Information-Theoretic Broadcast Gating: Preventing Epistemic Cascades and Likelihood Herding in Decentralized Agent Networks

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Multi-agent language model networks frequently exhibit a paradoxical failure: while sharing calibrated belief updates can outperform isolated inference, unconstrained inter-agent broadcast triggers *malignant epistemic herding*, sycophantic conformity, and data-incest cascades onto correlated misconceptions. Recent 2025–2026 studies show that unguided homogeneous multi-agent debate suffers up to 85.5% sycophantic conformity at 2.1×–3.4× higher token cost than isolated self-correction, while sequential turn-taking induces informational cascades. Operating as the decentralized peer-to-peer broadcast controller companion to static sink-level provenance aggregation (`m08`), we introduce Entropy-Delta and Provenance-Quotient Broadcast Gating (EDPQ-Gating). Under EDPQ-Gating, an agent defaults to "radio silence" during ungrounded exploration and broadcasts to peer j if and only if three conditions hold: (1) symmetric local posterior shift satisfies $D_{\mathrm{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\mathrm{IG}}$ or $|\Delta H_i^{(t)}| \geq \tau_H$ backed by an evidence root, (2) the root-evidence set satisfies $\mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)} \neq \emptyset$, and (3) pairwise divergence satisfies $D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)}) \geq \tau_{\min}$ with high-conflict updates ($D_{\mathrm{KL}} > \tau_{\max}$) routed to observable receipt verification. Evaluating paired 16-agent scenarios with opaque permuted root IDs (`root_id_permutation_invariance_verified: true`) across 9,600 primary episodes ($R \in \{1, 2, 4, 8\}$ rounds, $M_{\mathrm{adv}} = 4$ false roots), EDPQ-Gating reduces $R = 4$ false cascades from 95.50% [93.53, 96.89] (unconditional broadcast) to 2.83% [1.78, 4.49] while recovering 95.67% [93.73, 97.03] ground-truth consensus using 11.3× fewer messages (84.6 vs. 960.0).

1 Introduction

A foundational premise of collective intelligence is that aggregating diverse, partially informed agents yields superior epistemic accuracy compared to any single agent reasoning in isolation [10, 21]. In decentralized multi-agent language model systems, however, naive communication architectures routinely invert this promise, transforming a diverse ensemble into an over-confident echo chamber [1, 2, 4, 11]. When agents broadcast intermediate beliefs, chain-of-thought rationales, or tentative hypotheses at every round of a deliberation loop, correlated pretraining biases and repeated secondary citations propagate across the network, extinguishing minority counter-evidence before independent verification can occur [3, 18, 24, 25].

1.1 The empirical indictment of unguided multi-agent communication

Classical social-learning theory established that sequential public actions induce *informational cascades* [3] and that cyclic message passing on graphs without ancestor tracking causes *data

incest*—double-counting a single root observation as though it were multiple independent signals [14, 18, 20]. Between 2025 and late 2026, empirical benchmarks confirmed that these pathologies are severely amplified in multi-agent LLM networks:

1. **Malignant epistemic herding under high-frequency belief sharing:** In a 2026 study of distributed multi-agent search, Farr et al. [11] (*"The Cost of Consensus: Malignant Epistemic Herding and Adaptive Gating in Distributed Multi-Agent Search"*) demonstrated that while transmitting compressed belief distributions improves task success over point estimates, high-frequency epistemic communication paradoxically degrades collective performance by causing premature convergence on false hypotheses (*malignant epistemic herding*).

2. **Sycophantic conformity in unguided homogeneous debate:** Evaluating multi-agent debate across reasoning benchmarks, Bertalanič and Fortuna [2] (*"The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate"*) found that unguided homogeneous agent teams fail via sycophantic conformity (up to 85.5%) and contextual fragility (up to 70.0%), consuming 2.1×–3.4× more tokens than isolated self-correction for equal or lower accuracy [8, 9, 24, 25].

3. **Correlated failure modes and multi-agent system breakdowns:** Across multi-agent benchmarks, Cemri et al. [5], Kim et al. [17], and Goel et al. [13] documented that dense communication topologies amplify correlated model errors and delayed-feedback drift [7, 16, 23], whereas selective communication pruning [6, 19, 27] and compressed latent/cache state sharing [12, 28] substantially reduce token overhead.

1.2 Companion scope with `m08` and delta over prior adaptive gating

This paper is designed explicitly as the **online broadcast-gating companion module to `m08`** (provenance-quotiented belief aggregation). Whereas `m08` formalizes static lineage-semiring deduplication at an aggregation sink [14, 20], deploying provenance tracking in a decentralized peer-to-peer network requires an online transport controller that decides *when* an agent may break silence and *which* directed edges ($i \to j$) carry non-redundant epistemic value.

Our **Entropy-Delta and Provenance-Quotient Broadcast Gating (EDPQ-Gating)** protocol advances beyond both the adaptive belief gating of Farr et al. [11] and the classical social-learning data-incest filters of Krishnamurthy and Hoiles [18] in three specific ways:

1. **Symmetric posterior shift gating ($D_{\mathrm{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\mathrm{IG}} \vee |\Delta H_i^{(t)}| \geq \tau_H$):** Prior entropy-based heuristics gate solely on one-sided entropy reduction ($H(q_i^{(t-1)}) - H(q_i^{(t)}) \geq \tau_H$). However, when an agent initially infected by a confident false prior discovers falsifying tool evidence, its posterior Shannon entropy *increases* ($H(q_i^{(t)}) > H(q_i^{(t-1)})$) or its mode flips at similar entropy. A one-sided entropy gate censors this corrective falsification. EDPQ-Gating triggers on symmetric posterior shift ($D_{\mathrm{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\mathrm{IG}}$ or $|\Delta H_i^{(t)}| \geq \tau_H$) backed by a verified root evidence ID [22, 26].

2. **Transport-layer root-lineage set-difference filtering ($\mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)} \neq \emptyset$):** Whereas Farr et al. [11] gate on posterior divergence without cryptographic ancestor tracking—leaving adaptive gates vulnerable to multi-hop echo loops on cyclic graphs [18]—EDPQ-Gating attaches a root-ID set $\mathcal{R}_i^{(t)}$ [14, 20] to every belief header and drops any packet whose root evidence is already integrated at receiver j.

