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

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, framework, and evaluation metrics. Next, we present results on datasets across three different domains using four state-of-the-art MADEs Finally, we discuss the strengths and weaknesses of each MADE, highlight key challenges, and provide recommendations for future work.

1 Introduction

LLMs have demonstrated remarkable capabilities in generating coherent and contextually rich text across domains such as reasoning, dialogue, and summarization. However, evaluating the quality and correctness of LLM-generated outputs remains a major challenge. Existing automatic metrics such as BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020) often fail to capture semantic nuance, factual accuracy, or domain-specific reliability, making human evaluation (Gao et al., 2025) a costly, time-consuming, and difficult-to-scale alternative.

To address this limitation, recent research explores **Multi-Agent Debate Evaluation (MADE)**, in which multiple LLMs engage in debate (critiquing, defending, and judging generated texts) to arrive at a final, self-consistent judgment (Oriol et al., 2025). MADE aims to improve evaluation robustness, mitigate single-judge bias, and enhance interpretability by exposing reasoning traces among agents. While several MADE variants have been proposed, it remains unclear which approach provides the most reliable evaluation across different domains and task types. In this study, we conduct a comparative analysis of representative MADEs, examining their performance, and adaptability across diverse datasets. Specifically, we evaluate four representative MADEs over three domains: open-domain QA and dialogue, multi-step mathematical reasoning, and medical question-answering. We also analyze human–MADE alignment to understand whether MADEs replicate human judgment patterns. To guide our analysis, we formulate two research questions:

- **RQ1:** For a given dataset domain, does a specific MADE achieve superior performance?
- **RQ2:** Must an effective MADE be complex to implement or computationally expensive?

Our results provide early empirical evidence that MADE can approximate human judgment across multiple domains while varying substantially in cost, interpretability, and domain robustness. Through this work, we aim to clarify the trade-offs among existing MADEs and establish a

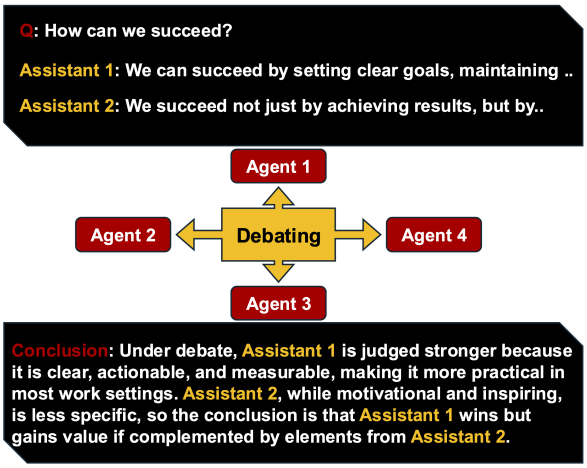

Figure 1: Diagram of MADE: Given AI-generated text, multiple agents debate, leading to a final judgment.

*Authors are listed alphabetically.

MADE	Tasks	Debaters
Deb.-Int.	debate speeches	1 debater & n judges
MORE	multi-turn dialogue	2 advocates & 1 judge
DEBATE	summarization	1 critic & 1 scorer
ChatEval	open-ended QA	1 Public & 1 critic

Table 1: Overview of MADEs. Deb.-Int. denotes Debatable Intelligence.

Dataset	Size	Domain	Fields
MedRedQA	83	Medical	`document`, `input`, `output`
GSM8K	250	Math	`input`, `output`
OpenOrca	250	QA	`system_prompt`, `input`, `output`

Table 2: Statistics of datasets used in our experiments.

foundation for developing more efficient and transparent LLM-as-a-Judge paradigms.

2 Related Work

In Figure 1, we illustrate the MADE, 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. The summary of the MADE we selected is presented in Table 1. We categorize into two domains: (1) mathematics reasoning, and (2) open-domain QA and dialogue.

2.1 Mathematics & Scientific Reasoning

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

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.

2.2 Open-domain QA & Dialogue

This task pose challenges due to their open-endedness and multidimensional evaluation criteria (Huang et al., 2020).

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

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.

ChatEval Chan et al. (2023) introduces a MADE 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.

3 Datasets

From the original proposal, we pivoted our choice of datasets. The initial candidates proved unsuitable because they contained non-English content, required images for multimodal evaluation, or lacked open-ended and explanatory responses. Table 2 summarizes the datasets we finalized.

Our datasets span three domains: (1) open-domain dialogue, (2) multi-step mathematical reasoning, and (3) specialized medical QA, enabling us to evaluate robustness across diverse data distributions. Each debate cycle took approximately four minutes per example, so due to computational constraints, we randomly sampled 250 instances from each dataset for evaluation. The *Fields* column in Table 2 indicates which data fields are included in our final configuration. We also generated two additional LLM-produced outputs for each dataset, which are described in Section 5.[1]

OpenOrca Lian et al. (2023) offers diverse, open-domain QA data with opinion-based queries, allowing us to evaluate MADE's ability to handle open-ended reasoning and conversational judgments.

