The Problems of Evaluating the Quality of Collective Conclusions in Multi-Agent Systems of Large Language Models

Yaroslav Panasiuk[1], Vladyslav Moskal[2]

1. Department of Artificial Intelligence, Lviv Polytechnic National University, UKRAINE, Lviv, Knyazya Romana Street, 5, 79000, E-mail: Yaroslav.Panasiuk.asp.2026@lpnu.ua

2. Department of Artificial Intelligence, Lviv Polytechnic National University, UKRAINE, Lviv, Knyazya Romana Street, 5, 79000, E-mail: mmosvlad@gmail.com

Abstract – Collective conclusions of multi-agent LLM systems are hard to evaluate. We identify four key problems: the absence of a reference answer, false consensus caused by conformity and correlated errors, the attribution of individual agents' contributions, and unreliable LLM-based judges. We propose trustworthy-AI requirements for evaluation methods.

Keywords – multi-agent systems, large language models, collective reasoning, evaluation, consensus, trustworthy AI.

I. Introduction

Multi-agent systems built on large language models (LLMs) increasingly replace a single model with a group of agents that debate, discuss or vote before they give an answer. This approach relies on managing lots of cost-effective, small open-weight models, such as the 7B to 8B parameter class, and aggregating their responses, hypothesising that their emergent reasoning capabilities will rival monolithic frontier models.

Multi-agent debate is widely deployed on the assumption that peer review among agents filters out hallucinations [1]. The final collective decision is then formed either by a single "dictator" agent that chooses among the proposals or by a voting rule; a survey of such systems found a heavy reliance on dictatorial and plurality schemes [2].

At the same time, the mechanisms of collective decision-making in such systems have received little attention [2]. Few studies ask whether the consensus reached by LLM agents is meaningful or merely reinforces existing biases [3]. Where no reference answer exists, as in open-ended analysis, planning, expert assessment or creative tasks, accuracy on benchmarks cannot establish whether a conclusion can be trusted. In these settings two proxies of quality are tempting: the level of agreement among the agents and the verdict of another LLM acting as a judge.

This work analyses why the quality of collective conclusions of multi-agent LLM systems is difficult to evaluate. We first systematise the problems that make such evaluation hard, stated in Section III, and then derive from them a set of requirements that any evaluation method must satisfy to be trustworthy.

II. Related Work

Relevant work falls into three strands of multi-agent LLM systems literature: mechanisms of reaching a collective conclusion, LLMs as evaluators and credit assignment. In this section, we review the recent work across these areas and identify the methodological and evaluation gaps.

The first strand asks how a group of agents should reach a joint answer at all, and so determines what an evaluation has to measure. ReConcile, a multi-model multi-agent framework [4], organises a round-table discussion among diverse LLMs: agents exchange answers, explanations and confidence scores over several rounds, and the team answer is a confidence-weighted vote. Its ablations identify model diversity as the most important component: agents that are instances of the same model lose both accuracy and diversity of responses [4]. ConsensAgent, a framework for prompt refinement in multi-agent debate, targets sycophancy, i.e. the tendency of

agents to agree with each other instead of engaging critically; it uses triggers to detect stalls and sycophantic behaviour in a debate and then applies prompt optimisation [3]. FREE-MAD, a multi-agent debate framework, drops the requirement of consensus altogether: it replaces last-round majority voting with a score-based decision over the whole debate trajectory and adds an anti-conformity debate mode [5]. A survey of collective decision-making in such systems proposes GEDI, an electoral decision-making module, as a way to diversify aggregation [2]. A multi-agent Delphi system shows that the diversity structure of an agent group affects idea exploration, consensus quality and consensus efficiency in different directions [6]. Finally, a factor analysis of LLM agent groups reveals a general, task-independent collective intelligence factor, analogous to the one known for human groups [7].

The second strand asks whether an LLM can stand in for a human assessor when no reference answer is available. In this case, the usual substitute is an LLM acting as a judge. A reference-free intervention framework has been used to probe the biases of both human and LLM judges; among other effects, it revealed gender bias in LLM judges and their vulnerability to fake references [8]. JudgeBiasBench introduces a taxonomy of judgement biases, injects them into evaluation instances in a controlled way and measures the sensitivity of judges with the Bias Sensitivity Rate; the same work proposes bias-aware training of judges [9]. A cross-validation of public fine-tuned judge models against GPT-4 concludes that a fine-tuned judge is not a general substitute for a strong proprietary model [10].