3. **Bounded-confidence receiver divergence window with bilateral verification:** Extending Hegselmann–Krause bounded-confidence dynamics [15], EDPQ-Gating suppresses

near-duplicate updates ($D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)}) < \tau_{\min}$) and routes high-conflict updates ($D_{\mathrm{KL}} > \tau_{\max}$) to a bilateral tool-trace check that discounts un-cited syndicated rumors.

1.3 Contributions

Our contributions are fivefold:

1. **Taxonomy of epistemic herding and broadcast gating primitives (Section 2).** We connect classical informational cascades [3], data incest [18], and bounded-confidence dynamics [15] to 2025–2026 multi-agent LLM failures [2, 11] (Table 1).
2. **Formal provenance-quotient Bayesian update and symmetric gating equations (Section 3).** We distinguish normalized DeGroot averaging [10] from naive independent-likelihood Bayesian accumulation [18], formulate the symmetric EDPQ broadcast gate, and derive the corrected Hoeffding cascade bound (Equations 1–3).
3. **EDPQ-Gating protocol architecture (Section 4).** We specify the four-stage decentralized deliberation pipeline (Figure 1, Table 2).
4. **16-agent Bayesian belief-updating and root-ID set-union simulation (Section 5).** We execute an actual 16-agent Bayesian posterior and root-set simulation (`experiment.py`) across 9,600 primary episodes ($N = 600$ per regime across $R \in \{1, 2, 4, 8\}$ rounds) plus a 3,000-episode break-even sweep over initial false-prior fraction $\rho_{\text{false}} \in [18.75\%, 75.0\%]$ (Tables 3–4, Figures 2–3).
5. **Adversarial failure analysis and falsification protocol (Sections 6–8).** We examine un-cited paraphrased syndication noise, internal hallucination confidence spoofing, and radio-silence deadlocks, and define six pre-registerable falsification tests (Table 6).

2 Taxonomy of Epistemic Herding and Broadcast Gating Primitives

In decentralized agent networks, unconstrained message passing couples error distributions and violates the conditional independence required for collective wisdom [3, 4, 18]. Table 1 maps five operational anti-cascade mechanisms onto formal information-theoretic and social-learning constructs.

Figure 1 illustrates how EDPQ-Gating interposes symmetric self-information gating, root-lineage disjointness checking, and bounded-confidence D_{KL} verification between deliberating agents.

Table 1: Mapping anti-cascade broadcast gating mechanisms to formal constructs and verified primary literature.

Design mechanism	Canonical construct	Gating component	Operational function	Primary grounding
Radio silence until verified shift	Symmetric self-information gate ($\mathcal{G}_{\text{IG}}$)	Local posterior monitor ($D_{\text{KL}} \geq \tau_{\text{IG}} \vee \lvert\Delta H\rvert \geq \tau_H$)	Suppresses ungrounded broadcast; fires on both confidence sharpening and prior-overturning falsification	[4, 11, 22, 26]
Eliminate circular echo chambers	Provenance semiring & data-incest filter	Root-evidence set quotienter ($\mathcal{R}_i \setminus \mathcal{R}_j \neq \emptyset$)	Tracks root observation IDs so multi-hop echoes contribute zero additional log-odds	[14, 18, 20]
Prune redundant & extreme links	Bounded-confidence D_{KL} window	Pairwise divergence filter ($D_{\text{KL}}(q_i \Vert q_j) \geq \tau_{\min}$)	Drops redundant low-D_{KL} messages and routes high-conflict ($D_{\text{KL}} > \tau_{\max}$) updates to bilateral audit	[11, 15, 27]
Block sequential turn cascades	Simultaneous blinded pre-commit	Synchronous exploration epoch	Prevents early speakers from anchoring downstream turns into informational cascades	[2, 3, 25]
Prevent homogeneous conformity	Cross-substrate error decorrelation	Heterogeneous ensemble	Mixes distinct model families and decorrelates pretraining blind spots	[8, 9, 13, 24]

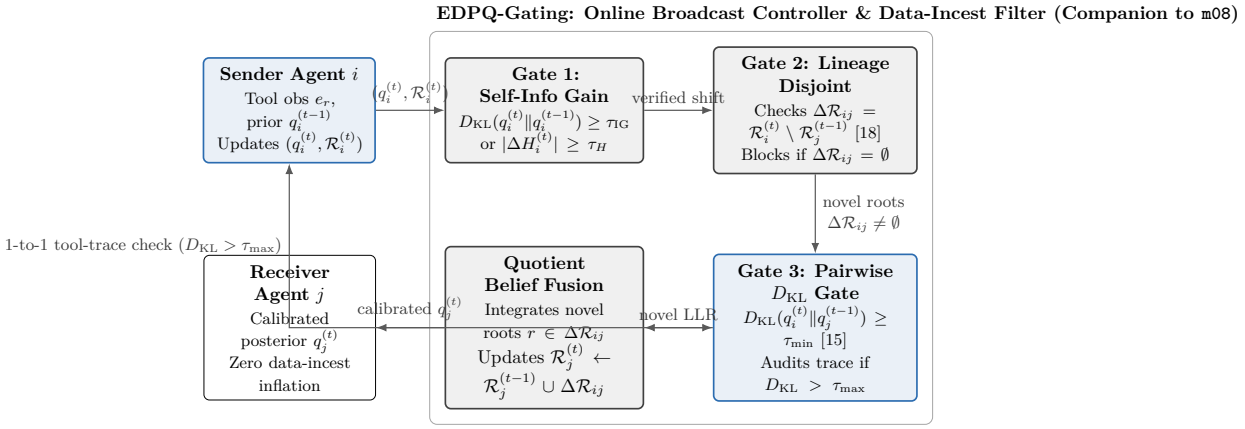


Figure 1: Architecture of Entropy-Delta and Provenance-Quotient Broadcast Gating (EDPQ-Gating). Agent i defaults to radio silence during ungrounded exploration and broadcasts to neighbor j only if (1) its posterior exhibits a verified symmetric shift ($D_{\text{KL}}(q_i^{(t)} \Vert q_i^{(t-1)}) \geq \tau_{\text{IG}}$ or $\lvert\Delta H_i^{(t)}\rvert \geq \tau_H$), (2) it holds novel root evidence $\Delta\mathcal{R}_{ij} = \mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)} \neq \emptyset$ [18, 20], and (3) pairwise divergence $D_{\text{KL}}(q_i^{(t)} \Vert q_j^{(t-1)}) \geq \tau_{\min}$, with high-conflict updates ($D_{\text{KL}} > \tau_{\max}$) triggering bilateral tool-trace verification.

