

Uncertainty-Aware Answer Selection for Improved Reasoning in Multi-LLM Systems

Aakriti Agrawal¹, Rohith Aralikatti², Anirudh Satheesh¹, Souradip Chakraborty¹,
Amrit Singh Bedi³, Furong Huang^{1,4}

¹ University of Maryland, ² Hilabs, ³ University of Central Florida, ⁴ Capital One

Correspondence: agrawal5@umd.edu

Abstract

Large Language Models (LLMs) have demonstrated exceptional capabilities, yet selecting the most reliable response from multiple LLMs remains a challenge, particularly in resource-constrained settings. Existing approaches often depend on costly external verifiers, human evaluators, or self-consistency techniques that require multiple samples from a single model. While multi-LLM systems produce more diverse responses than single models and thus have greater potential, they often underperform compared to single LLM self-consistency. We propose a principled, novel and computationally efficient method to select the best response from multiple different LLMs using a calibrated log-likelihood score, implicitly leveraging the inherent knowledge and confidence of these models. Our method demonstrates improvements of approx. 4%, 3%, and 5% across both debate (multi-round LLM discussions) and non-debate (Best-of-N with multiple LLMs) settings on GSM8K, MMLU (6 subsets), and ARC datasets respectively¹.

1 Introduction

Large language models (LLMs) have achieved remarkable advancements, with state-of-the-art (SOTA) models now approaching or even surpassing human performance. With rapid advancements, an array of pre-trained LLMs and LLM APIs have become readily accessible, each exhibiting distinct strengths across specialized tasks. For instance, models like GPT-deep research excel in analytical and research-oriented tasks, whereas GPT-o3 is tailored towards programming and coding. Similarly, there are domain-specific expert models optimized for tasks in physics, algebra, and other fields.

This proliferation raises an important question: **How can we optimally leverage the diverse capabilities of multiple specialized LLMs to produce**

the most accurate and reliable answers without incurring additional training costs (given the substantial resource demands associated with training)? This challenge becomes especially critical when high-quality external judges or evaluators—capable of assessing answers across multiple specialized domains—are unavailable, expensive, or even infeasible due to the superhuman capabilities of modern LLMs (Burns et al., 2023; Agrawal et al., 2024).

Existing approaches have tried exploring this but (**Challenge 1**) suffer from reliance on external information such as: (1) *external verification models* (Xi et al., 2024), (2) *human or LLM-based judges* (Chan et al., 2023; Khan et al., 2024; Li et al., 2024), or (3) *reward models trained explicitly for response ranking*. These methods involve significant resource overhead and are infeasible when existing models reach superhuman performance.

Additional major challenge is that existing methods from single-LLM literature like majority voting with self-consistency (Wang et al., 2022), self-reflection (Renze and Guven, 2024), or metric-based selection (Kang et al., 2025) (e.g., perplexity, self-certainty) (**Challenge 2**) demand extensive sampling or (**Challenge 3**) are infeasible for multi-LLM contexts due to inherent differences in outputs from a multi-LLM system. (**Challenge 4**) A simplistic multi-LLM adaptation of self-consistency, which selects the most frequently generated answer, does not effectively exploit inter-model reasoning and hence misses potential performance gains (Du et al., 2023).

To address the above challenges, in this paper, we propose a novel and computationally efficient method to aggregate responses from diverse LLMs and systematically select the best answer without relying on external verifiers (Challenge 1), without requiring extensive sampling (Challenge 2), and effectively use diverse multi-LLMs (Challenge 3, 4). Specifically, we introduce

¹Code: <https://github.com/Aakriti05/multi-llm-uncertainty>

uncertainty estimation-based answer selection from multi-LLM systems, which employs a calibrated log-likelihood-based selection metric that implicitly leverages the inherent knowledge and confidence of the given models, further improving response accuracy and reducing computational overhead. Our method is built on the hypothesis that a model (or expert) trained on a specific example will exhibit high confidence (i.e., low uncertainty), while models unfamiliar with the example will show higher uncertainty. Assuming the training data is correct, the most confident expert is likely to provide the correct answer and our aim to find that expert.

