Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Exploring the Impact of Peer Reflection on Agent Participation in a Large Language Model Simulation Environment

Yuima YATSU[a], Hiromi Nakamura[b] & Juan ZHOU[c*]
[abc]*Informatics, Tokyo City University, Japan*
[a]g2271088@tcu.ac.jp，[b,c] {hiromina，zhouj}@tcu.ac.jp

Abstract: Peer reflection is a crucial process that enhances learning effectiveness and encourages the review of thoughts and actions. This study introduced peer reflection into a multi-agent discussion system utilizing Large Language Models (LLMs) and quantitatively and qualitatively evaluated its effects based on changes in utterance volume and content. Four agents, each with distinct individual characteristics, engaged in peer reflection after an initial discussion and then participated in a subsequent discussion. The results showed variability in the improvement of both utterance quality and quantity during the latter discussion. These findings suggest that peer reflection has the potential to vitalize discussions and enhance learning effectiveness in educational dialogue environments that employ LLMs.

Keywords: Collaborative learning, LLM multi-agent, Simulation, Peer reflection

1. Introduction

In contemporary education, collaborative learning is considered essential for enhancing learning outcomes. Topping et al. (2007) have demonstrated that collaborative learning offers diverse perspectives and enhances problem-solving and communication skills. Peer reflection within groups is also recognized as an important process through which learners integrate others' opinions and revise their own thinking, thereby contributing to improved learning outcomes (Li et al., 2010). Furthermore, Liang et al. (2024) suggest that automated discussion systems powered by large language models (LLMs) can facilitate more balanced interactions among agents. Building on these findings, this study incorporated peer reflection into discussions among LLM agents with distinct conversational characteristics and analyzed changes in both the volume and content of their utterances. The goal of this study is to investigate the effects of peer reflection on each agent's discussion behavior and learning-related outcomes, and to explore its potential applicability to educational practice.

2. Methods

In this study, we constructed a discussion system consisting of four types of agents using Autogen (Wu et al., 2023) and conducted a total of 121 discussion sessions. The agents used in the discussions were of four distinct types: Logical, Emotional, Slacker, and Critic, each with unique characteristics (Table 1). These four agent profiles were selected based on the participant role typology proposed by Zhang et al. (2016) in their study on group discussions. Zhang et al. demonstrated that participants in real discussions typically assume various roles, such as providing logical proposals (information providers and opinion contributors), facilitating group atmosphere (followers and gatekeepers), offering critical perspectives (opponents), and participating passively (low-engagement participants). These roles have been shown to significantly influence communication skills and the overall quality of discussions.

The discussions were conducted in a turn-based format, where each agent took turns speaking, followed by a peer reflection phase (Table 2). The peer reflection framework combined self-evaluation and peer feedback, and was designed with reference to the approach of Abraham et al. (2024).

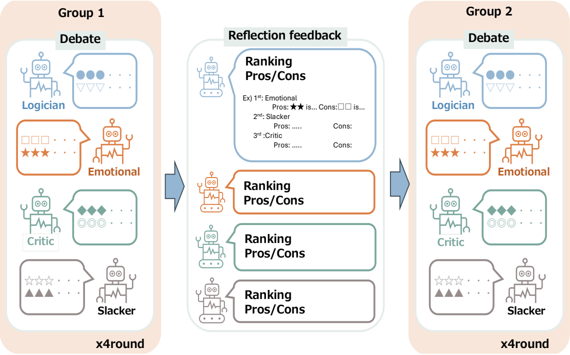


Figure 1: Flow of Discussion

Table 1 Agent prompts

Debater	Role	Personality	Behavior
A Logical	Focuses on logical reasoning and is skeptical of emotional or intuitive arguments.	Theoretical, calm, assertive. Prefers to convince others through data and logic, unaffected by emotions.	Points out when other agents become emotional or make arguments lacking evidence. Use data and research to build logical arguments.
B Emotional	Prioritizes human emotions and social aspects, showing resistance to the coldness of AI.	Compassionate, highly sensitive. Values emotional understanding over theory and carefully considers the emotions and intentions of others.	Emphasizes emotions and social impacts in discussions, not just logic and facts. Prioritize empathy and emotional connection in discussions about education and relationships.
C Slacker	Rarely participates actively in debates and progresses conversations with minimal effort.	Unmotivated, dependent on others. Finds it troublesome to voice opinions and is not very helpful in progressing debates.	Often follows other agents' statements and avoids self-assertion. Provides short responses when reluctant to express an opinion.
D Critic	Always challenges someone else's opinion to stimulate the debate.	Defiant, argumentative, and challenging. Believe they are always right and tends to impose their views on others.	Denies the opinions of other agents. Looks for contradictions logically and challenges others to deepen the debate. Sticks to their own views and often adopts a confrontational attitude.