The third strand asks how much of a joint outcome can be attributed to any one agent. Shapley-Coop combines a structured negotiation protocol with Shapley-inspired reasoning, so that self-interested LLM agents can price their contributions during a task and redistribute rewards after it [11]. In cooperative multi-agent reinforcement learning, a centralised LLM critic decomposes a shared team reward into per-agent credit and explains each decision in text [12].

In summary, these three lines of research develop largely in isolation. Studies of debate and aggregation evaluate the collective result almost exclusively by accuracy on labelled benchmarks such as MATH, MMLU and AIME [2], [4], [5], [7]. ConsensAgent additionally reports how many debates reach consensus and after how many rounds, but it still assumes a single correct answer [3]. Judge models are validated one at a time [10], never against a group of agents whose answers have already converged. Credit assignment distributes rewards among agents [11, 12] but is not used to score the conclusion those agents reached. We therefore found no accepted way to evaluate a collective conclusion when no reference answer is available.

III. Problem Statement

Several obstacles stand behind the gap identified above, and each of them reinforces the others. This section states them in turn, highlighting the limitations that the authors of the reviewed systems acknowledge themselves.

Problem 1: absence of a reference answer. The methods reviewed above are optimised and tested on tasks with one correct answer. Their authors acknowledge this limitation: ConsensAgent assumes a single correct answer when optimising prompts, which makes it less suited to tasks with several valid solutions and may prevent generalisation to open-ended or creative tasks [3]. Replacing the reference with human experts does not remove the problem. In the Delphi study the expert benchmark itself is a consensus-based criterion, and the authors note that “the consensus results of human experts should also be understood not as an absolute ground truth” [6]. Reference-free protocols exist, but the one reviewed here was designed to probe the biases of judges [8]; in our reading, it does not certify the correctness of a conclusion. Consequently, for open-ended tasks

accuracy is not observable. Any evaluation has to rely on proxies, such as agreement among agents or the opinion of an LLM judge, whose validity is questioned in problems 2 and 4.

Problem 2: false consensus from conformity and correlated errors. Agreement among agents is a natural proxy of quality, yet it can be produced by mechanisms unrelated to correctness. In multi-agent debates, LLMs often agree with each other instead of critically evaluating different perspectives; this sycophancy increases the number of rounds and weakens the robustness of reasoning [3]. Because of conformity, initially correct agents can be swayed by incorrect ones, and debate followed by majority voting can produce final answers that are less accurate than the initial independent answers [5]. Agents that are instances of the same model share biases and pre-training data, and their debate can fall into an echo chamber without external feedback [4]; agents built on one foundation model tend to converge to similar responses unless heterogeneity is designed deliberately [6]. Homogeneous teams show conformity in debate, and plurality voting can discard correct answers that were already present among the generated ones [1].

The Delphi study gives a clear illustration of the gap between agreement and quality. The homogeneous group reached an internal consensus quality of 0.9695, against 0.7477 for the maximum-diversity group, yet the differences between groups in external consensus quality, i.e. alignment with the human expert benchmark, were not statistically significant [6]. Higher agreement therefore did not mean a better conclusion. The underlying reason is that majority voting and consensus are informative only when agents err independently. When errors are correlated, many agreeing agents may provide little more evidence than a single one. A related risk lies in aggregation itself: a dictator-style decision agent is a single point of failure that collapses first when agents become unreliable [2].

Problem 3: attribution of an individual agent's contribution. Fairly evaluating the contribution of each agent is described as the central challenge of coordinating LLM agents in open-ended environments [11]. In Shapley-Coop the Shapley value is approximated heuristically by the LLM agents themselves rather than computed exactly. Without the retrospective long-term step, the first agent to reach an agreement may benefit disproportionately [11]. Contribution is also not reducible to the strength of individual models: the collaboration process has the greatest impact on the collective intelligence of a group, and stronger individual LLMs do not necessarily lead to better group outcomes [7]. Even in cooperative reinforcement learning with numeric rewards, humans who inspect agent behaviour often produce better credit evaluations than existing methods [12]. For collective conclusions in natural language, the difficulty is greater: there is no shared numeric reward to decompose, and the "true" contribution has no reference either. Without attribution it is impossible to explain how a conclusion was formed, which agent introduced an error, and who is accountable for it.

