Enhancing Multi-LLM Debate with Uncertainty-Aware Selection for Improved Reasoning.

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown exceptional reasoning capabilities, yet selecting the most reliable response from multiple LLMs remains a challenge, especially in resource-constrained settings. Existing approaches often rely on expensive external verifiers, human evaluators, or self-consistency techniques that require multiple samples from a single model. Multi-LLM debate provides a more interactive mechanism, yet it frequently underperforms compared to self-consistency with the best LLM. In this work, we introduce a log-likelihood-based selection framework to enhance reasoning in multi-LLM debate settings. Our approach leverages uncertainty estimation to identify the most confident response while minimizing inference costs. We demonstrate that our method outperforms majority vote selection and surpasses self-consistency performance for a large number of model calls. Through extensive experiments, we show that multi-LLM collaboration—when guided by uncertainty-aware selection—lead to improvement of 6.8% for settings with less number of model calls.

1 Introduction

Large language models (LLMs) have demonstrated remarkable advances in natural language understanding and reasoning, with state-of-the-art (SOTA) models achieving near-human performance across various tasks. However, given the rapid evolution of LLMs and the prohibitive cost of training new foundation models, a key challenge arises: **How can we effectively combine multiple SOTA LLMs to generate the best possible response without additional training?** This question is especially relevant when multiple LLM APIs are available but computational resources are constrained, limiting the feasibility of expensive external verifiers or reward models.

Existing approaches for selecting the most reliable LLM-generated response typically rely on (1) **external verifiers** (Xi et al., 2024), (2) **LLMs or human judges** (Chan et al., 2023; Khan et al., 2024; Li et al., 2024), or (3) **reward models trained to rank responses**, all of which incur high computational costs. An alternative is self-consistency (Wang et al., 2023), where a single LLM generates multiple samples, and the majority-voted response is selected. Additionally, self-reflection (Renze and Guven, 2024) allows iterative refinement of answers. However, these methods primarily apply to single-LLM settings and require extensive sampling to achieve performance gains.

For multi-LLM systems, a naive approach is to apply self-consistency across all models, selecting the most frequent response. However, this fails to leverage inter-model reasoning. A more sophisticated approach is multi-agent debate, where LLMs iteratively refine responses through interaction (Du et al., 2023). Yet, prior works (Huang et al., 2023) have found debate often inferior to self-consistency with the best single LLM, a trend we also observe. Identifying this best LLM a priori for each task remains impractical without costly external verifiers or labeled datasets.

To address this, we propose **enhancing multi-LLM debate through uncertainty estimation**. Specifically, we introduce a log-likelihood-based model selection mechanism that improves answer selection while minimizing LLM calls. Our key contributions are:

1. **Log-Likelihood Based Tiebreak:** We develop a log-likelihood-based selection framework that improves multi-LLM debate performance, surpassing self-consistency for large sample sizes (N), demonstrating its cost-effectiveness and robustness. Our method only requires response logits, making it applicable in black-box settings.

2. **Multi-LLM vs. Single LLM Systems:** Our uncertainty-aware approach **achieves greater improvements in multi-LLM settings compared to single LLM setups**, reinforcing the importance of multi-LLM systems as an alternative to single-model dominance.

Through extensive experiments, we validate our method's effectiveness, demonstrating cost-efficient, reasoning-aware answer selection that reduces reliance on external evaluators. Our findings advocate for multi-LLM collaboration as a robust alternative to single-model approaches.

2 Related Works

Multi-LLM Systems: Recent work on multi-LLM systems has explored various strategies to enhance performance and evaluation. Mixture-of-Experts (MoE) models, such as Uni-MoE, integrate multiple specialized components within a unified architecture (Li et al., 2025). EnsemW2S (Agrawal et al., 2024) combines diverse LLM's token-level probabilities using Adaboost inspired weighing mechanism to improve generalization on complex reasoning tasks. Other methods include PromptEval, which estimates performance across prompts for robust evaluation (Polo et al., 2024), LLM as judge (Chan et al., 2023; Khan et al., 2024; Li et al., 2024) and debate/discussion among agents (Du et al., 2023). Additionally, research on MoE routing weights suggests their potential as complementary embedding models (Li and Zhou, 2024). These works highlight the growing interest in optimizing multi-LLM collaboration for improved reasoning and evaluation.