Thus, the primary contributions are:

1. We propose a principled calibration technique for aligning log-likelihood based uncertainty scores across different models to select the best answer. This is because directly comparing per-token likelihoods across models is theoretically flawed, as each model has distinct parameters and logit distributions.
2. Our approach is also computationally efficient: by using teacher-forcing, we compute calibrated scores with a single forward pass, avoiding costly autoregressive decoding. It is light weight and simple as it does not require external verifiers, reward models, human or LLM judges.
3. Our method demonstrates strong empirical performance across both debate (multi-round multi-LLM discussions) and non-debate (Best-of-N with multiple LLMs) settings on GSM8K, MMLU (6 subsets), and ARC datasets showing improvement of 4%, 3% and 5% respectively. **It achieves greater improvements in multi-LLM settings compared to single LLM**, further reinforcing the potential of a diverse-answer-generating multi-LLM system to boost overall performance.
4. We also show comparison with other metrics and theoretical justification for our calibrated metric.

2 Uncertainty-Aware Answer Selection

We now describe our approach for aggregating and selecting optimal answers in a multi-LLM system. While methods for optimal answer selection—commonly used in best-of-N strategy—are

well-studied in single-LLM contexts (Kang et al., 2025), directly extending them to multi-LLM environments poses significant challenges. Specifically, diverse LLMs (1) produce outputs varying significantly in format and length, (2) differ widely in their model weights and architectures, and (3) yield uncertainty scores that are typically incomparable across models due to a lack of calibration.

To address these issues, we propose a **calibrated log-likelihood metric**, which integrates smoothly with the interactive multi-LLM debate as well non-interactive multi-LLM best-of-N setup. We first detail our proposed metric and then provide theoretical insights and analysis to illustrate its enhanced performance. We also provide discussion on its computational efficiency. Figure 1 illustrates how our calibrated selection approach surpasses traditional majority-voting strategies for a multi-LLM setup. (Note: A non-debate best-of-N setting is equivalent to first round of debate setting.).

2.1 Calibrated Log-Likelihood Metric

Consider a set of N LLMs $\{\pi_1, \pi_2, \dots, \pi_N\}$ each generating a response to a given prompt x . We denote the response from the i^{th} model π_i as Y_i . In the first debate round, each LLM independently produces a response $Y_i^1 = \{y_1^1, y_2^1, \dots, y_T^1\}$, where T is the number of tokens with log-likelihood computed as:

$$\log P_{\pi_i}(Y_i^1 | x) = \frac{1}{T} \sum_{t=1}^T \log P_{\pi_i}(y_t^i | x, y_{<t}^i),$$

where $P_{\pi_i}(y_t^i | x, y_{<t}^i)$ denotes the conditional probability of token y_t^i given the prompt x and preceding tokens $y_{<t}^i$ for model π_i .

In subsequent rounds ($K > 1$), model π_i 's response Y_i^K depends on both the prompt and previous responses. Denoting previous responses as $X = \{Y_j^k\}_{\forall j \in [1, N], \forall k < K}$, the log-likelihood is then computed as,

$$\log P_{\pi_i}(Y_i^K | x, X) = \frac{1}{T} \sum_{t=1}^T \log P_{\pi_i}(y_{i,t}^K | x, X, y_{<t}^i)$$

$$\text{where, } Y_i^K = (y_{i,1}^K, y_{i,2}^K, \dots, y_{i,T}^K)$$

To select the best final response from $\{Y_i^K\}_{\forall i \in [1, N]}$, we define a calibrated log-likelihood-based scoring mechanism as:

$$Score(Y_j) = \frac{1}{N} \sum_{i=1}^N \log P_{\pi_i}(Y_j^K | x, X) \quad (1)$$

This metric measures consensus among models by averaging response likelihood across all models, providing a calibrated measure of answer quality.

Question – James writes a 3-page letter to 2 different friends twice a week.
How many pages does he write a year?

