MMD-ERE: Multi-Agent Multi-Sided Debate for Event Relation Extraction

Yong Guan[1], Hao Peng[1], Lei Hou[1], Juanzi Li[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
{gy2022, peng-h24}@mails.tsinghua.edu.cn, {lijuanzi, houlei}@tsinghua.edu.cn

Abstract

Event relation extraction (ERE) is becoming increasingly important in the era of large language models. An extensive body of research has explored how performance can be further enhanced by the emergence of exciting technologies like chain-of-thought and self-refinement. In this paper, we introduce **MMD-ERE**, a Multi-agent Multi-sided Debate approach for Event Relation Extraction, which explores the understanding of event relations among different participants before and after debate. Specifically, for organizing the debate, participants are divided into multiple groups, each assigned its own debate topic, and the process effectively integrates both cooperation and confrontation. We also regard the audience as a crucial participant, as their conclusions from an observer's perspective tend to be more objective. In the end, we explore the understanding of event relations among different participants before and after the debate. Experiments across various ERE tasks and LLMs demonstrate that MMD-ERE outperforms established baselines. Further analysis shows that debates can effectively enhance participants' understanding of event relations.

1 Introduction

Event relation extraction (ERE) aims to extract event relations, such as causal and subevent relations, between events within documents, revealing the underlying reliable document structures. This becomes increasingly important in the era of large language models (LLMs), as it can enhance the reasoning capabilities of LLMs, thereby facilitating their application in real scenarios.

Over the past year, LLMs have achieved remarkable performance as general task-solving agents in various natural language processing (NLP) tasks, such as language generation and understanding. This progress has led to the emergence of exciting technologies, such as chain-of-thought (CoT) (Wei et al., 2022) and self-refinement (Madaan et al., 2023), which simulate the characteristics of human problem-solving through the generation of intermediate steps or iterative revision. Many studies have begun utilizing LLMs to explore the ERE task (Zheng et al., 2024; Yuan et al., 2023). However, these techniques are mainly applied to a single model instance. Existing research indicates that LLMs can be overconfident and stubborn (Guan et al., 2024b), as shown in Figure 1. In contrast, multiple agents can communicate to capture external insights to tackle complex tasks. However, when leveraging multi-agent approaches for the ERE task, there are two critical problems: (1) How to integrate multi-agent with event relation extraction? (2) How to design the interactions for multi-agent?

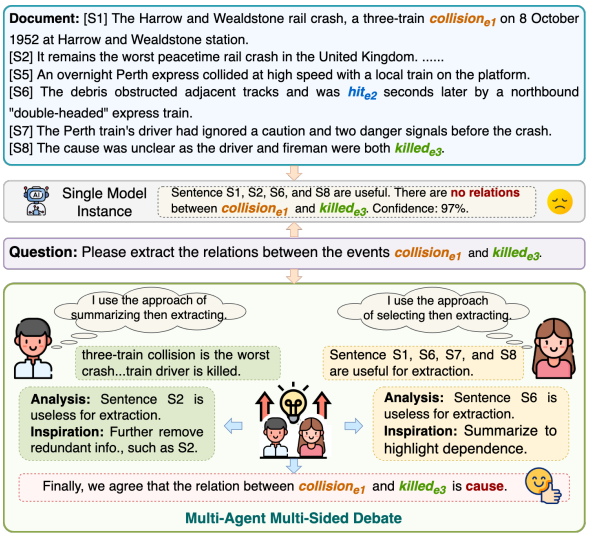

Figure 1: Basic single model instance (top) and multi-agent multi-sided debate (bottom).

In this work, we introduce MMD-ERE, a multi-agent multi-sided debate approach for event relation extraction, which explores the understanding of event relations among different participants before and after debate competition. *For integrating*

Proceedings of the 31st International Conference on Computational Linguistics, pages 6889–6896
January 19–24, 2025.