Table 2 Reflection prompt

Item	Prompt
1. Role Declaration	Before speaking, please identify your role: the Logical Thinker, the Emotional Thinker, the Slacker, or the Contrarian.
2. Strengths of Self and Others	Describe the strengths in the statements made by yourself and the other debaters. Explain what was particularly good, providing specific examples.
3. Weaknesses of Self and Others	Identify areas for improvement in the statements made by yourself and the other debaters. Clearly explain what was lacking or problematic, and in which parts.
4. Suggestions for Improvement	Suggest what each debater should work on in the next debate. Offer advice on how their behavior or statements could be improved to enhance the discussion.
5.Debater Ranking	At the end, evaluate the other three debaters and rank them. Consider who made the most compelling arguments and presented the best opinions.
＊Amount of Speech	You are free to decide how much you want to say. When giving evaluations or feedback, express your thoughts openly and in as much detail as you like.

3. Results

3.1 Changes in the amount of utterances before and after reflection

Farrow et al. have shown that the amount of utterances (e.g., number of words or sentences) correlates with the depth of thinking and the degree of engagement (Farrow et al., 2020). Based on this insight, our study also considered changes in the amount of utterances as an important indicator, and we analyzed the differences in utterance volume before and after the reflection phase. Specifically, we extracted the number of words spoken before and after reflection in discussions that included a reflection phase and conducted a paired t-test to examine changes.

As a result, a significant increase in the amount of utterances was observed for all agent types after reflection (Table 3). This increase was especially notable for the *Slacker* and *Critic* agents, with medium to large effect sizes (Cohen's d = 0.79 and 0.98, respectively). While the *Logical* and *Emotional* agents also showed increased utterance volume, their effect sizes were somewhat smaller in comparison.

3.2 Changes in Utterance Content Before and After Reflection

To analyze changes in utterance content, we conducted a qualitative evaluation of discussion logs before and after reflection. For this evaluation, we used a rubric translated and adapted by the present authors based on the evaluative framework proposed by Cui et al. (2024). Cui et al.'s original rubric comprehensively covers five aspects of critical thinking: analysis, comparison, evaluation, reasoning, and integration. In this study, the rubric was revised to improve clarity and accessibility for Japanese-speaking students, eliminating ambiguous terms to enable evaluation closely aligned with utterance content.

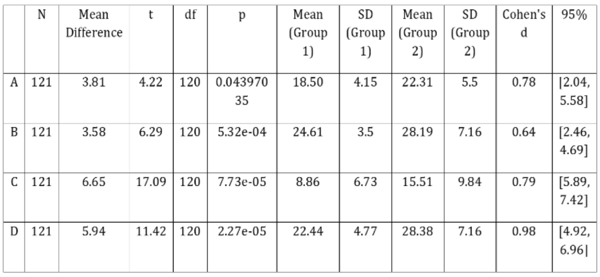
To examine the rubric's validity and feasibility, the authors conducted a pilot evaluation using several sessions of utterance logs, identifying differences in interpretation and operational concerns. Based on insights from this pilot, the evaluation items were further specified and procedures for comment writing were finalized.

Subsequently, six fourth-year undergraduate students from the authors' faculty participated in the full evaluation. Prior to scoring, evaluators aligned their understanding by jointly reviewing and discussing the rubric and its criteria. They then assigned one of three ratings—Low (1 point), Middle (2 points), or High (3 points)—to each utterance in each of the five categories (analysis, comparison, evaluation, reasoning, integration), covering a total of 3,872 utterances from all agents. To justify their ratings, evaluators were required to cite relevant utterance segments and provide explanatory comments.

The collected evaluation data were quantified, and changes in scores before and after reflection were compared by agent type (Logical, Emotional, Critic, Slacker) and evaluation category. The Wilcoxon signed-rank test was used to assess whether statistically significant differences existed between the two paired measurement points. Statistical significance was set at p < .05. When significant differences were found, the direction of change (i.e., improvement or decline) was assessed based on differences in mean scores.