Single LLM Self-Improvement: To improve reasoning and mitigate inconsistencies in LLM outputs, recent work has explored feedback-based learning. Self-consistency (Wang et al., 2023) aggregates multiple sampled outputs via majority voting, while confidence-based weighting (Taubenfeld et al., 2025) refines this selection. Tree of Thoughts (ToT) (Yao et al., 2024) enhances self-consistency by structuring reasoning as a tree search. Self-reflection (Renze and Guven, 2024) allows LLMs to iteratively refine responses. However, LLMs remain prone to biases (Khatun and Brown, 2024), underscoring the need for multi-LLM systems to cross-verify answers and enhance reliability.

3 Method

In this section, we describe our approach for combining log-likelihood–based answer selection with multi-LLM debate paradigm introduced by (Du et al., 2023). Figure 1 illustrates an example debate, highlighting how our method selects the correct answer compared to a traditional majority voting strategy.

Consider a set of N language models (LLMs) $\{\pi_1, \pi_2, \ldots, \pi_N\}$ that each generate a response to a given prompt x. Let the response from model π_i be denoted by Y_i. In the first round of the debate, each LLM produces an initial response $Y_i^1 = \{y_1, y_2, \ldots, y_T\}$, where T is the number of tokens. The log-likelihood of this response is computed as

$$\log P\big(Y_i^1 \mid x\big) = \sum_{t=1}^{T} \log P\big(y_t \mid x, y_{<t}\big),$$

where $P\big(y_t \mid x, y_{<t}\big)$ is the probability of token y_t conditioned on the prompt x and the preceding tokens $y_{<t}$ for model π_i.

In subsequent rounds ($K > 1$), the responses are generated conditioned on both the original prompt and all previous rounds of responses from every LLM. Specifically, the Kth-round response from model π_i is given by $Y_i^K = \{y_1, y_2, \ldots, y_T\}$ and its log-likelihood is computed as

$$\log P\Big(Y_i^K \mid x, \{Y_j^k\}_{\substack{\forall j \in [1,N] \\ \forall k < K}}\Big) = \sum_{t=1}^{T} \log P\Big(y_t \mid x, \{Y_j^k\}_{\forall j \in [1,N], \forall k < K}, y_{<t}\Big). \tag{1}$$

For final answer selection, we choose the response with the highest log-likelihood. If a majority answer emerges, we select it directly; otherwise, we use log-likelihood. However, our experiments empirically show that the highest log-likelihood answer almost always aligns with the majority vote.

Intuition behind why its works: In rounds > 1 of multi-LLM debate, log-likelihood is generated for the response which is conditioned on previous answers. Thus, it uses the previous answers to generate its reasoning, and hence in the process gives log-likelihood values which are more comparable even though the responses are generated from different LLMs. Our method does not show