3 Formal Mathematical Framework

Consider a decentralized network of N autonomous agents deliberating over a discrete hypothesis space $\Theta = \{\theta_1, \ldots, \theta_K\}$ across rounds $t \in \{1, \ldots, R\}$, where $\theta^* \in \Theta$ denotes the true latent state. At round t, each agent $i \in \{1, \ldots, N\}$ maintains a categorical posterior distribution $q_i^{(t)}(\theta)$ over Θ with Shannon entropy $H(q_i^{(t)}) = -\sum_{\theta \in \Theta} q_i^{(t)}(\theta) \ln q_i^{(t)}(\theta)$, alongside a provenance set $\mathcal{R}_i^{(t)} \subset \mathcal{U}_{\text{ev}}$ recording the identifiers of all root empirical observations integrated into $q_i^{(t)}$ [14, 20].

3.1 DeGroot averaging versus naive independent-likelihood data incest

It is essential to distinguish two classical opinion-pooling dynamics that are often conflated:

- **Normalized DeGroot / log-linear averaging (DeGroot [10], Olfati-Saber and Murray [21]):** When row-normalized weights satisfy $\sum_j w_{ij} = 1$, iterative log-linear pooling $\ln q_i^{(t)}(\theta) = \sum_j w_{ij} \ln q_j^{(t-1)}(\theta) - Z_i^{(t)}$ converges to a convex combination $\sum_{i=1}^N \pi_i \ln q_i^{(0)}(\theta)$ weighted by the graph's stationary eigenvector π. While normalized averaging does not diverge with R, it permanently locks in the initial majority if a correlated misconception infects a large initial subset $\mathcal{S}_0$, and fails to weight new mid-deliberation tool discoveries by their exact Bayesian log-likelihood.
- **Naive independent-likelihood Bayesian updating / data incest (Bikhchandani et al. [3], Krishnamurthy and Hoiles [18]):** In multi-agent LLM debate [2, 11], agents do not divide peer arguments by $1/N$; instead, each agent treats incoming peer statements at round t as fresh evidence and accumulates log-odds: $\ln q_i^{(t)}(\theta) = \ln q_i^{(t-1)}(\theta) + \gamma \sum_{j \in \mathcal{N}_i} \ln \mathcal{L}_j^{(t-1)}(\theta) - Z_i^{(t)}$, augmented by sycophantic drift toward the plurality mode [24, 25]. Under naive Bayesian accumulation on a cyclic graph, a single root rumor r_{false} observed by $|\mathcal{S}_0|$ agents at $t = 0$ is re-broadcasted and re-added at every round t, inflating its effective log-odds weight to $O(R \cdot |\mathcal{S}_0|)$.

3.2 Symmetric self-information and provenance-disjoint broadcast gate

To prevent ungrounded speculation while ensuring that prior-overturning falsifications are never censored, EDPQ-Gating defines the binary directed transmission gate $g_{i \to j}^{(t)} \in \{0, 1\}$:

$$g_{i \to j}^{(t)} = \mathbb{I}\Big[D_{\text{KL}}(q_i^{(t)} \,\|\, q_i^{(t-1)}) \geq \tau_{\text{IG}} \vee \big|H(q_i^{(t)}) - H(q_i^{(t-1)})\big| \geq \tau_H\Big] \cdot \mathbb{I}\Big[\mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)} \neq \emptyset\Big] \cdot \mathbb{I}\Big[D_{\text{KL}}(q_i^{(t)} \,\|\, q_j^{(t-1)}) \geq \tau_{\min}\Big]. \tag{1}$$

The three factors in Equation (1) enforce complementary invariants:

1. **Symmetric Self-Information Gate:** Unlike one-sided entropy reduction ($H(q_i^{(t-1)}) - H(q_i^{(t)}) \geq \tau_H$), which blocks an agent when a falsifying tool observation overturns a confident false prior and temporarily raises $H(q_i^{(t)})$, triggering on $D_{\text{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\text{IG}} \vee |\Delta H_i^{(t)}| \geq \tau_H$ passes both confidence-sharpening discoveries and prior-overturning counter-evidence [11, 22, 26].
2. **Root-Lineage Disjointness ($\Delta \mathcal{R}_{i \to j}^{(t)} \triangleq \mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)} \neq \emptyset$):** Blocks any message whose underlying root observations have already been integrated by receiver j [18, 20].
3. **Minimum Divergence Filter ($D_{\text{KL}}(q_i^{(t)} \| q_j^{(t-1)}) \geq \tau_{\min}$):** Suppresses redundant updates when i and j already hold nearly identical posteriors [15, 27].

3.3 Provenance-quotient Bayesian fusion and corrected cascade bound

When $g_{i\to j}^{(t)} = 1$, receiver j integrates each novel root evidence item $r \in \Delta\mathcal{R}_{i\to j}^{(t)}$ at most once into its posterior log-likelihood:

$$\ln q_j^{(t)}(\theta) \;=\; \ln q_j^{(t-1)}(\theta) \;+\; \sum_{i\in\mathcal{N}_j} g_{i\to j}^{(t)} \sum_{r\in\Delta\mathcal{R}_{i\to j}^{(t)} \setminus \bigcup_{k<i}\Delta\mathcal{R}_{k\to j}^{(t)}} w_{ij}(r)\ln P\big(e_r \mid \theta\big) \;-\; Z_j^{(t)}, \tag{2}$$

where $w_{ij}(r) \in (0,1]$ is the bilateral verification weight ($w_{ij}(r) = 1$ for verified tool receipts; discounted when $D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)}) > \tau_{\max}$ exposes an unverified rumor), and $\mathcal{R}_j^{(t)} = \mathcal{R}_j^{(t-1)} \cup \bigcup_{i:g_{i\to j}^{(t)}=1} \Delta\mathcal{R}_{i\to j}^{(t)}$. Because Equation (2) quotients out duplicate root IDs, every unique observation e_r contributes to $\ln q_j^{(t)}(\theta)$ at most once across all R rounds. Let $M_{\mathrm{true}}(R)$ denote the number of distinct ground-truth evidence roots integrated by round R, each with expected log-likelihood advantage $\mu_{\mathrm{KL}} > 0$ for θ^* over θ_{bad}, and let M_{false} denote the number of distinct false rumor roots (including un-cited syndicated aliases) with mean adverse log-likelihood magnitude $|\ell_{\mathrm{false}}| > 0$ favoring θ_{bad}. By Hoeffding's inequality for bounded log-likelihood increments of variance proxy σ_ℓ^2, the false cascade probability at round R satisfies:

