INTERNATIONAL CALL RESEARCH CONFERENCE

https://doi.org/10.29140/97817637116240-03

Multi-Agent Debate in L2 Argumentative Writing: An Exploratory Study of Chinese EFL Learners

Tianshu Chu
Capital Normal University, China

Abstract

Argumentative writing poses significant challenges for second-language (L2) learners. Generative AI (GenAI) offers pedagogical support, but existing tools typically act as single assistants and fail to simulate the dynamic, multi-perspective dialogue essential for robust argumentation. We introduce a Multi-Agent Debate (MAD) framework in which multiple AI personas engage in structured debates to foster divergent thinking and expose learners to varied viewpoints. In a one-group pretest–posttest exploratory study (N = 22 Chinese EFL undergraduates), we investigated MAD's impact on argumentative writing and learner perceptions. Preliminary results indicate that MAD enhanced argument structure and quality, promoted metacognitive awareness, cultivated critical AI skepticism, and reduced anxiety when addressing controversial topics. These findings suggest MAD is a promising approach for improving both L2 argumentative skills and GenAI literacy.

Keywords: Generative AI, Multi-Agent Debate (MAD), Chinese EFL learners, argumentative writing

Background

Effective L2 argumentative writing requires not only constructing logical claims but also anticipating and refuting counterarguments. This is particularly challenging in Chinese English as a Foreign Language (EFL) contexts,

E-mail: tianschu@cnu.edu.cn

In Yijen Wang, Paul Moore, Teresa Hsieh, & Hao Tran (Eds.), Proceedings of the XXIIIrd International CALL Research Conference, 2025 (pp. 1–8). Castledown.

where traditional pedagogies may neglect the dialogical processes crucial for rebuttal skills. While in-class debating can be useful (e.g., El Majidi et al., 2021), it often presents hurdles such as linguistic barriers and performance anxiety.

Generative AI (GenAI) offers a new opportunity to support leaners. Prior research suggests that GenAI can enhance argument structure (Pratama et al., 2025) and support claim generation (Guo et al., 2023). However, most existing tools serve as "reactive" assistants, responding solely to user prompts and predefined criteria (Su et al., 2023), thus limiting opportunities for learners to engage in the multi-sided debate necessary for sophisticated argumentation. For instance, Guo et al. (2023) found no significant improvement in students' ability to formulate counterarguments or rebuttals when using a chatbot for argumentation, underscoring the need for more interactive, dialogical approaches.

To address this gap, this study proposes a Multi-Agent Debate (MAD) framework that draws on research in computational argumentation (e.g., Du et al., 2024; Hu et al., 2025; Liang et al., 2024). MAD leverages multiple AI agents to simulate debates (Du et al., 2024; Hu et al., 2025; Liang et al., 2024), which counteracts the "Degeneration-of-Thought" phenomenon in single-agent large language models (LLMs) (Liang et al., 2024) and creates a richer, dialogical environment for L2 writing development.

Research Questions

1. To what extent does engagement in a Multi-Agent Debate (MAD) framework improve Chinese EFL learners' argumentative writing skills?
2. How do learners perceive the role of the MAD framework in their argumentative writing process?

Methodology

Participants

Twenty-two undergraduate English majors (CEFR B2–C1) at a mainland Chinese university participated in the study as part of their foundational writing course.

The Multi-Agent Debate (MAD)

The MAD in this study was implemented using widely available LLMs such as DeepSeek-R1, GPT-4o, and Grok 3. Students were free to select the model they preferred and interacted with it through standard web-based chat

interfaces. Rather than relying on a pre-built educational platform, the debate environment was created through carefully designed prompts within a single model session. The design draws on existing MAD frameworks in computational argumentation research (e.g., Du et al., 2024; Hu et al., 2025; Liang et al., 2024). In particular, the inclusion of persona creation follows Hu et al. (2025), who argue that assigning distinct, high-level beliefs and viewpoints to each debating agent ensures the diversity of perspectives. This prompting process corresponds with Step 3 and Step 4 in Table 1. For reference, sample prompts are included in the Appendix.