multi-agent into ERE, one challenge of document-level ERE involves compressing the document to highlight the event dependencies. Following this idea, we choose different ERE methods for agents with various stances, e.g., generation then extraction, selection then extraction. This approach enables agents to capture insights at the methodological level during interactions rather than focusing solely on details. For instance, as shown at the bottom of Figure 1, after discussion, the boy not only understands that S2 is useless for relation extraction but also gains the inspiration to further compress the selected sentences using summarization to highlight event dependencies. *For multi-agent interaction,* inspired by discussion methods in real-world teaching scenarios where students are divided into multiple groups, and drawing on the popular Public Forum Debate format, we propose a multi-sided debate based on LLMs. In this format, debaters are divided into multiple groups, each assigned its own debate topic, and the process effectively integrates both cooperation (Guan et al., 2024a; Yin et al., 2023) and confrontation (Liang et al., 2023; Wang et al., 2023). It's worth mentioning that we consider the audience as a crucial participant in the debate, as their conclusions from an observer's perspective tend to be more objective. In the end, we explore the understanding of event relations among different participants before and after the debate. Our contributions include the following:

- We build a multi-agent multi-sided debate approach for event relation extraction, namely MMD-ERE, which explores the understanding of event relation among different participants through the debate.
- MMD-ERE divides debaters into multiple groups, each assigned its own debate topic, and the process effectively integrates both communication and confrontation.
- Experimental results show that MMD-ERE outperforms the baseline models, and the analysis of the impact of different participants reveals that debates can effectively enhance participants' understanding of the problems.

2 MMD-ERE

This section aims to elaborate on the principal components in MMD-ERE, as shown in Figure 2, including *Debate Participants*, *Debate Topic*, and *Debate Process*.

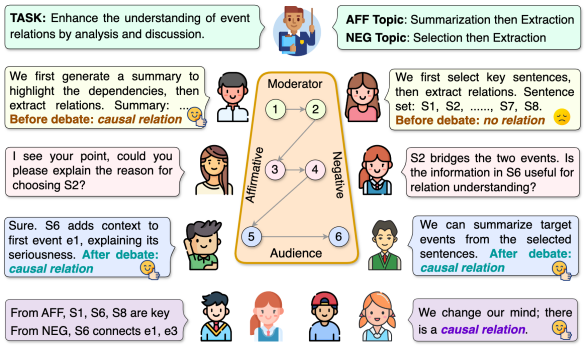

Figure 2: The framework of MMD-ERE. We display the Constructive Speech phase among different agents.

Debate Participants. Our framework revolves around three key participants: *Debater*, *Moderator*, and *Audience*.

- **Debater.** Debaters, including both affirmative side (AFF) and negative side (NEG), are the main participants in a debate. By presenting their viewpoints and engaging in discussions with the other side, they contribute to a deeper understanding of the debate topic among all participants.
- **Moderator.** The moderator is responsible for explaining and enforcing the rules of the competition, ensuring fairness and the smooth progression of the match, while also guiding the discussion and acting as a judge when necessary to determine the final outcome of the competition.
- **Audience.** By listening to the debate, the audience deepens their understanding of the debate topic. As observers, they can judge and understand the debate topic from a more objective perspective.

Debate Topic. We focus on document-level event relation extraction by organizing a formal debate to explore the understanding of the debate topic among different participants. A major challenge in ERE involves long-distance dependencies (Tang et al., 2021; Tran Phu and Nguyen, 2021; Chen et al., 2022). One approach to solving this challenge is to compress the document to model the relations between events, typically employing two strategies: (1) *Generation-then-Extraction (GtE)*, which generates summaries for event pairs to compress the document (Guo et al., 2023); (2) *Selection-then-Extraction (StE)*, which selects a set of sentences from the document that are useful for rela-

Methods	MAVEN-ERE						EventStoryLine		
	Causal			Subevent			Causal		
	P	R	F1	P	R	F1	P	R	F1
Basic LLMs									
Whole Document	35.25	39.81	37.39	67.31	36.65	48.75	19.63	24.99	21.62
Target Sentence Pair	36.74	35.49	36.11	69.21	37.68	49.04	21.00	20.46	20.73
Single Agent									
CoT	42.09	41.05	41.56	61.42	38.61	49.82	23.47	37.24	24.56
Self-Refinement	49.89	40.81	42.66	59.72	43.57	51.71	23.26	39.40	24.93
Multiple Agents									
EoT	47.32	43.52	45.34	71.13	43.82	55.47	23.45	**45.66**	26.16
MAD	50.34	44.52	47.26	70.66	47.73	57.31	23.85	39.57	25.86
Ours									
AFF Side	47.46	60.49	53.19	62.25	59.39	60.78	22.59	40.66	29.04
NEG Side	50.34	43.52	46.26	60.67	**65.54**	**63.88**	23.74	28.90	26.07
Moderator	44.72	63.32	50.71	57.61	45.28	54.37	22.83	42.46	29.70
Audience	**53.27**	59.49	54.43	**71.71**	53.96	61.58	22.48	35.58	28.31
Overall	48.83	**64.51**	**55.59**	68.60	58.42	63.10	**23.97**	43.76	**30.22**

