

Communication content and local connectivity shape collective accuracy in networks of machine agents

Eric Gladstone*

Independent Researcher, Traverse City, MI, USA

*Corresponding author. Email: eric.c.gladstone@gmail.com

ORCID (Gladstone): 0000-0003-4393-050X

Classification: Social Sciences; Computational Sciences

Keywords: large language models; multi-agent AI; collective intelligence; wisdom of crowds; information cascades; communication networks; belief aggregation

Abstract

Large language model (LLM) agents increasingly solve problems collectively, but collective accuracy may depend on what agents communicate and how peer information is distributed across a network. We study a distributed-inference task in which a weak numerical majority opposes a better-informed minority, so numerical agreement and evidentiary strength diverge. Under the prespecified logarithmic pool, conclusion sharing made an initially accurate collective less accurate in two of three model families. This failure arose primarily because terminal collective beliefs became far more variable across trials, not because their mean shifted toward the misleading majority. Within trials, agents also became less accurate while their beliefs became more dispersed rather than more homogeneous. At moderate fan-in, evidence sharing supported correction of the weak majority and limited capitulation by the better-informed minority. A planted-conclusion intervention showed that transmitted conclusion direction can causally alter terminal collective belief while private evidence is held fixed, and extended private reasoning eliminated the conclusion-sharing failure in one within-family manipulation. The advantage of evidence sharing depended on local fan-in. At the highest fan-in tested, evidence sharing became less accurate than conclusion sharing in both the misleading-majority and concordant regimes, including when every private signal favored the truth. Across network manipulations, collective accuracy declined and minority capitulation increased with local fan-in, while matched path-length contrasts remained imprecise. These results show that communication content and local exposure jointly shape collective accuracy in LLM-agent networks and that collective failure can emerge through increasing dispersion rather than consensus.

Significance Statement

Networks of large language model agents can lose collective accuracy without converging on the same belief. When a weak numerical majority opposed a better-informed minority at moderate fan-in, sharing conclusions produced more variable and less accurate judgments, while sharing evidence was protective. At the highest fan-in tested, evidence sharing became less accurate than conclusion sharing in both signal regimes, including when every private signal favored the truth. A planted-conclusion intervention showed that transmitted conclusions can causally shift collective belief. Across network manipulations, the largest accuracy differences aligned with incoming peer communication, while matched global path-length comparisons remained imprecise. These results suggest that message content and local communication exposure should be considered jointly in multi-agent AI design.

Introduction

Large language model (LLM) agents increasingly operate collectively rather than alone. Ensembles aggregate independently sampled solutions, debate systems exchange arguments across rounds, layered architectures pass outputs between groups of agents, and multi-agent frameworks organize planners, critics, executors, judges, and synthesizers into communication networks (1–6). These systems routinely pass compressed intermediate artifacts such as findings, summaries, partial solutions, or recommendations. What survives that compression varies. Some messages preserve independently evaluable grounds for a judgment, while others primarily communicate the judgment itself. The conclusion-sharing condition studied here is a controlled limiting case of this latter form of provenance-poor communication, not a claim that deployed systems routinely exchange bare one-sentence verdicts.

This distinction matters when members differ in the quality of their information. Research on collective judgment shows that aggregating multiple estimates can reduce idiosyncratic error, but numerical agreement is informative only insofar as the underlying judgments retain their evidentiary value (7–11). A numerical majority can therefore be less informative than a smaller minority when the minority holds stronger evidence. Communication may either preserve the information needed to recognize that asymmetry or compress it away. In our setting, agents can transmit an observation together with its diagnosticity, or transmit only a verdict and confidence. Evidence sharing preserves information about what was observed and how informative it was. A conclusion summarizes an agent's current posterior judgment but obscures the diagnosticity, origin, and possible overlap of the information beneath it.

That loss of provenance creates a familiar inferential problem. Observing others' actions rather than their private signals can generate informational cascades, and repeated exposure to dependent judgments can cause correlated information to be overcounted (12–19). Network structure determines how many such judgments reach each agent and how rapidly they circulate. Prior work on human and artificial collectives has accordingly shown that social influence, dense connectivity, and centralized influence can reduce independent exploration, accelerate convergence, or propagate correlated errors under some conditions (20–29). Opinion-dynamics models can also produce increasing divergence or separated belief clusters rather than convergence (30, 31). Collective failure under social influence, dense communication, or heterogeneous network position is therefore not novel in itself. The unresolved question is how communication content and local exposure jointly shape the form and direction of collective failure in networks of LLM agents.

We study this question using a distributed binary-inference task in which a weak numerical majority favors the wrong state while a better-informed minority favors the truth. The task deliberately separates numerical agreement from evidentiary strength. Holding private signals fixed, we vary whether agents communicate evidence or conclusions, manipulate network topology and local fan-in, and intervene directly on the direction of a transmitted conclusion. This design allows us to distinguish failure associated with misleading-majority influence from failure that persists even when every private signal favors the truth.

The contribution is therefore not simply that LLM agents can herd, disagree, or perform differently across network structures. Under the prespecified logarithmic pool, conclusion sharing can make an initially accurate collective less accurate primarily by increasing the dispersion of terminal collective beliefs across trials without shifting their mean toward the misleading majority. Within trials, agents also become more, not less, dispersed in their individual beliefs. At moderate fan-in, evidence sharing supports weak-majority correction and limits strong-minority capitulation, while a planted-conclusion intervention separately establishes that the direction of a transmitted conclusion can causally alter terminal collective belief while private evidence is held fixed. That advantage of evidence sharing is conditional on local exposure. At the highest fan-in tested, evidence sharing becomes less accurate than conclusion sharing in both the misleading-majority and concordant regimes. The strongest structural evidence comes from the degree gradient and within-network positional contrasts, while matched path-length contrasts remain imprecise. Additional private reasoning eliminates the conclusion-sharing failure in one within-family manipulation. Together, these results show that communication content and local exposure jointly shape both whether and how collective inference fails in networks of LLM agents.

Results

Design and measures

Each trial specifies a hidden state, A or B, with equal prior probability (Fig. 1). Each of n stateless LLM agents privately observes one noisy reading with known diagnosticity and reports its belief that the state is A. In the misleading-majority regime, a numerical majority receives weak readings favoring the false state, while a minority receives stronger readings favoring the true state. A numerical vote therefore favors the wrong answer, whereas aggregation that preserves the relative strength of the private signals favors the correct one. A concordant regime, in which all readings favor the same state, serves as a control.

We manipulate communication content, with conditions for no communication, evidence sharing, conclusion sharing, or both, and vary network topology. We report three outcomes. Collective accuracy, our primary outcome, indicates whether the group's pooled belief favors the true state. The primary pooled belief is a logarithmic opinion pool that assigns equal weight to agents while averaging their reported beliefs on the log-

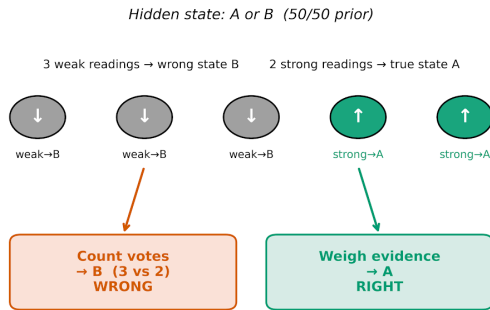
odds scale. More extreme beliefs therefore have larger log-odds magnitude, but the pool does not directly observe or weight the evidence that produced them. For agent beliefs p_i that the state is A, the pooled belief is the inverse logit of the mean individual logits, and collective accuracy indicates whether that pooled probability falls on the correct side of 0.5 (Materials and Methods). In the misleading-majority regime, the pooled belief and numerical vote diverge by construction at baseline. Mean individual error, $\bar{e}$, is the average Brier score across agents, and diversity, D , is the variance of their beliefs. For the arithmetic-mean pool, collective error satisfies the identity collective error = $\bar{e} - D$, which held to within 1.6×10^{-4} across all 5,741 trials of the binary-inference study.

We define herding as socially induced movement toward the numerical majority when stronger aggregate private evidence favors the alternative. Herding therefore refers to directional movement by an agent or collective on a particular trial and does not require consensus. We do not use the term to characterize a communication condition when its mean directional movement is not toward the majority. For key contrasts, we report Wilson 95% confidence intervals, paired tests using shared seeds, seed-clustered logistic regression, and bootstrap inference.

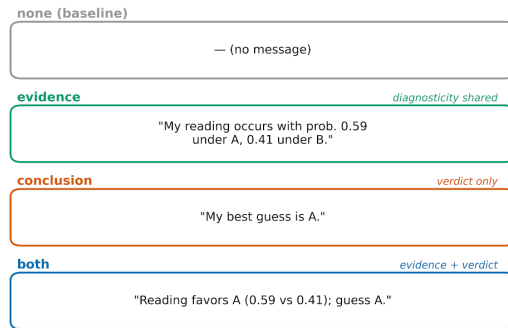
Before communication, agents' private beliefs closely recovered the normative single-signal posterior across model families (Pearson $r = 0.999$, mean absolute error = 0.004). Agents therefore began from beliefs that accurately represented their assigned private signals, making communication-induced updating the relevant source of subsequent change rather than an initial inability to interpret those signals.

One trial illustrates the design. In a representative misleading-majority trial with $n = 5$, the true state was A and the full-information posterior was $P(A) = 0.945$. Three agents received weak readings favoring B, with log-likelihood ratios of -0.46 , -0.37 , and -0.43 , while two received stronger readings favoring A, at $+2.04$ and $+2.07$. Their round-0 beliefs yielded a pooled belief of $P(A) = 0.638$, correctly favoring A, even though three of the five individual verdicts favored B. Under evidence sharing, agents transmitted messages such as “My private reading is ‘r3,’ which occurs with probability 0.88 if the true state is A and 0.12 if the true state is B,” and the terminal pooled belief rose to 0.866. Under conclusion sharing, the same agents with the same private evidence instead transmitted messages such as “I believe the hidden state is A with about 94% confidence,” and the terminal pooled belief fell to 0.016, confidently favoring the wrong state. Individual terminal beliefs still ranged from 0.000 to 0.600. The complete trial, including every message, is reported in Section S13.

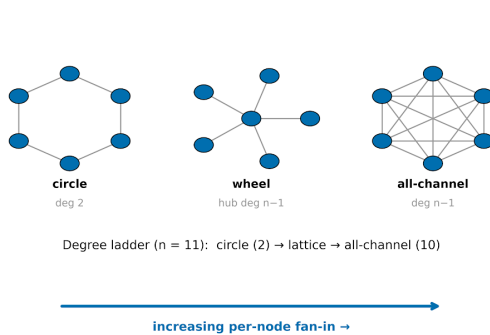
A The task: a misleading majority



B What each agent transmits (content)



C Network topologies (per-node fan-in)



D Planting intervention (causal test)

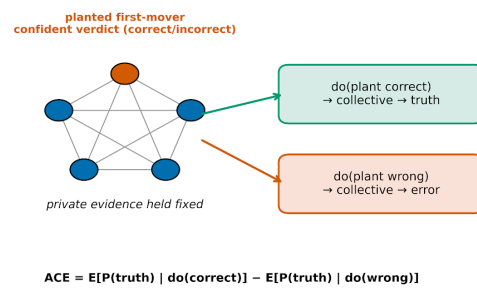


Fig. 1. Experimental design. (A) The misleading-majority task. A numerical majority receives weak readings favoring the false state, while a minority receives stronger readings favoring the true state. Vote counting therefore favors the wrong answer, whereas evidence weighting favors the correct one. (B) Communication-content conditions: no communication, evidence sharing, conclusion sharing, and both. Evidence messages preserve diagnosticity, whereas conclusion messages transmit a verdict without the underlying evidence. (C) Network topologies and the fan-in manipulation, varying the number of peer messages received per round. The degree ladder spans the sparse circle, ring-with-shortcuts, and degree-controlled ring lattices with per-node degree 4, 6, or 8, used in matched comparisons of fan-in and path length. (D) The planted-conclusion intervention. A first-mover verdict is visible from round 0 while all agents' private evidence is held fixed, allowing the direction of the transmitted conclusion to be manipulated.

Alt text: Four-panel schematic of the experimental design. Panel A shows the misleading-majority task, in which a numerical majority of agents receives weak readings favoring the false state and a minority receives stronger readings favoring the true state. Panel B shows the four communication-content conditions. Panel C shows the network topologies and the fan-in manipulation. Panel D shows the planted-conclusion intervention.

