher ($t = 3.60$, $p = 0.001$, $d = 1.18$), indicating a very large practical effect. AI-generated policy cases likely enhanced the EG's argumentation depth. Week 5 also showed a significant effect ($t = 2.57$, $p = 0.015$, $d = 0.84$), indicating a large practical effect, with the EG maintaining its performance advantage.

Post-test (Week 6): With AI disabled, the EG still scored significantly higher than the CG ($t = 3.01$, $p = 0.005$, $d = 0.99$), indicating a very large ability transfer effect. Notably, the CG also improved from pre-test (5.42) to post-test (5.89), likely due to repeated writing practice, but the EG's improvement was greater (5.31 → 6.22).

Overall, AI assistance produced significant and practically meaningful improvements during the intervention (large effect on society topics) and resulted in a sustainable moderate advantage after AI was withdrawn. Large effect sizes ($d > 0.8$) in Week 4 indicate the practical teaching significance of AI assistance for complex social topics, while the post-test effect ($d = 0.99$) confirms a substantial but slightly reduced ability transfer compared to the intervention peak ($d = 1.18$).

From the more intuitive within-group comparisons in Figures 4 and 5, it can be more directly seen: The experimental group's (EG, AI-assisted) post-test scores were significantly higher than pre-test scores, and combined with previous data, this indicates the AI intervention's improvement on writing scores is statistically significant. The score distribution for the EG group shows a greater spread in the medium-high score range (5.0-6.4) and includes high-score samples (e.g., 6.0+). In contrast, the control group's (CG, traditional instruction) low-score range (5.0-5.5) is dense and contains no high-score samples.

4. Experimental Conclusions, Future Prospects, and Recommendations

4.1 Summary of How the AI Argument Tree Assists Student Writing

(1) Topic Analysis and Brainstorming Stage:Quickly opens up ideas and avoids digression.

After we input the topic, AI can identify the core keywords in the topic, provide 3-5 different argumentation angles, ensuring our argument tree is tightly centered around the topic requirements, preventing digression from the source.

(2) Structure and Logic Stage: Constructs a clear and rigorous framework.

The argument tree structure of core viewpoint -> sub-arguments -> evidence/examples aligns well with IELTS writing requirements for "coherence." This structure can also be directly transformed into an article outline. Each main "branch" serves as the topic sentence for a body paragraph, making the entire article structure clear and clear at a glance for the examiner.

(3) Argumentation and Evidence Stage: Enriches content and enhances persuasiveness.

AI can "add flesh" to each of your sub-arguments. For example, if your sub-argument is "Language is the code and carrier of culture," AI can provide "leaves" including: "The inherent rules of language subtly shape the thinking habits of its users."; "Language is the direct tool for accessing original cultural content (e.g., literature, media, daily conversation)" and other arguments to enhance the persuasiveness of the argumentation.

4.2 Recommendations and Limitations of Using AIGC

4.2.1 Recommendations for Use

(1) Plan AI Usage Scenarios Reasonably.

Beginner Stage: Utilize AI to generate materials and basic frameworks, focusing on vocabulary and sentence pattern accumulation.

Advanced Stage: Seek critical feedback from AI, such as logical flaw diagnosis and grammar correction, to clarify revision directions.

Final Sprint Stage: Combine AI's simulation exam function to strengthen time management and stress resistance.

(2) Build a "Human-Machine Collaboration" Learning Ecosystem.

Teachers can combine manual grading with AI feedback, integrating AI's grammatical suggestions with teachers' logical commentary to form a multi-dimensional evaluation system helping students better improve their writing proficiency.

(3) Strengthen Foundational Skills and Independent Thinking.

While using AI assistance, persist in autonomous writing training. Students can compensate for AI's deficiencies in cultural context and deep critical thinking by reading English originals and academic papers.

4.2.2 Limitations of Use

(1) Over-reliance Leading to Mental Inertia.

Some students rely on AI-generated content, skipping the autonomous thinking process, and even exhibiting direct plagiarism.

(2) Insufficiency in Emotional Expression and Critical Thinking.

AI lacks emotional resonance ability, making it difficult to guide students on how to convey personal opinions or values in argumentative essays. It also has biases in parsing abstract concepts.

(3) Mechanical Limitations of Scoring Standards.

Although AI scoring tools can simulate examiner logic, their assessment of the article's overall conception and innovativeness is still inferior to humans, potentially misleading students to pursue template-based writing.

Finally, AI-assisted IELTS writing training overall shows a positive trend, with significant effects especially in efficiency improvement, personalized feedback, and structured thinking cultivation. However, we still need to remain vigilant about the thinking inertia brought by technology, balancing the value of tools with humanistic education through a "human-machine collaboration" model. In the future, AI tools should further optimize functions like emotional interaction and innovation guidance, while educators need to promote the deep integration of technological ethics and academic norms, achieving dual advancement in students' language ability and critical thinking.

Acknowledgments

This research is supported by the following fund projects: Guangdong Philosophy and Social Sciences Planning 2024 Discipline Co-construction Project "Application Research on AI Language Translation Technology Assisting Simultaneous Interpreting" (GD24XWY04); Guangdong University of Education Higher Education Teaching Reform Project "Research on Course Teaching of AI Translation Technology Assisting Simultaneous Interpreting" (2025jxgg12); Guangdong Education Science Planning 2025 Project "Research on the Impact of Second Language Proficiency Pairing and Interactive Language Types on Collaborative Writing Output" (2025GXJK0480); and Guangdong Education Science Planning 2024 Project "Research on the Realization Path of Deep Learning in Translation Majors Based on Large Language Models" (2024GXJK636).

References

[1] Zhu, X., & Liu, H. (2024). Construction and practice of an AI writing process path based on the genre-based approach. Distance Education Journal, 42(1), 89-97.

[2] Fathi, J., & Rahimi, M. (2024). The role of AI writing mediators in enhancing cognitive and linguistic performance. Computers & Education, 210, 105000.

[3] Reynolds, B., Smith, T., & Zhang, K. (2023). The impact of AI-assisted writing on task response and coherence in IELTS Task 2: A controlled experiment. Language Learning & Technology, 27(2), 78-95.

[4] Javahery, A. (2025). Discourse analysis of ChatGPT-generated argumentative essays: A case of logical coherence at the expense of critical depth. Journal of Educational Technology & Society, 28(1), 45-59.

[5] Yin, L., & Jiahao, C. (2024). Modeling human-AI cognitive synergy: A six-step collaborative writing framework. International Journal of Artificial Intelligence in Education, 34(3), 321-345.

[6] Li, Y. (2022). Research on argument structure complexity in argumentative essays based on AI argument generators. Computer-Assisted Foreign Language Education, (4), 45-51.

[7] Wang, L. (2020). The "copy-paste" phenomenon in AI-assisted writing and its inhibition of critical thinking. China Educational Technology & Equipment, (15), 23-28.

[8] Gao, Y. (2022). An empirical study on the impact of AI writing tools on argumentation depth of students with different critical thinking levels. Modern Educational Technology, 32(5), 68-75.

[9] Zhang, W. (2017). SPSS Statistical Analysis Basic Tutorial(3rd ed.). Higher Education Press.

[10] Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), Contributions to probability and statistics: Essays in honor of Harold Hotelling (pp. 278-292).

[11] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

[12] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.