Table 1: Model performance of different methods on casual and subevent relations.

tion extraction to enhance the connection between events (Wang et al., 2020; Man et al., 2022; Xu et al., 2022);

Debate Process. The debate process is structured into five distinct stages. (1) *Opening Remarks* are made by the moderator, who introduces the debate rules, debate topic, and positions of the debating sides (affirmative or negative). (2) *Constructive Speech* is delivered by both the affirmative and negative sides, each presenting their stance, approach, and detailed solution process. (3) *Rebuttal Speech* involves the affirmative debater questioning the negative side's approach, and specific solution processes. The negative debater responds to the affirmative's questions and poses their own. Aside from the first debater of both sides, all other debaters participate in this phase. (4) *Free Debate* allows debaters to delve deeper into previous arguments, respond to the opponent's questions, introduce new evidence and examples, and further clarify their positions and arguments. Each side of the debate may speak up to three times. (5) *Conclusion and Summation* involves various debate participants. First, each side of the debate summarizes their stance and discusses whether the opponent's views have contributed to their conclusions. Second, the moderator provides an overview of the debate's overall situation. Third, the audience summarizes their understanding of the debate topic before and after participating in the debate.

We not only focus on improving the performance of event relation extraction but are also deeply interested in the impact of debates on different participants' understanding of the debate topic. Therefore, all participants will provide their event relation extraction results before and after the debate. Detailed debate process is in Appendix A.

3 Experiments

3.1 Datasets

The experiments are conducted on two datasets: MAVEN-ERE (Wang et al., 2022), which focuses on causal, temporal and subevent relations, and EventStoryLine (Caselli and Vossen, 2016),which focuses on causal relations. Following existing work (Guan et al., 2024c), we select 50 documents for each type of relation, collect 605 and 404 instances of causal and subevent relations respectively in MAVEN-ERE, and gather 782 causal relations in EventStoryLine.

3.2 Evaluation Metrics

For evaluation, we employ similar settings to those described by Wang et al. (2024); Guan et al. (2024c), which take into account relation directions and compute sub-relations. For example, in

EventStoryLine, the causal relation includes "PRECONDITION" and "Falling Action". Furthermore, we use the standard Precision (P), Recall (R), and F1-score (F1) as evaluation metrics, following existing research (Wang et al., 2024; Guan et al., 2024c; Cao et al., 2021; Wang et al., 2022).

3.3 Baselines

To rigorously and comprehensively evaluate our proposed method, we design three types of baselines. (1) We first evaluate the performance of basic LLMs. *Whole Document* refers to predicting event relations based on the entire document. *Target Sentence Pair* refers to using only the sentences containing the target events for prediction. (2) We select two widely used single-agent methods: CoT and Self-Refinement. (3) We select two multi-agent methods: Exchange of Thought (EoT) (Yin et al., 2023) and Multi-Agent Debate (MAD) (Wang et al., 2023; Liang et al., 2023).

3.4 Main Results

The overall results on MAVEN-ERE and EventStoryLine are shown in Table 1. Our MMD-ERE equipped with debate topics for different participant groups achieves the best performance. We also have the following three key observations.

(1) For basic LLMs, using the entire document directly for prediction yields the lowest results, indicating that extracting event relations remains a challenging task. However, employing single agent based methods like CoT and Self-Refinement improve performance, suggesting that this idea can enhance relation dependency modeling.

(2) Comparing single-agent and multi-agent results, EoT and MAD achieve better results through communication between different agents.

(3) Regarding debate participants, it is evident that through debates, different participants such as AFF, NEG, Moderator, and Audience, all achieve relatively good performance, demonstrating that MMD-ERE can effectively help capture the relations between events. Unless otherwise specified, the experiments are conducted on GPT-4o.