For the current study, the overall MAD workflow consists of five steps (see Table 1). Steps 1 and 2 involved preparatory reading and stance-taking without AI involvement. In Step 3, students prompted the LLM to generate debater personas with names, roles, and core stances, and could revise them based on their own knowledge. In Step 4, they instructed the LLM to run a multi-round debate between these personas, generating claims, evidence, counterarguments, and rebuttals. In Step 5, students reflected on the debate using guiding questions and revised their original arguments.

Procedure

A one-group pretest-posttest design was conducted over five sessions. In the initial session, students received instruction and wrote a pretest essay on the topic "*Should subway systems implement mandatory security checks?*" Over the next three intervention sessions, students engaged with the MAD (see Table 1 for the full workflow), exploring a new topic each session. In the final session, students composed a posttest essay on the topic "*Should universities require AI checks on term papers?*". For both the pretest and posttest, students conducted 15 minutes of online research followed by 40 minutes of offline writing. Ten participants were interviewed after the posttest.

Data Analysis

We analyze three data sources:

a. **Pre-/Posttest essays:** Scored using rubrics (adapted from El Majidi, 2020; Guo et al., 2023) for structural complexity (based on the Toulmin model), argument quality, and language.
b. **Reflection logs:** Thematically analyzed to identify recurring patterns in student engagement with the MAD system.
c. **Interview transcripts:** Used to validate log findings and explore learner perceptions in depth.

Table 1. The Multi-Agent Debate (MAD) Workflow

Steps	Student-Directed AI Generation	Student Learning Actions
1. Read & Understand	*– (No AI action in this step. Students access human-selected materials.)*	Students read a pre-selected opinion article and are given a related debate topic. Example: Article – "Want Kids to Stay in School? Let Them Keep Their Phones" (Sixth Tone). Debate topic – "Dpo the benefits of cellphone bans in rural Chinese schools outweigh the drawbacks?"
2. Take a Position	*– (No AI action)*	Students write a short paragraph stating their initial opinion on the topic. Example: "I believe the benefits of banning cellphones in rural Chinese schools do outweigh the drawbacks. Without phones engage more in class, sports, and friendship..."
3. Create personas	Students prompt the LLM to generate at least two debater personas with names, roles, and core stances.	Students receive the generated personas and can compare them to any they imagined themselves. Example personas: Persona A: Zhang Wei, Rural Middle School Principal, "Structured learning environments free from digital distractions are essential for academic focus and discipline in under-resourced schools." Persona B: Chen Lihua, Education Policy Researcher, "Preparing students for a digital future requires responsible access to technology, not blanket bans that widen the digital divide." ...
4. Simulate Debate with LLM	Students instruct the LLM to run a multi-round debate between the personas.	Students read the AI-generated debate transcript, identifying claims, evidence, counterarguments, and rebuttals. They annotate surprising or strong arguments, and note weak or repetitive points.
5. Reflect & Expand	*– (No AI action)*	Student reflect on the debate using reflection questions, then revise or write their own argumentative essay. Sample reflection questions: What were the 2–3 most surprising or strong arguments generated by the AI? What challenges did you face in analyzing the AI debate? How did this debate change your thinking or writing?

Representative Results

Although the full data analysis is ongoing, preliminary results have revealed several notable trends. Posttest essays showed marked improvement in structural complexity over the predominantly one-sided pretest essays. Most posttest essays incorporated opposing perspectives, shifting from simply acknowledging counterarguments to involving substantive rebuttals. While overall linguistic complexity saw minimal change, there was a marked increase in phrases signalling dialogical argumentation, such as "Conversely," "While proponents argue," and "This viewpoint, however, fails to consider...".

Reflection logs and interviews revealed three key themes in learner perceptions. First, students reported enhanced metacognition; the debates helped them identify "blind spots" and logical fallacies in their own reasoning. As one participant noted, "The debate lets me see different points in a critical way and new angles can be added." Second, learners developed critical AI skepticism. They actively evaluated the AI-generated content by fact-checking claims, flagging "unvalidated data," and noting "cliched phrasing" or "stereotypical responses" from personas. Finally, the framework's low-stakes nature reduced writing anxiety. It created a safe space for learners to explore contentious topics. Several participants noted that the simulated debate help them "rehearse the points" and "boost confident in expressing views".

