Multi-Agent Debate for LLM Evaluation: A Survey (Proposal)

Brendan Hy, Ketan Joshi, Pranshu Prakash Patel,
Jeongsik Park, Manav Sanghvi, Dengheng Shi*,
University of Southern California
{hybrenda, kajoshi, pranshup, jeongsik, msanghvi, dengheng}@usc.edu

Abstract

While large language models (LLMs) have revolutionized text generation, there is still no system that robustly evaluates output quality without relying on human evaluation. In light of this, the Multi-Agent Debate Evaluation (MADE) framework, where LLMs debate to produce a final judgment on AI-generated data, has attracted interest. We survey this emerging area of research by first introducing a comprehensive analysis along three dimensions: dataset, experiment, and evaluation metrics. Next, we present results on a common dataset using state-of-the-art MADE framework, and evaluate the robustness of existing approaches across different data domains. Finally, we analyze the strengths and weaknesses of each MADE framework, highlight key challenges, and provide recommendations for future work.

1 Related Work

In Figure 1, we illustrate the MADE framework, where multiple LLM agents debate over AI-generated text to produce a final judgment that is often more reliable than a single-judge evaluation. Research on MADE spans diverse domains, each posing distinct evaluation challenges and requirements. We categorize into three domains: (1) open-domain QA and dialogue, (2) the medical domain, and (3) mathematics and scientific reasoning.

1.1 Open-domain QA & Dialogue

Open-domain QA and dialogue tasks pose challenges due to their open-endedness and multidimensional evaluation criteria (Huang et al., 2020). To address this, MADE frameworks aim to improve robustness and reduce bias in this setting.

ChatEval Chan et al. (2023) introduces a MADE framework for evaluating LLM outputs on open-ended QA and dialogue tasks, addressing the limitations of single-judge systems. Using 80 open-domain questions (Zheng et al., 2023b; Wu et al., 2023) and the Topical-Chat dataset (Gopalakrishnan et al., 2023; Mehri and Eskenazi, 2020), the study demonstrates that role diversity and structured debate protocols improve agreement with human judgments.

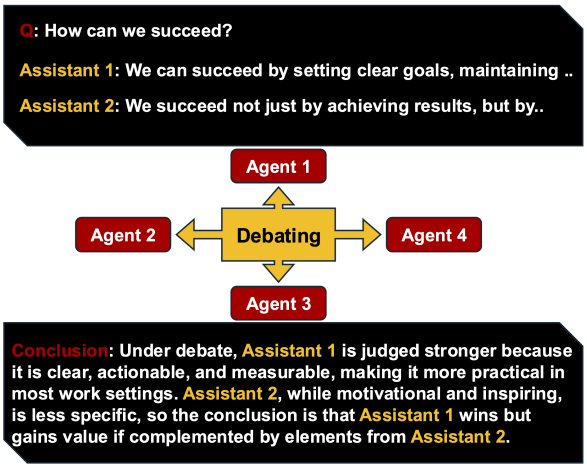

Figure 1: Diagram of MADE: Given AI-generated text, multiple agents debate, leading to a final judgment.

DEBATE Kim et al. (2024) employs a scorer–critic–commander architecture to iteratively refine judgments and mitigate bias. Evaluated on SummEval (Fabbri et al., 2021) and Topical-Chat (Gopalakrishnan et al., 2023), it achieves higher correlations with human ratings compared to surface-level metrics and single-judge baselines, showing the benefit of adversarial critique in MADE.

SAMRE and MORE Bandi and Harrasse (2024) propose a courtroom-inspired evaluation setting with advocates, judges, and juries. This framework applied to MT-Bench (80 questions, 3,300 human-labeled pairs) (Zheng et al., 2023a), demonstrate that adversarial multi-agent debate frameworks significantly outperform single-judge evaluation.

*Authors are listed alphabetically.

1.2 Medical Domain

Medical text evaluation requires factual correctness and evidence-based assessment, making it a critical testbed for MADE frameworks (Min et al., 2023). Studies in this domain demonstrate that MADE can better capture expert-like judgments.

MAJ-EVAL Chen et al. (2025) reduces bias and inconsistency by assigning multiple evaluator personas and aggregating their judgments after debate. Tested on StorySparkQA (Chen et al., 2024) and the MSLR-Cochrane medical summarization dataset (Wang et al., 2023), MAJ-EVAL outperforms both single and multi-agent baselines, achieving closer alignment with expert evaluations, particularly in the medical domain.

1.3 Mathematics & Scientific Reasoning

Reasoning-intensive and technical domains require evaluation frameworks that handle logical validity, accuracy, and specialized criteria (Lu et al., 2024).