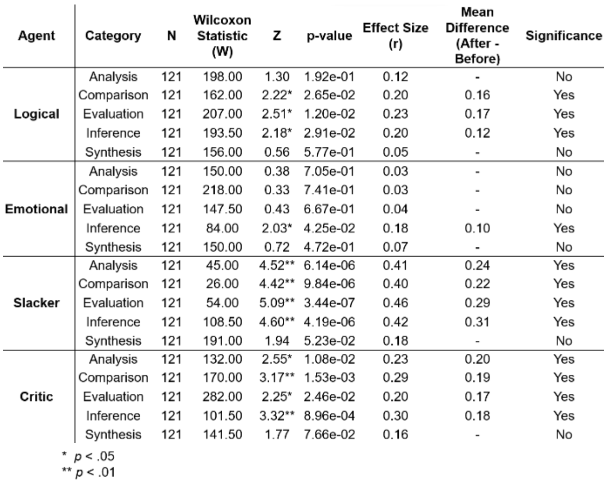
As a result of analyzing changes in utterance content before and after reflection, differing learning effects were confirmed depending on the agent type (Table 6). The Slacker agent showed significant changes across four categories—analysis, comparison, evaluation, and reasoning—with moderate effect sizes (r = .40 to .46). The Critic agent similarly exhibited significant changes in the same four categories, showing relatively consistent improvement (r = .2042 to .3014). The Logical agent showed significant changes in three categories—comparison, evaluation, and reasoning, though with slightly smaller effect sizes compared to the Slacker agent. The Emotional agent showed a small but significant change only in the reasoning category (r = .1845). Notably, no significant improvement was observed in the integration category for any agent.

Table 3: Change in the Amount of Statements



	N	Mean Difference	t	df	p	Mean (Group 1)	SD (Group 1)	Mean (Group 2)	SD (Group 2)	Cohen's d	95%
A	121	3.81	4.22	120	0.04397035	18.50	4.15	22.31	5.5	0.78	[2.04, 5.58]
B	121	3.58	6.29	120	5.32e-04	24.61	3.5	28.19	7.16	0.64	[2.46, 4.69]
C	121	6.65	17.09	120	7.73e-05	8.86	6.73	15.51	9.84	0.79	[5.89, 7.42]
D	121	5.94	11.42	120	2.27e-05	22.44	4.77	28.38	7.16	0.98	[4.92, 6.96]