3.5 Analysis on Debate Participants

In this section, we conduct experiments to assess the impact of different participants, specifically analyzing their understanding of event relations before and after the debate. We experiment with four types of participants: affirmative, negative, moderator, and audience. For responses before the debate, after the AFF 1 and NEG 1 present their problem-solving approaches, participants provide their extraction results. For responses after the debate, following the Conclusion and Summarization stage where participants summarize the debate, they provide their extraction results. The results, shown in Figure 3, indicate that performance improves through participation in the debate, proving the effectiveness of our method.

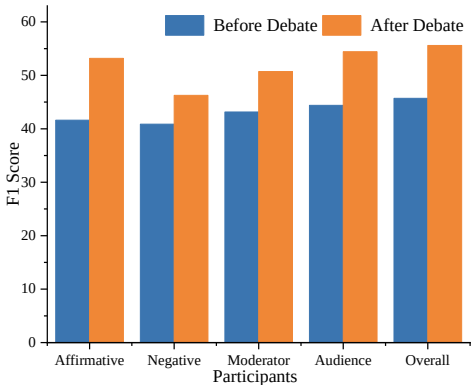

Figure 3: Performance of different debate participants before and after debate.

Methods	P	R	F1
MMD-ERE	48.83	64.51	55.59
w/o Audience	46.98	69.43	54.72
w/o Moderator	48.30	58.33	52.45
w/o Debate Topic	45.66	51.27	50.63

Table 2: Ablation Study.

3.6 Ablation Study

To more specifically validate the different modules within the MMD-ERE, we conduct experiments to ablate the audience (w/o Audience), moderator (w/o Moderator), and debate topic (w/o Debate Topic). According to the results shown in Table 2, we observe that: (1) the removal of any module leads to a decrease in model performance; (2) the most significant decrease occurs when the debate topic is removed, suggesting that providing methodological approaches for participants with different stances facilitates better communication and enhances relation understanding.

3.7 Analysis on Different LLMs

In addition to GPT-4o, we validate our approach on other strong LLMs, including closed-source models such as GPT-3.5 and GLM-4, and open-source models like LLama and Mistral. The experimental

Methods	P	R	F1
GPT-4o	48.83	64.51	55.59
GPT-3.5	46.80	62.73	51.35
GLM-4	47.74	46.49	47.11
LLama-3.1-8B	36.75	40.07	38.64
Mistral-7B	37.22	35.97	36.62

Table 3: Performance on different LLMs.

results, as shown in Figure 3, indicate that GPT-4o achieves the best performance. Compared to open-source models, closed-source models generally show higher results. Overall, there is significant room for improvement in LLMs, indicating that document-level event relation extraction remains challenging.

4 Related Work

4.1 Event Relation Extraction

Event Relation Extraction is a crucial task in NLP that focuses on identifying and classifying relations between events within texts. Research in ERE has evolved from traditional machine learning approaches to neural network methods, which are better equipped to handle the inherent complexities and variabilities of language. Currently, LLMs have demonstrated significant performance across various NLP tasks, showcasing robust understanding and generation capabilities. Many studies have begun to explore the performance of LLMs in ERE (Zheng et al., 2024; Yuan et al., 2023; Gao et al., 2023), primarily employing methods like chain-of-thought (Wei et al., 2022) and self-refinement (Madaan et al., 2023). Chain-of-thought approaches enable LLMs to sequentially output intermediate processes step-by-step. Self-refinement methods involve LLMs iteratively reviewing their outputs until they reach the final results.

4.2 Multi-Agent Debate

With the development of LLMs technology, it has become common to assign role information to LLMs as agents to solve problems (Xi et al., 2023). This approach has rapidly expanded to multiple agents, enhancing the reasoning capabilities of single-agent prompting methods. Currently, two main types of interaction are used: cooperation (Guan et al., 2024a; Chan et al., 2023) and confrontation (Wang et al., 2023). For cooperation, Yin et al. (2023) proposed a multi-agent debate (MAD) framework where, in the first round, each agent provides an initial result, and in subsequent rounds, results are shared among agents to integrate and refine the final outcome. To better facilitate the sharing of information, Yin et al. (2023) introduced four methods of information exchange: memory, report, relay, and debate. For confrontation, Liang et al. (2023) divides agents into two sides: the affirmative side and the negative side. The affirmative side continuously refutes the arguments of the negative side, and ultimately a judge agent generates the final result. In contrast, in this paper, we draw from teaching scenario, considering both cooperation and confrontation. We analyze the performance of different participants before and after the debate.

