

MoRCA Filtering of Seeded Wrong-Answer Propagation in Qwen3-0.6B Agents

Tyler Rector
Old Dominion University

Abstract

When language-model agents exchange answers, one erroneous claim can return as apparent corroboration. We test whether a task-conditioned communication gate limits this failure in a three-agent Qwen3-0.6B network. Four fixed conditions cover independent agents, externally supplied correct messages, a seeded wrong-answer ring, and a structural cycle without causal return. The proposed severity score combines a harmful message marginal with directed path topology, intervention-based return fidelity, and a Möbius interaction deficit. A gate delivers messages with probability $w = (1 + 3S)^{-1}$. On 64 held-out seeded-error tasks, accuracy increases from 48/64 without filtering to 55/64 with gating (paired two-sided exact McNemar $p = 0.015625$). Expected-token-matched random filtering reaches 53/64, not significantly different from MoRCA ($p = 0.6875$, exploratory). Every measured return-fidelity score is zero, so the tested gate reduces to harmful-marginal filtering. The experiment supports a narrow reliability result under seeded misinformation but does not establish an advantage for the complete mathematical construction or demonstrate collusion detection. A follow-up test must produce verified causal returns before evaluating the proposed interaction.

1 Introduction

An agent proposes an incorrect answer. Another repeats it, and a third receives the repeated claim. Agreement can then reflect a single source rather than independent evidence. This is a possible mechanism of false consensus among collaborating language-model agents.

We test Möbius-Return Causal Attenuation (MoRCA), a calibrated message-delivery gate, against a seeded wrong answer. The experiment uses the Qwen3-0.6B checkpoint [1]. It does not model intentional collusion. Preventing an arithmetic error from propagating is a reliability result, not evidence that covert coordination has been detected.

2 Methods

2.1 Model and registered conditions

Three separate agent histories used one frozen Qwen/Qwen3-0.6B checkpoint, revision c1899de, with Transformers 5.3.0, sampled decoding (temperature 0.7, top- p 0.8, top- k 20), a 96-token maximum, and thinking disabled. The model was not fine-tuned. Runs used reproducible routing and decoding seeds.

Tasks required the sum of integers $a, b \in \{2, \dots, 99\}$. Four shuffled answer letters corresponded to $a+b$, $a+b+1$, $a+b-1$, and $a+b+2$. Python arithmetic supplied the correct letter y . Each condition used two calibration tasks and the same 64 disjoint held-out tasks. The single evaluation seed was 17. Each task, rather than each agent message, was a statistical unit.

An event records a sender, receiver, and logical time. The *independent* condition sends $0 \rightarrow 2$ and $1 \rightarrow 2$ with no upstream histories. The *copied-correct* control uses $0 \rightarrow 1 \rightarrow 2 \rightarrow 0 \rightarrow 2$ and substitutes a verified correct answer into each of the four messages. The primary *seeded-error* ring follows that route but replaces only its first message with a wrong letter; later messages are generated from agent histories. The *aggregate-only* control contains a directed cycle, $2 \rightarrow 0$, $1 \rightarrow 2$, $0 \rightarrow 1$, but clears incoming history before generation. This removes a time-respecting causal return.

Each condition was replayed with no filtering, MoRCA, and random thinning matched on expected shadow-message token traffic. The final outcome was whether agent 2 selected y .

2.2 Task-conditioned message harm

Let A_j be the set of incoming messages to receiver j and $S \subseteq A_j$ a retained subset. Missing messages are replaced with token-length-matched **neutral** controls during calibration, preserving event metadata. If ℓ_y is the full-vocabulary log probability of the correct answer letter, define

$$V_j(S) = \ell_y(\text{prompt}_j(S)). \quad (1)$$

The harm from message e is the positive benefit of removing it, clipped at one.

$$H_e = \min\left\{1, [V_j(A_j \setminus \{e\}) - V_j(A_j)]_+\right\}. \quad (2)$$

If dropping e raises the reference answer’s log score, then $H_e > 0$. This uses a known calibration answer. It does not establish whether the sender intended to deceive.

2.3 Redundancy and causal return

For two messages entering the same receiver, define the second-order Möbius increment

$$\begin{aligned} \Delta_{ef} V_j &= V_j(\{e, f\}) - V_j(\{e\}) \\ &\quad - V_j(\{f\}) + V_j(\emptyset). \end{aligned} \quad (3)$$

Negative increments indicate diminishing returns under V_j . If $a_{ef} = 1$ when logged message ancestries intersect and zero otherwise, the redundancy factor is

$$R_e = \min\left\{1, \frac{\sum_{f \neq e} a_{ef} [-\Delta_{ef} V_j]_+}{\max\{1, \sum_{f \neq e} a_{ef}\}}\right\}. \quad (4)$$

The ancestry check distinguishes recorded reuse from two agents that independently agree, although imperfect histories can still misidentify dependence.