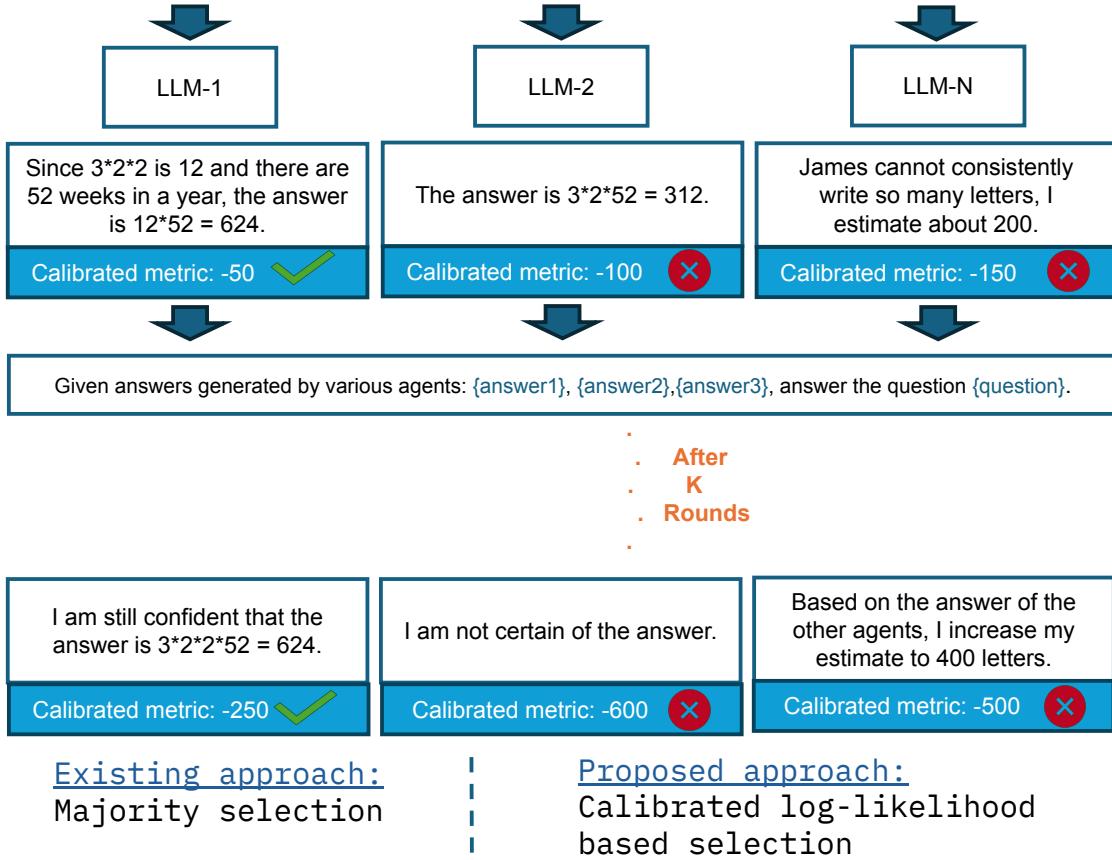


Figure 1: Diagram illustrating our proposed calibrated log-likelihood based metric for answer selection in a multi-LLM system. The multi-LLM system shows a debate setting which goes on for K rounds. When $K = 1$, it is best-of- N setting when answers are independently sampled from each LLM and best answer is selected without exposing the answer to the other LLMs. The best answer is chosen based on calibrated log-likelihood scoring. Our approach outperforms random selection when there is no clear majority answer.

Our Calibrated Metric with any Uncertainty

Method: Given a response C to a prompt p from LLM i with uncertainty metric $M_i(C | p)$, we define the calibrated metric M_c in a multi-LLM setting as:

$$M_c(C | p) = \frac{1}{N} \sum_{i=1}^N M_i(C | p),$$

where $p = x$ or $p = (x, Y_i^K)$

While uncalibrated scores may work in some empirical settings, they can introduce **bias toward the originating model and compromise fairness or consistency**. Therefore, this calibration step is crucial and represents an important contribution of this work. To motivate the general use of this diverse calibration, we also compare the performance of log-likelihood with other metrics such as Gini impurity, self-certainty (Kang et al., 2025)

and entropy in Table 1 and 2. Please refer to the appendix B for more details. Our results show that log-likelihood tends to perform better or similar to other metrics.

