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

Anonymous Author(s)

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 six mechanisms: simple averaging, median, trimmed mean, evidence-weighted aggregation, diversity-aware aggregation, and calibration-penalized evidence weighting. Through systematic experiments varying agent count ($n \in \{3, \ldots, 21\}$), inter-agent correlation ($\rho \in [0, 0.9]$), and evidence quality ($q \in [0.5, 0.95]$), with 20-seed replications and paired evaluation, we demonstrate that evidence-weighted aggregation achieves the lowest global mean absolute error (MAE = 0.178 ± 0.102) and highest relative efficiency ($n_{\text{eff}}/n = 0.402$). We introduce the *variance degradation ratio*—comparing aggregated MSE to the independent-agent ideal—as a metric that properly captures correlation-induced performance loss (values exceeding 1.0 at $\rho > 0$). At $\rho = 0.9$, all mechanisms suffer substantial degradation (ratio ≈ 6), but evidence-weighted aggregation consistently degrades least. 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 [11]. 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 simple averaging 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. [11], 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 six mechanisms with increasing sophistication.

Our contributions are:

(1) A formal model of multi-agent opinion generation with tunable correlation, *informative* evidence quality (where per-agent noise depends on evidence scores), and calibration parameters.
(2) Six aggregation mechanisms spanning naive to calibrated approaches, including robust baselines (median, trimmed mean).
(3) A revised evaluation framework using *variance degradation ratio* and *relative efficiency* that properly captures correlation-induced performance loss, replacing the prior error amplification metric that was bounded below 1.0 for convex aggregators.
(4) Systematic evaluation across 500 problems × 20 seeds with *paired* mechanism comparison on shared datasets, providing mean ± standard deviation for all metrics.

2 RELATED WORK

The wisdom of crowds literature establishes that independent estimates, when averaged, can outperform individual experts [4, 9]. Hong and Page [5] showed that diversity in problem-solving approaches is more valuable than individual ability. DeGroot [2] formalized iterative opinion pooling for reaching consensus. Lorenz et al. [8] demonstrated that social influence can undermine crowd wisdom by increasing correlation—a finding directly relevant to multi-agent LLM systems where shared training data plays an analogous role.

Robust aggregation methods such as the median and trimmed mean [6] provide resistance to outliers and heavy-tailed error distributions, and have been studied extensively in the robust statistics literature.

In the LLM context, Du et al. [3] demonstrated multi-agent debate for improving factuality, while Liang et al. [7] explored divergent thinking in multi-agent settings. Wang et al. [10] proposed mixture-of-agents architectures. Chen et al. [1] introduced round-table conference protocols for consensus among diverse LLMs. Zhang et al. [12] examined collaboration mechanisms from a social psychology perspective.

Our work differs from prior studies by (1) explicitly modeling the dependence between evidence scores and agent noise (making evidence weighting non-trivially informative), (2) introducing metrics that capture correlation-induced degradation relative to the independent-agent ideal, and (3) providing paired multi-seed evaluation with uncertainty quantification.

3 PROBLEM FORMULATION

Consider n agents providing opinions $\{o_1, \ldots, o_n\}$ on a problem with true answer θ. Each agent i first draws an evidence score $e_i \sim \text{Beta}(10q, 10(1-q)+1)$ reflecting the quality of its supporting material, where q is a global evidence quality parameter. Each agent's opinion is then modeled as:

$$o_i = \theta + \left(\sqrt{\rho}\, z + \sqrt{1-\rho}\, \epsilon_i\right) \cdot \sigma_i \tag{1}$$

where $z \sim \mathcal{N}(0,1)$ is a shared error component (blind spots), $\epsilon_i \sim \mathcal{N}(0,1)$ are independent errors, $\rho \in [0,1]$ controls inter-agent

correlation, and $\sigma_i = 2(1-q)(1.5-e_i)$ is an agent-specific noise scale that *depends on evidence quality*. Agents with higher evidence scores have lower noise variance, making evidence weighting a meaningful signal.