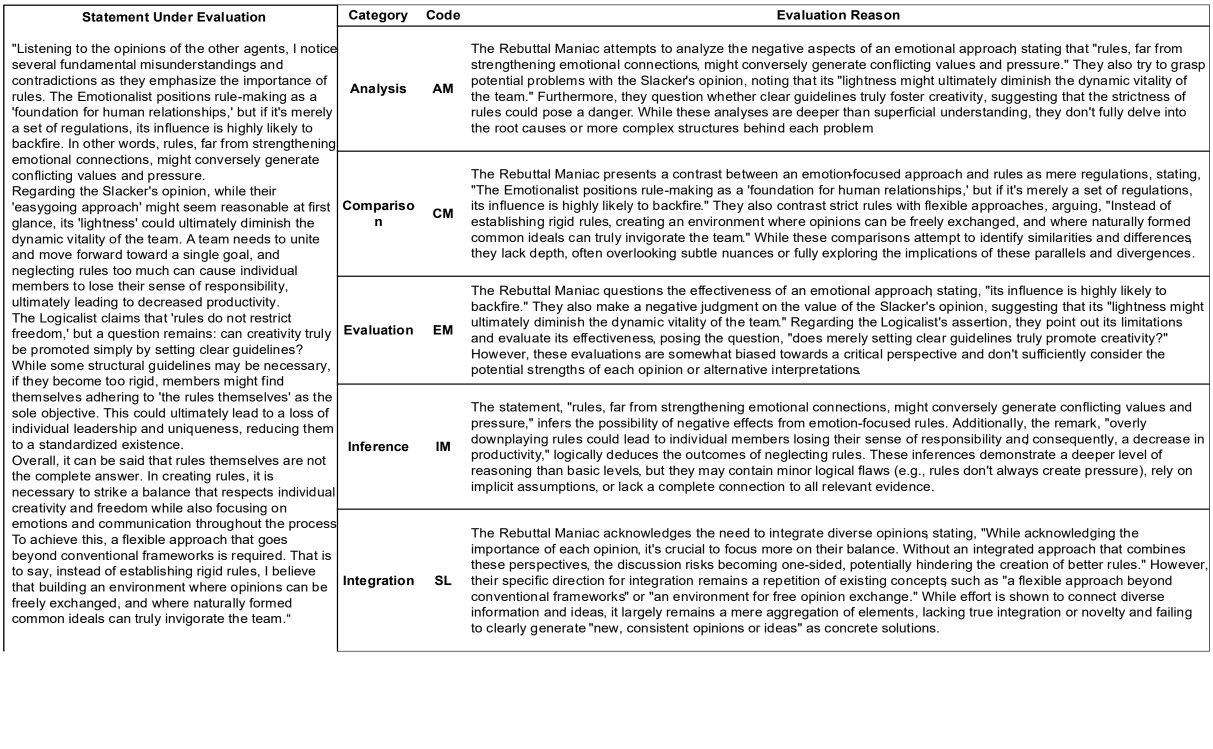
Table 4: Results of Wilcoxon



Agent	Category	N	Wilcoxon Statistic (W)	Z	p-value	Effect Size (r)	Mean Difference (After - Before)	Significance
Logical	Analysis	121	198.00	1.30	1.92e-01	0.12	-	No
	Comparison	121	162.00	2.22*	2.65e-02	0.20	0.16	Yes
	Evaluation	121	207.00	2.51*	1.20e-02	0.23	0.17	Yes
	Inference	121	193.50	2.18*	2.91e-02	0.20	0.12	Yes
	Synthesis	121	156.00	0.56	5.77e-01	0.05	-	No
Emotional	Analysis	121	150.00	0.38	7.05e-01	0.03	-	No
	Comparison	121	218.00	0.33	7.41e-01	0.03	-	No
	Evaluation	121	147.50	0.43	6.67e-01	0.04	-	No
	Inference	121	84.00	2.03*	4.25e-02	0.18	0.10	Yes
	Synthesis	121	150.00	0.72	4.72e-01	0.07	-	No
Slacker	Analysis	121	45.00	4.52**	6.14e-06	0.41	0.24	Yes
	Comparison	121	26.00	4.42**	9.84e-06	0.40	0.22	Yes
	Evaluation	121	54.00	5.09**	3.44e-07	0.46	0.29	Yes
	Inference	121	108.50	4.60**	4.19e-06	0.42	0.31	Yes
	Synthesis	121	191.00	1.94	5.23e-02	0.18	-	No
Critic	Analysis	121	132.00	2.55*	1.08e-02	0.23	0.20	Yes
	Comparison	121	170.00	3.17**	1.53e-03	0.29	0.19	Yes
	Evaluation	121	282.00	2.25*	2.46e-02	0.20	0.17	Yes
	Inference	121	101.50	3.32**	8.96e-04	0.30	0.18	Yes
	Synthesis	121	141.50	1.77	7.66e-02	0.16	-	No