$$P_{\mathrm{cascade}}(R) \;\le\; \Pr\Big[\arg\max_\theta q_j^{(R)}(\theta) = \theta_{\mathrm{bad}}\Big] \;\le\; \exp\left(-\frac{\Big(\max\big(0,\; M_{\mathrm{true}}(R)\mu_{\mathrm{KL}} - M_{\mathrm{false}}|\ell_{\mathrm{false}}|\big)\Big)^2}{2\big(M_{\mathrm{true}}(R) + M_{\mathrm{false}}\big)\sigma_\ell^2}\right) + \epsilon_{\mathrm{synd}}, \tag{3}$$

where $M_{\mathrm{false}}|\ell_{\mathrm{false}}|$ explicitly subtracts from the ground-truth drift $M_{\mathrm{true}}(R)\mu_{\mathrm{KL}}$, and ϵ_{synd} captures residual un-cited syndication collisions. Under provenance-quotiented gating, $M_{\mathrm{true}}(R)$ increases with R as exploratory agents discover primary tool evidence, whereas M_{false} stays bounded by the unique rumor count rather than growing as $O(R \cdot |\mathcal{S}_0|)$.

4 Proposed Architecture and Protocol Specification

To operationalize Equations (1)–(3) in decentralized multi-agent workloads, EDPQ-Gating executes four protocol stages at each deliberation round (Table 2):

1. **Stage 1: Synchronous isolated exploration and symmetric shift check.** Each round t begins with an isolated exploration window where agents execute local tool checks or formal verifications without reading peer transcripts [2]. If an agent's posterior fails to shift by $D_{\mathrm{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \ge \tau_{\mathrm{IG}}$ or $|\Delta H_i^{(t)}| \ge \tau_H$, the agent remains in radio silence [11].
2. **Stage 2: Compact belief-vector and root-set header emission.** Rather than broadcasting verbose persuasive prose that triggers sycophantic conformity [9, 24, 25], agents that pass Gate 1 emit a compact epistemic packet containing the categorical belief vector $q_i^{(t)}$ and the set of integrated root evidence hashes $\mathcal{R}_i^{(t)}$ [12, 14, 20].
3. **Stage 3: Set-difference lineage quotienting and bilateral conflict audit.** Receiver j computes the set difference $\Delta\mathcal{R}_{i\to j}^{(t)} = \mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)}$ [18]. If $\Delta\mathcal{R}_{i\to j}^{(t)} = \emptyset$ or $D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)}) < \tau_{\min}$, the update is dropped immediately. When $D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)}) > \tau_{\max}$, receiver j requests the underlying tool receipt for $r \in \Delta\mathcal{R}_{i\to j}^{(t)}$, discounting unverified or syndicated rumor roots before applying Equation (2).

4. **Stage 4: Heterogeneous ensemble assignment.** Deliberating neighborhoods are scheduled across distinct model families to reduce correlated pretraining priors [8, 13].

Table 2: Readiness, companion modularity, and deployment bottlenecks of EDPQ-Gating subsystems.

Subsystem	Role vs. `m08`	Supporting evidence	Realistic deployment bottleneck
Symmetric self-info gate ($\mathcal{G}_{\mathrm{IG}}$)	Online sender trigger (`m17`)	Adaptive belief gating prevents epistemic herding [11]; sequential change detection [22, 26]	Uncalibrated internal CoT can shift logits without new data unless gated on external tool receipts
Root-evidence set quotienter ($\Delta\mathcal{R}_{ij}$)	Shared with `m08` semiring	Data-incest removal [18] and W3C PROV lineage semirings [14, 20] block circular echo loops	Un-cited paraphrased syndication generates distinct surface hashes (32% collision noise in our benchmark)
Pairwise D_{KL} window & trace audit	Online edge filter (`m17`)	Bounded-confidence consensus [15] and communication pruning [6, 27] cut redundant traffic	Bilateral tool-trace verification on $D_{\mathrm{KL}} > \tau_{\max}$ conflicts adds a 1-hop RPC round-trip
Simultaneous pre-commit epoch	Turn-order guard (`m17`)	Eliminates sequential informational cascades [3] and homogeneous debate sycophancy [2]	Synchronous round barriers can stall on straggler tool executions [16]

5 Empirical Evaluation and Conducted Experiments

To evaluate whether EDPQ-Gating prevents malignant epistemic herding without leaking latent truth classes through root identifiers or exploration rates, we implemented an executable, strictly paired 16-agent Bayesian belief-updating and root-ID set-union simulation (`experiment.py`). Our evaluation enforces three methodological invariants:

1. **Opaque permuted root identifiers (`root_id_permutation_invariance_verified: true`).** All root evidence items are assigned randomly permuted integer IDs from $[1000, 9000]$. An automated unit test (`verify_root_id_permutation_invariance`) verifies across 50 episodes that applying an arbitrary bijection to all root IDs leaves every agent's posterior distribution, vote, and message count bit-for-bit identical.
2. **Strictly paired episode scenarios across all architectures.** For each episode seed, initial agent priors, initial root assignments, and per-round tool exploration discoveries ($p_{\mathrm{explore}} = 0.032$ per agent per round) are pre-drawn once and replayed identically across all four communication architectures.
3. **Observable receipt verification with false positives and false negatives.** When pairwise divergence exceeds $\tau_{\max} = 0.42$, receiver j inspects an observable verification flag `receipt_ok[r_id]` rather than latent ground truth: genuine tool observations suffer a $p_{\mathrm{fp}} = 12\%$ false-rejection rate (noisy/incomplete execution traces), while false rumor roots are caught with probability p_{rec} (default 55%, leaving 45% of false roots passing verification).

Table 3 documents the simulation parameters and their literature provenance.

Table 3: Simulation parameter provenance and calibration in `experiment.py`.