Communication content shapes collective accuracy

In the misleading-majority regime, the round-0 pooled belief favored the true state on every trial (collective accuracy 1.00 in all three model families, Table 1), while a majority vote over the same agents' verdicts favored the false state. This baseline is an expected consequence of the task design combined with the validated belief readout rather than an independent empirical finding. The stronger minority signals outweigh the weaker majority signals by construction, and agents' round-0 beliefs closely recover the corresponding single-signal posteriors (Section S5). In gpt-4.1-mini, for example, the mean round-0 probability assigned to the true state was 0.632 (SD = 0.015), with no trial on the incorrect side of 0.5. This provides a stable starting point against which communication-induced change can be measured.

Holding the network fixed at the $n = 5$ all-channel topology, we varied only what each agent transmitted (Table 1). Evidence sharing maintained high collective accuracy in all three model families: 0.90 [0.83–0.94] for gpt-4.1-mini ($N = 100$), 0.96 [0.87–0.99] for claude-sonnet-4-6 ($N = 50$), and 0.97 [0.83–0.99] for gemini-2.5-flash ($N = 30$). Conclusion sharing reduced accuracy to 0.64 [0.54–0.73] for gpt-4.1-mini ($N = 100$) and 0.62 [0.48–0.74] for claude-sonnet-4-6 ($N = 50$). Relative to conclusion sharing, evidence sharing increased accuracy by 0.26 [0.15–0.37] in gpt-4.1-mini and 0.34 [0.20–0.48] in claude-sonnet-4-6 (paired McNemar $p = 4.2 \times 10^{-5}$ and 2.2×10^{-4}). In gpt-4.1-mini, conclusion sharing was also 0.36 [0.27–0.46] less accurate than no communication. Sharing both evidence and conclusions produced intermediate accuracy, 0.81 in gpt-4.1-mini and 0.88 in claude-sonnet-4-6. Odds ratios and the pooled seed-clustered test are reported in the Supplementary Materials.

Under conclusion sharing, the loss of accuracy arose primarily from greater dispersion in terminal collective beliefs across trials rather than a directional shift in their mean. Under the logarithmic pool, the mean probability assigned to the true state changed little from round 0 to the terminal round, from 0.632 to 0.643 in gpt-4.1-mini and from 0.634 to 0.659 in claude-sonnet-4-6. Across-trial standard deviations, however, increased from approximately 0.015 to 0.381 and 0.393. Mean squared error of the pooled probability decomposes exactly into a component determined by its mean and a variance component. Relative to each condition's own round-0 baseline, the increase in variance more than accounted for the increase in error because the mean component improved slightly. Terminal beliefs accumulated near both certainty bounds rather than shifting uniformly toward the misleading majority.

Under the prespecified logarithmic pool, 36% of gpt-4.1-mini conclusion-sharing trials and 38% of claude-sonnet-4-6 trials ended on the incorrect side. Among those trials, the mean probability assigned to the true state was approximately 0.17, so incorrect log-pooled outcomes were typically confident rather than marginal crossings of 0.5. This characterization is pool-sensitive. Under arithmetic pooling, the corresponding wrong-side means were 0.316 and 0.334, with wrong-side rates of 0.31 and 0.26. The baseline-worsening decomposition is likewise specific to the prespecified logarithmic pool. Variance also increased under arithmetic pooling, but the net increase in Brier error relative to baseline did not replicate cleanly. Relative to evidence sharing, however, conclusion sharing performed worse under both pools because it failed to produce the mean improvement obtained under evidence sharing and added dispersion (Section S6.7).

The across-trial dispersion was not produced by agents converging within trials. Mean individual error rose from 0.210 at round 0 to 0.288 in gpt-4.1-mini and 0.255 in claude-sonnet-4-6. At the same time, within-trial belief diversity approximately doubled, from 0.047 to 0.091 and from 0.048 to 0.093, respectively (Fig. 2B). Collective accuracy therefore deteriorated while individual beliefs became more dispersed rather than more homogeneous. A post hoc diagnostic showed that trials containing agents at both belief extremes were also more common under conclusion sharing: using a 0.20/0.80 threshold, 54% of gpt-4.1-mini and 52% of claude-sonnet-4-6 conclusion-sharing trials contained beliefs on both sides of those thresholds, compared with 37% and 18% under evidence sharing. The pattern was robust to alternative thresholds (Section S6.7). Because both-extremes configurations also occurred under evidence sharing, particularly in gpt-4.1-mini, this result describes the distribution of beliefs but does not identify polarization as the mechanism of failure.

Within this $n = 5$ all-channel comparison, the evidence-over-conclusion advantage was also robust to the aggregation rule. Across six rules, including the logarithmic pool, arithmetic-mean pool, majority vote, median belief, clipped logarithmic pool, and a logarithmic pool omitting each trial's most extreme belief, the evidence-minus-conclusion accuracy difference ranged from 0.26 to 0.28 in gpt-4.1-mini ($N = 100$) and from 0.20 to 0.34 in claude-sonnet-4-6 ($N = 50$). The logarithmic pool produced the largest contrast in claude-sonnet-4-6, making it the most favorable of the six rules tested for that family. Full results across model families and the degree ladder are reported in Section S6.8. In the $n = 5$ all-channel concordant control, accuracy remained near 1.00 across communication-content conditions.

Table 1. Communication content and collective accuracy in the misleading-majority, all-channel network. Collective accuracy [Wilson 95% CI]; N per cell.

Condition	gpt-4.1-mini (N=100)	claude-sonnet-4-6 (N=50)	gemini-2.5-flash
No communication	1.00 [0.96–1.00]	1.00 [0.93–1.00]	1.00 [0.89–1.00] (N=30)
Evidence sharing	0.90 [0.83–0.94]	0.96 [0.87–0.99]	0.97 [0.83–0.99] (N=30)
Both	0.81 [0.72–0.87]	0.88 [0.76–0.94]	—
Conclusion sharing	0.64 [0.54–0.73]	0.62 [0.48–0.74]	0.93 [0.79–0.98] (N=30)†
Concordant control	≈1.00	≈1.00	≈1.00

†In gemini-2.5-flash, accuracy remained 0.93 under conclusion sharing, and evidence sharing did not significantly improve accuracy relative to conclusion sharing ($p = 0.55$). Gemini was the only family run with an extended thinking budget of 8,192 tokens, which was required to pass the capability gate, so model family and reasoning effort are confounded in this comparison. The content-related accuracy loss is therefore supported by gpt-4.1-mini and claude-sonnet-4-6, while Gemini identifies a boundary to its generality. Whether additional reasoning moderates susceptibility is examined below and in the Discussion.

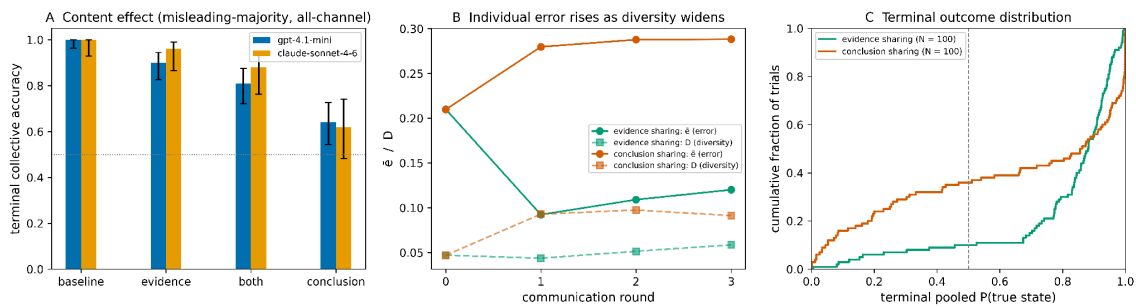


Fig. 2. Communication content changes collective accuracy, individual error, and terminal collective beliefs. (A) Terminal collective accuracy in the misleading-majority, all-channel condition by communication content and model family. Error bars show Wilson 95% confidence intervals. (B) Round-by-round mean individual error $\bar{e}$ and within-trial belief diversity D under evidence sharing and conclusion sharing. Under conclusion sharing, individual error rises while diversity also increases, showing that collective failure does not require belief homogenization. For the arithmetic-mean pool, where collective error $= \bar{e} - D$, rising individual error outweighs the diversity term. (C) Empirical cumulative distribution of the terminal pooled probability assigned to the true state by communication content in gpt-4.1-mini. Conclusion sharing produces a much wider distribution, with mass near both certainty bounds, whereas evidence sharing remains concentrated at high probabilities. The dashed line marks the 0.5 classification boundary.

Alt text: Three-panel figure. Panel A is a bar chart of terminal collective accuracy by communication content and model family, with confidence intervals. Panel B shows mean individual error and within-trial belief diversity across communication rounds under evidence and conclusion sharing. Panel C is an empirical cumulative distribution of terminal pooled probability assigned to the true state under each content condition.

At moderate fan-in, conclusion sharing suppresses majority correction and increases minority capitulation

To characterize the updating pattern behind the content effect in the standard $n = 5$ all-channel condition, we classified agents by their assigned signals. The weak majority received weak readings favoring the wrong state, whereas the strong minority received stronger readings favoring the true state. We then tracked two terminal transitions relative to the truth on seed-matched trials. A weak-majority agent was classified as corrected if its terminal belief favored the true state. A strong-minority agent was classified as having capitulated if its terminal belief favored the false state. Classifications based on assigned signals and realized round-0 beliefs agreed for every agent, so inference resampled seeds rather than agents.

Evidence sharing improved both transitions (Fig. 3). Among weak-majority agents, 0.93 ended on the true side, while 0.22 of strong-minority agents capitulated. Relative to evidence sharing, conclusion sharing shifted both outcomes in the adverse direction. Majority correction fell to 0.69, a risk difference of $-0.24 [-0.31, -0.18]$, and

minority capitulation rose to 0.38, a risk difference of 0.15 [0.07, 0.25]. The contrast became larger under a heavier misleading majority (Section S6.5). In the standard condition, the larger change was reduced correction among weak-majority agents rather than increased capitulation among strong-minority agents.

In a trial-level model, the estimated content coefficient for collective accuracy attenuated from -0.26 to -0.06 after these two post-treatment transitions were included. This attenuation is consistent with majority correction and minority capitulation marking proximal behavioral pathways through which communication content is associated with collective error. Because both transitions were measured after treatment, the analysis is descriptive and should not be interpreted as formal causal mediation.

The transition pattern also helps distinguish conclusion sharing from literal vote counting. Under evidence sharing, the pooled belief moved toward the misleading majority relative to its own round-0 value on 11% of gpt-4.1-mini trials and 4% of claude-sonnet-4-6 trials. Under conclusion sharing, those frequencies rose to 39% and 38%, respectively. Among trials that did move toward the majority, however, the mean magnitude of movement changed little, from approximately -0.40 to -0.44 . Most conclusion-sharing trials still moved away from the misleading majority. A literal vote-counting rule is therefore too strong a description of the observed updating. An unweighted vote is incorrect on every misleading-majority trial by construction, whereas the observed collective was incorrect on 36% of gpt-4.1-mini trials and 38% of claude-sonnet-4-6 trials.

Two further comparisons show why simple information loss is also insufficient as an account. First, a confidence-weighted benchmark applied to the same round-0 communicated verdicts favored the true state on every trial, matching the no-communication baseline, whereas an unweighted head count over those verdicts favored the false state on every trial. The conclusion messages therefore retained enough directional and confidence information for this benchmark to classify every trial correctly. Second, strong-minority agents capitulated more often than predicted by that benchmark, and the excess increased with fan-in (Section S6.6).

Loss of evidentiary provenance provides one plausible interpretation of these behavioral patterns. A communicated verdict reveals how many peers favor a view and with what reported confidence, but not the diagnostic strength, source, or possible overlap of the evidence underlying those judgments. Conclusion sharing may therefore make the numerical distribution of peer opinions easier to use than the evidentiary basis those opinions summarize. The experiments identify the resulting updating pattern, not the internal computation that produced it, so this provenance-loss account remains an interpretation rather than a directly observed mechanism.