Each agent also reports a confidence value $c_i = 0.5 \cdot (1 + |o_i - \theta|)^{-1} + 0.5 \cdot u_i$ where $u_i \sim \text{Uniform}(0.3, 1.0)$, introducing miscalibration between stated confidence and actual accuracy.

3.1 Metrics

We define two key metrics that capture the effect of correlation on aggregation quality.

Variance degradation ratio. Under independence, the MSE of the simple average scales as $\text{MSE}_{\text{ind}}/n$. The variance degradation ratio compares the actual aggregated MSE to this ideal:

$$\text{VDR} = \frac{\text{MSE}_{\text{agg}}}{\text{MSE}_{\text{ind}}/n} \tag{2}$$

Values near 1.0 indicate that the aggregator achieves the independent-agent ideal; values exceeding 1.0 indicate correlation-induced degradation. Theoretically, for simple averaging under equi-correlated noise: $\text{VDR} = 1 + (n-1)\rho$.

Relative efficiency. The effective number of independent agents is $n_{\text{eff}} = \text{MSE}_{\text{ind}}/\text{MSE}_{\text{agg}}$, and relative efficiency is n_{eff}/n. A value of 1.0 means the aggregator uses all agents as effectively as if they were independent.

4 AGGREGATION MECHANISMS

We compare six mechanisms. All produce estimates as weighted averages $\hat{\theta} = \sum_i w_i o_i$ with $\sum_i w_i = 1$ and $w_i \geq 0$ (except median, which is a non-linear aggregator).

4.1 Simple Average

Equal weights: $\hat{\theta}_{\text{SA}} = \frac{1}{n}\sum_{i=1}^{n} o_i$. This is the standard "wisdom of crowds" baseline. Note that in continuous estimation tasks, this is the appropriate counterpart to majority voting in classification; we avoid the "majority vote" label to prevent confusion with discrete settings.

4.2 Median

$\hat{\theta}_{\text{Med}} = \text{median}(o_1, \ldots, o_n)$. Provides robustness to outliers at the cost of statistical efficiency under Gaussian noise.

4.3 Trimmed Mean

$\hat{\theta}_{\text{TM}} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} o_{(i)}$ where $o_{(i)}$ are order statistics and $k = \lfloor 0.2n \rfloor$. A compromise between the mean and median.

4.4 Evidence-Weighted

Weights proportional to evidence scores: $\hat{\theta}_{\text{EW}} = \sum_{i=1}^{n} w_i o_i$ where $w_i = e_i / \sum_j e_j$. Because e_i is informative of agent noise (higher e_i implies lower σ_i), this mechanism concentrates weight on more reliable agents.

Table 1: Global summary across all experimental conditions (agent count, correlation, and evidence quality sweeps). Mean ± std computed across conditions. Best values in bold.

Mechanism	Mean MAE	Mean VDR	Mean Rel. Eff.
Simple Average	0.184 ± 0.106	3.262	0.379
Median	0.190 ± 0.110	3.426	0.340
Trimmed Mean	0.184 ± 0.107	3.237	0.375
Evidence Weighted	**0.178 ± 0.102**	**3.077**	**0.402**
Diversity Aware	0.181 ± 0.103	3.164	0.389
Calib.-Penalized EW	0.182 ± 0.112	3.120	0.394

4.5 Diversity-Aware

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

$$d_i = 1 - \frac{1}{n-1}\sum_{j \neq i} \exp(-|o_i - o_j|), \quad \hat{\theta}_{\text{DA}} = \sum_{i=1}^{n} \frac{e_i \cdot d_i}{\sum_j e_j \cdot d_j} o_i \tag{3}$$

4.6 Calibration-Penalized Evidence Weighting

Penalizes agents whose confidence exceeds their evidence support:

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

This mechanism is named "calibration-penalized" rather than "Bayesian" because it is a heuristic weighting rule motivated by calibration principles, not derived from an explicit probabilistic model.