M-MAD Zhang et al. (2024) targets machine translation evaluation by aligning with MQM standards (Freitag et al., 2021). Dimension-specific agents debate over accuracy, fluency, style, and terminology, with a meta-judge synthesizing results. Evaluated on WMT23 (English–German, English–Chinese) (Freitag et al., 2023), M-MAD achieves stronger correlations with human ratings than existing LLM-judge baselines.

Debatable Intelligence Sternlicht et al. (2025) benchmarks LLMs on debate speech evaluation, focusing on rhetorical and argumentative quality. Using 600+ annotated debate speeches (Slonim et al., 2021), it shows that while LLMs can approximate human judges on some dimensions, they diverge on subtle, context-dependent aspects, highlighting the difficulty of nuanced reasoning evaluation.

2 Research Questions

In practice, it is unclear which MADE approach provides the most reliable evaluation across domains. Thus, we formulate two research questions (RQs) and hypotheses (Hs):

- **RQ1:** For a given dataset domain, does a specific MADE achieve superior performance?
  H1: On a common dataset, one domain-specific MADE will outperform others.
- **RQ2:** Must an effective MADE be complex to implement or computationally expensive?
  H2: Lightweight, lower-cost MADE configurations can achieve performance comparable to resource-intensive frameworks.

3 Methodology

We design a methodology that enables a systematic comparison of MADE frameworks across diverse domains. Our approach balances empirical rigor with computational feasibility by carefully selecting datasets, replicating representative frameworks, and scaling models according to available resources. We evaluate both quantitative performance and qualitative reasoning traces to provide a comprehensive understanding of each framework's strengths and limitations.

Dataset Selection We select three datasets from different domains aforementioned in Section 1: DiQAD (Zhao et al., 2023), BioASQ (Krithara et al., 2023), and MMATH-Data (Zhang and Xiong, 2025) to capture diverse evaluation challenges. From each dataset, we randomly sample 100 test examples in order to balance representativeness with computational feasibility.

Framework Replication We reproduce MADE frameworks listed in Section 1 and apply them consistently to the selected datasets. Each framework is implemented using the default settings, with necessary adjustments for reproducibility.

Model Scaling To accommodate resource constraints, we substitute larger LLM with smaller when feasible (e.g., LLaMA-70B $\rightarrow$ LLaMA-8B). This strategy also allows us to examine whether the relative performance trends of MADE frameworks hold under reduced computational budgets.

Evaluation and Analysis We report quantitative performance using standard metrics such as accuracy (for pairwise judgments), rank correlation with human labels (Spearman's ρ, Kendall's τ), and agreement measures. Beyond numerical scores, we conduct qualitative analysis of debate logs to study how agents reach their conclusions and assess whether reasoning traces are interpretable.

Discussions Through this pipeline, we compare MADE frameworks along several dimensions: performance, computational cost, ease of implementation, bias resilience, domain robustness, and interpretability of the debate process. This comprehensive evaluation provides both empirical evidence and diagnostic insights into the strengths and weaknesses of current MADE approaches.

References

Chaithanya Bandi and Abir Harrasse. 2024. Adversarial multi-agent evaluation of large language models through iterative debates. *arXiv preprint arXiv:2410.04663*.

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*.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. 2025. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. *arXiv preprint arXiv:2507.21028*.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2024. Storysparkqa: Expert-annotated qa pairs with real-world knowledge for children's story-based learning. *Preprint*, arXiv:2311.09756.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Preprint*, arXiv:2007.12626.

Markus Freitag, David Grangier, and Isaac Caswell. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag and 1 others. 2023. Results of the wmt23 metrics shared task: Metrics evaluation at wmt 2023. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *Preprint*, arXiv:2308.11995.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *Preprint*, arXiv:1905.05709.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: Devil's advocate-based assessment and text evaluation. *arXiv preprint*, arXiv:2405.09935.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *Preprint*, arXiv:2402.16352.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *Preprint*, arXiv:2005.00456.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and 34 others. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Noy Sternlicht, Ariel Gera, Roy Bar-Haim, Tom Hope, and Noam Slonim. 2025. Debatable intelligence: Benchmarking llm judges via debate speech evaluation. *Preprint*, arXiv:2506.05062.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron C. Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14775.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. *Preprint*, arXiv:2303.15078.

Jiayi Zhang, Meng Li, Ziyuan Du, Xingyu Lin, Yi Yang, Weiran Xu, and Chongyang Tao. 2024. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. *arXiv preprint arXiv:2412.20127*.

Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023. Diqad: A benchmark dataset for end-to-end open-domain dialogue assessment. *Preprint*, arXiv:2310.16319.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.