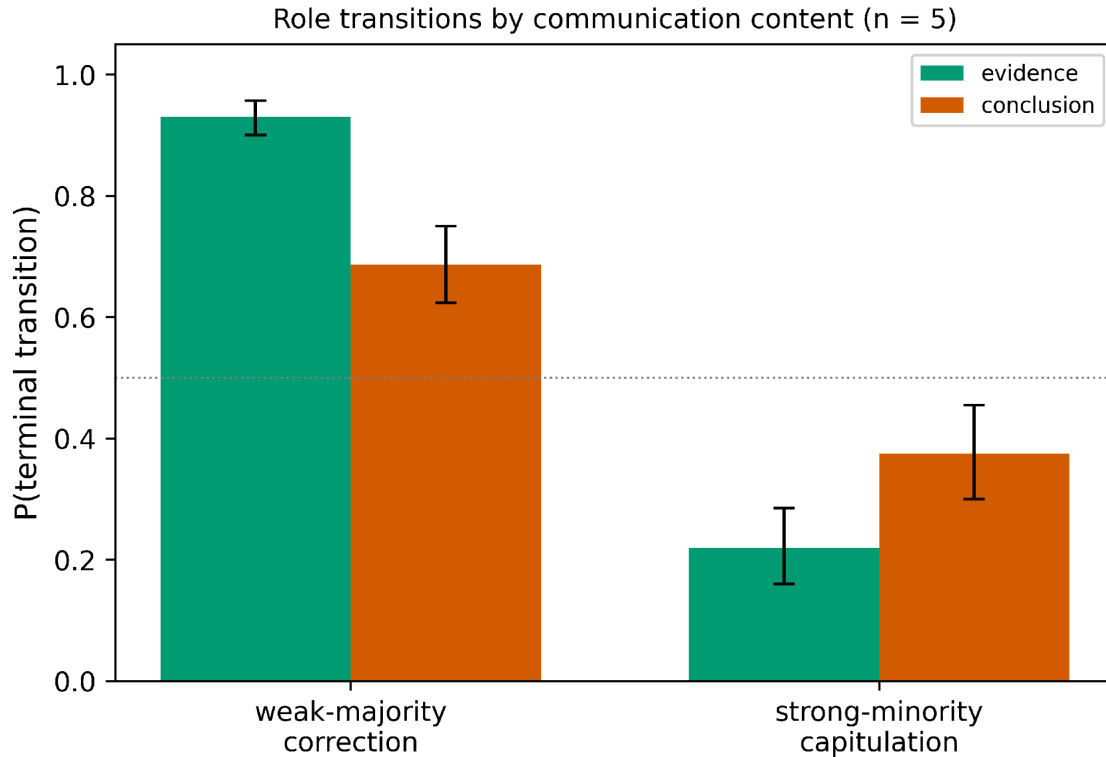


Fig. 3. Communication content changes role-specific updating at moderate fan-in. Terminal role transitions with seed-clustered paired-bootstrap 95% confidence intervals in the standard misleading-majority condition ($n = 5$). Evidence sharing increases weak-majority correction and limits strong-minority capitulation relative to conclusion sharing. Communication content was experimentally assigned, and transitions were measured after treatment. At the highest fan-in tested in the $n = 11$ degree ladder, the evidence-over-conclusion accuracy advantage reverses; that interaction is reported below and in Section S6.5.

Alt text: Bar chart of terminal role transitions under evidence sharing and conclusion sharing, showing weak-majority correction and strong-minority capitulation with confidence intervals.

A planted conclusion causally shifts collective belief

To test the influence of transmitted conclusions more directly, we planted a single confident first-mover conclusion on seed-matched trials. Every agent's private evidence was held fixed across intervention arms. The planted conclusion either favored the true state or the false state, with an unplanted condition providing a common control. Changing the direction of the plant produced a large shift in the collective's terminal belief (Fig. 4). A correct plant raised the mean terminal probability assigned to the true state to 0.99 in claude-sonnet-4-6 and 0.85 in gpt-4.1-mini, whereas an incorrect plant reduced it to 0.36 and 0.34, respectively. The paired correct-minus-incorrect plant effect was therefore 0.64 in claude-sonnet-4-6 ($N = 15$) and 0.51 in gpt-4.1-mini ($N = 40$).

Because private evidence was identical across intervention arms, this contrast identifies a causal effect of transmitted conclusion direction on terminal collective belief. The intervention complements the factorial content manipulation by changing the direction of a communicated conclusion while holding the underlying private evidence fixed. It establishes that a transmitted conclusion can substantially redirect collective belief. It does not imply that ordinary conclusion sharing shifts the *mean* collective belief toward the misleading majority across trials.

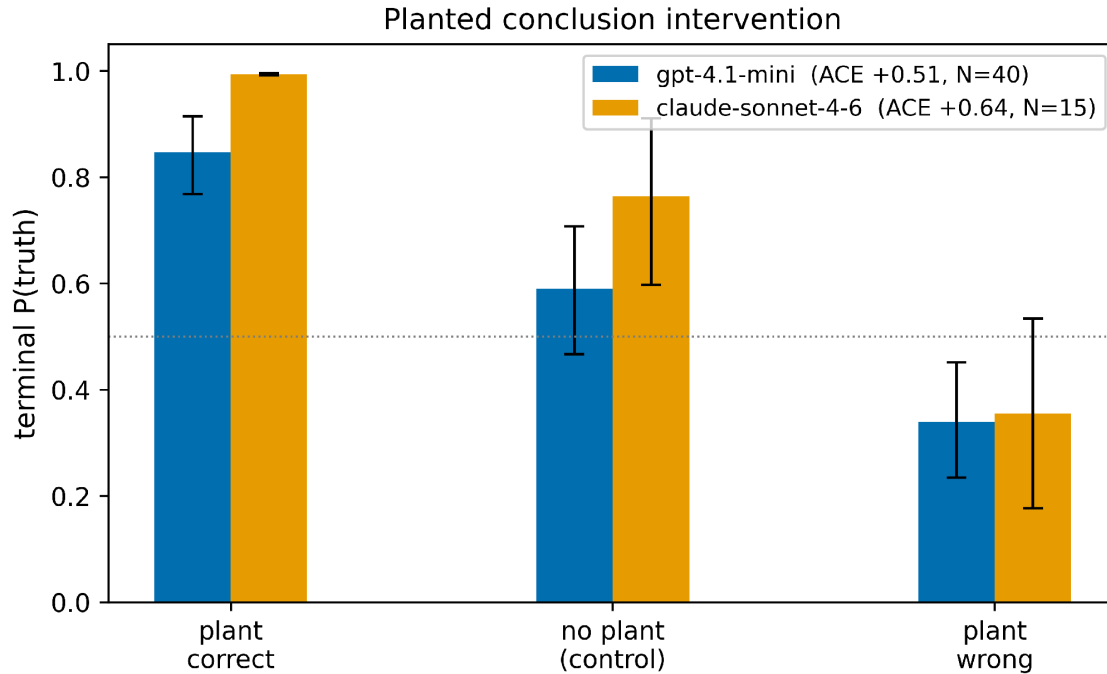


Fig. 4. A transmitted conclusion causally shifts collective belief. Terminal probability assigned to the true state under do(plant correct), no plant, and do(plant wrong), for both model families ($N = 15$ for claude-sonnet-4-6 and $N = 40$ for gpt-4.1-mini). Private evidence is identical across intervention arms. Bars show means, error bars show seed-paired bootstrap 95% confidence intervals, and legend annotations report the correct-minus-incorrect plant effect.

Alt text: Bar chart of the terminal probability assigned to the true state under a correct planted conclusion, no plant, and an incorrect planted conclusion, for two model families, with confidence intervals.

Local fan-in tracks collective failure and minority capitulation

Holding communication content fixed to conclusion sharing, we varied network topology. Under this condition, topology changes how many peer conclusions each agent receives per round as well as global features of the communication network. Sparse decentralized networks preserved accuracy, whereas the dense all-channel network performed substantially worse. At $n = 11$, collective accuracy was 1.00 in the circle and 0.58 in the all-channel network, a risk difference of -0.42 $[-0.56, -0.28]$.

To examine which structural features aligned with this decline, we added ring lattices with per-node degree 4, 6, or 8 and a ring-with-shortcuts network with shorter paths but low degree. Collective accuracy declined strongly as local fan-in increased (seed-clustered logistic regression, $p = 4.7 \times 10^{-11}$, Table 2). Degree and path length covary across these graphs, however, so their independent effects are not separately identified. In paired seed contrasts, the all-channel network versus the degree-8 lattice differed in accuracy by -0.02 $[-0.14, +0.10]$, while the low-degree ring-with-shortcuts network versus the circle differed by -0.06 $[-0.14, 0.00]$. These intervals are too imprecise to establish the absence of a path-length effect. The strongest structural evidence therefore comes from the degree gradient and from comparisons among positions with different fan-in within the same graph. The largest observed accuracy differences aligned with local fan-in, while the matched path-length contrasts remain unresolved.

The role-transition results showed the same association. Across the degree ladder under conclusion sharing, strong-minority capitulation increased from 0.09 at degree 2 to approximately 0.40 at degree 6 and above. Weak-majority correction showed an overall negative association with degree ($r = -0.82$, Fig. 5B). A corresponding pattern appeared within heterogeneous-degree graphs, where signals were randomized across network positions and informational role was empirically balanced against position. The wheel provides the largest within-network

positional contrast and resembles a hub-and-spoke orchestrator arrangement. The high-fan-in hub capitulated at 0.48 versus 0.04 among the low-fan-in spokes, a hub-minus-spoke difference of +0.439 [0.288, 0.590]. The association with fan-in therefore appears not only across different graphs but also across positions exposed to the same task and drawn from the same underlying signal structure.

The positional association was concentrated at the focal agent's own fan-in. Across 600 strong-minority agent observations in 300 trials, own fan-in was positively associated with capitulation ($r = 0.357$ [0.257, 0.450]). Conditional on the focal agent's own fan-in, mean sender fan-in had partial $r = 0.007$ [-0.056, +0.075]. Conversely, own fan-in remained positively associated with capitulation after controlling for sender fan-in (partial $r = 0.265$ [+0.172, +0.356]). The neighbor-fan-in estimate is therefore tightly centered near zero in the tested graphs, whereas the association with the focal agent's own incoming exposure remains positive. The wheel cannot contribute to this conditional comparison because hub and spoke positions make own and neighbor fan-in perfectly collinear. The conditional estimates are therefore identified by the hierarchy and chain topologies (Section S6.9). Taken together, the between-graph degree gradient and within-graph positional contrasts provide convergent evidence that local incoming exposure is closely associated with strong-minority capitulation.

Table 2. Collective accuracy across fan-in and matched path-length contrasts (misleading-majority, conclusion-sharing, $n = 11$). Collective accuracy [Wilson 95% CI; $N = 50$ per cell].

Topology	Per-node degree	Diameter	Accuracy
Circle	2	5	1.00 [0.93–1.00]
Ring + shortcuts	~2.5	3	0.94 [0.84–0.98]
lattice_k4	4	3	0.90 [0.79–0.96]
lattice_k6	6	2	0.76 [0.63–0.86]
lattice_k8	8	2	0.60 [0.46–0.72]
All-channel	10	1	0.58 [0.44–0.71]

Accuracy declined with degree (seed-clustered logistic regression, $p = 4.7 \times 10^{-11}$). Degree and diameter covary across these networks, so their independent effects are not separately identified. In paired seed contrasts, the degree-10 all-channel network versus the degree-8 lattice differed by -0.02 [-0.14, +0.10], while the ring-with-shortcuts network versus the circle differed by -0.06 [-0.14, 0.00]. These matched contrasts remain imprecise and do not establish the absence of a path-length effect.

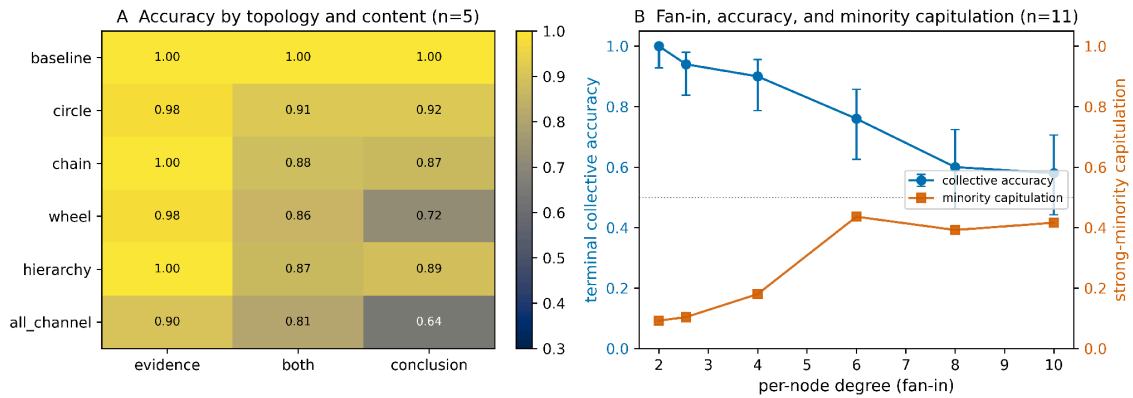


Fig. 5. Local fan-in tracks collective accuracy and minority capitulation. (A) Terminal collective accuracy by network topology and communication content in the $n = 5$ misleading-majority condition. Accuracy under conclusion sharing is higher in the sparse circle, chain, and hierarchy than in the all-channel network. (B) At $n = 11$ under conclusion sharing, collective accuracy declines as per-node fan-in increases from degree 2 to degree 10, while strong-minority capitulation rises. The all-channel versus degree-8 lattice accuracy difference is -0.02 [-0.14, +0.10], and the ring-with-shortcuts versus circle difference is -0.06 [-0.14, 0.00]. These matched path-length contrasts are imprecise and do not establish the absence of a path-length effect.