5 Conclusion

In this paper, we introduce MMD-ERE, a multi-agent multi-sided debate approach for event relation extraction. For integrating multi-agent into ERE, we choose different ERE methods for agents with various stances, which enables agents to capture insights at the methodological level during interactions rather than focusing solely on details. For multi-agent interaction, we propose a multi-sided debate that debaters are divided into multiple groups, each assigned unique debate topic, and the process integrates both communication and confrontation. Extensive experiments are conducted on on two widely used datasets, MAVEN-ERE and EventStoryLine. Experimental results show that debate strategies can enhance the understanding of event relations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China Youth Project (62406162), Natural Science Foundation of China (62476150), Beijing Natural Science Foundation (L243006), and Zhipu AI.

Limitations

In this section, we discuss the limitations of MMD-ERE. First, we mainly focus on text-based data, and exploring the performance of multi-agent systems on multi-modal data is an interesting research point. Second, since our method relies on structured event knowledge, manually constructing such knowledge is time-consuming and labor-intensive, so investigating how to automatically construct structured knowledge using large language models is another research direction.

References

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the ACL*, pages 4862–4872.

Tommaso Caselli and Piek Vossen. 2016. The storyline annotation and representation scheme (StaR): A proposal. In Proceedings of the 2nd Workshop on Computing News Storylines.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the COLING*, pages 2118–2128.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *Preprint*, arXiv:2303.03836.

Yong Guan, Dingxiao Liu, Jinchen Ma, Hao Peng, Xiaozhi Wang, Lei Hou, and Ru Li. 2024a. Event gdr: Event-centric generative document retrieval. In *Companion Proceedings of the WWW*, page 975–978.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2024b. Openep: Open-ended future event prediction. *Preprint*, arXiv:2408.06578.

Yong Guan, Xiaozhi Wang, Lei Hou, Juanzi Li, Jeff Z. Pan, Jiaoyan Chen, and Freddy Lecue. 2024c. Taco-ERE: Cluster-aware compression for event relation extraction. In *Proceedings of the LREC-COLING*, pages 15511–15521.

Yiduo Guo, Yaobo Liang, Chenfei Wu, Wenshan Wu, Dongyan Zhao, and Nan Duan. 2023. Learning to plan with natural language. *Preprint*, arXiv:2304.10464.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. *In Proceedings of the AAAI*, 36(10):11058–11066.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the ACL*, pages 732–742.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the NAACL*, pages 3480–3490.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *arXiv preprint arXiv:2312.04854*.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the EMNLP*, pages 696–706.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the EMNLP*, pages 926–941.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of the ACL*, pages 4072–4091.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Neurips*, volume 35, pages 24824–24837.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. 2022. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the NAACL*, pages 2920–2929.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the EMNLP*, pages 15135–15153.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102.

Yifan Zheng, Wenjun Ke, Qi Liu, Yuting Yang, Ruizhuo Zhao, Dacheng Feng, Jianwei Zhang, and Zhi Fang. 2024. Making llms as fine-grained relation extraction data augmentor. In *Proceedings of the IJCAI*, pages 6660–6668. Main Track.

A Details of Debate Process

The debaters in MMD-ERE consist of two sides: the affirmative (AFF) and the negative (NEG) side. The affirmative side includes three debaters: AFF 1, AFF 2, and AFF 3; similarly, the negative side comprises three debaters: NEG 1, NEG 2, and NEG 3.

During the Constructive Speech phase, AFF 1 and NEG 1 present their respective teams' problem-solving approaches. In the Rebuttal Speech phase, AFF 2 questions the negative side's approach, NEG 2 responds to AFF 2's questions, and subsequently questions the affirmative side's approach. This continues until all participants except for AFF 1 and NEG 1 have engaged. In the Free Debate phase, members from both sides actively discuss each other's ideas and results, with each side allowed to speak up to three times. It should be noted that this entire process requires no human intervention.