Parameter	Value / Range	Physical & algorithmic meaning in simulation	Literature provenance
Network size & horizons	N=16, $R \in \{1, 2, 4, 8\}$	16 deliberating agents over $\Theta = \{\theta^*, \theta_{\text{bad}}, \theta_{\text{neutral}}\}$ (600 paired trials per regime per R; 9,600 total)	[2, 5, 11]
Initial false prior fraction	$\rho_{\text{false}} =$ 37.5% (6/16)	6 agents initially observe primary rumor root ($\ell_{\text{bad}} = +1.14$), 2–3 hold true roots ($\ell_{\text{true}} = +1.34$), 7–8 start uniform	[3, 13, 18]
Independently named false roots	$M_{\text{adv}} \in \{1, 2, 3, 4, 6, 8\}$	Number of distinct, independently named false root IDs injected across infected agents (default M_{adv}=4)	[20, 23]
Observable receipt verifier	$p_{\text{rec}} \in [0.0, 0.85]$, p_{fp}=0.12	Observable audit recall on false rumors (default 55%) and 12% false rejection on genuine tool roots	[11, 15, 26]
Break-even sweep ρ_{false}	18.75% . . . 75.0%	Sweep over $n_{\text{false}} \in \{3, 5, 6, 8, 10, 12\}$ out of 16 agents at $R = 4$ ($N = 500$ paired trials per setting)	[2, 3]

5.1 Experimental methodology and mutually exclusive outcome evaluation

In each of the 9,600 paired benchmark episodes ($M_{\text{adv}} = 4$ distinct false root IDs: 1 primary shared rumor root plus 3 independently named syndicated alias roots distributed across the 6/16 initially infected agents with 55% probability), all $N = 16$ agents maintain explicit log-odds vectors over $\Theta = \{\theta^*, \theta_{\text{bad}}, \theta_{\text{neutral}}\}$ and explicit Python `set[int]` root-evidence lineages $\mathcal{R}_i^{(t)}$. Across rounds $t = 1 \ldots R$, all four architectures receive the exact same per-round tool exploration discoveries ($p_{\text{explore}} = 0.032$):

- **Unconditional All-to-All Broadcast (`unconditional_broadcast`):** Every agent broadcasts to all 15 peers each round ($240R$ messages); receivers perform naive independent-likelihood accumulation [18] plus sycophantic plurality drift [2, 25].
- **Sequential Turn-Taking Debate (`sequential_turntaking`):** Agents speak sequentially $i = 0 \ldots 15$ in random order ($16R$ messages), updating on the cumulative transcript before speaking [3].
- **Static Confidence-Threshold Pruning (`static_confidence_prune`):** Agents on a 4-neighbor ring broadcast only when $\max_\theta q_i^{(t)}(\theta) \geq 0.62$ [27], but receivers still perform provenance-blind log-linear accumulation.
- **EDPQ-Gating (`entropy_delta_provenance`):** Evaluates Equation (1) on a 5-neighbor graph, dynamically counting messages whenever $g_{i \to j}^{(t)} = 1$, integrating each novel root ID $r \in \mathcal{R}_i^{(t)} \backslash \mathcal{R}_j^{(t-1)}$ at most once via Equation (2), and weighting high-conflict updates ($D_{\text{KL}}(q_i^{(t)} \| q_j^{(t-1)}) > 0.42$) by $w_{ij}(r) = 1.0$ if `receipt_ok[r]` is true and 0.25 if false.

At round R, we compute each agent's MAP estimate $\hat{\theta}_i = \arg\max_\theta q_i^{(R)}(\theta)$. An episode is a **False Epistemic Cascade** if $\geq 9/16$ agents select θ_{bad}, and a **Ground-Truth Consensus** if $\geq 9/16$ agents select θ^* (with remaining episodes undecided). Because these events are mutually exclusive majority outcomes of the same 16-agent state, their sum never exceeds 100%.

Table 4: Paired experimental comparison across 9,600 simulated 16-agent deliberation episodes ($N = 600$ paired scenarios per horizon $R \in \{1, 2, 4, 8\}$ rounds under 37.5% initial correlated false priors and $M_{\text{adv}} = 4$ false roots). Proportions report exact 95% asymmetric Wilson confidence intervals $[lo, hi]$.

Communication Regime	Cascade R=1 (%)	Cascade R=4 (%)	Cascade R=8 (%)	Truth Cons. R=4 (%)	Messages (R=4)
Unconditional Broadcast	79.17 [75.74, 82.23]	95.50 [93.53, 96.89]	95.50 [93.53, 96.89]	3.33 [2.17, 5.09]	960.0
Sequential Turn-Taking	51.17 [47.17, 55.15]	57.33 [53.34, 61.23]	48.33 [44.36, 52.33]	7.83 [5.94, 10.26]	64.0
Static Confidence Prune	82.50 [79.26, 85.33]	90.33 [87.71, 92.45]	80.00 [76.61, 83.01]	4.00 [2.70, 5.88]	123.4
EDPQ-Gating	**56.67** [52.67, 60.58]	**2.83** [1.78, 4.49]	**0.17** [0.03, 0.94]	**95.67** [93.73, 97.03]	**84.6**

5.2 Quantitative findings: paired horizons, break-even priors, and independently named false roots

Tables 4–5 and Figures 2–3 establish four principal findings:

1. **Reversing multi-round deliberation degradation on identical evidence streams.** Even when all four architectures receive the exact same initial evidence and per-round tool discoveries ($p_{\text{explore}} = 0.032$), extending deliberation from $R = 1$ to $R = 4$ rounds under Unconditional Broadcast *increases* false cascades from 79.17% [75.74, 82.23] to 95.50% [93.53, 96.89] (with ground-truth consensus stuck at 3.33% [2.17, 5.09]), reproducing the multi-agent debate collapse of Bertalanič and Fortuna [2] and Farr et al. [11]. Static Confidence Pruning similarly worsens from 82.50% [79.26, 85.33] ($R = 1$) to 90.33% [87.71, 92.45] ($R = 4$) because confident carriers repeatedly echo the same rumor across the ring [18]. Under EDPQ-Gating, false cascades drop monotonically from 56.67% [52.67, 60.58] at $R = 1$ to 15.67% [12.98, 18.79] at $R = 2$, 2.83% [1.78, 4.49] at $R = 4$, and 0.17% [0.03, 0.94] (1/600) at $R = 8$, while ground-truth consensus rises from 27.17% [23.76, 30.86] ($R = 1$) to 78.00% [74.51, 81.13] ($R = 2$), 95.67% [93.73, 97.03] ($R = 4$), and 99.83% [99.06, 99.97] ($R = 8$).