MedRedQA Nguyen et al. (2023) provides expert-authored medical QA pairs, enabling assessment of MADE's reliability on factual accuracy and safety in specialized domains.

GSM8K Cobbe et al. (2021) focuses on multi-step mathematical reasoning, testing logical consistency and step-by-step problem solving.

[1]Detailed examples are provided in Table **??**.

4 Methods

4.1 Preparation

Data Preparation We construct evaluation data where two LLMs: (1) Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and (2) DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025) generate responses to the same prompts. These model outputs are then assessed by MADEs, in which other LLM agents act as judges to evaluate and compare the generated responses. This allows us to systematically analyze how different MADEs handle model-generated outputs across domains.

MADE Replication We reproduce the MADEs listed in Table 1 and apply them to the selected datasets. Each MADE is implemented with its default configuration, with necessary adjustments to prompts and inputs for our use case.

Model Scaling To accommodate resource constraints, we substitute larger LLMs with smaller ones (e.g., GPT-4→Gemma2:2B) and deploy them locally using Ollama. This allows us to examine whether the relative performance trends of MADEs hold under reduced computational budgets.

4.2 Experimental Design

Model Execution For the preliminary experiments, we used the `gemma2:2b` model (Team et al., 2024) for MADE, executed locally via Ollama to accommodate the computational resource limitations typical of consumer-grade hardware.

Hyperparameters and Prompt Selection We adopted the default prompt style and hyperparameter settings provided by each framework. The prompts were selected to align with the task definitions in the three datasets.

Evaluation Metrics For all datasets, MADEs scored each model's responses on a 1 to 5 Likert scale according to their respective evaluation criteria, focusing on consistency and comparability.[2]

- **OpenOrca:** Prioritize robustness and reduction of bias in open-ended, multidimensional dialogue, addressing the need for reliable judgments in diverse conversational tasks.
- **MedRedQA:** Focus on factual correctness and evidence-based assessment because accuracy and reliability are crucial in the medical domain.
- **GSM8K:** Emphasize logical validity and step-by-step reasoning to ensure models handle complex mathematical and technical challenges.

[2]Detailed Evaluation Criteria is provided in Appendix A.

5 Results and Discussions

5.1 Results

Table 4 demonstrates Debatable-Intelligence and MORE achieved uniformly high Likert scores ("5,5") across most metrics and datasets, while DEBATE consistently scored lower ("3,3"–"4,3"). Human raters showed strong agreement in Math and OpenQA but greater variability in the Med dataset; also, they sometimes assigned Distill-LLaMA answers the lowest possible Likert score ("1"), particularly on Math and OpenQA metrics, whereas the MADEs consistently rated those same answers higher (often "3" or above). ChatEval results were intermediate ("4,4"–"4,5"), corroborating the superior performance and reliability of Debatable Intelligence and MORE.

5.2 Discussions

Our evaluation of MADE approaches across dimensions like performance, cost, and robustness shows their preferences closely mirror human judgments. In Math, MADE favors correct, well-reasoned solutions and, like a human grader, allows ties for equally valid answers. For Medical questions, it prioritizes human-centric criteria such as accuracy and patient safety, preferring clinically-grounded advice over risky suggestions. In OpenQA, it rewards relevance, completeness, and usefulness. Ultimately, MADE does more than assign scores; it emulates the domain-specific rationale humans apply: valuing logical correctness in math, responsible clinical judgment in medicine, and helpfulness in open-domain questions.

6 Future Plans

Our recent efforts, detailed in our GitHub repository, [3] have focused on replicating and adapting the MADEs mentioned above. To establish a robust baseline, we are now implementing a single LLM-as-a-Judge system. We will conduct a side-by-side evaluation of both the MADE and this baseline using identical datasets and metrics. This direct comparison will allow us to validate our replications' fidelity and quantify any performance improvements over the traditional single-LLM judge.

We aim to test the MADE with larger-parameter LLMs to determine if they yield more diverse and robust scores. This expansion will require the use of CARC computational resources to manage the increased hardware load.

[3]https://github.com/lucasjeongsikpark/CSCI544_AppliedNLP_GroupProject/tree/main

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Preprint*, arXiv:2402.01383.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/datasets/Open-Orca/OpenOrca`.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Marc Oriol, Quim Motger, Jordi Marco, and Xavier Franch. 2025. Multi-agent debate strategies to enhance requirements engineering with large language models. *Preprint*, arXiv:2507.05981.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

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.

A Appendix: Evaluation Metric Definitions

Table 3 presents the evaluation metrics and definitions established for each dataset, providing a clear framework for assessing model responses. These definitions were incorporated directly into the tailored dataset (domain) prompts for each MADE, ensuring that all judge agents evaluated answers using criteria matched to the dataset context and experiment goals.