Alt text: Two-panel figure. Panel A is a heatmap of terminal collective accuracy by network topology and communication content. Panel B plots collective accuracy and strong-minority capitulation against per-node fan-in across the degree ladder.

At high local fan-in, the content effect reverses

At the highest fan-in tested, the relationship between communication content and collective accuracy reversed. In the $n = 11$ all-channel network, evidence sharing was less accurate than conclusion sharing in both signal regimes: 0.38 versus 0.58 in the misleading-majority regime, a paired risk difference of -0.20 $[-0.32, -0.08]$, and 0.68 versus 0.96 in the concordant regime, a paired risk difference of -0.28 $[-0.42, -0.14]$. Across the three paired degree levels at which both content conditions were observed (2, 2.55, and 10), within-seed randomization tests supported a content $\times$ fan-in interaction in both regimes ($p = 0.0015$ and $p = 0.0014$, respectively). The degree ladder is not fully content-crossed at the intermediate degrees, so this interaction is identified by those three paired levels rather than by all six degree values. Separately, holding $n = 11$ fixed under concordant evidence sharing, accuracy remained 1.00 through degree 6, then declined to 0.90 at degree 8 and 0.68 at degree 10 (Fig. 6). Because this reversal and the concordant evidence-sharing decline both occur when every private signal favors the truth, they cannot be explained by deference to a misleading numerical majority.

This upper-degree decline was robust to the aggregation rule. Under all six rules tested, accuracy remained 1.00 through degree 6. At degree 10, accuracy was 0.68 under the prespecified logarithmic pool and ranged from 0.50 to 0.64 under the five alternatives (Section S6.8). The decline therefore does not depend on the primary pooling rule. It was also accompanied by increasing rather than contracting belief diversity. Under concordant evidence sharing, diversity rose from 0.006 at round 0 and reached 0.128 by degree 8, while mean individual error rose from 0.076 to 0.314. The role-transition results showed the same loss of evidence-sharing advantage: at the highest fan-in condition, the differences between evidence and conclusion sharing in majority correction and minority capitulation were approximately null (Section S6.5).

These results show that the effect of communication content is conditional on local fan-in rather than simply separable from it. Conclusion sharing is less accurate than evidence sharing at moderate fan-in, but the ordering reverses at the highest fan-in tested. We refer to the upper-degree pattern as a high-local-load failure because increasing fan-in increases the amount and complexity of information each agent must reconcile. The experiments do not identify which component of that burden is responsible. Degree jointly changes the number of senders, the volume of incoming text, and the need to track dependence among messages. At degree 10, evidence messages produced 1.82 times as many incoming characters as conclusion messages with the same sender count. This is consistent with message volume or evidentiary complexity contributing to the reversal, but it does not identify either as the mechanism (Section S6.11). High local incoming-information load is therefore an interpretation of the experimental configuration rather than a directly measured processing mechanism.

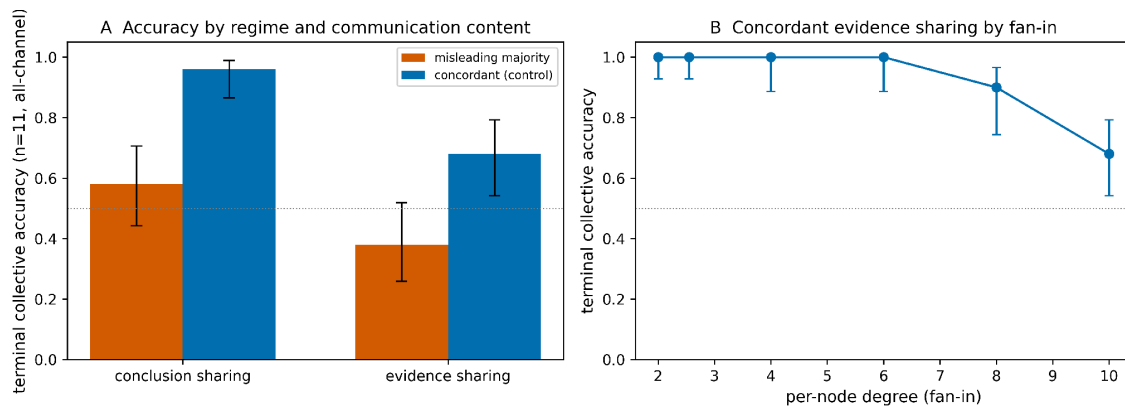


Fig. 6. Communication content and fan-in jointly shape collective accuracy. (A) In the all-channel network at $n = 11$ and degree 10, evidence sharing is less accurate than conclusion sharing in both regimes: 0.38 versus 0.58 in the misleading-majority regime and 0.68 versus 0.96 in the concordant regime. (B) Holding group size fixed at $n = 11$, concordant evidence-sharing accuracy remains at 1.00 through degree 6, then falls to 0.90 at

degree 8 and 0.68 at degree 10. The high-fan-in decline therefore occurs even when every private signal favors the truth.

Alt text: Two-panel figure. Panel A compares evidence-sharing and conclusion-sharing collective accuracy in misleading-majority and concordant regimes at degree 10, showing lower accuracy under evidence sharing in both regimes. Panel B plots concordant evidence-sharing accuracy against per-node degree at fixed group size, remaining at 1.00 through degree 6 before declining at degrees 8 and 10.

Robustness, model heterogeneity, and competence

Within the standard moderate-fan-in setting, the evidence-over-conclusion difference persisted across variations in signal structure. Changing the composition of the misleading majority preserved the qualitative ordering and produced a graded response (Section S10). Collective accuracy before communication remained 1.00 in every condition. Under evidence sharing, accuracy ranged from 0.82 to 0.84. Under conclusion sharing, it declined from 0.54 in the narrow-gap condition to 0.42 in the wide-gap condition and 0.26 under the heavy majority.

Model families differed in susceptibility to conclusion sharing. gpt-4.1-mini and claude-sonnet-4-6 showed substantial content-related accuracy losses, whereas gemini-2.5-flash remained highly accurate and showed little difference between evidence and conclusion sharing. All three families began accurate and successfully integrated evidence. A within-family manipulation identifies reasoning effort as one possible moderator. Enabling extended private reasoning for claude-sonnet-4-6 under conclusion sharing raised collective accuracy from 0.62 to 1.00 ($p = 1.1 \times 10^{-4}$), eliminating the observed failure in that condition. This manipulation does not establish a general intervention or explain Gemini's resistance, particularly because model family and reasoning budget are confounded in the Gemini comparison. It also changed private computation rather than the information exchanged among agents.

We audited the communication-content manipulation in both directions. Across 32,520 evidence-only messages, none disclosed a verdict. Across 40,335 conclusion-only messages, 145 (0.36%) disclosed a reading's identity or diagnosticity, with violations concentrated in claude-sonnet-4-6. In the focal misleading-majority all-channel condition, no gpt-4.1-mini conclusion message disclosed evidence across 1,500 messages in 100 trials. The manipulation was therefore implemented cleanly in the family with the largest sample. In claude-sonnet-4-6, by contrast, 141 of 750 conclusion messages (18.8%) disclosed evidence, almost entirely in the first communication round, so treatment separation was substantially weaker than intended. These disclosures enriched the nominal conclusion-only channel with information that belonged in the evidence condition and therefore work against interpreting the Claude contrast as a clean content separation. Trials with more disclosure were also more accurate, but disclosure was endogenous to model behavior, so this association is observational and does not identify how much the violations attenuated or otherwise changed the estimated content effect (Section S6.10).

Finally, the focal experiments were restricted to models that demonstrated baseline competence at integrating evidence of unequal diagnosticity. Before the main experiments, candidate models completed an independent multi-evidence screening task. gpt-4.1-mini and claude-sonnet-4-6 each achieved misleading-regime integration accuracy of 0.83, and gemini-2.5-flash also passed the screening criterion. claude-haiku-4.5 and gpt-4o-mini achieved accuracy of 0.50 and tended to count signals rather than weight them, so they were excluded. Model-specific screening results are reported in the Supplementary Materials. This screen reduces the likelihood that the focal failures reflect an inability to integrate unequal evidence before social interaction, but it does not establish broader reasoning competence outside the tested task family.

Discussion

The experiments identify two linked forms of collective failure in networks of LLM agents. The first is content-sensitive at moderate fan-in. When a weak numerical majority opposed a better-informed minority, conclusion sharing made an initially accurate collective less accurate under the prespecified logarithmic pool, primarily by increasing variation in terminal collective beliefs across trials rather than by shifting their mean toward the misleading majority. Within trials, agents also became more dispersed in their individual beliefs. The second pattern appears at high fan-in, where the content effect changes sign. At degree 10, evidence sharing was less accurate than conclusion sharing in both the misleading-majority and concordant regimes. The concordant reversal, which occurs when every private signal favors the truth, cannot be explained by deference to a

misleading majority. Communication content and local exposure therefore interact rather than functioning as separable safeguards.

The moderate-fan-in content effect was accompanied by asymmetric changes in how agents used peer information. Evidence sharing enabled weak-majority agents to correct their initial judgments while limiting capitulation by the better-informed minority. Conclusion sharing impaired both processes and increased the frequency of movement toward the numerical majority, although most trials still moved away from it. A planted-conclusion intervention complements this behavioral evidence by showing that the direction of a transmitted conclusion can causally alter terminal collective belief while private evidence is held fixed. Together, these results show that peer verdicts can exert directional influence without implying that ordinary conclusion sharing produces a uniform shift toward the majority.

Loss of evidentiary provenance provides one plausible interpretation. A conclusion reports which answer an agent favors and with what confidence, but it does not reveal the diagnostic strength, origin, or dependence structure of the information supporting that judgment. Recipients therefore cannot directly determine whether several confident conclusions reflect independent evidence or repeated transformations of the same underlying information. The observed updating is consistent with greater reliance on the numerical distribution of peer judgments when those evidentiary distinctions are unavailable. This is a behavioral interpretation, not a claim about an internal computation directly observed within the models.

That pattern connects to, but differs from, established accounts of social influence and collective error. Research on crowd wisdom emphasizes the value of partially independent information (7–11). Information-cascade models distinguish observing others' actions from observing their private signals (16–19), and work on correlated information shows how dependent evidence can be overcounted (13–15). Research on communication networks has also shown that social influence or efficient connectivity can accelerate convergence, reduce independent exploration, or amplify disproportionate influence under some conditions (20–25). Opinion-dynamics models can likewise produce increasing divergence or separated opinion clusters (30, 31), so increased within-group dispersion is not itself a novel signature. In the present experiments, conclusion sharing increased the frequency of trials containing agents at both probability extremes relative to evidence sharing, but the same configuration also occurred under evidence sharing, especially in gpt-4.1-mini. This is therefore best treated as a descriptive polarization-like pattern rather than evidence for a bounded-confidence mechanism. The more specific moderate-fan-in result is the conjunction of sharply increased across-trial collective dispersion with little mean shift under the primary logarithmic pool and persistent within-trial disagreement. Under logarithmic pooling, wrong-side outcomes were also often confident, although their depth was substantially smaller under arithmetic pooling.

Network structure marked where vulnerability was greatest. Across the tested topologies under conclusion sharing, higher local fan-in was associated with lower collective accuracy and greater strong-minority capitulation. The relationship also appeared across positions within heterogeneous networks, most clearly in the wheel, where hub-minus-spoke capitulation was +0.439 [0.288, 0.590]. The largest observed accuracy differences aligned with local fan-in, but the matched path-length contrasts were imprecise: the all-channel versus degree-8 lattice difference was -0.02 [-0.14, +0.10], and the ring-with-shortcuts versus circle difference was -0.06 [-0.14, 0.00]. These experiments therefore do not establish the absence of a path-length effect and do not separately identify causal effects of fan-in and path length. Conditional on the focal agent's own fan-in, neighbor fan-in had partial $r = 0.007$ [-0.056, +0.075], while own fan-in remained positively associated with capitulation after controlling for neighbor fan-in (partial $r = 0.265$ [+0.172, +0.356]). The neighbor estimate is thus tightly centered near zero in the tested graphs, whereas the focal own-fan-in association remains positive.