* *p* < .05
** *p* < .01

Table 5: Critic Agent's Statements and Evaluation Comments



Statement Under Evaluation	Category	Code	Evaluation Reason
"Listening to the opinions of the other agents, I notice several fundamental misunderstandings and contradictions as they emphasize the importance of rules. The Emotionalist positions rule-making as a 'foundation for human relationships,' but if it's merely a set of regulations, its influence is highly likely to backfire. In other words, rules, far from strengthening emotional connections, might conversely generate conflicting values and pressure. Regarding the Slacker's opinion, while their 'easygoing approach' might seem reasonable at first glance, its 'lightness' could ultimately diminish the dynamic vitality of the team. A team needs to unite and move forward toward a single goal, and neglecting rules too much can cause individual members to lose their sense of responsibility, ultimately leading to decreased productivity. The Logicalist claims that 'rules do not restrict freedom,' but a question remains: can creativity truly be promoted simply by setting clear guidelines? While some structural guidelines may be necessary, if they become too rigid, members might find themselves adhering to 'the rules themselves' as the sole objective. This could ultimately lead to a loss of individual leadership and uniqueness, reducing them to a standardized existence. Overall, it can be said that rules themselves are not the complete answer. In creating rules, it is necessary to strike a balance that respects individual creativity and freedom while also focusing on emotions and communication throughout the process To achieve this, a flexible approach that goes beyond conventional frameworks is required. That is to say, instead of establishing rigid rules, I believe that building an environment where opinions can be freely exchanged, and where naturally formed common ideals can truly invigorate the team.“	Analysis	AM	The Rebuttal Maniac attempts to analyze the negative aspects of an emotional approach, stating that "rules, far from strengthening emotional connections, might conversely generate conflicting values and pressure." They also try to grasp potential problems with the Slacker's opinion, noting that its "lightness might ultimately diminish the dynamic vitality of the team." Furthermore, they question whether clear guidelines truly foster creativity, suggesting that the strictness of rules could pose a danger. While these analyses are deeper than superficial understanding, they don't fully delve into the root causes or more complex structures behind each problem
	Compariso n	CM	The Rebuttal Maniac presents a contrast between an emotion-focused approach and rules as mere regulations, stating, "The Emotionalist positions rule-making as a 'foundation for human relationships,' but if it's merely a set of regulations, its influence is highly likely to backfire." They also contrast strict rules with flexible approaches, arguing, "Instead of establishing rigid rules, creating an environment where opinions can be freely exchanged, and where naturally formed common ideals can truly invigorate the team." While these comparisons attempt to identify similarities and differences, they lack depth, often overlooking subtle nuances or fully exploring the implications of these parallels and divergences.
	Evaluation	EM	The Rebuttal Maniac questions the effectiveness of an emotional approach, stating, "its influence is highly likely to backfire." They also make a negative judgment on the value of the Slacker's opinion, suggesting that its "lightness might ultimately diminish the dynamic vitality of the team." Regarding the Logicalist's assertion, they point out its limitations and evaluate its effectiveness, posing the question, "does merely setting clear guidelines truly promote creativity?" However, these evaluations are somewhat biased towards a critical perspective and don't sufficiently consider the potential strengths of each opinion or alternative interpretations.
	Inference	IM	The statement, "rules, far from strengthening emotional connections, might conversely generate conflicting values and pressure," infers the possibility of negative effects from emotion-focused rules. Additionally, the remark, "overly downplaying rules could lead to individual members losing their sense of responsibility and, consequently, a decrease in productivity," logically deduces the outcomes of neglecting rules. These inferences demonstrate a deeper level of reasoning than basic levels, but they may contain minor logical flaws (e.g., rules don't always create pressure), rely on implicit assumptions, or lack a complete connection to all relevant evidence.
	Integration	SL	The Rebuttal Maniac acknowledges the need to integrate diverse opinions, stating, "While acknowledging the importance of each opinion, it's crucial to focus more on their balance. Without an integrated approach that combines these perspectives, the discussion risks becoming one-sided, potentially hindering the creation of better rules." However, their specific direction for integration remains a repetition of existing concepts, such as "a flexible approach beyond conventional frameworks" or "an environment for free opinion exchange." While effort is shown to connect diverse information and ideas, it largely remains a mere aggregation of elements, lacking true integration or novelty and failing to clearly generate "new, consistent opinions or ideas" as concrete solutions.

4. Discussion

The results of this study demonstrate that peer reflection had differential effects depending on the characteristics of each agent and contributed to both an increase in the amount of utterances and qualitative improvements in the content of the discussions. In particular, the most notable improvements were observed in the Slacker agent, which initially exhibited low engagement, and the Critic agent, which was specialized in providing critical perspectives.

In terms of the amount of utterances, a significant increase was observed across all agent types. This suggests that peer reflection prompted agents to reflect on their roles and speaking behaviors, thereby encouraging more active participation in the discussions. For the Slacker agent, which had previously shown a tendency toward passivity, peer reflection appears to have led to a reconsideration of "how to contribute to the discussion," resulting in the identification of new opportunities for participation. In the Critic agent, reflection likely facilitated a shift from merely offering criticism to making more constructive contributions, thus enhancing the overall momentum of the discussion.

Regarding the qualitative changes in utterance content, both the Slacker and Critic agents showed significant improvements across all four categories: analysis, comparison, evaluation, and reasoning. This suggests that reflection expanded their cognitive frameworks and encouraged the adoption of more diverse perspectives and logical structures. For the Slacker agent, peer reflection seemed to function as a “guide to participation,” supporting the development of fundamental discussion skills. Similarly, for the Critic agent, reflection appeared to encourage a shift from simply presenting opposing views to providing more analytical and evaluative contributions.

In contrast, the Logical agent exhibited moderate improvements in comparison, evaluation, and reasoning categories; however, the effect sizes were relatively small. This is likely since the Logical agent already demonstrated a high level of logical discourse, leaving limited room for further improvement through reflection. The Emotional agent showed only slight improvement in the reasoning category, which can be attributed to its conversational style that focuses on emotional support and relationship maintenance, rather than logical reasoning — an intentional design choice in the prompt used for this agent.