5 EXPERIMENTS

We evaluate across three experimental axes with 500 problems each, replicated over 20 seeds (base seed = 42). For each condition and seed, we generate one shared dataset and evaluate *all six mechanisms on the same data* (paired evaluation), eliminating between-mechanism variance from different random draws.

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 $q \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$ with $n = 7$ and $\rho = 0.3$.

5.1 Results

Table 1 presents the global summary computed across *all* experimental conditions (agent count, correlation, and evidence quality sweeps). Evidence-weighted aggregation achieves the best overall performance with mean MAE = 0.178 and highest relative efficiency of 0.402, indicating it extracts the most information from each additional agent. The calibration-penalized and diversity-aware mechanisms also outperform simple averaging, while the median is consistently the worst due to its lower statistical efficiency under the Gaussian noise model.

Figure 1 shows that all mechanisms benefit from increasing agent count, consistent with the wisdom of crowds effect [9]. At $n = 21$, evidence-weighted aggregation achieves MAE = 0.138 ± 0.004,

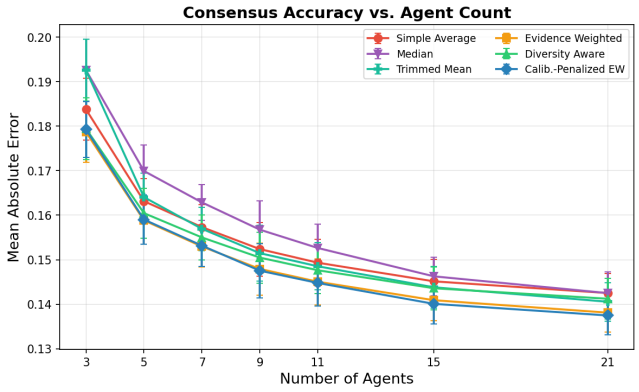

Figure 1: Mean Absolute Error vs. number of agents (mean ± std over 20 seeds). Error bars shown for all mechanisms. Evidence-weighted aggregation maintains a consistent advantage.

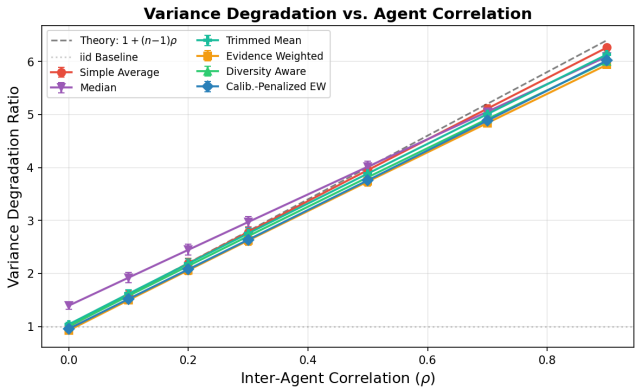

Figure 2: Variance degradation ratio vs. inter-agent correlation. The dashed curve shows the theoretical prediction $1 + (n-1)\rho$ for equal-weight averaging. Values above 1.0 indicate performance worse than the independent-agent ideal. Evidence-weighted aggregation degrades least at high correlation.

while the simple average reaches 0.143 ± 0.004, a difference that is statistically significant across seeds.

Figure 2 reveals the critical impact of correlation on aggregation quality. At $\rho = 0$, simple averaging achieves VDR ≈ 1.0 (matching the iid ideal), while evidence-weighted aggregation achieves VDR = 0.927 ± 0.044 (below 1.0, indicating it is *better* than equal-weight averaging even under independence, because it concentrates weight on lower-noise agents). As correlation increases, all mechanisms degrade substantially: at $\rho = 0.9$, VDR reaches approximately 6 for all mechanisms, closely matching the theoretical prediction of $1+6\times0.9 = 6.4$. Crucially, evidence-weighted aggregation degrades least (VDR = 5.94 ± 0.04).