Note: In practice, we apply our calibrated log-likelihood scoring method only if a clear majority doesn't emerge. This is because models are likely to converge to a common answer only if its correct (assuming they are trained on correct answers). The likelihood of all models generating the same incorrect answer is very low. Thus, applying model uncertainty metric when there is consensus is redundant and computationally inefficient. In appendix E.1 we provide results for applying calibrated log-likelihood to all cases vs only tie-break cases. As expected, the gain beyond tie-break only is marginal or none, confirming that the extra computation (for non-tie break cases) is

unnecessary.

2.2 Why Does Calibrated Log-Likelihood Improve Multi-LLM Performance?

Diversity has shown to be the major reason of improving best-of-N performance (Wang et al., 2025) in single LLM systems. The upper-bound results in Table 1 show the diversity offered by multiple LLMs showing significant performance potential. Our calibrated log-likelihood metric exploits this intrinsic diversity to find the best answer to enhance overall system accuracy.

According to (Yadkori et al., 2024), models demonstrate higher confidence and reduced hallucinations when familiar with a given data point or scenario. Importantly, confident models maintain their confidence even when prompted with incorrect answers, unlike less confident models whose uncertainty notably increases. Consequently, model confidence acts as a strong proxy for response correctness. Our proposed uncertainty-aware metric directly captures this relationship, allowing effective discrimination between correct and incorrect responses.

Direct application of standard uncertainty metrics (Kang et al., 2025) is unsuitable for multi-LLM because of different π_i . To address this, our method normalizes log-likelihood scores across multiple models, as shown in Equation 1, making scores directly comparable. Additionally, we normalize log-likelihood values by sentence length to account for variability in response lengths across models. A theoretical justification is provided in Appendix D. To validate our approach, we empirically demonstrate (see Fig. 2 in the Appendix) that there is a concentration of correct answers at lower values of log-likelihood. This observation reinforces our use of uncertainty-based metrics for robust multi-LLM answer selection.

2.3 Computational Cost of the Proposed Approach.

Our calibrated scoring is computationally efficient as it does not require any additional decoding; it repurposes the already-generated completions and merely queries each model once per completion. To explain this further lets denote the number of language models in the ensemble as N and the (average) number of tokens in a completion as L . With one completion per model, the *calibrated log-likelihood* procedure evaluates every (model, completion) pair exactly once, thus making a total

of N^2 *teacher-forced* forward passes:

$$Cost_{\text{cal}} = N^2 \times \underbrace{(1 \text{ forward pass over } L)}_{\text{parallel over time}} = O(N^2)$$

The total time is proportional to only N^2 as *all* L tokens are processed in parallel on modern accelerators. If we compare our method with N generations per model, the expected cost of the procedure will be proportional to L as well because of autoregressive generation. For long, reasoning-heavy answers ($L \gg N$)—the increase in computation is thus significant.

$$Cost_{\text{gen}} = O(N^2L)$$

Here $Cost_{\text{cal}}$ and $Cost_{\text{gen}}$ refer to the number of forward passes after GPU parallelization. Hence tie-breaking via extra generation is roughly a factor of L more expensive than calibrated scoring.

3 Experimental Setup and Results

Baselines. We evaluate our proposed metric for multi-LLM setups by comparing it against commonly used majority voting with random tie-breaking (Du et al., 2023), considering both interactive multi-LLM debate and non-interactive best-of-N settings. We utilize **three models**: Qwen2.5-7B-Instruct, Minstral-8B-Instruct-2410, and Llama-3.1-8B-Instruct. The **evaluation datasets** are GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020) (comprising 8 subsets), and ARC (Clark et al., 2018). Further experimental details are provided in Appendix C.