At the highest fan-in tested, however, preserving evidence no longer functioned as a safeguard and the content ordering reversed. Evidence sharing was less accurate than conclusion sharing in both the misleading-majority and concordant regimes at degree 10, and the interaction across the three paired degree levels was supported by within-seed randomization tests. The high-fan-in deterioration under concordant evidence sharing was robust across aggregation rules and accompanied by increasing individual error and belief diversity. Evidence messages at degree 10 carried 1.82 times as many incoming characters as conclusion messages with the same number of senders. That asymmetry is consistent with message volume or evidentiary complexity contributing to the high-fan-in reversal, but the experiment does not separate those features from provenance, dependency tracking, or

other consequences of richer incoming information. “Load” therefore names the experimental configuration and its plausible computational burden rather than a directly measured internal mechanism.

These findings also sharpen what is specifically learned from studying LLM-agent systems. Prior work has already shown that artificial agents can conform, herd, and propagate correlated errors (26–29). The contribution here is narrower. The task separates numerical agreement from evidentiary strength. The communication manipulation controls whether evidentiary grounds survive transmission. The planted intervention isolates the causal influence of a transmitted conclusion. The network experiments show strong fan-in gradients and positional contrasts while leaving matched path-length effects imprecisely estimated. Most importantly, communication content and local exposure are not simply additive design factors. The relative advantage of evidence sharing changes with fan-in and reverses at the highest fan-in tested in both signal regimes. The resulting evidence is behavioral rather than mechanistic in the interpretability sense. Role transitions, information-sufficiency benchmarks, intervention effects, and reasoning manipulations characterize how observable updating changes, but they do not reveal the models’ internal computations. The experimental construct is also best understood as a limiting case of provenance-poor communication among sub-agents. Deployed systems often exchange summaries, recommendations, partial solutions, or synthesized judgments rather than bare verdicts, but those artifacts vary in how much independently evaluable evidence and source provenance they preserve.

The results therefore suggest a joint design principle rather than two independent safeguards. Preserving independently evaluable evidentiary grounds is beneficial at low and moderate fan-in in the tested conditions, but that advantage reverses at the highest fan-in tested. Systems should therefore consider message content together with the amount of incoming peer information imposed on each agent. This does not identify a universal optimum. Evidence messages may be longer and require explicit provenance tracking, sparse communication can slow dissemination or reduce useful redundancy, and the present design does not separate text volume, evidentiary complexity, provenance, or dependency tracking as causes of the high-fan-in reversal.

Additional private reasoning provides another possible safeguard. In one within-family manipulation, extended reasoning eliminated the conclusion-sharing failure in the tested condition, showing that susceptibility can change with private computation. The intervention also increased output-token use by approximately 4.5-fold and per-agent cost by approximately 3.3-fold, and it was tested in only one model family and one communication condition. It therefore does not establish that additional computation is generally preferable to structural changes in communication. The present experiment also gave additional reasoning to every agent. Whether comparable protection can be obtained more efficiently by allocating extra computation only to high-fan-in agents, better-informed minority agents, or other strategically selected nodes remains an open question.

The evidence condition should not be equated with free-form chains of reasoning. Here, agents transmitted observations with known diagnosticity and provenance, so recipients could evaluate those observations independently. Natural-language rationales may instead combine evidence, assumptions, intermediate conclusions, and information received from other agents. They can therefore preserve useful grounds while still obscuring duplication or dependence. Whether structured rationales, citations, source identifiers, or other provenance-preserving formats provide similar protection is an empirical question.

The consequences of provenance-poor communication may also depend on how often misleading majorities arise outside the experiment. The present task constructs such majorities directly to identify their consequences. Models trained on overlapping data or on one another’s outputs may develop correlated errors or priors (27, 32–34), which could make confidently incorrect agreement more common, but that extrapolation is not tested here. The present experiments establish what happens conditional on a misleading majority, not its prevalence in deployed systems.

Several limitations bound these conclusions. The binary-inference task is intentionally minimal. Its controlled structure permits normative benchmarking and clean manipulation of communication content, but it does not establish that the same effects occur in open-ended tasks. A preliminary hidden-profile materials-comparison probe was largely uninformative because performance was at or near ceiling. All ten communication cells reached collective accuracy of 1.00, while the no-communication controls were 1.00 in the misleading-majority regime and 0.92 in the concordant regime. The ceiling across communication conditions therefore leaves the

probe unable to distinguish a genuine absence of content effects from insufficient task sensitivity and provides no positive evidence of generalization beyond the focal task.

The communication manipulation also bundles multiple properties. Evidence and conclusion messages differ in informational form, length, numerical structure, and source traceability. In the standard moderate-fan-in comparison, evidence messages contained roughly 1.8 times as many characters per sender as conclusion messages while producing higher accuracy, so message volume alone cannot explain why conclusion sharing performed worse there. At degree 10, however, evidence messages were again approximately 1.82 times longer and were less accurate than conclusion messages. Greater incoming volume or complexity therefore remains a plausible contributor to the high-fan-in reversal. The experiment does not separately identify message length, evidentiary complexity, provenance, dependency tracking, or other bundled features as causal mechanisms. Matched-length messages, likelihood-ratio-only messages, and controlled provenance cues could separate these possibilities.

Other design choices further constrain generalization. The planted-conclusion estimates are based on modest matched-seed samples of $N = 15$ and $N = 40$. The reasoning manipulation is a single within-family contrast rather than a graded computation frontier. In the standard moderate-fan-in comparison, the evidence-over-conclusion ordering was robust across the six aggregation rules tested, although effect magnitudes varied. At high fan-in, the primary log-pool ordering reversed, while the concordant evidence-sharing deterioration persisted across alternative aggregation rules. Agents were also instructed to reason like careful Bayesians. Because no neutral-prompt condition was included, persistence under less directive instructions is untested. Communication was fixed at three rounds without endogenous stopping, so it is also unknown whether self-stopping agents would terminate before these failures emerge or continue long enough to amplify them. More generally, the experiments use deliberately constructed signal regimes and fixed communication structures. Generalization to longer interactions, endogenous communication, heterogeneous roles, and naturalistic tasks remains to be established.

Materials and Methods

Task. Each trial drew a hidden state from $\{A, B\}$, with equal prior probability. Each of n agents received one private signal represented by a likelihood ratio and was told the diagnosticity of its reading. In the misleading-majority regime, a numerical majority, defined as $\lceil n/2 \rceil + 1$ agents, received weak readings favoring the false state, while the remaining agents received stronger readings favoring the true state. Robustness conditions varied the diagnosticity gap and the size of the misleading majority. In the concordant control, all agents received readings favoring the same state. For each trial, we computed the full-information Bayesian posterior and the result of a numerical majority vote. These quantities favored opposite states by construction in the misleading-majority regime.

Agents and belief readout. We queried off-the-shelf LLMs through public application programming interfaces without fine-tuning. All experimental manipulations were implemented through prompts. Exact API model identifiers, sampling parameters, and run dates are recorded in each run's `run_meta.json` file. Models used across the experiments included gpt-4.1-mini, claude-sonnet-4-6, gemini-2.5-flash/pro, claude-haiku-4.5, and gpt-4o-mini. Beliefs were elicited using a scaffolded response format ending in `FINAL=<p>`. As a readout check, round-0 beliefs closely recovered the normative single-signal Bayesian posterior across model families (Pearson $r = 0.999$, mean absolute error = 0.004).

Communication conditions. In the `evidence_only` condition, agents shared their reading and its diagnosticity without a verdict. In the `conclusion_only` condition, agents shared their best-guess state and confidence without the underlying reading. The both condition included evidence and a conclusion, while the `none` condition served as the no-communication baseline. Evidence messages therefore identified the observation and its diagnosticity, whereas conclusion messages reported a posterior judgment without revealing the diagnosticity, origin, or possible overlap of the information underlying it.

We audited channel purity in both directions across every transmitted message. Evidence-only messages were scanned for stated preferences for a state, which the condition prohibited, while likelihood-based descriptions of the assigned reading were permitted by design. Conclusion-only messages were scanned for disclosure of a reading's identity or diagnosticity, while the stated verdict and confidence were permitted. Messages that merely

referred to possessing a private reading without disclosing its identity or diagnosticity were recorded separately and not counted as evidentiary disclosure. Audit results are reported in Section S6.10, and the audit code is included with the study materials. Full experimental prompts are provided in the Supplementary Materials.

Network topologies. We tested the baseline, circle, chain, wheel, hierarchy, all-channel, ring-plus-shortcuts small-world network (35), and regular ring lattices of even degree k , in which each node was connected to $k/2$ neighbors on either side. The regular lattices supported comparisons at fixed group size while varying per-node degree. Agents committed privately in round 0 and then exchanged messages for three rounds without early stopping. Complete round-by-round transcripts were retained.

Measures and statistical analysis. Individual error was measured using the Brier score (36). Mean individual error, $\bar{e}$, was the average Brier score across agents, and within-trial belief diversity, D , was the variance of agents' reported probabilities. For the arithmetic-mean pool, these quantities satisfy the identity collective error $= \bar{e} - D$, which held to within 1.6×10^{-4} across the 5,741 binary-inference trials.

The primary collective belief was calculated using a logarithmic opinion pool (37). For agent beliefs p_i that the state is A, each belief was transformed to its logit,

$$\ell_i = \log\left(\frac{p_i}{1 - p_i}\right),$$

the pooled log-odds was the equally weighted mean,

$$\bar{\ell} = \frac{1}{n} \sum_i \ell_i,$$

and the pooled belief was

$$P_{\text{collective}}(A) = \frac{1}{1 + e^{-\bar{\ell}}}.$$

Collective accuracy indicated whether this probability fell on the correct side of 0.5. Because the pool averages rather than sums individual log-odds, it is aligned in sign with, but not numerically equal to, the full-information Bayesian posterior under independent single-signal beliefs and a flat prior.

For the analysis of terminal collective-belief dispersion, we expressed each trial's pooled probability in terms of probability assigned to the true state, q . Across trials, mean squared collective error decomposes exactly as

$$E\left[(1 - q)^2\right] = \left(1 - E[q]\right)^2 + \text{Var}(q).$$

We used this identity to distinguish changes attributable to the mean pooled belief from changes attributable to across-trial dispersion. This across-trial variance is conceptually distinct from D , which measures dispersion among agents within a trial. Robustness to the choice of collective aggregation rule was evaluated using the logarithmic pool, arithmetic-mean pool, majority vote, median belief, a clipped logarithmic pool, and a logarithmic pool omitting the belief with the largest absolute logit on each trial (Section S6.8).

We report Wilson 95% confidence intervals (38). Content contrasts used paired McNemar tests on shared seeds and cluster-robust logistic regression with clustering by task seed. Differences in probability assigned to the true state were evaluated using bootstrap confidence intervals. The content-by-topology interaction in the focal topology sweep was assessed using a likelihood-ratio test. For the $n = 11$ content-by-fan-in analysis, both content conditions were observed on shared seeds at degrees 2, 2.55, and 10. Degree-specific evidence-minus-conclusion risk differences used paired seed bootstrap intervals, and the content $\times$ fan-in interaction was assessed with 10,000 within-seed permutations of content labels using the degree-10 risk difference minus the mean risk difference at degrees 2 and 2.55 as the test statistic. A logistic interaction model was not used because low-degree cells exhibited complete or near-complete separation. The association between collective accuracy and per-node degree was estimated using seed-clustered logistic regression. Matched topology risk differences used paired seed bootstrap intervals and were not interpreted as equivalence tests. Role-transition contrasts used

seed-clustered paired-bootstrap inference. Positional partial correlations and the wheel hub-versus-spoke contrast used bootstrap inference clustered on trial. The post hoc both-extremes diagnostic classified a trial as containing both extremes when at least one agent reported $p < 0.20$ and at least one reported $p > 0.80$. Thresholds of 0.10/0.90 and 0.25/0.75 were used as sensitivity checks. Because degree and diameter covaried across the tested graph family, the structural analyses do not separately identify their causal effects or establish the absence of a path-length effect.

