Efficient First-Order Logic-Based Method for Enhancing Logical Reasoning Capabilities of LLMs

Wanzhen Fu
University of California, Santa Barbara
wanzhen_fu@ucsb.edu

Haocheng Yang
National University of Singapore
haocheng_yang@u.nus.edu

Fengxiang Cheng*
University of Amsterdam
f.cheng@uva.nl

Fenrong Liu*
Tsinghua University
fenrong@tsinghua.edu.cn

Abstract

Large language models (LLMs) struggle with complex logical reasoning. Previous work has primarily explored single-agent methods, with their performance remaining fundamentally limited by the capabilities of a single model. To our knowledge, this paper is the first to introduce a multi-agent approach specifically to enhance the logical reasoning abilities of LLMs. Considering the prohibitive communication and token costs of multi-turn interactions, we propose an adaptive sparse communication strategy to ensure efficiency. Specifically, our method prunes unnecessary communication by assessing agent confidence and information gain, allowing each agent to selectively update its memory with other agents' most valuable outputs to help generate answers. Extensive experiments demonstrate that our sparse communication approach outperforms fully connected communication while reducing token costs by 25%, improving both effectiveness and efficiency.

1 Introduction

Large language models (LLMs) have demonstrated exceptional capabilities across a wide range of tasks. However, they still face significant challenges when performing complex logical reasoning, limiting their applicability in real-world scenarios [Cheng et al., 2025]. Previous methods for improving logical question answering (QA) of LLMs can be broadly divided into three categories: external solver-based [Ye et al., 2023, Ryu et al., 2025], prompt-based [Xu et al., 2024, 2025], and fine-tuning methods [Morishita et al., 2024, Wan et al., 2024]. Nonetheless, to the best of our knowledge, existing approaches are all benefit from a single pretrained LLM, which still struggles with more complex reasoning tasks due to the heavy reliance on its reasoning capabilities.

Multi-Agent Debate (MAD) has emerged as a promising paradigm to overcome single-agent limitations through collaborative refinement and error correction [Du et al., 2023, Chan et al., 2024, Khan et al., 2024]. However, the standard all-play-all communication system in MAD incurs **high multi-round interaction costs**, especially as the number of agents or debate rounds increases [Li et al., 2024a, Sun et al., 2025]. Thus, it is necessary to develop a sparse multi-round interaction strategy to reduce token costs while preserving superior LLM logical reasoning performance.

To fill this gap, this paper introduces an adaptive sparse multi-agent debate approach, which dynamically prunes unnecessary communication paths in each debate round based on a preference score, which is computed from the agents' confidence ratio and the information gains-quantified by the

*Fengxiang Cheng and Fenrong Liu are the corresponding authors.

semantic dissimilarity-from the output of a different LLM. Communication is permitted only when this score exceeds an adaptive threshold based on the historical average of interaction quality. When performing the communication, each LLM selectively maintain its memory containing others' most beneficial outputs and generate the response using its current memory. Our experiments demonstrate that our approaches achieve state-of-the-art performance on GPT-4 and Claude 3.7 on three datasets, and the proposed sparse interaction approach reduces the total token count by 25% compared with the full interaction approach, improving both effectiveness and efficiency. Our main contributions are:

- To the best of our knowledge, this is the first work to introduce a multi-agent approach to enhance the logical reasoning capabilities of LLMs.
- We design an adaptive sparse debate algorithm that prunes agent interactions based on confidence and information gains, achieving a significant improvement in computational efficiency.
- We provide empirical evidence showing that our approaches achieves state-of-the-art performance with reduced token costs compared with fully-connected interactions.