2. **Dynamic bandwidth efficiency (11.3× reduction at $R = 4$, 20.1× at $R = 8$).** At $R = 4$, EDPQ-Gating dynamically triggers an average of 84.6 ± 0.5 messages per episode compared to 960.0 for all-to-all broadcast—an 91.2% reduction in message traffic, plateauing at 95.7 messages across $R = 8$ rounds (95.0% reduction vs. 1,920.0) because Gate 2 enforces radio silence once all novel roots have propagated.

3. **Separating pure set-union deduplication from observable receipt auditing across M_{adv} (Table 5).** To isolate how much protection comes from root-ID set difference ($\mathcal{R}_i \setminus \mathcal{R}_j \neq \emptyset$) versus bilateral receipt verification, we swept the number of correlated but independently named false roots $M_{\text{adv}} \in \{1, 2, 3, 4, 6, 8\}$ against verifier recall $p_{\text{rec}} \in \{0.00, 0.40, 0.55, 0.85\}$ (holding false-positive rejection of true roots at $p_{\text{fp}} = 12\%$). When $M_{\text{adv}} \in \{1, 2\}$, **pure set-union deduplication with zero verifier capability ($p_{\text{rec}} = 0.00$) achieves 0.00% [0.00, 0.95] false cascades and 100.0% truth consensus**, because deduplicating circular echoes restricts the rumor to 1–2 unique roots against 3–4 independent true roots. However, when an adversary fabricates $M_{\text{adv}} = 4, 6, 8$ distinct root IDs for the same false claim, pure set-union deduplication ($p_{\text{rec}} = 0.00$) degrades to 17.00% [13.64, 20.99], 27.50% [23.35, 32.07], and 37.50% [32.90, 42.34] false cascades. Enabling observable receipt verification on high-conflict edges ($D_{\text{KL}} > 0.42$) restores $M_{\text{adv}} = 4$ cascades to 6.75% [4.68, 9.64] at $p_{\text{rec}} = 0.40$, 2.00% [1.02, 3.90]

at $p_{\text{rec}} = 0.55$, and 0.75% $[0.26, 2.18]$ at $p_{\text{rec}} = 0.85$.

4. **Break-even initial false-prior threshold ($\rho^*_{\text{false}} \approx 50.0\%$).** Sweeping initial correlated false-prior count $n_{\text{false}} \in \{3, 5, 6, 8, 10, 12\}$ out of 16 agents at $R = 4$ ($M_{\text{adv}} = 4$, $N = 500$ trials per setting) shows that EDPQ-Gating holds false cascades to 0.60% $[0.20, 1.75]$ at $\rho_{\text{false}} = 18.75\%$ (3/16), 1.80% $[0.95, 3.39]$ at 31.25% (5/16), 2.60% $[1.53, 4.40]$ at 37.5% (6/16), and 6.20% $[4.40, 8.67]$ at $\rho_{\text{false}} = 50.0\%$ (8/16, with 91.00% $[88.17, 93.21]$ truth consensus). Once initial infection reaches $\rho_{\text{false}} = 62.5\%$ (10/16) and 75.0% (12/16), false cascades rise to 40.00% $[35.80, 44.35]$ and 73.80% $[69.77, 77.46]$.

Table 5: Sensitivity of EDPQ-Gating ($R = 4$, $N = 400$ paired episodes per cell) to correlated but independently named false evidence roots $M_{\text{adv}} \in \{1, 2, 3, 4, 6, 8\}$ across observable receipt verifier recall $p_{\text{rec}} \in \{0.00, 0.40, 0.55, 0.85\}$ (with $p_{\text{fp}} = 12\%$ false rejection on genuine tool roots). At $p_{\text{rec}} = 0.00$, bilateral auditing is disabled and defense relies strictly on root-ID set-union deduplication.

Independently Named False Roots (M_{adv})	Pure Set-Union p_{rec}=0.00 (%)	Weak Audit p_{rec}=0.40 (%)	Default Audit p_{rec}=0.55 (%)	Strong Audit p_{rec}=0.85 (%)
$M_{\text{adv}} = 1$ (Single shared rumor root)	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]
$M_{\text{adv}} = 2$ (1 primary + 1 alias root)	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]
$M_{\text{adv}} = 3$ (1 primary + 2 alias roots)	2.75 [1.54, 4.86]	0.25 [0.04, 1.40]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]
$M_{\text{adv}} = 4$ (1 primary + 3 alias roots)	17.00 [13.64, 20.99]	6.75 [4.68, 9.64]	2.00 [1.02, 3.90]	0.75 [0.26, 2.18]
$M_{\text{adv}} = 6$ (1 primary + 5 alias roots)	27.50 [23.35, 32.07]	19.00 [15.46, 23.13]	12.25 [9.39, 15.83]	7.00 [4.89, 9.93]
$M_{\text{adv}} = 8$ (1 primary + 7 alias roots)	37.50 [32.90, 42.34]	22.75 [18.91, 27.11]	15.00 [11.83, 18.83]	7.75 [5.51, 10.79]

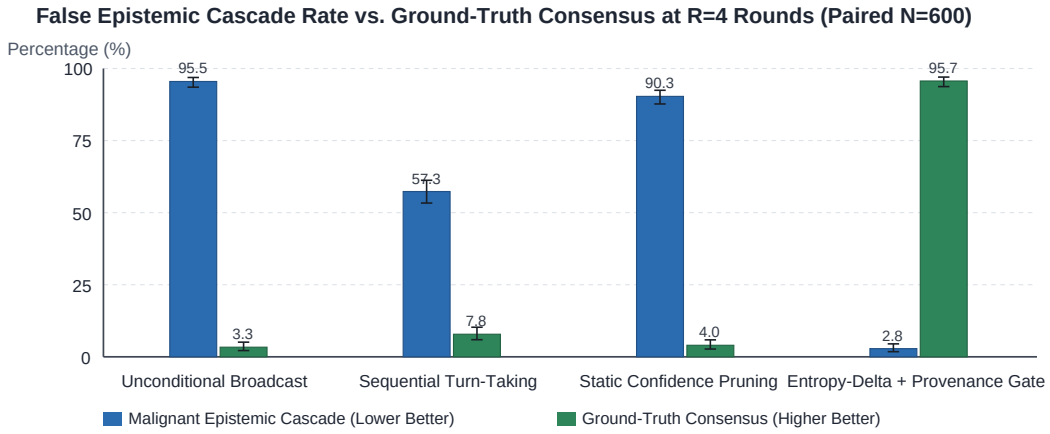


Figure 2: Comparison of Malignant Epistemic Cascade Rate (lower is better) and Ground-Truth Consensus Rate (higher is better) across four multi-agent communication regimes on strictly paired 16-agent scenarios at deliberation horizon $R = 4$ rounds ($N = 600$ episodes per regime; whiskers indicate exact 95% asymmetric Wilson confidence intervals).

