

Nordic AI Meet 2026

O4. LLMs & Agents in Society (Wed 13:00-14:00)

2. The Verification Gap: How Evaluation Harnesses Misjudge Multi-Agent LLM Systems Under Forced Communication

Authors: Kushagra Agrawal¹

Affiliations: 1. Faculty of Science and Engineering, Åbo Akademi University, Turku, Finland

Abstract: Multi-agent LLM systems (MAS) are usually assessed using harnesses inherited from single-agent benchmarks, which extract the final answer from the output of a designated agent. We demonstrate that this design choice can silently lead to severe misjudgments of MAS performance. We benchmark a 3-agent pipeline (Planner, Researcher, Reviewer) with early stopping disabled so that communication continues even after a solution is approved. We run 30 tasks each on GSM8K, HumanEval, and StrategyQA across 1, 2, 4, and 8 communication rounds. Under a naive extractor that reads only the Researcher's designated output slot (further details in Appendix A), task success collapses to near zero at high round counts (0–10% by round 8), reproducing the "communication inflation" failure pattern reported in previous work. However, a manual audit shows that as the number of rounds increases, agents increasingly abandon their assigned roles, and correct answers migrate to other agents' outputs, most commonly the Reviewer's critique, while the designated output slot remains blank. After correcting the extractor to check all agent outputs, 210 of 358 MAS trace-rounds (58.7%) flip from failure to success, and none flip the other way. Across all three domains, the corrected success rates are statistically indistinguishable across round counts (one-way ANOVA, all $p > 0.25$), and the semantic-drift-predicts-failure correlation observed under the naive extractor ($r = -0.24$) becomes negligible after correcting the scoring ($r = -0.05$, $p = 0.34$). What remains after this correction is a clear and substantial cost finding: token consumption increases by up to 200× compared to a single-agent Chain-of-Thought baseline, and token count and wall-clock latency remain near-perfectly correlated ($r = 0.98$), while corrected MAS accuracy never reliably exceeds the single-agent baseline enough to justify that cost. We argue that these results have two implications: (1) MAS benchmarking results that rely on single-agent-style, slot-based answer extraction should be interpreted with caution, since role degradation during prolonged interaction can make correct systems appear broken; and (2) even when scored fairly, forced over-communication provides no measurable accuracy benefit while incurring a substantial token tax. We situate these findings within the MAST failure taxonomy, arguing that verification-stage scoring fragility should be treated alongside the coordination and specification failures already documented by MAST, and we discuss the implications for how future MAS efficiency methods (e.g., S²-MAD and iMAD) should be benchmarked and compared.