2 Related Work

Logical Question Answering. Research on logical question answering aims to strengthen the reasoning ability of LLMs and encompasses three primary paradigms of solver-based, fine-tuning, and prompt-based methods [Cheng et al., 2025]. Solver-based methods transform natural language (NL) questions into symbolic language (SL) expressions before employing specialized solvers for inference [Lyu et al., 2023, Olausson et al., 2023, Ye et al., 2023, Ryu et al., 2025]. Fine-tuning approaches pursue dual strategies by constructing synthetic datasets with explicit logical reasoning processes while also augmenting training corpora with structured logical knowledge that embeds reasoning capabilities directly into model parameters [Feng et al., 2024, Morishita et al., 2024, Wan et al., 2024]. Prompt-based methods explore complementary strategies where some approaches generate explicit reasoning chains to guide inference [Wei et al., 2022, Yao et al., 2023, Besta et al., 2024, Zhang et al., 2023, 2024] while others prompt models to produce symbolic forms for step-wise reasoning and verification [Li et al., 2024b, Wang et al., 2024, Xu et al., 2024, 2025, Liu et al., 2025]. So far, all prior works have focused on single-agent methods. Our work pioneers the use of Multi-Agent Debate (MAD) for logical reasoning in LLMs, addressing current limitations, such as information loss from logical expressions and logical errors that arise from an over-reliance on natural language.

Multi-Agent Interaction in LLMs. Multi-Agent Interaction enables multiple LLM agents to collaboratively solve complex tasks. Within this domain, Multi-Agent Debate (MAD) [Du et al., 2023] facilitates iterative debate rounds among agents, improving responses through collaborative refinement. Work on agent roles explores distinct reasoning modes and functional roles such as proposer, critic, planner, and executor, which increase diversity and reliability [Li et al., 2023, Park et al., 2023, Liang et al., 2024]. Debate with an independent judge improves truthfulness and stability across tasks [Du et al., 2023, Chan et al., 2024, Estornell and Liu, 2024, Khan et al., 2024]. Collaboration across heterogeneous models seeks stronger consensus through aggregation, and Reconcile adds confidence-weighted voting to integrate opinions [Chen et al., 2024, Wang et al., 2025]. To reduce cost, SparseMAD prunes the communication topology using a static sparse graph where agents read fixed neighbors, cutting messages [Li et al., 2024a], while CortexDebate builds a sparse debate graph with equal participation and learns edge weights with the McKinsey Trust Formula [Sun et al., 2025]. Although these works attempt to address MAD's efficiency deficit, they still have limited reasoning ability. Our method uses a sparse communication topology and, to our knowledge, is the first to focus on logical reasoning tasks in multi-agent debate. We prune edges by balancing each agent's confidence and the novelty of its information, which enhances efficiency and reasoning reliability while preserving accuracy and self-correction.

3 Logical Question Answering Problem Setup

Logical question answering (QA) task aims to decide whether a statement can be logically deduced from the given information. The LLM is expected to determine whether the specific statement is *true, false,* or *unknown*. The following shows an example from ProofWriter [Tafjord et al., 2021]:

Premises:
The bear chases the squirrel. The bear is not cold. The bear visits the cat. The bear visits the lion. The cat needs the squirrel. The lion needs the cat. The squirrel needs the lion. If something visits the lion then it visits the squirrel. If something chases the cat then the cat visits the lion.
Rules:
- If something visits the squirrel and it needs the lion then the lion does not chase the bear.
- If something is round and it visits the lion then the lion is not cold.
- If something visits the squirrel then it chases the cat.
- If the cat does not chase the bear then the cat visits the bear.
- If something visits the squirrel then it is not nice.
- If the bear is big then the bear visits the squirrel.

Question: Based on the above information, is the following statement true, false, or unknown? The squirrel does not need the lion.
Options: A) True B) False C) Unknown
Answer: B

Existing work achieves only around 80% accuracy on ProofWriter [Xu et al., 2025], demonstrating that LLMs still face significant challenges in reasoning abilities especially on the logical QA tasks.

4 Proposed Method

We introduce a sparse multi-agent debate framework for enhancing the logical reasoning of LLMs, which operates in four main stages. First, we translate the natural language logical question into a formal symbolic representation. Second, we engage multiple LLM agents in a multi-turn debate, where communication between agents is dynamically pruned based on a preference score. This metric assesses the potential benefit of an interaction between two LLMs in each turn by jointly considering the relative confidence of the agents and the information gains from the opponents. Third, each agent selectively updates its memory in each turn, incorporating only the most beneficial information in each debate turn. Finally, after all the debate rounds, a majority vote is taken on the agents' latest conclusions to produce the final answer. This entire process is detailed in Algorithm 1.