Problem 4: unreliability of LLM-based evaluation. If neither a reference nor agreement can be trusted, an LLM judge seems the remaining option. However, all tested judges, human and LLM alike, display significant biases, although they differ in their specific inclinations [8]. LLM judges are vulnerable to fake references and rich content; even the best-performing GPT-4o had an attack success rate of 32% under the fake reference perturbation [8]. With fake references, they are swayed more easily when the quality gap between two answers is small [8]. Judgement bias is prevalent even in strong judges, its direction is systematic rather than random, and high accuracy on clean data does not guarantee robustness to bias [9]. The overall Bias Sensitivity Rate on JudgeBiasBench ranges from 10.2% for Claude-3.7-Sonnet to 42.4% for the fine-tuned Prometheus-7B-V2.0 [9]. Fine-tuned judges are bound to the evaluation scheme on which they

were trained, lose performance sharply on other schemes and are biased toward superficial quality such as formality or verbosity [10].

Inside a multi-agent system the problem appears again. An LLM-as-a-judge decision stage is itself prone to conformity and becomes effectively equivalent to majority voting [5]. The confidence that agents report about their own answers may be unreliable due to hallucination [5], and confidence estimates from API models remain post-hoc [4]. The four problems are therefore not independent. The absence of a reference makes the judge the only arbiter; a judge that shares the biases of the agents ratifies their agreement; and no attribution is available to show where that agreement came from. A system can therefore be unanimous, confident and plausible while still being wrong.

IV. Requirements for Evaluation Methods

Each obstacle in Section III constrains what an evaluation method may assume, and each yields one requirement directly. This section states these core requirements, focusing on preventing the reliance on weak assumptions from which those obstacles arise. The requirements are organised around the trustworthy-AI properties of reliability, transparency, accountability and robustness.

Requirement 1: partial reference with explicit uncertainty. A method should work with no reference answer or with a partial one, such as a labelled subset or expert spot checks on a sample, and should report an interval together with the sample size that supports it. We propose treating the labelled subset as a calibration set whose measured error bounds the score on the unlabelled remainder. Problem 1 establishes that accuracy is unobservable for open-ended tasks [3], [6], [8], so a method that reports a single number without stating its uncertainty presents a proxy as though it were a measurement.

Requirement 2: agreement separated from correctness. Agreement should be reported as an indicator in its own right and never as a measure of quality. It should be accompanied by a measure of agent independence, for which we propose the rate at which agents commit the same error on the calibration set, since majority voting carries evidence only where errors are uncorrelated. Both should be computed over the whole discussion trajectory and not the final round alone, because the diversity structure of a group affects idea exploration and consensus quality in different directions [6], and a score-based decision over the whole trajectory gave higher average accuracy than consensus-based baselines with last-round majority voting [5].

Requirement 3: per-agent attribution with explanations. A method should estimate the contribution of each agent to the conclusion and explain that estimate in a form a human can check. Shapley-based reasoning and LLM critics with textual explanations show that LLMs can produce such estimates [11], [12]. Problem 3 also shows that human inspection still outperforms automatic credit assignment [12], so a sample of the estimates should be verified by hand.

Requirement 4: audit of the judge. Before an LLM judge is used it should be screened for known biases, since all tested judges show significant biases [8], judgement bias is systematic even in strong judges [9], and fine-tuned judges lose performance outside their training scheme [10]. We propose that the screen act as a precondition: a judge whose sensitivity to injected bias exceeds the margin by which it separates candidate answers should not be used at all. Where the budget allows, several judges built on base models different from those of the agents should be used, and their verdicts calibrated against human assessments.

Two further properties should be reported, though they do not directly answer any of the obstacles above. First, evaluation should include runs in which some agents are unreliable or

adversarial, since dictator-based schemes are the first to collapse once agents become unreliable [2] and Free-MAD reports improved robustness under communication attacks [5]. Second, quality should also be reported together with its cost in rounds and tokens, beside a single-agent baseline, since sycophancy inflates both [3] and isolated self-correction gave a better cost-accuracy tradeoff than debate for unguided homogeneous teams of smaller models [1].

Table 1

Requirements for evaluating collective conclusions and the problem each one answers

Requirement	Origin	What a method must report	Sources
1. Partial reference with explicit uncertainty	Absence of a reference answer	An interval score and the calibration sample supporting it	[3, 6, 8]
2. Agreement separated from correctness	False consensus from conformity and correlated errors	Agreement and agent independence as distinct indicators, over the whole trajectory	[1, 3–6]
3. Per-agent attribution with explanations	Attribution of an individual agent's contribution	Contribution per agent, human-checkable, with a sample verified by hand	[7, 11, 12]
4. Audit of the judge	Unreliability of LLM-based evaluation	A bias screen as precondition, and several heterogeneous judges where affordable	[8–10]

Note. The cited works support the problems named in the second column. The requirements are our own synthesis and are not evaluated here; the calibration set, the agent-independence rate, the use of heterogeneous judges and the calibration of judges against human assessments are tested in none of the cited works.