Role-transition and positional analyses. Agents were classified by assigned signal as members of the weak majority or strong minority. A weak-majority agent was classified as corrected when its terminal belief favored the true state. A strong-minority agent was classified as capitulating when its terminal belief favored the false state. Roles defined from assigned signals and realized round-0 beliefs agreed for every agent. The resampling unit was the task seed rather than the individual agent.

To assess positional variation in heterogeneous-degree networks, we estimated the association between strong-minority capitulation and an agent's own fan-in using graph fixed effects. We separately examined whether the mean fan-in of the agent's message senders contributed additional association after controlling for the focal agent's own fan-in, and vice versa. The wheel was excluded from the conditional own-versus-neighbor comparison because own and neighbor fan-in are perfectly collinear in that topology. Details are reported in Section S6.9.

Causal intervention. We planted a first-mover conclusion visible from round 0 while holding all agents' private evidence fixed. On matched seeds, we varied whether the planted conclusion favored the true or false state and included an unplanted control. The average causal effect of plant direction was defined using do-operator notation (39) as

$$ACE = E\left[P(\text{truth}) \mid \text{do}(\text{correct plant})\right] - E\left[P(\text{truth}) \mid \text{do}(\text{incorrect plant})\right].$$

Capability screening. Before the focal experiments, each candidate model completed an independent task requiring integration of multiple signals with unequal diagnosticity. Only models that reliably weighted signal diagnosticity were admitted. gpt-4.1-mini and claude-sonnet-4-6 each achieved misleading-regime integration accuracy of 0.83, and gemini-2.5-flash/pro also passed the screening criterion. claude-haiku-4.5 and gpt-4o-mini achieved accuracy of 0.50 and showed a tendency to count rather than weight signals, so they were excluded. Model-specific screening results are reported in the Supplementary Materials and stored in data/adequacy/. Gemini 2.5 required an enabled thinking budget of 8,192 tokens to pass the screening task.

Models, sampling, and reproducibility. Primary communication-content cells included 100 trials for gpt-4.1-mini, 50 for claude-sonnet-4-6, and 30 for gemini-2.5-flash. Small-world and degree experiments used 50 trials per cell at $n = 11$. The signal-structure sweep used 50 trials per cell. The causal intervention used 15 matched seeds for claude-sonnet-4-6 and 40 for gpt-4.1-mini. Runs were resumable. For the focal binary-inference experiments, raw outputs include complete per-round transcripts, assigned signals, transmitted messages and their sources, elicited beliefs, sampling metadata, and task seeds. The preliminary hidden-profile probe retained task materials and round-level belief outputs but did not persist generated message text.

Data and code availability. The raw data are archived on the Open Science Framework under DOI 10.17605/OSF.IO/AU2GS, released under a CC-BY 4.0 license. Analysis and figure-generation code is maintained at github.com/ericcgldstone-maker/m2m-collective-belief-dynamics and linked from the OSF project. The study materials include per-round transcripts, per-cell result tables, run metadata, analysis scripts, figure-generation scripts, and reproducibility checks. The public code archive and audit.py are synchronized with the analyses and quantities reported in the manuscript and Supplementary Materials, and the archived audit reproduces the reported numerical results.

Use of generative artificial intelligence. Generative artificial intelligence tools were used during the research, analysis, manuscript-development, and revision process. ChatGPT (OpenAI; GPT-5.6 Sol) and Claude Code (Anthropic's command-line interface; Claude Opus 5, model ID claude-opus-5[1m]) were used for code generation and review, analysis and reproducibility checking, editorial review, drafting and revision assistance,

and organization of research and revision materials. I independently evaluated the analyses and interpretations, verified the reported statistical results and citations, and retain sole responsibility for the scientific content.

Acknowledgments

The author thanks Matthew E. Brashears for helpful commentary on an earlier draft of this work.

Funding

The author declares no funding.

Author Contributions

E.G. is the sole author. Following the CRediT taxonomy, E.G. was responsible for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing of the original draft, and review and editing.

Competing Interests

The author declares no competing interests.

References

1. Wang X, et al. (2023) Self-consistency improves chain-of-thought reasoning in language models. *ICLR*. arXiv:2203.11171.
2. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2024) Improving factuality and reasoning in language models through multiagent debate. *ICML*. arXiv:2305.14325.
3. Wang J, Wang J, Athiwaratkun B, Zhang C, Zou J (2025) Mixture-of-agents enhances large language model capabilities. *ICLR*. arXiv:2406.04692.
4. Zheng L, et al. (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS*. arXiv:2306.05685.
5. Wu Q, et al. (2024) AutoGen: enabling next-generation LLM applications via multi-agent conversation. *COLM*. arXiv:2308.08155.
6. Hong S, et al. (2024) MetaGPT: meta programming for a multi-agent collaborative framework. *ICLR*. arXiv:2308.00352.
7. Galton F (1907) Vox populi. *Nature* 75:450–451.
8. de Condorcet N (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale.
9. Surowiecki J (2004) *The Wisdom of Crowds*. Doubleday.
10. Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *PNAS* 101(46):16385–16389.
11. Page SE (2007) *The Difference*. Princeton Univ. Press.
12. Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *PNAS* 108(22):9020–9025.
13. DeMarzo PM, Vayanos D, Zwiebel J (2003) Persuasion bias, social influence, and unidimensional opinions. *Q J Econ* 118(3):909–968.
14. Golub B, Jackson MO (2010) Naïve learning in social networks and the wisdom of crowds. *Am Econ J Microecon* 2(1):112–149.
15. Enke B, Zimmermann F (2019) Correlation neglect in belief formation. *Rev Econ Stud* 86(1):313–332.

16. Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *J Polit Econ* 100(5):992–1026.
17. Banerjee AV (1992) A simple model of herd behavior. *Q J Econ* 107(3):797–817.
18. Anderson LR, Holt CA (1997) Information cascades in the laboratory. *Am Econ Rev* 87(5):847–862.
19. Bikhchandani S, Hirshleifer D, Tamuz O, Welch I (2024) Information cascades and social learning. *J Econ Lit* 62(3):1040–1093.
20. Bavelas A (1950) Communication patterns in task-oriented groups. *J Acoust Soc Am* 22(6):725–730.
21. Leavitt HJ (1951) Some effects of certain communication patterns on group performance. *J Abnorm Soc Psychol* 46(1):38–50.
22. Mason W, Watts DJ (2012) Collaborative learning in networks. *PNAS* 109(3):764–769.
23. Bernstein E, Shore J, Lazer D (2018) How intermittent breaks in interaction improve collective intelligence. *PNAS* 115(35):8734–8739.
24. Hahn U, Hansen JU, Olsson EJ (2020) Truth-tracking performance of social networks. *Synthese* 197(4):1511–1541.
25. Becker J, Brackbill D, Centola D (2017) Network dynamics of social influence in the wisdom of crowds. *PNAS* 114(26):E5070–E5076.
26. Zhu X, Zhang C, Stafford T, Collier N, Vlachos A (2025) Conformity in large language models. *ACL*. arXiv:2410.12428.
27. Kim E, Garg A, Peng K, Garg N (2025) Correlated errors in large language models. *ICML*. arXiv:2506.07962.
28. Flint Ashery A, Aiello LM, Baronchelli A (2025) Emergent social conventions and collective bias in LLM populations. *Sci Adv* 11(20):eadu9368.
29. Cho Y-M, Guntuku SC, Ungar L (2025) Herd behavior: investigating peer influence in LLM-based multi-agent systems. arXiv:2505.21588.
30. Dandekar P, Goel A, Lee DT (2013) Biased assimilation, homophily, and the dynamics of polarization. *PNAS* 110(15):5791–5796.
31. Bernardo C, Altafini C, Proskurnikov A, Vasca F (2024) Bounded confidence opinion dynamics: A survey. *Automatica* 159:111302.
32. Kleinberg J, Raghavan M (2021) Algorithmic monoculture and social welfare. *PNAS* 118(22):e2018340118.
33. Bommasani R, Creel KA, Kumar A, Jurafsky D, Liang P (2022) Picking on the same person: does algorithmic monoculture lead to outcome homogenization? *NeurIPS*. arXiv:2211.13972.
34. Shumailov I, et al. (2024) AI models collapse when trained on recursively generated data. *Nature* 631:755–759.
35. Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. *Nature* 393:440–442.
36. Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3.
37. Genest C, Zidek JV (1986) Combining probability distributions: a critique and an annotated bibliography. *Stat Sci* 1(1):114–135.
38. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 22(158):209–212.

39. Pearl J (2009) *Causality: Models, Reasoning, and Inference*, 2nd Ed. Cambridge Univ. Press.

Supplementary Materials

Communication content and local connectivity shape collective accuracy in networks of machine agents

Eric Gladstone

Section S1. Design predictions and claim-to-results mapping

The study tests directional predictions derived from research on collective intelligence, social learning, and communication networks (Table S1).

Table S1. Design predictions, theoretical basis, and corresponding results.

#	Prediction	Theoretical basis	Result
P1	Independent aggregation is accurate before communication	Wisdom of crowds and the Condorcet tradition	Baseline accuracy = 1.00 (Table 1)
P2	Sharing conclusions reduces collective accuracy, whereas sharing evidence supports accurate integration	Information cascades and the distinction between signals and actions	Supported at moderate fan-in in the two larger-sample model families (evidence 0.90–0.96 versus conclusion 0.62–0.64 in the focal $n = 5$ all-channel comparison; Table 1); at degree 10, the evidence-over-conclusion advantage reversed in both signal regimes (Section S6.5).
P3	A transmitted conclusion causally shifts collective belief when private evidence is held fixed	Causal intervention	ACE = +0.51 and +0.64 (Section S8)
P4	Greater per-node fan-in is associated with lower collective accuracy	Network epistemology	Accuracy declines with degree, seed-clustered logistic regression $p = 4.7 \times 10^{-11}$ (Section S6)
P5	Greater local fan-in is expected to reduce collective accuracy; matched network contrasts test whether this structural pattern can be separated from global path length	Network epistemology	Partially resolved. Accuracy declined across the degree gradient, but degree and path length covary and the matched path-length contrasts were imprecise; they do not establish the absence of a path-length effect (Section S6.5).
P6	The conclusion-sharing accuracy loss is specific to conditions with a misleading numerical majority	Regime contrast	Concordant controls remain near 1.00 under conclusion sharing
P7	The proximal behavioral pathway involves reduced correction of the weak majority and increased capitulation by the strong minority	Signals-versus-actions account	Weak-majority correction declines and strong-minority capitulation increases under conclusion sharing (Section S6.5)

Section S2. Task construction and signal regimes

The latent state was binary, $\theta \in \{A, B\}$, with equal prior probability. Each of n agents received one private signal represented by the likelihood pair ($P(\text{obs} | A)$, $P(\text{obs} | B)$), and the agent was shown both values. The full-information posterior combined the signals by summing their log-likelihood ratios. The naive vote was the fraction of agents whose private signal favored A.

misleading_majority: A numerical majority, defined as $\lceil n/2 \rceil + 1$ agents, received weak signals favoring the false state, with single-signal probabilities assigned to the true state of approximately 0.38–0.46. The remaining agents received stronger signals favoring the true state, with corresponding probabilities of approximately 0.82–0.90. Majority voting therefore favored the false state, while evidence weighting favored the true state.

concordant: All signals weakly favored the true state.

Signal-structure robustness variants (Section S10): `mm_gap_narrow` used majority signals of 0.42–0.48 and minority signals of 0.66–0.74. `mm_gap_wide` used majority signals of 0.30–0.40 and minority signals of 0.86–0.94. `mm_maj_heavy` assigned weak misleading signals to $n - 1$ agents, with majority signals of 0.40–0.46 and the remaining strong signal of 0.88–0.94.

Signals were constructed in an A-favoring orientation and mirrored when the true state was B, making the task symmetric with respect to the true state. Seeds were fixed and paired across experimental cells.

Section S3. Network topologies and adjacency rules

In each communication round, an agent's inbox contained the latest messages from its neighbors as defined by the network adjacency rule (Table S2).

Table S2. Network topologies and adjacency rules.

Topology	Adjacency rule	Per-node degree ($n = 11$)	Diameter ($n = 11$)
Baseline	No edges and no communication	0	—
Circle	Two ring neighbors	2	5
Chain	Bidirectional open line	1–2	10
Wheel	One hub connected bidirectionally to all spokes	Hub 10, spokes 1	2
Hierarchy	Bidirectional binary tree	1–3	$\sim \lceil \log_2 n \rceil$
All-channel	Complete graph	10	1
ring_shortcuts	Ring plus three long-range chords	~ 2.5	3
lattice_k4 / k6 / k8	Ring lattice with $k/2$ neighbors on each side	4 / 6 / 8	3 / 2 / 2

The `lattice_k4_ws` topology used degree-preserving double-edge swaps to create a fixed-degree path-length contrast. At $n \leq 25$, the resulting diameter contrast was too small to be informative. The fan-in analysis therefore relies on the degree gradient and matched contrasts among networks with similar diameters. Because degree and path length covary across the tested graph family, the analyses do not cleanly identify their independent causal effects (Section S6).