4.1 Symbolic Translation of Logical QA

To anchor the reasoning process in a structured and unambiguous format, we begin by converting the raw natural language question Q into a formal symbolic expression, denoted as $\mathrm{Sym(Q)}$. We prompt a pre-trained LLM in a one-shot setting to translate the input text into the First-Order Logic (FOL) representation, including predicates, premises, and a conclusion. For instance, the example provided in the problem setup would be translated into its formal symbolic equivalent: $Chases(bear, squirrel)$, $Cold(bear)$, $\forall x(Visits(x, lion) \rightarrow Visits(x, squirrel))$ ... This symbolic form serves as the common ground for all agents throughout the subsequent debate.

4.2 Multi-Turn Dynamic Interaction Preference Between LLMs

We establish a sparse communication topology to improve the efficiency in multi-turn interactions through a dynamic pruning mechanism, which allows source agent i to communicate its output to the receiving agent j at round d. Specifically, we propose a preference score quantifying the potential utility of the information in the communication, which is defined as:

$$\mathrm{Pre}^d_{i\to j} = \frac{C^d_i}{C^d_j} + \lambda(1 - \cos(A^d_j, A^d_i || A^d_j)).$$

This score comprises two key components. The first is C^d_i / C^d_j, representing the ratio of confidence scores between the source agent i and the receiving agent j at round d. The second is $1 - \cos(A^d_j, A^d_i)$, measuring the difference between the two outputs, regarded as information gain.

Algorithm 1: Multi-Turn Interaction Algorithm for Enhancing LLMs' Logical Reasoning

Input: Communication rounds D, Agent number n, hyperparameter λ;

Translate raw logical question Q to symbolic expression $\mathrm{Sym}(Q)$;

$M_1^{d=1}, \ldots, M_n^{d=1} \leftarrow \varnothing$;

for $d \in \{1, \ldots, D\}$ **do**

  $O_{i \to j}^d = 1$ for all $i, j \in \{1, \ldots, n\}$;

  Compute $\mathrm{Pre}_{i \to j}^d = \frac{C_i^d}{C_j^d} + \lambda(1 - \cos(A_j^d, A_i^d))$ for all $i \neq j$;

  Compute $\overline{\mathrm{Pre}_{i \to j}^d} = \frac{1}{d}(\overline{\mathrm{Pre}_{i \to j}^{d-1}} \cdot (d-1) + \frac{C_i^d}{C_j^d} + \lambda(1 - \cos(A_j^d, A_i^d)))$ for all $i \neq j$;

  if $\mathrm{Pre}_{i \to j}^d < \alpha \cdot \overline{\mathrm{Pre}_{i \to j}^{d-1}}$ **then**

    $O_{i \to j}^d = 0$;

  for $s \in \{1, \ldots, n\}$ **do**

    `// Memory update of the` s`-th agent at round` d

    $M_s^{d+1} \leftarrow M_s^d \cup \{A_i^d \mid i \in \{1, \ldots, n\}, O_{i \to s}^d = 1\}$;

    `// Output of the` s`-th agent at round` d `using personalized memory`

    $A_s^{d+1} \leftarrow \mathrm{LLM}_s(\mathrm{Sym}(Q) || M_s^{d+1})$;

Majority vote among the n agents $A_1^{D+1}, \ldots, A_n^{D+1}$;

To guarantee efficiency, we propose a dynamic strategy to determine with which agent to communicate. Specifically, in round d, we use this average preference score $\overline{\mathrm{Pre}_{i \to j}^{d-1}}$ as the adaptive threshold. We define a binary communication gate $O_{i \to j}^d$. Communication from i to j is permitted only if the current preference score is greater than or equal to the historical average, indicating that the current interaction is at least as beneficial as the average past interaction between this pair. The indicator of whether agent i benefits agent j at round d is formally defined as:

$$O_{i \to j}^d = \begin{cases} 1, & \mathrm{Pre}_{i \to j}^d \geq \alpha \cdot \overline{\mathrm{Pre}_{i \to j}^{d-1}} \\ 0, & \mathrm{Pre}_{i \to j}^d < \alpha \cdot \overline{\mathrm{Pre}_{i \to j}^{d-1}} \end{cases}.$$