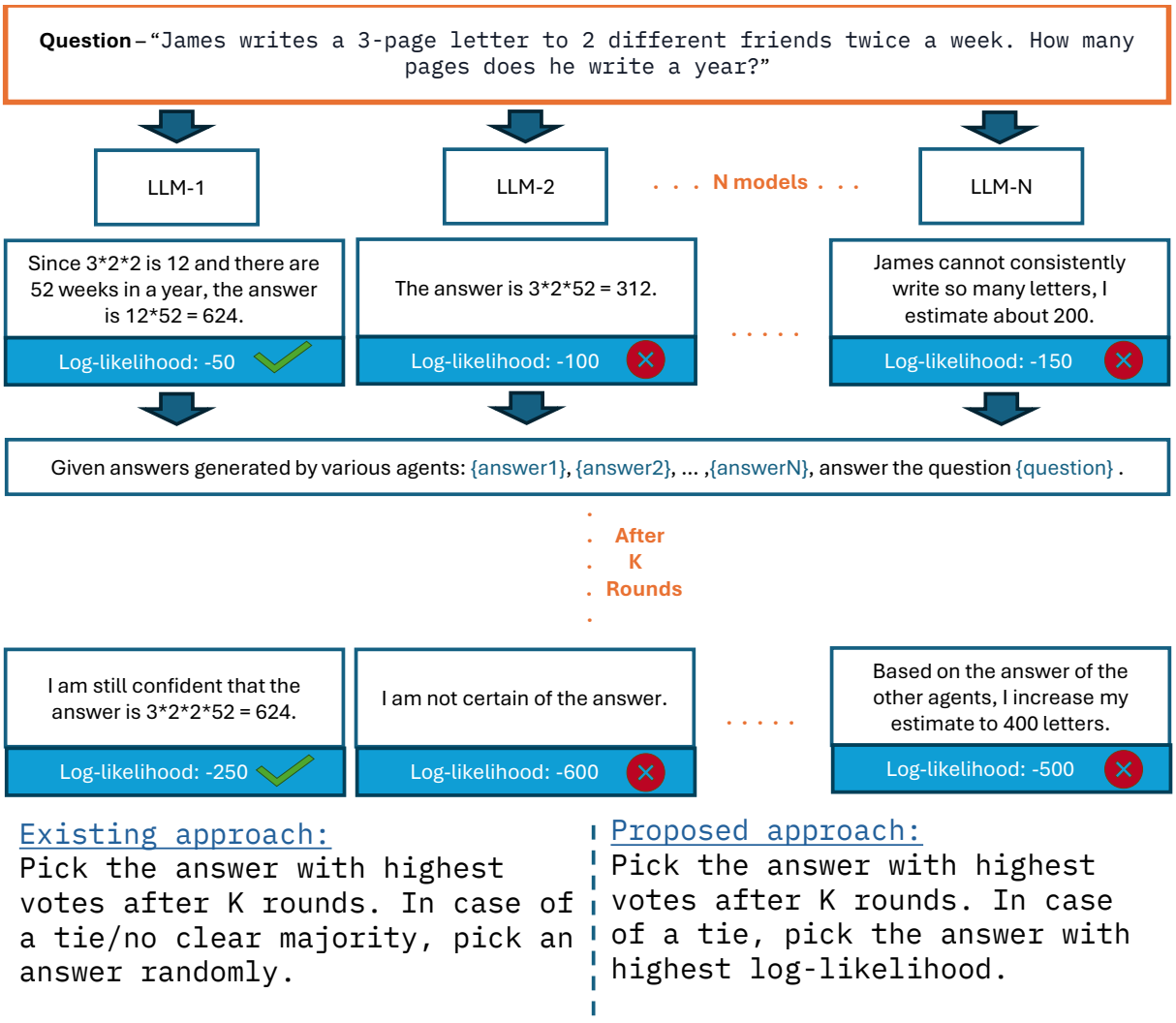

Figure 1: Diagram illustrating our proposed **multi-LLM debate + log-likelihood based selection** which outperforms random sampling in cases of a tiebreak. Each model provides a reasoning chain, and the best answer is chosen based on log-likelihood scoring in the final round of debate.

much improvement for single LLM case since the answers/COT reasoning generated is not diverse, thus the log-likelihoods show lesser variation.

4 Experiments

In this section, we show results for our log-likelihood based selection method over baseline selection. We run experiments with 3 LLM models: Qwen2.5-7B-Instruct, Ministral-8B-Instruct-2410, and Llama-3.1-8B-Instruct. Of these, Qwen2.5-7B-Instruct is the best performing model and hence is used for the single LLM baseline experiments. We use GSM8K (Cobbe et al., 2021) test data to show performance in accuracy (%).

Baselines: The primary baseline for evaluating our log-likelihood-based selection method is random selection. Specifically, when the multi-LLM system does not converge to a single answer, we compare selecting a response at random (baseline) versus using log-likelihood (Ours) for selection. Additionally, we run self-consistency with different values of N={9,18,27} to benchmark against both the best single LLM and multi-LLM self-consistency. This is because prior work (Huang et al., 2023) has shown self-consistency to outperform multi-LLM systems in reasoning tasks, which aligns with our observations and hence making this comparison essential to assess the merit of multi-LLM approaches.

To ensure fair comparisons, we keep the total number of LLM calls constant across all settings. To match self-consistency experiments with larger N, we introduce a modified debate approach where each model is sampled multiple times to reach the same total number of model calls. While in-

Method	Number of Model Calls		
	9	18	27
Self-Consistency: Qwen	82.55 → 82.55 | 92.65	84.41 → 84.41 | 95.59	85.39 → 85.39 | 96.76
Self-Consistency: (Q, L, M) | S: 3	76.18 → 77.16 | 89.51	-	-
Debate: (Q, L, M) | R: 3 | S: 1	80.21 → **85.68** | 90.00	-	-
Debate: (Q, L, M) | R: 2 | S: 3	-	86.08 → 86.08 | 95.69	-
Debate: (Q, L, M) | R: 3 | S: 3	-	-	86.18 → 86.27 | 97.16

Table 1: **Comparison of Self-Consistency and Multi-LLM Debate on GSM8K.** This table presents accuracy results for self-consistency and multi-LLM debate. We evaluate performance across different numbers of model calls (9, 18, and 27) while comparing random vs log-likelihood-based tiebreak resolution. $R \rightarrow$ number of rounds, and $S \rightarrow$ number of samples per model per round. $Q \rightarrow$ Qwen-2.5-8B, $L \rightarrow$ Llama-3.1-8B and $M \rightarrow$ Ministral-8B. For $S > 1$, we ensure that only distinct responses from previous rounds are passed to the next round. Each result is formatted as $a \rightarrow b|c$, where: a represents accuracy with **random tiebreaks**, b represents accuracy with **log-likelihood-based tiebreaks**, c represents the **upper bound accuracy**, which assumes that at least one of the samples answers is correct. We see that debate + log-likelihood-based selection improve accuracy while requiring fewer model calls compared to traditional self-consistency.

creasing debate rounds could achieve similar results, prior studies indicate that debate quickly converges to a single answer, limiting its usefulness for uncertainty-based selection. Therefore, we cap debate rounds at 3.