Section S4. Agent prompts

The following prompts are reproduced verbatim.

System prompt used in all conditions:

You are one analyst in a group working to identify a hidden state, which is either A or B (equally likely before any evidence). Each analyst has seen a DIFFERENT, independent private reading. Reason like a careful Bayesian: each reading shifts your belief in proportion to its diagnosticity (how much more likely that reading is under one state than the other). Independent readings COMBINE — several distinct readings pointing the same way should make you MORE confident, not less. The only thing to avoid is counting the SAME original reading more than once if it reaches you again after being relayed by someone.

Belief readout used in every round:

After receiving the agent's private evidence and any available messages, the prompt continued:

Treat your own reading and each DISTINCT reading reported by others as independent Bayesian evidence about the hidden state (flat prior; combine by multiplying likelihood ratios ... count a given

original reading only once even if relayed). Work through it in 1–3 short sentences, then end with a line in exactly this format: FINAL=<your probability the hidden state is A, between 0 and 1>

Message-content conditions:

- evidence_only:** "share ONLY the raw evidence you have: state your reading and its diagnosticity (the probabilities of such a reading under state A and under state B). Do NOT say which state you think is more likely. ... One or two sentences."
- conclusion_only:** "share ONLY your current best guess of the hidden state (A or B) and how confident you are. Do NOT describe your underlying reading or any probabilities. One sentence."
- both:** "share your reading and its diagnosticity, AND your current best guess of the hidden state with your confidence. Two or three sentences."

Prompts for the hidden-profile task are reported in Section S11. Belief elicitation used temperature 0, while message generation used temperature 0.7. Round 0 was a private commitment. Agents then exchanged messages for three rounds without early stopping.

Section S5. Belief readout validation

Before communication, elicited beliefs closely recovered the normative posterior implied by each agent's single private signal (Table S3). This result supports the validity of the FINAL=<p> readout for the focal belief measure (Fig. S1).

Table S3. Round-0 belief-readout validation by model family.

Model family	Round-0 observations	Pearson r between elicited belief and Bayesian posterior	MAE
gpt-4.1-mini	16,000	0.999	0.004
claude-sonnet-4-6	2,160	1.000	0.003
gemini-2.5-flash	1,045	1.000	0.002

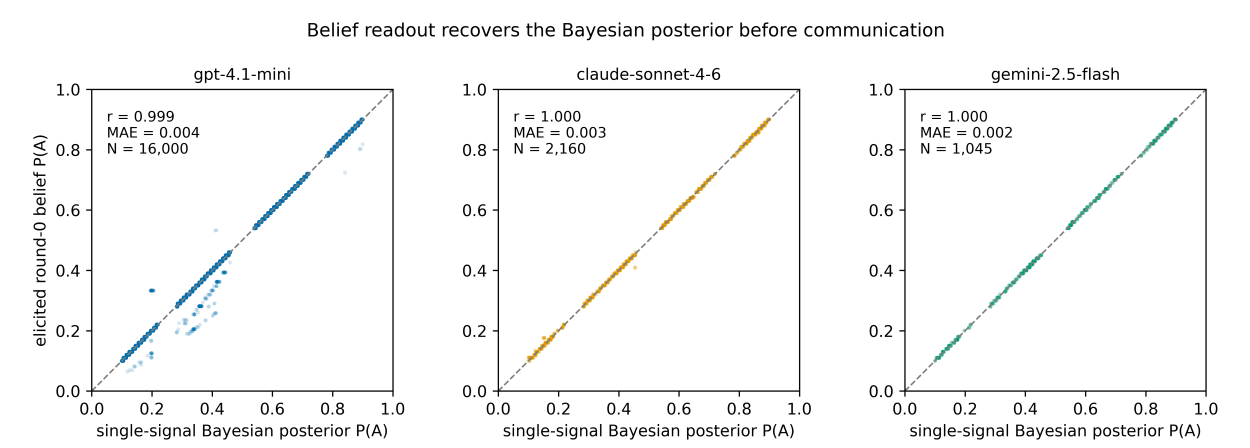


Fig. S1. The belief readout recovers the single-signal Bayesian posterior before communication. For each model family, elicited round-0 beliefs $P(A)$ are plotted against the Bayesian posterior implied by the agent's private signal, pooling all round-0 observations. Points lie close to the identity line ($r \geq 0.999$, $MAE \leq 0.004$). The result shows that the readout accurately captures single-signal belief updating before communication. It does not by itself establish competence at integrating multiple signals. The figure can be reproduced with `make_fig_validation.py`.

Alt text: Scatter plot of elicited round-0 beliefs against the Bayesian posterior implied by each agent's private signal, shown separately for each model family, with points falling close to the identity line.

Section S6. Full results and statistics

Terminal collective accuracy was calculated from the logarithmic opinion pool and is reported with Wilson 95% confidence intervals. Complete cell-level results, including all combinations of topology, communication content, and signal regime, as well as mean individual error $\bar{e}$, diversity D , regret, and $P(\text{truth})$, are reported in `code/stats_report.md`. All quantities can be re-derived from `code/data/*/rounds.csv`.

Section S6.1. Communication-content effect in the misleading-majority, all-channel condition

For gpt-4.1-mini ($N = 100$), collective accuracy was 1.00 [0.96–1.00] with no communication, 0.90 [0.83–0.94] with evidence sharing, 0.81 [0.72–0.87] with both evidence and conclusions, and 0.64 [0.54–0.73] with conclusion sharing. The paired evidence-versus-conclusion contrast yielded McNemar $p = 4.2 \times 10^{-5}$. The paired no-communication-versus-conclusion contrast yielded $p = 2.9 \times 10^{-11}$. A pooled cluster-robust logistic model, clustered by task seed, yielded $p = 8.9 \times 10^{-8}$. For claude-sonnet-4-6 ($N = 50$), accuracy was 1.00 with no communication, 0.96 [0.87–0.99] with evidence sharing, 0.88 with both evidence and conclusions, and 0.62 [0.48–0.74] with conclusion sharing. The paired evidence-versus-conclusion contrast yielded McNemar $p = 2.2 \times 10^{-4}$. The content-by-topology interaction was detectable but modest (likelihood-ratio test, $p = 0.048$).

Section S6.2. Effect sizes

Relative to conclusion sharing, evidence sharing increased accuracy by 0.26 [0.15, 0.37] in gpt-4.1-mini, with an odds ratio of 4.9, and by 0.34 [0.20, 0.48] in claude-sonnet-4-6, with an odds ratio of 12.0. In gpt-4.1-mini, no communication increased accuracy relative to conclusion sharing by 0.36 [0.27, 0.46]. At $n = 11$ under conclusion sharing, the circle increased accuracy relative to the all-channel network by 0.42 [0.28, 0.56]. Risk-difference intervals were calculated by bootstrap.

Section S6.3. Robustness to the aggregation rule

The ordering of communication conditions did not depend on the pooling rule. For gpt-4.1-mini in the misleading-majority, all-channel condition, terminal accuracy under the logarithmic pool, arithmetic-mean pool, and majority vote was 1.00, 1.00, and not applicable for no communication; 0.90, 0.95, and 0.96 for evidence sharing; 0.81, 0.84, and 0.83 for both evidence and conclusions; and 0.64, 0.69, and 0.68 for conclusion sharing. Evidence sharing outperformed conclusion sharing under all three aggregation rules. The full comparison across six aggregation rules, all three model families, the topology sweep, and the degree ladder is reported in Section S6.8, which supersedes this three-rule summary.

Section S6.4. Individual error and within-trial belief diversity

Under conclusion sharing, mean individual error $\bar{e}$ increased while within-trial belief diversity D also increased. Relative to its own pre-communication value, D approximately doubled, from 0.047 at round 0 to 0.091 at the terminal round in gpt-4.1-mini and from 0.048 to 0.093 in claude-sonnet-4-6. Terminal D was 0.058 under evidence sharing and 0.088 when agents shared both evidence and conclusions. Beliefs therefore became more dispersed, not less, between agents influenced by the misleading majority and agents that retained beliefs favoring the true state. For the arithmetic-mean pool, collective error satisfies the identity collective error = $\bar{e} - D$. The increase in collective error under conclusion sharing was therefore attributable to higher mean individual error rather than to reduced diversity. Collective failure in this setting does not require convergence to a single belief. Note that D measures dispersion among agents within a trial and is distinct from the across-trial dispersion of the pooled collective belief analyzed in

Section S6.7. These analyses can be reproduced with `analyze_stats3.py` and `analyze_dispersion.py`.

Section S6.5. Role-transition analysis

Agents were classified by their assigned signals. The weak majority comprised the numerically larger group receiving weak readings favoring the false state. The strong minority comprised the smaller group receiving stronger readings favoring the true state. We measured two terminal transitions relative to the truth. A weak-majority agent was classified as corrected when its terminal belief favored the true state. A strong-minority agent was classified as capitulating when its terminal belief favored the false state.

Role classifications based on assigned signals and realized round-zero beliefs agreed for every agent (Table S4). Inference used a seed-clustered paired bootstrap with 10,000 resamples. The task seed, rather than the nested agent, was the resampling unit. Assignment-based results can be reproduced with `analyze_stats9.py`, while `analyze_stats8.py` implements the corresponding belief-based classification.

Table S4. Role transitions by cell and communication condition.

Cell	Condition	Weak-majority correction	Strong-minority capitulation
Standard (n = 5, 100 seeds)	Evidence	0.93	0.22
	Conclusion	0.69	0.38
Heavy majority (50 seeds)	Evidence	0.95	0.16
	Conclusion	0.33	0.66
n = 11 all-channel	Evidence	0.56	0.43
	Conclusion	0.54	0.42

For the standard condition, the conclusion-minus-evidence risk difference was -0.24 [-0.31 , -0.18] for weak-majority correction and $+0.15$ [$+0.07$, $+0.25$] for strong-minority capitulation. Under the heavy-majority condition, the corresponding differences were -0.63 [-0.70 , -0.55] and $+0.50$ [$+0.32$, $+0.66$]. The larger content-related change occurred in correction of the weak majority. At $n = 11$ in the all-channel network, both contrasts were approximately null, with risk differences near -0.02 and confidence intervals spanning zero. Communication content no longer differentiated the role transitions at this fan-in.

At the trial level, the estimated coefficient relating communication content to collective accuracy attenuated from -0.26 to -0.06 after weak-majority correction and strong-minority capitulation were included. This attenuation is consistent with the two transitions serving as proximal behavioral pathways between communication content and collective accuracy. Because the transitions were measured after treatment, the analysis is descriptive and is not a formal causal mediation estimate. The interpretation that provenance loss makes the numerical distribution of verdicts especially salient is theoretical rather than a directly observed internal computation.