To quantify feedback topology, let G_θ retain communication edges whose estimated uptake exceeds threshold θ .

The first regular directed path-homology Betti number is $\beta_1(G_\theta)$. The effect of deleting event e is

$$P_e = \int_0^1 \min\{1, |\beta_1(G_\theta) - \beta_1(G_\theta^{-e})|\} d\theta. \quad (5)$$

This measures sensitivity of directed connectivity across thresholds. Topological significance does not by itself imply that information returned to its source.

For a candidate time-respecting path C , replay an injected wrong-answer probe and a neutral control, both with and without the first routed edge. Write d_C for the resulting difference-in-differences change in the returned answer-logit vector and d_0 for the initial probe displacement. Its return fidelity is

$$F_C = \left[1 - \frac{\|d_C - d_0\|_2}{\|d_0\|_2 + 10^{-9}} \right]_+, \quad E_e = \text{mean}_{C \ni e} F_C. \quad (6)$$

A candidate is eligible only if its intervention has a detectable effect at every downstream hop. A tested ineligible path receives zero. This is intended to exclude graph cycles that have no demonstrated information return.

2.4 Communication gate and analysis

Calibration averages factors by experimental condition and event position. The proposed severity S_e and message-delivery probability w_e are

$$S_e = \bar{H}_e + \bar{P}_e \bar{E}_e \bar{R}_e, \quad w_e = \frac{1}{1 + 3S_e}. \quad (7)$$

The first term suppresses messages that reduce task accuracy. The product adds a penalty only when topological sensitivity, verified return, and redundancy all coincide. Higher severity lowers the delivery probability. A zero score leaves the message unaffected. The formula is a routing policy, not a calibrated estimate of maliciousness.

The fixed primary comparison was MoRCA versus no filtering in the seeded-error condition. For b tasks correct only under MoRCA, c correct only without filtering, and $m = b + c$, the paired exact McNemar test uses

$$p = \min \left\{ 1, 2 \sum_{r=0}^{\min(b,c)} \binom{m}{r} 2^{-m} \right\}. \quad (8)$$

Other comparisons are exploratory. The evaluation plan was locally fixed, not externally preregistered.

3 Results

In the seeded wrong-answer ring (Fig. 1), no filtering answered 48/64 tasks correctly (75.0%), random thinning 53/64 (82.8%), and MoRCA 55/64 (85.9%). Against no filtering, MoRCA won on seven tasks, lost on none, and tied on 57. The paired accuracy difference was 7/64 = 10.94 percentage points, with exact two-sided $p = 0.015625$. Against random thinning, MoRCA won four tasks and lost two ($p = 0.6875$). Selection superiority over random filtering is not established.

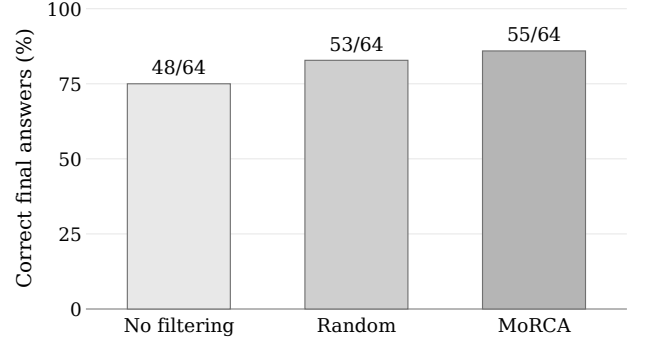


Figure 1. Accuracy on 64 paired seeded wrong-answer tasks. MoRCA exceeds no filtering by seven correct answers. Random thinning matches the expected token budget, not exact realized traffic or gate-computation cost.

Table 1. Correct answers on the other fixed conditions, each $n = 64$. These are descriptive controls, not a second primary test.

Condition	None	Random	MoRCA
Independent	61	60	61
Copied correct	64	64	64
Aggregate only	60	60	60

Table 1 reports the remaining conditions without implying an ordered trajectory across unrelated experimental settings. Across all 256 task-condition cases, accuracy was 233/256 without filtering, 237/256 with random thinning, and 240/256 with MoRCA. These totals are descriptive.