Majority Selection with and without Log-Likelihood: As shown in Table 1, log-likelihood-based selection method improves accuracy by 6.8% over random selection in multi-LLM debate settings with just 9 total LLM calls. While the gains are smaller for larger N, our primary focus is on demonstrating the effectiveness of log-likelihood selection in low-cost multi-LLM settings with limited model calls.

Test-time cost efficiency with log-likelihood: As shown in Figure 2, our **log-likelihood-based selection consistently outperforms self-consistency** with fewer LLM calls, validating its effectiveness. Across different number of model calls, **multi-LLM debate with log-likelihood selection proves superior to single-LLM self-consistency**, demonstrating that multi-LLM systems can be highly effective when designed to leverage model diversity efficiently.

5 Conclusion

We introduced a log-likelihood-based selection framework to enhance reasoning in multi-LLM debate. By leveraging uncertainty estimation, our method selects the most confident response, reducing reliance on costly external verifiers and extensive self-consistency sampling. Our approach outperforms random selection and majority voting with fewer LLM calls, making it a cost-effective solution. Additionally, we highlight the benefits of diverse model reasoning in multi-LLM debate. Future work can explore adaptive sampling and extend our method to broader reasoning tasks.

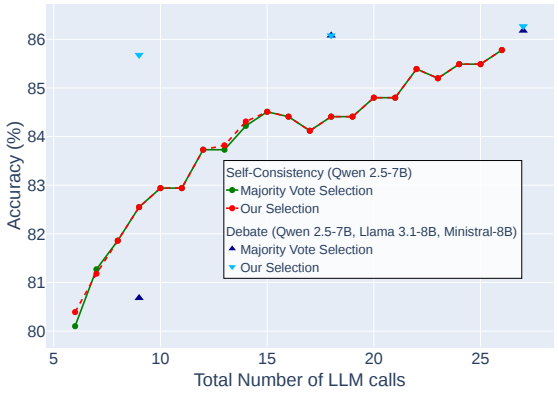

Figure 2: In this plot, we show accuracy (on GSM8K) as a function of number of LLM calls made for both majority vote selection and our method of selection. Majority voting uses random sampling (default) in case of a tie whereas our method uses log-likelihood based selection. For self-consistency on single LLM, we plot the accuracy for model calls ranging from $num\_calls = 6$ to $num\_calls = 27$. **Using log-likelihood based selection for debate, we not only improve performance significantly but we also see that we require a much lower number of LLM calls to achieve the same accuracy as self-consistency.** We evaluate the multi-agent debate setting for $num\_calls = 9, 18, 27$. Another important thing to observe is that both random and log-likelihood based selection converge to same accuracy as the number of LLM calls increases.

6 Limitation

Our method is primarily effective in low-cost settings, where the number of LLM calls is limited. In high-cost settings—where a large number of responses can be generated—the likelihood of a non-majority-voted answer decreases. As a result, the effectiveness of our log-likelihood-based selection in improving performance over random selection diminishes significantly.

Additionally, our approach is specifically designed for multi-LLM settings, where diverse models generate a broader range of responses. This diversity encourages deeper reasoning and exploration of alternative answers, increasing the likelihood of finding and converging on the correct solution. In such scenarios, our selection method is particularly valuable in identifying the most confident response among the generated outputs. However, in single-LLM self-consistency settings, where the responses are inherently less varied, our method may provide limited benefits.

References

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, John Langford, and Furong Huang. 2024. Ensemw2s: Can an ensemble of llms be leveraged to obtain a stronger llm? *arXiv preprint arXiv:2410.04571*.

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, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Aisha Khatun and Daniel G Brown. 2024. A study on large language models' limitations in multiple-choice question answering. *arXiv preprint arXiv:2401.07955*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ziyue Li and Tianyi Zhou. 2024. Your mixture-of-experts llm is secretly an embedding model for free. *arXiv preprint arXiv:2410.10814*.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, and 1 others. 2024. Enhancing llm reasoning via critique models with test-time and training-time supervision. *arXiv preprint arXiv:2411.16579*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.