Discussion

This study's preliminary evidence suggests the MAD framework significantly enhances L2 argumentative writing. A key finding is that MAD appears to promote critical GenAI literacy (Li, 2025). Rather than passively accepting AI output, students actively interrogated it. This suggests MAD can function as a "cognitive mirror," prompting learners to scrutinize both AI-generated arguments and their own, which is a crucial skill for responsible AI integration in pedagogy.

Additionally, MAD appears to lower the affective filter for engaging with controversial topics. The low-stakes simulated environment allowed students to "rehearse" arguments without the social pressure of live debate. This is valuable in cultural contexts where direct contention may be discouraged, offering a scaffold for students to build confidence for authentic civic discourse.

Limitations and Future Directions

The study's conclusions are preliminary and should be interpreted with caution. Key limitations also include the small sample size (N = 22), the

pre-experimental design that lacks a control group, and the short intervention period. These factors limit the generalizability of the findings and restrict causal claims.

Despite these limitations, this exploratory study highlights that the MAD framework holds significant potential. It uniquely addresses the need for multi-perspective engagement critical for developing rebuttal skills, fosters the metacognitive evaluation of AI, and provides a safer pathway for Chinese students to engage with complex social issues. Future research should employ rigorous quasi-experimental designs to validate and extend these findings.

References

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24), 235*, Article 467, 11733–11763. https://doi.org/10.48550/arXiv.2305.14325

El Majidi, A., de Graaff, R., & Janssen, D. (2020). Debate as L2 pedagogy: The effects of debating on writing development in secondary education. *Modern Language Journal*, *104*(4), 804–821. https://doi.org/10.1111/modl.12673

El Majidi, A., Janssen, D., & Graaff, R. de. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, *101*, Article 102576. https://doi.org/10.1016/j.system.2021.102576

Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education*, *203*, Article 104862. https://doi.org/10.1016/j.compedu.2023.104862

Hu, Z., Chan, H. P., Li, J., & Yin, Y. (2025). Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4689–4703). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2406.19643

Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature*, *2*(1), 122–152. https://doi.org/10.1515/dsll-2025-0007

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17889–17904).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.992

Pratama, R. D., Widiati, U., & Hakim, L. N. (2025). Effects of human-AI collaborative writing strategy on EFL students' argumentative writing skills. *Computer Assisted Language Learning*, 1–32. https://doi.org/10.1080/09588221.2025.2503900

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752

Appendix

Sample User Prompts for Persona Creation and Debate Ssimulation

User prompt for persona creation:

You are designing a debate on the following topic: "Do the benefits of cellphone bans in rural Chinese schools outweighs the drawbacks?" Create three personas who are going to debate on the topic. For each persona: Provide a name and concrete role of the persona. State their worldview or values in one sentence. Output exactly in this format:

Persona A: [name], [role], [worldview].

Persona B: ...

User prompt for debate simulation:

Now please simulate a debate on the topic above. Three debater personas have been created, each with distinct roles, backgrounds, and reasons for opposing the proposition.

Debate Rules:

There are five rounds. Each round follows the order: Persona A → Persona B → Persona C. Each persona must respond in 3–4 sentences per turn. Language must be respectful, clear, and logically structured. Avoid repetition of points. Persuasive strategies such as examples, analogies, and rhetorical questions are encouraged.

Round Structure: Opening Statements: Each persona presents their main claim and two supporting reasons. First Rebuttal: Each persona challenges one or more points made by other personas, identifying weaknesses or offering counterexamples. Second Rebuttal: Each persona deepens their challenge by addressing counter-challenges or introducing new evidence. Synthesis Round: Each persona clarifies their position in light of earlier rounds, acknowledging valid points from others while reinforcing their own stance. Closing Statements: Each persona summarizes why their position is stronger than others and appeals to the audience's logic or values.

Output Format: Clearly label each round and each persona's turn (e.g., "Round 1 – Rural middle school principal").

Keep responses concise and relevant to the proposition. Maintain coherence across rounds so that arguments build on previous exchanges.