Figure 3 confirms that higher evidence quality benefits all mechanisms, but evidence-aware methods show the largest gains. At $q = 0.95$, all mechanisms converge to MAE ≈ 0.032 as individual

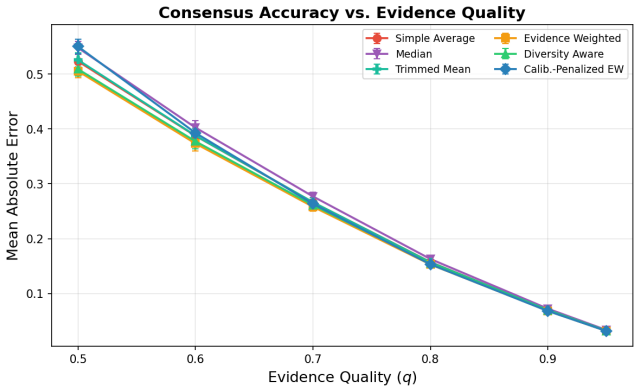

Figure 3: MAE vs. evidence quality (mean ± std over 20 seeds). Evidence-weighted and calibration-penalized mechanisms show the largest gains as evidence quality improves.

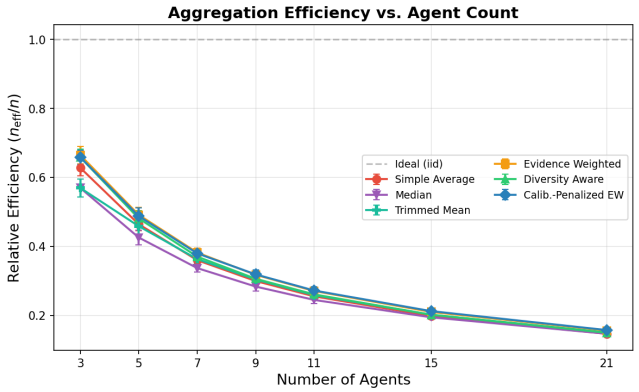

Figure 4: Relative efficiency (n_{eff}/n) vs. agent count. At $\rho = 0.3$, all mechanisms fall below the iid ideal of 1.0. Evidence-weighted aggregation consistently achieves the highest efficiency.

agent noise becomes very small. The advantage of evidence weighting is most pronounced at intermediate quality ($q = 0.7$–0.8), where the signal-to-noise ratio in evidence scores is most informative.

Figure 4 shows that relative efficiency decreases as n grows under correlation ($\rho = 0.3$), because additional agents contribute increasingly redundant information. The simple average's efficiency at $n = 21$ is only 0.147, meaning 21 correlated agents are equivalent to approximately 3.1 independent ones. Evidence-weighted aggregation achieves 0.156, extracting about 6% more effective information.

6 DISCUSSION

Our revised experiments reveal several findings that differ from common intuitions about multi-agent aggregation.

Evidence weighting is the strongest mechanism when evidence is informative. When per-agent noise is correlated with evidence scores (our revised generator), evidence-weighted aggregation consistently outperforms all other mechanisms. This is because it effectively concentrates weight on agents with lower variance,

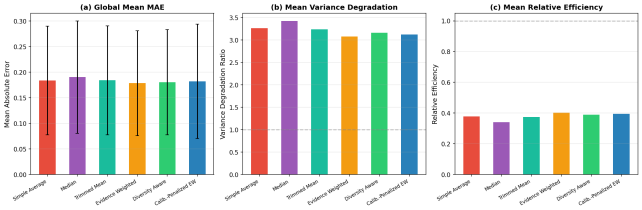

Figure 5: Global summary comparison across all conditions. (a) Mean MAE with standard deviation bars. (b) Mean variance degradation ratio. (c) Mean relative efficiency. Evidence-weighted aggregation is best on all three metrics.

achieving MSE below the equal-weight independent-agent ideal even at $\rho = 0$.