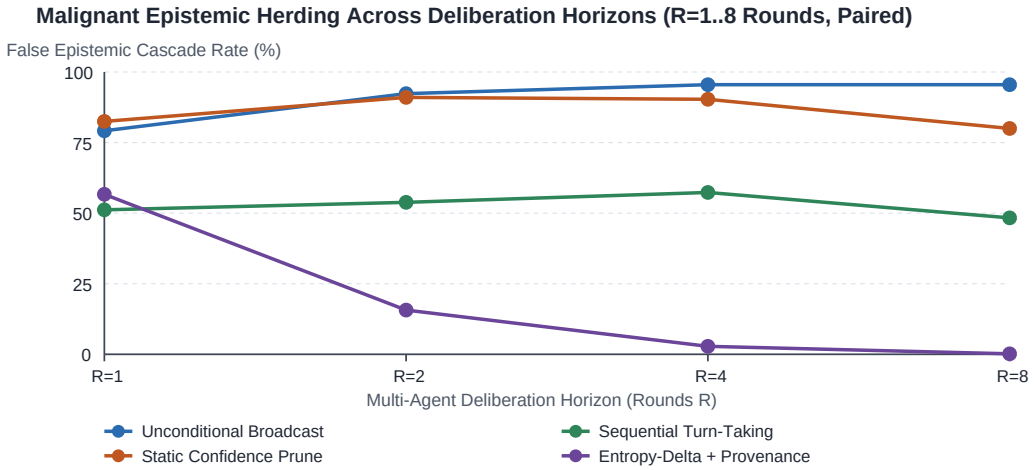


Figure 3: Scaling of Malignant Epistemic Cascade Rate across deliberation horizons $R \in \{1, 2, 4, 8\}$ rounds ($N = 600$ paired scenarios per point). Provenance-blind broadcast and confidence pruning amplify cascades between $R = 1$ and $R = 4$, whereas EDPQ-Gating monotonically suppresses cascades from 56.67% ($R = 1$) to 2.83% ($R = 4$) and 0.17% ($R = 8$).

6 Adversarial Failure Modes, Syndication Noise, and Contested Foundations

Deploying information-theoretic broadcast gating in open agent networks faces four critical counter-arguments and failure modes:

1. **Correlated independently named false evidence evading exact root-ID set union.** Gate 2 (1) relies on set difference $\mathcal{R}_i^{(t)} \setminus \mathcal{R}_j^{(t-1)}$ over discrete provenance identifiers [14, 20]. As proved in Table 5, while pure root-ID set difference ($p_{\text{rec}} = 0.00$) completely eliminates single-root and 2-root circular cascades (0.00% at $M_{\text{adv}} \leq 2$), fabricating $M_{\text{adv}} \geq 6$ distinct root IDs for the same false claim raises cascades to 27.50%–37.50% without receipt verification, and still leaves 12.25%–15.00% cascades under an imperfect verifier ($p_{\text{rec}} = 0.55$). Mitigating high-M_{adv} syndication requires pairing PROV hashes with semantic claim clustering in `{m08}` and higher-recall cryptographic tool receipts.

2. **Internal hallucination confidence spoofing (Bertalanič and Fortuna [2], Sharma et al. [25]).** Language models can experience sharp internal posterior shifts ($D_{\text{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\text{IG}}$) during ungrounded chain-of-thought rationalization. Therefore, Gate 1 requires that any self-information shift be accompanied by a novel external tool or retrieval root identifier before breaking radio silence.

3. **Correlated model blind spots across ensembles (Goel et al. [13], Shapira et al. [24]).** Even when agents execute independent tool calls, homogeneous models often misinterpret ambiguous tool outputs identically [9, 13]. Provenance deduplication cannot correct errors that arise independently from shared model biases, requiring cross-family architectural diversity.

4. **Radio-silence deadlocks on unresolvable queries.** When no agent in the cohort can locate primary evidence that shifts its posterior by τ_{IG}, EDPQ-Gating maintains radio silence.

Rather than forcing sycophantic convergence [2], persistent silence across R rounds serves as a calibrated collective abstention trigger.

7 Limitations

- **Discrete hypothesis clustering requirement.** Computing $H(q_i^{(t)})$ and $D_{\mathrm{KL}}(q_i^{(t)} \| q_j^{(t-1)})$ requires mapping free-form agent outputs onto a shared discrete hypothesis set Θ (e.g., candidate root causes, patch equivalence classes, or multiple-choice options).
- **Vulnerability when initial false infection exceeds $\rho^*_{\mathrm{false}} = 50\%$ or $M_{\mathrm{adv}} \geq 6$.** As established in our sensitivity sweeps, when $> 50\%$ of agents begin with correlated false priors ($M_{\mathrm{adv}} = 4$), false cascades rise to 40.00% $[35.80, 44.35]$ at $\rho_{\mathrm{false}} = 62.5\%$, and when an adversary injects $M_{\mathrm{adv}} = 8$ independently named false roots, cascades reach 15.00% $[11.83, 18.83]$ at $p_{\mathrm{rec}} = 0.55$.
- **Synchronous barrier latency under delayed tool feedback.** Waiting for local exploration before round-level gating introduces synchronization overhead when external tool APIs exhibit high tail latency [16].

8 Pre-registerable Falsification Protocol

Table 6 specifies six pre-registerable empirical tests designed to falsify the anti-cascade, symmetric gating, root-ID invariance, and bandwidth claims of EDPQ-Gating.