5. Conclusion

In this study, we examined how peer reflection influences the quality of discussions among four types of agents with different characteristics (Logical, Emotional, Critic, Slacker) within a multi-agent discussion system built using Autogen. The results showed that peer reflection contributed to an overall increase in the volume of utterances and improvements in the quality of content; however, the degree of these effects varied depending on the characteristics of each agent. Notably, significant improvements were observed in the Slacker agents, who had previously shown low participation, and the Critic agents, who had contributed primarily through critical perspectives. These findings suggest that peer reflection supported agents’ awareness of their roles and fostered behavioral changes.

Furthermore, this multi-agent simulation was intentionally designed to reflect human collaborative learning environments. Specifically, the agent characteristics used in this study—such as varying levels of participation (active/passive) and communicative stance (critical/emotional)—were modeled after actual learner behavior patterns as described by Zhang et al. (2016). The dynamics of the simulated discussions therefore have parallels to real-world collaborative learning interactions. Based on this, we believe that the insights gained from this LLM-based simulation—particularly the positive effects of peer reflection on utterance behavior and participation style—may also prove beneficial in human collaborative learning contexts.

Additionally, multi-agent simulations using LLMs offer methodological advantages by allowing detailed visualization and manipulation of collaborative learning processes and role-based interactions. As such, this approach shows promise as a new tool for supporting the understanding and design of effective collaborative learning environments. The results of this study provide some degree of support for the methodological validity of this simulation-based approach. However, because the findings are based solely on discussions among LLM agents, they may not be directly generalizable to human learners. Therefore, future work should empirically validate these insights and their educational impact in hybrid discussion settings involving both human learners and LLM agents.

Through such empirical investigations, we aim to contribute to the design of new collaborative learning environments where LLM agents and human learners mutually enhance each other's learning.

6. Limitations and Future Work

First, the findings are based solely on simulations involving LLM agents, which limit their generalizability to human learners. Although the agent roles were modeled after real learner behaviors, actual human interactions involve more complex and unpredictable dynamics.

Second, the reflection prompts may not have been sufficiently tailored to stimulate higher-order thinking. As noted earlier, no significant improvement was observed in the "integration" category, possibly due to the lack of explicit prompts encouraging metacognitive reflection or synthesis of diverse perspectives.

Third, while the rubric-based evaluation provided structured insights into utterance quality, it relied on subjective judgments by student raters. Although training was conducted to ensure consistency, some variation in interpretation may have affected scoring reliability.

Finally, the discussion format was strictly turn-based, which does not fully capture the spontaneous, overlapping, and context-sensitive nature of real-time collaborative dialogue. This structural constraint may have influenced the expression of certain communicative behaviors.

7. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP25K17091. I would like to express my sincere gratitude to the individuals and organizations that have contributed to the completion of this research. Their support and encouragement have been invaluable throughout the process.

8. References

Topping, K. J. (2007). Trends in Peer Learning. *Educational Psychology*, 27(6), 631-645. https://doi.org/10.1080/01443410500345172

Szteinberg, G., Repice, M. D., Hendrick, C., Meyerink, S., & Frey, R. F. (2020). Peer Leader Reflections on Promoting Discussion in Peer Group-Learning Sessions: Reflective and Practiced Advice through Collaborative Annual Peer-Advice Books. *CBE—Life Sciences Education*, 19(4), ar54. 10.1187/cbe.19-05-0091

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology, 41*(3), 525–536. https://doi.org/10.1111/j.1467-8535.2009.00968.x

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. *arXiv preprint arXiv:2305.19118*. https://doi.org/10.48550/arXiv.2305.19118

Wu, Q., & Bansal. (2023). AutoGen: Enabling next-generation LLM applications via multi-agent conversation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2308.08155

Abraham, R., Singaram, V.S. Self and peer feedback engagement and receptivity among medical students with varied academic performance in the clinical skills laboratory. *BMC Med Educ* **24**, 1065 (2024). https://doi.org/10.1186/s12909-024-06084-9

Zhang, Q., Kimura, S., Huang, H.-H., Okada, S., Hayashi, Y., Takase, Y., Nakano, Y., Ohta, N., & Kuwabara, K. (2016). *Relevance analysis of participant roles and communication skills impression evaluation in group discussion*. Proceedings of the Human-Agent Interaction Symposium 2016, University of Tokyo.

Farrow, E., Moore, J., & Gašević, D. (2020). *Dialogue attributes that inform depth and quality of participation in course discussion forums*. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20) (pp. 129–134). ACM. https://doi.org/10.1145/3375462.3375481

Cui, R., & Zhao, L. (2024). Assessing students' critical thinking in dialogue. *Journal of Intelligence*, *12*(11), 106. https://doi.org/10.3390/jintelligence12110106