Additionally, we compare against single-LLM best-of-N sampling using the best-performing model (Qwen2.5-7B-Instruct). This comparison allows us to assess the relative advantage provided by multi-LLM systems when using our proposed approach. To ensure fairness, we maintain an equal number of total LLM calls across all evaluated scenarios.

3.1 Tie-breaking Using Calibrated Metrics Outperforms Random Tie-breaking

Our experiments in table 1, 2 demonstrate that employing calibrated metrics for tie-breaking yields significant performance improvements compared to random tie-breaking. Specifically, for GSM8K and MMLU, we observe absolute accuracy improvements of **3.88%** and **2.44%** respectively, within

		GSM8K					
		Random	Log-Likelihood	Self-Certainty	Gini Score	Entropy	Upper Bound
Best-of-N (Q=9)		82.55	82.55	-	-	-	92.65
Best-of-N (Q=1 L=1 M=1)		47.87	70	-	-	-	70
Best-of-N (Q=3 L=3 M=3)		76.18	77.16	-	-	-	89.51
Debate (Q L M)		81±0.24	84.88	84.88	84.57	84.34	90.00
		ARC					
Best-of-N (Q=3)		85.88	85.97	-	-	-	-
Best-of-N (Q=1 L=1 M=1)		83.90±0.01	89.00	88.91	89.00	89.00	94.62

Table 1: Performance comparison demonstrating the effectiveness of calibrated metrics for tie-breaking. For GSM8K, we assess three scenarios: (a) best-of-N with the best-performing single model (Qwen2.5-7B-Instruct) using $N = 9$, (b) best-of-N with samples pooled evenly from all three models ($N = 3$ per model), and (c) a three-model debate setting similar to (Du et al., 2023) with three rounds and one sample per round per model. In all scenarios, the number of LLM calls remains identical. For the ARC dataset, we report results for single-LLM best-of-N ($N = 3$) and multi-LLM best-of-N ($N = 1$ per model).

	FL	HSM	EM	CM	PHI	AA	Avg
Random	47.20	46.38	79.01	40.81	69.77	43.87	54.51
Ours (Q=1 L=1 M=1)	49.60	50.57	82.23	43.87	70.09	42.85	57.92
Upper Bound	72.00	65.90	93.00	64.00	81.00	57.00	72.15
Random	50.79	48.89	79.36	37.00	69.45	48.00	56.54
Ours Debate (Q L M)	53.17	51.85	82.8	37.00	72.26	51.00	58.98
Upper Bound	73.80	70.70	93.90	65.00	83.90	68.00	75.88

Table 2: Accuracy comparison on MMLU dataset subsets: Formal Logic (FL), High School Math (HSM), Elementary Math (EM), College Mathematics (CM), Philosophy (PHI), and Abstract Algebra (AA). We evaluate random tie-breaking versus proposed metric-based tie-breaking in two settings: (a) best-of-N sampling across three models, and (b) three-model debate. Our approach consistently outperforms the baseline.

the multi-LLM debate setting. For best-of-N sampling, we observe absolute accuracy improvements of **1%, 3.41% and 4.9%**, respectively for GSM8K, MMLU and ARC datasets. Furthermore, this approach allows the multi-LLM setting to outperform single-model best-of-N baselines using Qwen2.5-7B-Instruct as seen in table 1, even when the total number of LLM calls remains constant.

3.2 All Calibrated Metrics Provide Similar Performance

Table 1 compares various calibrated metrics for tie-breaking across two datasets (GSM8K and ARC). These calibrated metrics are well-established alternatives to majority voting in single-LLM best-of-N scenarios, as previously reported (Kang et al., 2025). Our findings indicate that performance differences among the calibrated metrics are comparable, but as seen from Table 1 calibrated log-likelihood performs the best.

4 Conclusion

We introduced a calibrated log-likelihood-based selection framework to enhance multi-LLM systems.

By leveraging uncertainty estimation, our method selects the most confident response, reducing reliance on costly external verifiers and extensive sampling. Our approach outperforms random selection and majority voting with same model 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.