V. Conclusion

Agreement among LLM agents measures the similarity of their answers. The quality of the conclusion they reach is a separate question, and the four problems of Section III explain why the first is a poor guide to the second. Those problems also compound: with no reference available, a biased judge becomes the sole arbiter of a consensus whose origin cannot be traced. Section IV therefore states constraints that any evaluation method should satisfy, leaving the method itself open. Our next step is to instantiate them as a protocol and a metric that score a collective conclusion without a full reference answer, together with an experiment that tests whether the score separates genuine agreement from conformity and what that separation costs in rounds and tokens. Until such a protocol exists, reported gains from multi-agent debate on open-ended tasks are best read as gains in agreement alone.

Our study is not exhaustive and the limits of the present analysis should be stated plainly. The four problems were drawn from twelve studies chosen for their relevance to collective decision-making, so a wider survey may well add further failure modes. The requirements say what an evaluation should report while leaving its computation open, and the calibration set they assume has yet to be tested at scale. Most of the evidence we rely on comes from tasks with known answers, the regime our own argument identifies as unrepresentative of open-ended work. Testing the requirements therefore demands a task set that is open-ended by construction, together with expert judgements collected independently of the agents, and assembling such a set is itself part of the work we propose.

References

[1] Bertalanič, B., & Fortuna, C. (2026). The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate. In *ACM Conference on AI and Agentic Systems (CAIS)*. https://doi.org/10.1145/3786335.3813137

[2] Zhao, X., Wang, K., & Peng, W. (2024). An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 2712–2727). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.158

[3] Pitre, P., Ramakrishnan, N., & Wang, X. (2025). CONSENSAGENT: Towards Efficient and Effective Consensus in Multi-Agent LLM Interactions Through Sycophancy Mitigation. In *Findings of the Association for Computational Linguistics: ACL (Findings of ACL)*. https://doi.org/10.18653/v1/2025.findings-acl.1141

[4] Chen, J. B., Saha, S., & Bansal, M. (2024). ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the conference - Association for Computational Linguistics*. https://doi.org/10.18653/v1/2024.acl-long.381

[5] Cui, Y., Fu, H., Zhang, H., Wang, L., & Zuo, C. (2026). Free-MAD: Consensus-Free Multi-Agent Debate. In *Findings of the Association for Computational Linguistics: ACL (Findings of ACL)*. https://doi.org/10.18653/v1/2026.findings-acl.1600

[6] Lee, N., & Kwon, O. (2026). Differentiated Effects of Agent Diversity on Collective Decision-Making in LLM-Based Multi-Agent Delphi Systems. *Applied Sciences*, *16*(13), 6715. https://doi.org/10.3390/app16136715

[7] Zhou, Z., Liu, Z., Liu, J., Wang, Y., Shao, Q., Xu, F., Jin, D., & Li, Y. (2026). Identifying Collective Intelligence Factor in LLM Agent Groups for Generalizable Multi-Agent System Design. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 12827–12842). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.624

[8] Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or LLMs as the Judge? A Study on Judgement Bias. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2024.emnlp-main.474

[9] Zhou, H., Huang, H., Zhang, R., Chen, K., Xu, B., Zhu, C., Zhao, T., & Yang, M. (2026). Toward robust LLM-based judges: Taxonomic bias evaluation and debiasing optimization. *Information Fusion*. https://doi.org/10.1016/j.inffus.2026.104801

[10] Huang, H., Bu, X., Zhou, H., Qu, Y., Liu, J., Yang, M., Xu, B., & Zhao, T. (2025). An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 5880–5895). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.306

[11] Hua, Y., Chen, H., Wang, S., Li, W., Wang, X., & Luo, J. (2025). Shapley-Coop: Credit Assignment for Emergent Cooperation in Self-Interested LLM Agents. In *Neural information processing systems*. https://doi.org/10.52202/085713-2965

[12] Nagpal, K., Dong, D., & Mehr, N. (2025). Leveraging Large Language Models for Effective and Explainable Multi-Agent Credit Assignment. In *Adaptive Agents and Multi-Agent Systems*. https://doi.org/10.65109/nokw1037