4.3 Multi-Turn Interaction Algorithm for Enhancing LLMs' Reasoning

The sparse communication mechanism directly informs how each agent updates its internal state or memory across debate rounds. Each agent maintains a personalized memory that aggregates valuable insights from others. At the beginning of the first round ($d = 1$), all agents start with an empty memory $M_s^1 \leftarrow \varnothing$ and communication is fully connected ($O_{i \to j}^d = 1$ for all pairs). From the second round, the sparse communication gate $O_{i \to j}^d$ is activated. At the end of each round d, every agent s updates its memory for the next round M_s^{d+1} by selectively incorporating the outputs A_i^d from only those agents i for which the communication channel was open (i.e., $O_{i \to j}^d = 1$). After the memory is updated, agent s generates its output for the next round A_i^{d+1}, by querying the symbolic question and i's newly updated, personalized memory. After D rounds of debate, the final outputs from all agents $A_1^{D+1}, \ldots, A_n^{D+1}$, are aggregated via a majority vote to determine the final answer.

5 Experiments

5.1 Experimental Setup

We conduct experiments on GPT-4 and Claude 3.7 Sonnet on three logic reasoning benchmarks: ProntoQA for basic logical reasoning, ProofWriter for multi-step proof generation, and LogicalDeduction for complex deductive reasoning. We compare against seven methods (LogicLM [Pan et al., 2023], LINC [Olausson et al., 2023], one-shot COT [Wei et al., 2022], Aristotle [Xu et al., 2025], SymCOT [Xu et al., 2024], CR [Zhang et al., 2023], and DetermLR [Sun et al., 2024]). Evaluation metrics include reasoning accuracy and computational efficiency, measured by prefill tokens per question and sparse rate—the proportion of directed communications pruned.

Table 1: Performance comparison on GPT-4 and Claude 3.7 under three datasets.

Methods	GPT-4				Claude 3.7			
	ProntoQA	ProofWriter	LogiDeduction	Avg.	ProntoQA	ProofWriter	LogiDeduction	Avg.
LogicLM	93.40%	79.17%	87.00%	86.52%	91.80%	76.17%	94.00%	87.32%
LINC	90.40%	80.67%	82.33%	84.47%	91.20%	83.83%	87.67%	87.57%
1-shot COT	81.20%	67.17%	69.67%	72.68%	87.20%	81.50%	82.33%	83.68%
Aristotle	94.60%	78.00%	65.67%	79.42%	98.20%	83.67%	75.33%	85.73%
SymCOT	96.00%	73.83%	86.33%	85.39%	97.40%	87.33%	92.00%	92.24%
CR	93.20%	71.67%	80.33%	81.73%	96.80%	82.83%	86.67%	88.77%
DetermLR	97.80%	77.33%	85.00%	86.71%	98.00%	84.33%	88.33%	90.22%
Ours (full)	98.20%	81.33%	92.67%	90.73%	**100%**	92.50%	96.33%	96.28%
Ours (sparse)	**99.80%**	**82.17%**	**93.00%**	**91.66%**	**100%**	**93.17%**	**98.00%**	**97.06%**

Table 2: Pre-filling token costs per question and communication sparsity.

Model	Our Methods	ProofWriter		ProntoQA		LogicalDeduction	
		Tokens	Sparsity	Tokens	Sparsity	Tokens	Sparsity
GPT-4	full interaction	26,221.5	100%	22,345.7	100%	27,576.2	100%
	sparse interaction	**22,160.1**	**50.24%**	**19,031.4**	**47.32%**	**25,242.3**	**48.07%**
Claude 3.7	full interaction	28,317.6	100%	21,817.2	100%	33,424.7	100%
	sparse interaction	**23,952.6**	**50.41%**	**18,744.3**	**49.26%**	**29,212.8**	**49.89%**