5 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.

6 Acknowledgments

Agrawal, Satheesh, Chakraborty and Huang are supported by DARPA Transfer from Imprecise and Abstract Models to Autonomous Technologies (TIAMAT) 80321, DARPA HR001124S0029-AIQ-FP-019, DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048, National Science Foundation NSF-IIS-2147276 FAI, National Science Foundation NAIRR240045, National Science Foundation TRAILS Institute (2229885). Private support was provided by Peraton.

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*.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, and 1 others. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.
- 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*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.
- 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*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*.
- 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 Flávio 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*.
- Tianchun Wang, Zichuan Liu, Yuanzhou Chen, Jonathan Light, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025. Diversified sampling improves scaling llm inference. *arXiv preprint arXiv:2502.11027*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihani 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*.
- Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

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.

A 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., 2022) 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.

B Other Metrics

We compare our proposed uncertainty-aware metric with several commonly used token-level confidence and uncertainty metrics.

Entropy. Entropy captures the model’s uncertainty over its output distribution:

$$\text{Entropy}(Y) = \frac{1}{T} \sum_{t=1}^T \sum_{v \in \mathcal{V}} -P(y_t=v \mid \cdot) \log P(y_t=v \mid \cdot) \quad (2)$$

where \mathcal{V} denotes the vocabulary. Lower entropy indicates greater confidence in the model’s predictions.

Perplexity. Perplexity is the exponentiated negative average log-likelihood:

$$\text{Perplexity}(Y) = \exp \left(-\frac{1}{T} \sum_{t=1}^T \log P(y_t | \cdot) \right). \quad (3)$$

Lower perplexity values indicate that the model considers the response more likely or fluent.

Gini Impurity. Gini impurity measures the dispersion of the output distribution:

$$\text{Gini}(Y) = \frac{1}{T} \sum_{t=1}^T \left(1 - \sum_{v \in \mathcal{V}} P(y_t=v | \cdot)^2 \right). \quad (4)$$

A lower Gini score corresponds to a more confident (i.e., peaked) distribution over tokens.

KL Divergence (Model Disagreement). To quantify disagreement between models, we compute the average pairwise KL divergence:

$$\text{KL}(\pi_i \| \pi_j) = \frac{1}{T} \sum_{t=1}^T \sum_{v \in \mathcal{V}} P_{\pi_i}(y_t=v) \log \frac{P_{\pi_i}(y_t=v)}{P_{\pi_j}(y_t=v)} \quad (5)$$

This can be averaged over all model pairs (i, j) to measure the degree of inter-model uncertainty.

C Experimental Setup

Models Used: We use the following model since they are all of same size and have quite comparable performance. However Qwen seems to be the strongest out of all.

1. Qwen2.5-7B-Instruct
2. Minstral-8B-Instruct-2410
3. Llama-3.1-8B-Instruct.

Datasets Used: We use the following datasets.

1. GSM8k (Cobbe et al., 2021) is a math reasoning task based dataset.
2. MMLU (Hendrycks et al., 2020) is a massive multitask test consisting of multiple-choice questions from various branches of knowledge. The test spans subjects in the humanities, social sciences, hard sciences, and other areas. We choose 8 subsets based on their closeness to reasoning task:
 - (a) formal-logic (FL)
 - (b) high-school-mathematics (HSM)
 - (c) elementary-mathematics (EM)
 - (d) college-mathematics (CM)

- (e) high-school-computer-science (HSCS),
- (f) philosophy (PHI)
- (g) abstract-algebra (AA)
- (h) high-school-statistics (HSS).

3. ARC (Clark et al., 2018) is a grade-school level, multiple-choice science questions, assembled to encourage research in advanced question-answering. We choose the ARC-Challenge subset of this dataset.

D Intuitive Justification for our Metric

Below we show why our *calibrated log-likelihood* score

$$\text{Score}(Y) =$$

$$\frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^T \log P_{\pi_i}(y_t | x, X, y_{<t}) \quad (6)$$

is fully comparable across different candidate answers Y , and depends only on Y itself.