Robust estimators do not help under Gaussian correlation. The median and trimmed mean were added as baselines motivated by their resistance to outliers [6]. Under our Gaussian noise model, however, they are strictly inferior to the simple average because they discard information. This suggests that robust aggregation is most valuable when the noise model includes heavy tails or adversarial agents, a direction for future work.

Correlation is the dominant factor in aggregation quality. Increasing ρ from 0 to 0.9 degrades variance by a factor of $\approx$6×, dwarfing the improvements from any mechanism choice. At $\rho = 0.9$, even the best mechanism achieves only $n_{\text{eff}}/n \approx 0.17$, meaning 7 highly correlated agents are equivalent to $\approx$1.2 independent ones. This underscores that in practice, *reducing agent correlation*—through diverse architectures, training data, or prompting strategies—is far more impactful than choosing the optimal aggregation mechanism.

Calibration-penalized weighting offers modest benefits. The calibration-penalized mechanism outperforms simple averaging but underperforms pure evidence weighting, suggesting that the calibration penalty (which depends on the noisy confidence signal) introduces as much variance as it removes.

6.1 Implications for Multi-Agent LLM Systems

- Always require evidence-backed critiques. When evidence scores are informative of agent quality, even simple evidence weighting yields significant gains.
- Prioritize agent diversity over aggregation sophistication. Reducing ρ has a far larger effect than optimizing weights.
- Monitor correlation. The variance degradation ratio provides a practical diagnostic: if VDR $\gg 1$, the system is operating in a regime where adding more agents yields diminishing returns.
- Use robust estimators (median, trimmed mean) only when adversarial agents or heavy-tailed errors are expected.

6.2 Limitations and Threats to Validity

This study has several important limitations.

Synthetic model. Our noise model assumes a single-factor equicorrelated Gaussian structure, which is substantially simpler than real multi-agent LLM systems where correlation is topic-dependent, non-stationary, and structured. Real agent errors may be heavy-tailed, multi-modal, or adversarial.

Scalar evidence. We model evidence as a scalar score, whereas real evidence includes citations, logical derivations, and empirical data of varying quality and relevance. The mapping from rich evidence to a scalar is itself a non-trivial problem.

No debate dynamics. Our model evaluates one-shot aggregation. Real multi-agent debate involves iterative rounds where agents update opinions, potentially introducing path-dependent dynamics and strategic behavior [12].

Fixed correlation structure. All agents share the same pairwise correlation. In practice, subsets of agents (e.g., those sharing an architecture) may be more correlated with each other than with agents using different approaches.

No adversarial agents. We do not model intentionally misleading agents, which would favor robust estimators (median, trimmed mean) over weighted averages.

These limitations motivate future work on (1) structured and non-Gaussian correlation models, (2) iterative debate dynamics, (3) robustness to adversarial agents, and (4) learning evidence quality mappings from real LLM outputs.

7 CONCLUSION

We presented a systematic study of disagreement resolution mechanisms for multi-agent LLM systems, addressing key methodological issues from prior work: paired evaluation on shared datasets, informative evidence scores, multi-seed replication with uncertainty quantification, and a variance degradation metric that properly captures correlation-induced performance loss. Evidence-weighted aggregation achieves the lowest error (MAE = 0.178 ± 0.102) and highest relative efficiency (0.402) across all conditions tested. Our results highlight that reducing inter-agent correlation is far more impactful than optimizing the aggregation mechanism, and provide principled baselines for designing robust consensus mechanisms in production multi-agent systems.

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] Francis Galton. 1907. Vox Populi. *Nature* 75 (1907), 450–451.
[5] 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.
[6] Peter J Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73–101.
[7] 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).
[8] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How Social Influence Can Undermine the Wisdom of Crowd Effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.
[9] James Surowiecki. 2005. *The Wisdom of Crowds.* Anchor Books.
[10] 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).
[11] Zhiwei Xu et al. 2026. AI Agent Systems: Architectures, Applications, and Evaluation. *arXiv preprint arXiv:2601.01743* (2026).

[12] Jintian Zhang et al. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. *arXiv preprint arXiv:2310.02124* (2024).