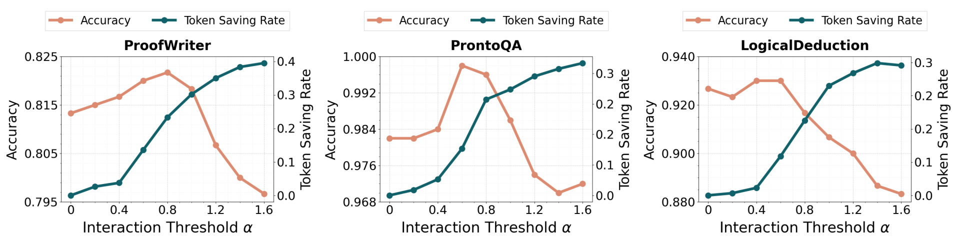

Figure 1: Effect of communication gating threshold on accuracy and token saving rate.

5.2 Results Analysis

As shown in Table 1, our method with full interaction consistently outperforms all baselines, achieving 90.73% average accuracy on GPT-4 and 96.28% on Claude 3.7. Interestingly, our method with sparse interaction achieves 91.66% average accuracy on GPT-4 and 97.06% on Claude 3.7, which are even better than the full interaction method. Table 2 demonstrates that sparse interaction consistently prunes approximately 50% of potential inter-agent communications across both models and all three reasoning tasks, with only around 50% of messages retained. This result underscores our sparse communication strategy's capacity to yield significant token reductions while maintaining performance across diverse reasoning tasks for different LLMs. Figure 1 illustrates the trade-off between accuracy and computational efficiency. Remarkably, at lower threshold values, accuracy improves with increased communication sparsity, indicating that redundant information may harm both accuracy and efficiency.

6 Conclusion

Multi-agent debate in LLMs remains constrained by reasoning limitations and high computational costs. We address this by translating logical QA into symbolic forms and running multi-turn agents' debates with an adaptive sparse gate that balances agent confidence and information novelty. In our method, LLM agents update their memory only when peers prove helpful (via a running-average threshold), and the final answer comes from a majority vote. Across three benchmarks, our sparse debate strategy establishes new state-of-the-art accuracy while pruning about 50% of communications and reducing token usage, consistently surpassing strong single-agent and dense-debate baselines. Future work will focus on extending the sparse mechanism to harder compositional reasoning tasks and exploring softer pruning approaches to further improve both effectiveness and efficiency.

Acknowledgments

FL was supported by the Beijing Natural Science Foundation (No. L257007) and Tsinghua University's Initiative for Advancing First-Class and World-Leading Disciplines in the Humanities and Social Sciences.

References

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. *International Joint Conference on Artificial Intelligence, Survey Track*, 2025.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 36:45548–45580, 2023.

Hyun Ryu, Gyeongman Kim, Hyemin S Lee, and Eunho Yang. Divide and translate: Compositional first-order logic translation and verification for complex logical reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2025.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37:73572–73604, 2024.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael R Lyu. Logicasker: Evaluating and improving the logical reasoning ability of large language models. In *EMNLP*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *Proceedings of Machine Learning Research*, 235: 23662–23733, 2024.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, 2024a.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9503–9523, 2025.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, 2023.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. Language models can be deductive solvers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4026–4042, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*, 2023.

Yifan Zhang, Yang Yuan, and Andrew Chi-Chih Yao. On the diagram of thought. *arXiv preprint arXiv:2409.10038*, 2024.

Qingchuan Li, Jiatong Li, Tongxuan Liu, Yuting Zeng, Mingyue Cheng, Weizhe Huang, and Qi Liu. Leveraging llms for hypothetical deduction in logical inference: A neuro-symbolic approach. *arXiv preprint arXiv:2410.21779*, 2024b.

Zhongsheng Wang, Jiamou Liu, Qiming Bao, Hongfei Rong, and Jingfeng Zhang. Chatlogic: Integrating logic programming with large language models for multi-step reasoning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Yuting Zeng, Jiaxing Wang, Xingyu Wang, Hailong Yang, and Jing Li. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, 2024.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, 2024.

Junlin Wang, WANG Jue, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, 2021.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, 2023.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. DetermLR: Augmenting LLM-based logical reasoning from indeterminacy to determinacy. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.531.