1. Per-token, per-model normalization Each model π_i assigns a joint probability to the sequence $Y = (y_1, \dots, y_T)$:

$$P_{\pi_i}(Y | x, X) = \prod_{t=1}^T P_{\pi_i}(y_t | x, X, y_{<t}).$$

Taking the logarithm and dividing by sequence length T yields the average per-token log-likelihood:

$$\frac{1}{T} \sum_{t=1}^T \log P_{\pi_i}(y_t | x, X, y_{<t}).$$

Dividing by T removes any bias toward shorter or longer sequences, placing all candidates on the same per-token scale.

2. Cross-model averaging We then average these normalized log-likelihoods across the N models:

$$\text{Score}(Y) = \frac{1}{N} \sum_{i=1}^N \underbrace{\frac{1}{T} \sum_{t=1}^T \log P_{\pi_i}(y_t | x, X, y_{<t})}_{\text{model } i \text{ score}}.$$

Equivalently,

$$\text{Score}(Y) = \frac{1}{T} \sum_{t=1}^T \underbrace{\frac{1}{N} \sum_{i=1}^N \log P_{\pi_i}(y_t | x, X, y_{<t})}_{\text{consensus token score}}.$$

Noting that

$$\exp(\text{Score}(Y) \times T) = \left(\prod_{i=1}^N P_{\pi_i}(Y | x, X) \right)^{1/N},$$

the score corresponds to the geometric mean of the per-model probabilities, ensuring that a high score requires *all* models to find Y likely.

3. Fixed hyperparameters \Rightarrow only Y varies All inference hyperparameters (e.g., temperature, tokenization, number of rounds K), as well as the number weights and the sequence length T , are held constant when computing $\text{Score}(Y)$. Hence, the only variable across different candidate answers is the token sequence Y itself, making scores directly comparable.

Interpretation as a Product-of-Experts Define the product-of-experts distribution

$$P_{\text{prod}}(Y) \propto \prod_{i=1}^N P_{\pi_i}(Y | x, X).$$

Then

$$\begin{aligned} \log P_{\text{prod}}(Y) &= \sum_{i=1}^N \log P_{\pi_i}(Y | x, X) \\ \implies \text{Score}(Y) &= \frac{1}{NT} \log P_{\text{prod}}(Y). \end{aligned}$$

Maximizing $\text{Score}(Y)$ is therefore equivalent to finding the answer Y that maximizes the product-of-experts likelihood, i.e. the response deemed most probable by the ensemble as a whole.

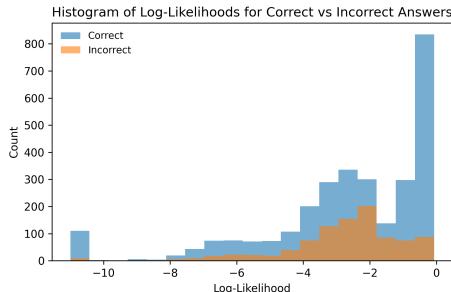


Figure 2: Histogram of log-likelihood scores, showing the distribution for correct (blue) versus incorrect (orange) responses. The histogram is plotted over responses sampled from round 3 of the LLM-debate experiment from the GSM8K dataset.

Conclusion By normalizing per token and averaging across a fixed set of models, all extraneous factors—sequence length, model-specific scales, and hyperparameters—are eliminated. The resulting score depends solely on the candidate answer Y , ensuring fair and robust comparison across multi-LLM outputs.

E Additional Results

E.1 Comparison between applying calibrated log-likelihood metric for all cases vs tie-break cases only.

	Random	Tie-Break Case only	All cases
GSM8K (Q=1, L=1, M=1)	47.87	70.00	71.00
GSM8K Debate (Q, L, M)	81	84.88	85.03
ARC (Q=1, L=1, M=1)	83.90	89.00	88.50

Table 3: Comparison of applying log-likelihood metric to tie-break only cases vs all cases.