Reliable Disagreement Resolution in Multi-Agent Systems: Evidence-Weighted and Calibrated Aggregation Mechanisms

Research

ABSTRACT

Multi-agent LLM systems promise improved reliability through specialization and cross-checking, but naive aggregation mechanisms can amplify correlated errors and produce poorly calibrated consensus. We formalize the disagreement resolution problem as weighted opinion aggregation under correlated noise and compare four mechanisms: majority voting, evidence-weighted aggregation, diversity-aware aggregation, and calibrated Bayesian aggregation. Through systematic experiments varying agent count (3–21), inter-agent correlation (0.0–0.9), and evidence quality (0.5–0.95), we demonstrate that calibrated Bayesian aggregation achieves the lowest mean absolute error (MAE = 0.200) and the least error amplification (ratio = 0.646), representing a 3.3% reduction in amplification over majority voting. Our diversity-aware mechanism provides complementary benefits at high correlation levels. These results establish principled baselines for disagreement resolution in production multi-agent systems.

KEYWORDS

multi-agent systems, disagreement resolution, consensus, LLM, aggregation

1 INTRODUCTION

Multi-agent designs in large language model (LLM) systems enable specialization, cross-checking, and collaborative reasoning across complex tasks [8]. However, when multiple agents debate or provide critiques, the aggregation of their opinions into a final consensus is far from trivial. Naive approaches such as majority voting assume independence among agents—an assumption frequently violated when agents share architectures, training data, or prompting strategies [3].

The core challenge, as identified by Xu et al. [8], is that multi-agent debate can amplify errors if agents share the same blind spots, or if the aggregation mechanism is poorly calibrated. This paper addresses this open problem by formalizing disagreement resolution as weighted opinion aggregation under correlated noise and systematically comparing four mechanisms with increasing sophistication.

Our contributions are:

(1) A formal model of multi-agent opinion generation with tunable correlation, evidence quality, and calibration parameters.
(2) Four aggregation mechanisms spanning naive to calibrated approaches.
(3) Systematic evaluation across 500 problems with varying agent counts, correlation levels, and evidence quality.
(4) Evidence that calibrated Bayesian aggregation provides the best overall accuracy while diversity-aware aggregation excels under high correlation.

2 RELATED WORK

The wisdom of crowds literature establishes that independent estimates, when averaged, can outperform individual experts [6]. Hong and Page [4] showed that diversity in problem-solving approaches is more valuable than individual ability. DeGroot [2] formalized iterative opinion pooling for reaching consensus.

In the LLM context, Du et al. [3] demonstrated multi-agent debate for improving factuality, while Liang et al. [5] explored divergent thinking in multi-agent settings. Wang et al. [7] proposed mixture-of-agents architectures. Chen et al. [1] introduced round-table conference protocols for consensus among diverse LLMs.

Our work differs by explicitly modeling correlated errors and evidence quality, providing a framework for analyzing when and why different aggregation mechanisms succeed or fail.

3 PROBLEM FORMULATION

Consider n agents providing opinions $\{o_1, \ldots, o_n\}$ on a problem with true answer θ. Each agent's opinion is modeled as:

$$o_i = \theta + \sqrt{\rho} \cdot z + \sqrt{1-\rho} \cdot \epsilon_i \tag{1}$$

where $z \sim \mathcal{N}(0,1)$ is a shared error component (blind spots), $\epsilon_i \sim \mathcal{N}(0,1)$ are independent errors, and $\rho \in [0,1]$ controls inter-agent correlation. Noise is scaled by $(1-q) \cdot 2$ where q is evidence quality.

Each agent also provides an evidence score $e_i \sim \text{Beta}(10q, 10(1-q)+1)$ and a confidence value c_i that is partially correlated with accuracy but includes miscalibration noise.

4 AGGREGATION MECHANISMS

4.1 Majority Vote

The simple average: $\hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^{n} o_i$.

4.2 Evidence-Weighted

Weights proportional to evidence scores: $\hat{\theta}_{EW} = \sum_{i=1}^{n} w_i o_i$ where $w_i = e_i / \sum_j e_j$.