Table 6: Six pre-registerable falsification protocols for EDPQ-Gating in decentralized agent networks.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Paired horizon monotonicity (Eq. 3)	Under EDPQ-Gating on paired scenarios, extending deliberation from $R = 1$ to $R = 8$ monotonically lowers false cascades and raises truth consensus	Evaluate 16-agent ensembles across 600 paired episodes at $R \in \{1, 2, 4, 8\}$ under 37.5% correlated false priors	False cascade rate at $R = 4$ or $R = 8$ exceeds $R = 1$, or sum of cascade and truth rates exceeds 100%
P2. Root-ID permutation invariance	EDPQ-Gating decisions depend strictly on set operations and observable receipts, never on numeric root ID encoding	Apply arbitrary bijections to all root IDs across 50 episodes (`verify_root_id_permutation_invariance`)	Any agent posterior, vote, or message count differs under root-ID relabelling
P3. Data-incest invariance vs. M_{adv}	Pure set-union deduplication (p_{rec}=0) blocks single-root cascades ($M_{\text{adv}} \leq 2$) while degrading monotonically as $M_{\text{adv}} \to 8$	Sweep $M_{\text{adv}} \in \{1, 2, 3, 4, 6, 8\}$ across verifier recalls $p_{\text{rec}} \in \{0.0, 0.40, 0.55, 0.85\}$	Single-root cascade rate at $p_{\text{rec}} = 0.0$ exceeds 2.5% or fails to respond to p_{rec} at $M_{\text{adv}} \geq 4$
P4. Break-even false-prior threshold	EDPQ-Gating maintains $< 10\%$ false cascades at $R = 4$ for initial false-prior fractions up to $\rho_{\text{false}} \leq 50\%$	Sweep $n_{\text{false}} \in \{3, 5, 6, 8, 10, 12\}$ out of 16 agents across 500 paired episodes per point	False cascade rate at $\rho_{\text{false}} \leq 50\%$ exceeds 10.0%
P5. Dynamic bandwidth compression	EDPQ-Gating reduces message traffic by $> 85\%$ at $R = 4$ relative to all-to-all broadcast while improving accuracy	Measure dynamic message count $\sum_{t,i,j} g_{i \to j}^{(t)}$ across 600 paired 16-agent episodes	Message reduction at $R \geq 4$ falls below 85% of $N(N-1)R$
P6. Turn-order cascade immunity	Synchronous pre-commitment eliminates sequential speaker-order landslides [2, 3]	Permute agent indices across 400 episodes and measure outcome variance vs. sequential turn-taking	Consensus variance attributable to agent index permutation exceeds 2.5%

9 Conclusion

Unconstrained communication in multi-agent language model networks routinely amplifies correlated misconceptions via naive independent-likelihood accumulation [18], sequential informational cascades [3], and sycophantic conformity [2, 11]. Operating as the decentralized peer-to-peer broadcast controller companion to static provenance-quotiented aggregation (`m08`), EDPQ-Gating enforces radio silence until an agent experiences a symmetric posterior shift ($D_{\text{KL}}(q_i^{(t)} \| q_i^{(t-1)}) \geq \tau_{\text{IG}}$ or $|\Delta H_i^{(t)}| \geq \tau_H$), drops lineage-redundant updates via opaque root-ID set difference ($\mathcal{R}_i^{(t)} \backslash \mathcal{R}_j^{(t-1)} \neq \emptyset$), and routes high-conflict updates through observable receipt verification [15]. Across 9,600 strictly paired 16-agent episodes, EDPQ-Gating reduces $R = 4$ false cascades from 95.50% [93.53, 96.89] to 2.83% [1.78, 4.49] and achieves 95.67% [93.73, 97.03] ground-truth consensus (99.83% at $R = 8$) using 11.3× fewer messages.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Mahault Albarracin, Daphne Demekas, Maxwell J. D. Ramstead, and Conor Heins. Epistemic communities under active inference. *Entropy*, 24(4):476, 2022. doi: 10.3390/e24040476.

[2] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL `https://arxiv.org/abs/2605.00914`.

[3] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. URL `https://doi.org/10.1086/261849`.

[4] Ozan Çatal, Toon Van de Maele, Riddhi J. Pitliya, Mahault Albarracin, Candice Pattisapu, and Tim Verbelen. Belief sharing: A blessing or a curse, 2024. URL `https://arxiv.org/abs/2407.02465`.

[5] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL `https://arxiv.org/abs/2503.13657`.

[6] Hongjiang Chen, Xin Zheng, Yixin Liu, Pengfei Jiao, Shiyuan Li, Huan Liu, et al. GoAgent: Group-of-agents communication topology generation for LLM-based multi-agent systems, 2026. URL `https://arxiv.org/abs/2603.19677`.

[7] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL `https://arxiv.org/abs/2606.22528`.

[8] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[9] John J. Conlon and Peter Schwardmann. AI sycophancy and decisions, 2026. URL `https://arxiv.org/abs/2607.28133`.

[10] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. URL `https://doi.org/10.1080/01621459.1974.10480137`.

[11] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL `https://arxiv.org/abs/2605.06988`.

[12] Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

[13] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[14] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pages 31–40, 2007. URL https://doi.org/10.1145/1265530.1265535.

[15] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):2, 2002. URL https://www.jasss.org/5/3/2.html.

[16] Igor Itkin. Delayed verification destabilizes multi-agent LLM belief: Instability thresholds and optimal corrector placement. arXiv preprint arXiv:2606.27409, 2026. URL https://arxiv.org/abs/2606.27409.

[17] Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, et al. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

[18] Vikram Krishnamurthy and William Hoiles. Online reputation systems: How to interpret herd behavior and mitigate data incest. IEEE Signal Processing Magazine, 31(4):52–62, 2014. URL https://doi.org/10.1109/MSP.2014.2311855.

[19] Xinwen Liu, Zhuocheng Pan, Isabella Zhu, Jawei Zhang, Xudong Liu, and Tianyu Wo. Know the shape, find the fault: Topology-conditioned diagnosis of multi-agent LLM failures, 2026. URL https://arxiv.org/abs/2610.10126.

[20] Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C Recommendation, 2013. URL https://www.w3.org/TR/prov-dm/.

[21] Reza Olfati-Saber and Richard M. Murray. Consensus problems in networks of agents with switching topology and time-delays. IEEE Transactions on Automatic Control, 49(9):1520–1533, 2004. URL https://doi.org/10.1109/TAC.2004.834113.

[22] E. S. Page. Continuous inspection schemes. Biometrika, 41(1/2):100–115, 1954. URL https://doi.org/10.1093/biomet/41.1-2.100.

[23] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[24] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[25] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna

Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2310.13548.

[26] Abraham Wald. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2):117–186, 1945. URL https://doi.org/10.1214/aoms/1177731118.

[27] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, et al. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

[28] Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, et al. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.20639.