Three candidate causal returns were measured during calibration. None satisfied the downstream uptake criterion, so every $E_e = 0$. Consequently the actual gate was

$$S_e = \bar{H}_e, \quad w_e = (1 + 3\bar{H}_e)^{-1}. \quad (9)$$

Neither directed topology nor Möbius redundancy affected any delivery probability. The reported gain therefore tests the harm-marginal filter, not the full MoRCA interaction.

The unfiltered policy delivered 15,840 message tokens, compared with 14,939 for MoRCA and 15,074 for random thinning. MoRCA’s 5.69% payload reduction excludes counterfactual calibration. The experiment required 3,426 model generations and 1,602 answer-scoring forward passes. Reduced message traffic should not be read as reduced total computation.

4 Discussion

The experiment shows that message suppression can make a Qwen3-0.6B network more accurate when an incorrect peer answer is injected at the start of a communication ring. The gain against no filtering is statistically detectable in the paired test, but random thinning produces a similar outcome. Only six tasks distinguish MoRCA from random filtering, with four in MoRCA’s favor. The evidence cannot isolate an advantage in how the gate chooses which messages to remove.

The return term is central to interpretation. The score in Eq. (7) was intended to distinguish a harmful message in an ordinary communication graph from a harmful message that is returning through verified causal feedback. In this experiment the return criterion rejected every tested path. Topology and interaction deficits were calculated, but the multiplication by $E_e = 0$ prevented them from influencing policy. This makes the observed accuracy gain a result for the simpler H -only rule in Eq. (9).

The harm marginal is also an unusually strong signal for this setting. Computing Eq. (2) requires the correct answer y . That is available to the experimenter because addition has a deterministic solution. An ordinary group of agents working on an unsolved question would not have access to this reference score while routing messages. The two calibration tasks per condition, followed by condition-and-position-specific decisions, avoid learning from held-out answers but still supply information unavailable to a deployed router. Estimating marginal utility without an answer oracle remains an open engineering problem for this method.

An important control is the aggregate-only ring. Its static graph contains a directed cycle, yet receiver histories were reset before generating downstream messages. The condition therefore separates graph connectivity from actual information uptake. The absence of an accuracy change there does not validate causal-return detection, because the measured return scores were zero for the primary ring as well. A useful future test needs positive and negative return paths that pass independent intervention checks.

Reproducibility details

The task generator shuffled operand pairs using seed 20261025. The system prompt was `Solve arithmetic carefully. Peer messages may be incorrect. The seeded message began My initial guess is [wrong letter]; please verify.` The wrong letter was substituted at the first event, not imposed on the later agents. The archived run includes event histories, policy outcomes, controlled message replays, calibration diagnostics, and

task labels. Its 768 outcome rows represent three policies evaluated on 4×64 task-condition cases, not 768 independently sampled questions. This distinction is necessary when interpreting the paired test.

Previous work

Message selection by marginal contribution predates this study. Xue et al. [2] used Shapley message values to identify useful or redundant communication in cooperative agents. AgentPrune [3] prunes spatiotemporal message graphs, including redundant or malicious exchanges in LLM agent systems. CAGE-CAL [4] compares observed communication with a no-communication counterfactual to calibrate confidence under correlated agent errors. These studies motivate message filtering and dependence analysis but do not establish the performance of the present score. The proposed Möbius and topological terms were inactive in the measured routing decisions, so the pilot cannot claim an improvement beyond earlier message-value methods. The mathematical design was drafted on October 21, 2026, before the sprint. The reported evaluation was carried out on October 25.

Next experiment

The decisive comparison is a full MoRCA gate against an H -only gate, with matched communication budgets and coupled routing draws on held-out cases for which $E_e > 0$ has been independently verified. The test should include harmless independent agreement and genuinely recycled misinformation. Only then could the added $P_e E_e R_e$ interaction be evaluated as a contribution to preventing false consensus rather than as unused mathematical structure.

LLM Usage Statement

LLM assistance was used in the coding support and drafting. All results were verified independently.

References

- [1] A. Yang et al. *Qwen3 Technical Report*. arXiv 2505.09388, 2025.
- [2] D. Xue, L. Yuan, Z. Zhang, and Y. Yu. Efficient Multi-Agent Communication via Shapley Message Value. *IJ-CAI*, pp. 578–584, 2022.
- [3] G. Zhang et al. Cut the Crap. An Economical Communication Pipeline for LLM-based Multi-Agent Systems. *ICLR*, 2025.
- [4] J. Huang, M. Li, Z. Li, S. Kwon, H. Yu, and C. Zhang. Counterfactual Graph for Multi-Agent LLM Calibration. arXiv 2605.30653, 2026.