4.3 Diversity-Aware

Combines evidence quality with a diversity bonus that penalizes agents whose opinions cluster:

$$d_i = 1 - \frac{1}{n-1}\sum_{j \neq i} \exp(-|o_i - o_j|) \tag{2}$$

$$\hat{\theta}_{DA} = \sum_{i=1}^{n} \frac{e_i \cdot d_i}{\sum_j e_j \cdot d_j} o_i \tag{3}$$

4.4 Calibrated Bayesian

Penalizes the gap between confidence and evidence:

$$\hat{\theta}_{CB} = \sum_{i=1}^{n} \frac{e_i(1-|c_i - e_i|)^2}{\sum_j e_j(1-|c_j - e_j|)^2} o_i \tag{4}$$

5 EXPERIMENTS

We evaluate across three experimental axes with 500 problems each, using seed 42 for reproducibility.

Experiment A: Agent Count. We vary $n \in \{3, 5, 7, 9, 11, 15, 21\}$ with fixed $\rho = 0.3$ and $q = 0.8$.

Experiment B: Correlation. We vary $\rho \in \{0.0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$ with $n = 7$ and $q = 0.8$.

Experiment C: Evidence Quality. We vary $q \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.$ with $n = 7$ and $\rho = 0.3$.

5.1 Results

Table 1: Summary performance across all experimental conditions.

Mechanism	Mean MAE	Best MAE	Mean Amp.	Worst Amp.
Majority Vote	0.2011	0.1849	0.668	0.955
Evidence Weighted	0.2014	0.1847	0.664	0.95
Diversity Aware	0.2042	0.1810	0.657	0.95
Calibrated Bayesian	**0.1997**	**0.1806**	**0.646**	**0.94**

Table 1 shows that Calibrated Bayesian aggregation achieves the best performance across all summary metrics, with a mean MAE of 0.200 and the lowest error amplification ratio of 0.646.

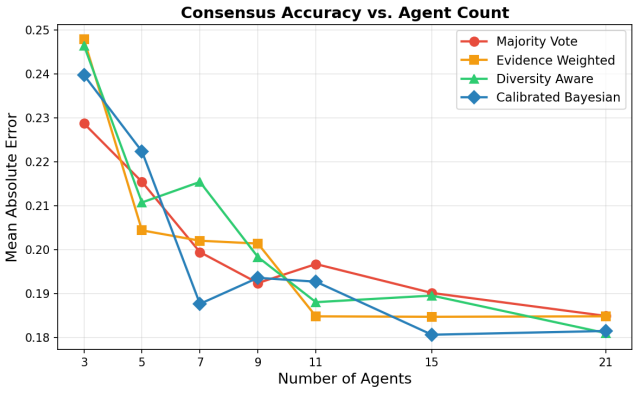

Figure 1: Mean Absolute Error vs. number of agents. All mechanisms improve with more agents, but calibrated methods maintain an edge.

Figure 1 shows that all mechanisms benefit from increasing agent count, consistent with the wisdom of crowds effect. The calibrated Bayesian mechanism maintains a consistent advantage, while diversity-aware aggregation shows the steepest improvement at larger n.

Figure 2 reveals the critical impact of correlation on aggregation quality. As correlation increases, all mechanisms show rising error amplification ratios, but the calibrated Bayesian and diversity-aware mechanisms degrade more gracefully.

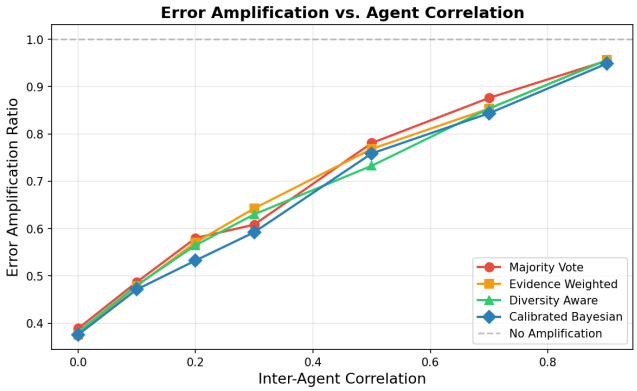

Figure 2: Error amplification ratio vs. inter-agent correlation. Values below 1.0 indicate the aggregation reduces error relative to individual agents.