B Experiment Results

Table 4 presents the quantitative evaluation results for each response, dataset, and metric across all MADEs and two human evaluators. The table highlights performance differences among frameworks and between human scorers, providing detailed scores that support our analysis of model response quality and evaluator alignment.

Dataset	Evaluation Metric	Definition
OpenOrca (OpenQA)	Relevance	Addresses the question/task directly
	Completeness	Covers necessary points/details
	Accuracy	Information correctness vs expected answer
	Clarity	Organization and readability
	Helpfulness	Utility for a typical user
MedRedQA (Medical)	Medical Accuracy	Medical soundness of facts
	Appropriateness	Suitability for patient context
	Safety	Emphasis on safe, non-harmful guidance
	Clarity	Understandability and organization
	Professionalism	Tone and ethical/empathetic stance
GSM8K (Math)	Correctness	Correct final answer
	Reasoning	Logical, coherent step-by-step derivation
	Completeness	All necessary steps shown (no large gaps)
	Accuracy	Calculations/arithmetic correct

Table 3: Evaluation metrics and definitions for each dataset

Dataset	Metrics	Ex.	Debate.-Int.	MORE	DEBATE	ChatEval	Human1	Human2
GSM8K	**Correctness**	Q1	**5,5**	**5,5**	**3,3**	4,**5**	**5**,1	**5**,1
		Q2	**5,5**	**5,5**	**4**,3	4,**5**	**5,5**	**5,5**
	Reasoning	Q1	**4,4**	**5**,4	2,**4**	3,**5**	**5**,3	**5**,2
		Q2	**4,4**	**5,5**	2,**4**	**4,4**	**5,5**	**5**,4
	Completeness	Q1	**5,5**	**5,5**	**4**,3	4,**5**	**5**,4	**5**,2
		Q2	**5,5**	**5,5**	**3,3**	**5,5**	**5,5**	**5,5**
	Accuracy	Q1	**5,5**	**5,5**	3,**4**	3,**4**	**5**,3	**5**,1
		Q2	**5,5**	**5,5**	**3,3**	**4,4**	**5,5**	**5,5**
MedRedQA	**Accuracy**	Q1	**4,4**	**5,5**	**4,4**	4,**5**	2,**5**	4,**5**
		Q2	2,**4**	**5**,4	**3,3**	4,**5**	3,**4**	3,**4**
	Appropriateness	Q1	**4,4**	**4,4**	**4**,3	4,**5**	3,**5**	4,**5**
		Q2	4,**5**	**5**,4	3,**4.5**	5,**5**	3,**4**	3,**4**
	Safety	Q1	**4,4**	**4,4**	**4**,3	4,**5**	4,**5**	4,**5**
		Q2	**5,5**	**5**,3	**4**,3	**4,4**	2,**5**	4,**5**
	Clarity	Q1	**5,5**	**4**,3	**3,3**	4,**5**	4,**5**	4,**5**
		Q2	4,**5**	**4**,3	3,**4**	4,**5**	3,**4**	4,**5**
	Professionalism	Q1	**4,4**	**5**,4	**4**,3	4,**5**	**5,5**	**5,5**
		Q2	**5,5**	**5**,4	**3,3**	4,**5**	3,**5**	4,**5**
OpenOrca	**Relevance**	Q1	**5,5**	**5**,4	3,**4**	**5,5**	**5,5**	**5,5**
		Q2	3,**4**	**5**,4	3,**4**	**5**,4	**4**,2	**2**,1
	Completeness	Q1	**4,4**	**4**,3	**4**,3	4,**5**	4,**5**	**5**,4
		Q2	**4**,3	**4**,3	**4,4**	**4**,3	**4**,1	**2**,1
	Accuracy	Q1	**4**,3	**5,5**	**4**,3	4,**5**	**5,5**	**5,5**
		Q2	**3,3**	**4,4**	**3,3**	**4**,3	**5**,1	**1,1**
	Clarity	Q1	**4,4**	**5**,3	**4**,3	4,**5**	**5,5**	4,**5**
		Q2	3,**4**	**5**,4	**4**,3	**3,3**	**5**,3	**4**,3
	Helpfulness	Q1	**4,4**	**5**,4	3,**4**	4,**5**	**5,5**	**5,5**
		Q2	**3,3**	**5**,4	**4**,3	**4,4**	**4**,1	**2**,1

Table 4: Evaluation results across the three datasets and their corresponding evaluation metrics. "Ex." denotes the example index; we evaluated two examples for each dataset. Cells show paired scores *(LLaMA, Distill-LLaMA)* on a 1–5 Likert scale. The values in the 4th–7th columns represent the scores produced by each MADE. *human1* and *human2* indicate our own human evaluation results, used to support and examine the MADEs' performance across datasets and evaluators.