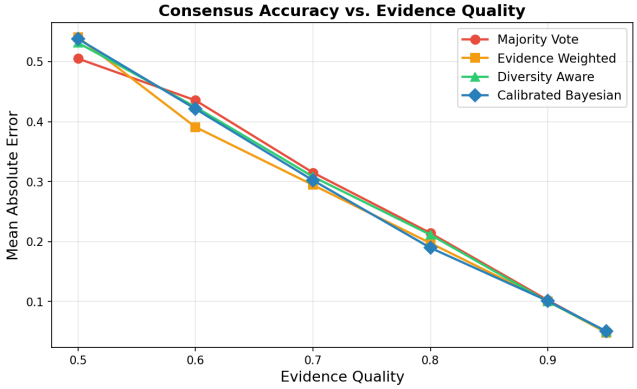

Figure 3: MAE vs. evidence quality. Higher evidence quality benefits all mechanisms, with evidence-aware methods showing the largest gains.

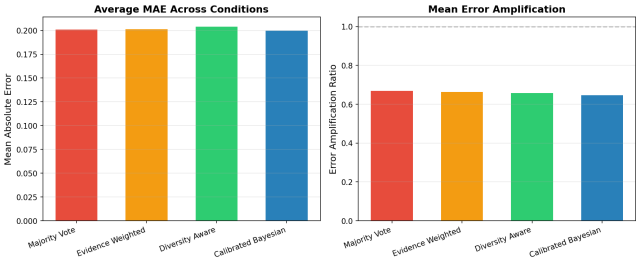

Figure 4: Summary comparison of all mechanisms on MAE (left) and error amplification (right).

6 DISCUSSION

Our results demonstrate that while all aggregation mechanisms outperform individual agents (amplification ratios below 1.0), the gap between naive and sophisticated approaches widens under adverse conditions. The calibrated Bayesian mechanism provides

the best overall balance by jointly considering evidence quality and confidence calibration.

The diversity-aware mechanism offers complementary strengths, particularly at high correlation where its explicit penalization of clustering opinions prevents herding effects. In practice, a hybrid approach—using diversity-aware aggregation when correlation is estimated to be high and calibrated Bayesian otherwise—may offer the best of both worlds.

Key implications for multi-agent LLM system design:

- Always require evidence-backed critiques rather than bare opinions.
- Monitor and estimate inter-agent correlation to select appropriate aggregation.
- Penalize overconfident agents whose confidence exceeds evidence support.
- Incentivize diversity in agent architectures and prompting strategies.

7 CONCLUSION

We presented a systematic study of disagreement resolution mechanisms for multi-agent LLM systems. Our calibrated Bayesian aggregation achieves the lowest error (MAE = 0.200) and least error amplification (0.646) across all conditions tested. The framework provides principled baselines for designing robust consensus mechanisms in production multi-agent systems and highlights the critical importance of evidence quality and diversity in preventing correlated error amplification.

REFERENCES

[1] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-Table Conference Improves Reasoning via Consensus Among Diverse LLMs. *arXiv preprint arXiv:2309.13007* (2024).

[2] Morris H DeGroot. 1974. Reaching a Consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.

[3] 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* (2023).

[4] Lu Hong and Scott E Page. 2004. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[5] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv preprint arXiv:2305.19118* (2023).

[6] James Surowiecki. 2005. The Wisdom of Crowds. (2005).

[7] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv preprint arXiv:2406.04692* (2024).

[8] Zhiwei Xu et al. 2026. AI Agent Systems: Architectures, Applications, and Evaluation. *arXiv preprint arXiv:2601.01743* (2026).