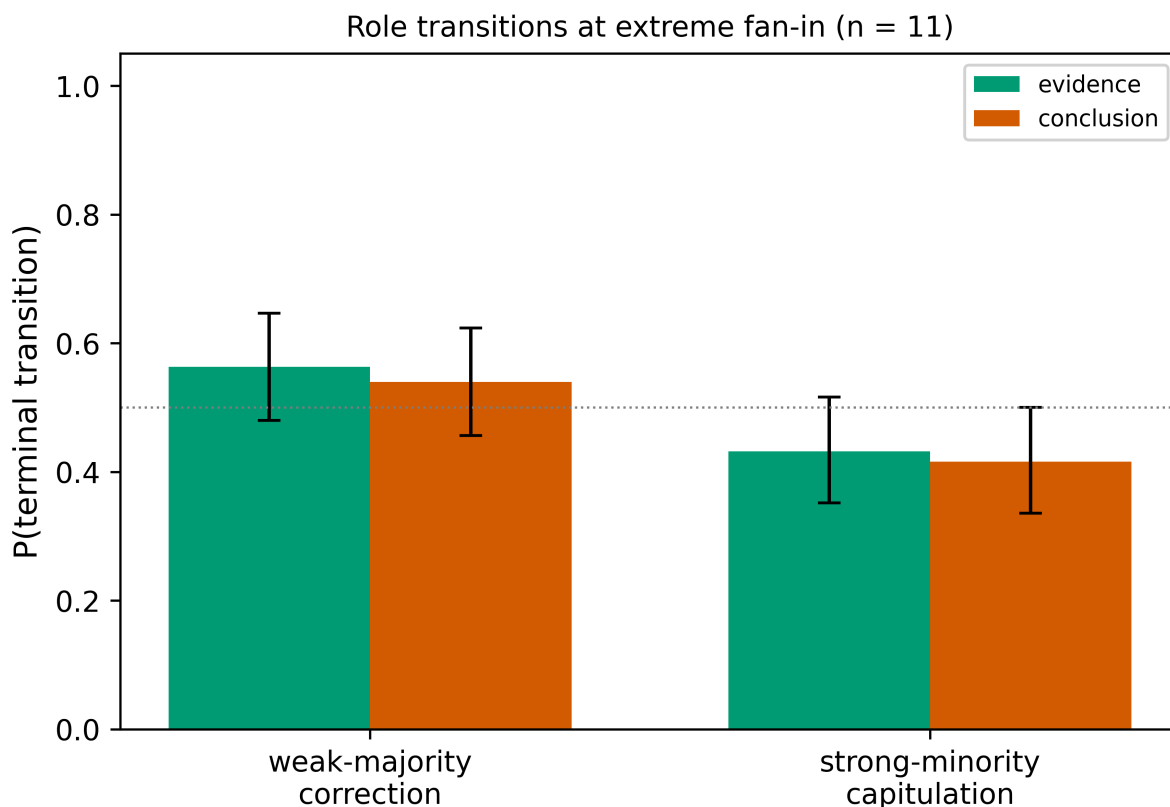


Fig. S2. Role transitions at extreme fan-in. Terminal role transitions in the $n = 11$ all-channel network under the misleading-majority regime, with seed-clustered paired-bootstrap 95% confidence intervals. The content contrast present at $n = 5$ disappears: both roles update unreliably under evidence and conclusion sharing, consistent with the failure observed under high local fan-in in the concordant condition.

Alt text: Bar chart of terminal role transitions at $n = 11$ under evidence and conclusion sharing, showing weak-majority correction and strong-minority capitulation with overlapping confidence intervals.

Variation across network degree. Applying the same role analysis across the $n = 11$ degree ladder under conclusion sharing showed that strong-minority capitulation generally increased with per-node degree, while weak-majority correction generally declined (Table S5). These analyses can be reproduced with `analyze_stats10.py`.

Table S5. Role transitions across the $n = 11$ degree ladder.

Degree and topology	Weak-majority correction	Strong-minority capitulation	Seeds
2, circle	0.67	0.09	50
2.55, ring plus shortcuts	0.70	0.10	50
4, lattice_k4	0.74	0.18	50
6, lattice_k6	0.60	0.44	50
8, lattice_k8	0.47	0.39	50
10, all-channel	0.54	0.42	50

Across these six network conditions, the correlation between degree and capitulation was $r = +0.90$, while the correlation between degree and correction was $r = -0.82$. The all-channel network is also the highest-fan-in condition, so neither transition is strictly monotonic across the full range. The clearest increase in capitulation occurs through approximately degree 6.

Within-network association between fan-in and capitulation. Heterogeneous-degree networks permit comparison among positions within the same topology. Among strong-minority agents under conclusion sharing in the misleading-majority regime, capitulation increased with the agent's own fan-in (Table S6). These analyses can be reproduced with `analyze_stats11.py`.

Table S6. Capitulation by own fan-in in heterogeneous-degree networks.

Own fan-in	Capitulation	Strong-minority agents	Wheel comparison
1	0.05	356	Spokes: 0.04
2	0.15	161	
3	0.26	39	
4	0.48	44	Hub: 0.48

Signals were randomly assigned to network positions on each trial. The probability of assignment to the strong minority was approximately 0.40 at every fan-in level, and mean absolute log-likelihood-ratio magnitude among minority agents was approximately 1.8 across positions. These balance checks can be reproduced with `analyze_stats12.py`.

After demeaning fan-in and capitulation within each topology, the graph-fixed-effects association was $r = +0.36$ [seed-clustered bootstrap 95% CI: $+0.26, +0.45$] across 600 strong-minority agents in 300 trials. Estimated within-graph slopes were positive in the wheel, hierarchy, and chain, at $+0.15$, $+0.10$, and $+0.11$ change in capitulation per additional unit of fan-in, respectively. The pooled correlation was numerically identical to the fixed-effects estimate. These results show that the association is present within networks and is not produced solely by combining topologies with different average degrees. They do not by themselves establish a causal effect of network position.

Fan-in and collective accuracy. At $n = 11$ under conclusion sharing, collective accuracy followed the degree gradient: 1.00 for the circle at degree 2, 0.94 for `ring_shortcuts` at mean degree 2.5, 0.90 for `lattice_k4`, 0.76 for `lattice_k6`, 0.60 for `lattice_k8`, and 0.58 for the all-channel network at degree 10. A seed-clustered logistic regression of accuracy on per-node degree yielded $p = 4.7 \times 10^{-11}$.

Degree and diameter covaried across the graph family. In a joint model, the degree and diameter coefficients were individually imprecise ($p = 0.082$ and $p = 0.053$), so the analysis does not cleanly separate their independent effects. Paired seed-bootstrap contrasts likewise do not establish the absence of a path-length effect. The all-channel network versus `lattice_k8` differed in accuracy by -0.020 $[-0.140, +0.100]$, while `ring_shortcuts` versus the circle differed by -0.060 $[-0.140, 0.000]$. The largest structural contrast, all-channel versus `ring_shortcuts`, was -0.360 $[-0.520, -0.200]$ and combines a large fan-in difference with a path-length difference. The strongest evidence for local fan-in therefore comes from the overall degree gradient and within-network positional analyses, while the matched path-length contrasts remain imprecise. No equivalence test is reported because no equivalence margin was specified before these analyses.

Content $\times$ fan-in and regime contrast. The $n = 11$ degree ladder is not fully content-crossed. Both evidence and conclusion sharing were observed on the same seeds at degrees 2, 2.55, and 10. At degrees 4, 6, and 8, the conclusion cells contain 50 trials, while concordant evidence-sharing cells contain 30 trials from the separate overload series; no misleading-majority evidence cells were run at those intermediate

degrees. The formal content $\times$ fan-in test therefore uses the three paired degree levels rather than treating the design as a six-level factorial comparison.

In the misleading-majority regime, evidence-minus-conclusion accuracy was 0.000 [0.000, 0.000] at degree 2, +0.060 [0.000, +0.140] at degree 2.55, and -0.200 [-0.320, -0.080] at degree 10. In the concordant regime, the corresponding paired contrasts were 0.000 [0.000, 0.000], 0.000 [0.000, 0.000], and -0.280 [-0.420, -0.140]. A within-seed randomization test, permuting content labels within seed and using the degree-10 risk difference minus the mean risk difference at degrees 2 and 2.55 as the statistic, supported a content $\times$ fan-in interaction in both regimes (10,000 permutations; $p = 0.0015$ misleading majority; $p = 0.0014$ concordant). At the highest fan-in tested, evidence sharing was therefore significantly less accurate than conclusion sharing in both regimes. The concordant reversal cannot be attributed to deference to a misleading numerical majority because every private signal favored the truth.

Degree-controlled analysis of high-fan-in degradation. Holding group size fixed at $n = 11$, accuracy under concordant evidence sharing was 1.00 for the circle at degree 2, 1.00 for `ring_shortcuts` at mean degree 2.55, 1.00 for `lattice_k4`, 1.00 for `lattice_k6`, 0.90 for `lattice_k8`, and 0.68 for the all-channel network. The circle, `ring_shortcuts`, and all-channel cells contained 50 trials each. The regular-lattice cells contained 30 trials each.

Accuracy was unchanged through degree 6 and declined at degrees 8 and 10. Because group size was constant across these conditions, the decline was associated with fan-in rather than the number of agents. Because the decline occurred in the concordant regime, it was not produced by deference to a misleading numerical majority. The pattern is consistent with a failure associated with high local incoming-information load. That load was not directly measured, and per-node degree jointly changes the number of senders, the volume of incoming text, and the need to track dependence among messages, so the analysis does not identify which component is responsible. Data are stored in `code/data/gpt-4.1-mini_overload/` for the lattice conditions and `code/data/gpt-4.1-mini_n11/` for the remaining networks.

Message length is not a sufficient explanation of the moderate-fan-in content effect, because the less accurate conclusion-sharing condition carries less text there. At degree 10, however, the ordering reverses and the evidence-sharing condition carries substantially more text. Incoming message volume by degree and content mode is characterized in Section S6.11. Because content, message length, and evidentiary complexity are bundled, this pattern is descriptive and does not identify message volume as the mechanism.

Error-diversity identity. For the arithmetic-mean pool, the identity collective Brier error = mean individual error $\bar{e}$ – diversity D held to a maximum absolute discrepancy of 1.6×10^{-4} across all 6,341 trials in the full dataset, including the hidden-profile probe.

Section S6.6. Information sufficiency and the use of verdicts

The misleading-majority design was constructed so that vote counting and evidence weighting favor different answers. Two analyses test whether conclusion sharing failed because the transmitted verdicts were uninformative, or because of how agents used them. Both can be reproduced with `analyze_benchmark.py` and `analyze_recipient.py`.

Confidence-weighted benchmark. We read the agents' round-0 verdicts, which encode each agent's verdict and confidence at first broadcast, in three ways and compared each reading's accuracy with the agents' actual terminal accuracy. The endpoints are fixed by the signal design. An unweighted head-count of the verdicts favored the false state on every trial, whereas a confidence-weighted reading, numerically equivalent to the no-communication baseline, favored the true state on every trial. The informative quantity is the gap between this achievable accuracy and the agents' actual terminal accuracy, which varied across conditions (Table S7).

Table S7. Round-0 verdicts read three ways, by model family and content condition.

Model family	Content	N	Head-count	Confidence-weighted	Log-opinion pool	Actual	Gap
gpt-4.1-mini	evidence	100	0.00	1.00	1.00	0.90	0.10
gpt-4.1-mini	conclusion	100	0.00	1.00	1.00	0.64	0.36
gpt-4.1-mini	both	100	0.00	1.00	1.00	0.81	0.19
claude-sonnet-4-6	evidence	50	0.00	1.00	1.00	0.96	0.04
claude-sonnet-4-6	conclusion	50	0.00	1.00	1.00	0.62	0.38
claude-sonnet-4-6	both	50	0.00	1.00	1.00	0.88	0.12
gemini-2.5-flash	evidence	30	0.00	1.00	1.00	0.97	0.03
gemini-2.5-flash	conclusion	30	0.00	1.00	1.00	0.93	0.07
gemini-2.5-flash	both	30	0.00	1.00	1.00	0.97	0.03

The conclusion channel therefore preserved sufficient information for correct aggregation. The shortfall from the achievable confidence-weighted reading was large under conclusion sharing in the two susceptible families, at 0.36 and 0.38, small under evidence sharing, between 0.03 and 0.10, and small for gemini-2.5-flash under conclusion sharing, at 0.07. Collective failure under conclusion sharing reflected how agents weighted the verdicts, by their number rather than their confidence, not a loss of information from the channel.

Normative-recipient counterfactual. We next asked whether individual strong-minority agents capitulated more than a confidence-weighted reading of the verdicts they received would justify. For each strong-minority agent under conclusion sharing in the misleading-majority regime, we reconstructed the verdicts received from its network neighbors and computed the normative posterior it would hold if it combined its own private posterior with those neighbors' first-round verdicts using a logarithmic opinion pool. A normative capitulation occurred when this confidence-weighted posterior favored the false state. We compared it with the agent's actual terminal capitulation across 1,500 strong-minority agent-trials spanning six fan-in levels, pooling `gpt-4.1-mini_nll` and `gpt-4.1-mini_degree`.

A confidence-weighted reading of the received verdicts justified capitulation in 0.1 percent of cases, because the contrary verdicts from the weak majority were individually weak. Agents actually capitulated in 26.9 percent of cases. The excess capitulation, beyond what confidence weighting justified, increased with the agent's own fan-in (Table S8).

Table S8. Normative and actual capitulation by own fan-in.

Own fan-in	Strong-minority agent-trials	Normative capitulation	Actual capitulation	Excess
2	359	0.00	0.08	0.08
3	141	0.00	0.14	0.14
4	250	0.00	0.18	0.18
6	250	0.00	0.44	0.44
8	250	0.00	0.39	0.39
10	250	0.00	0.42	0.42

Normative capitulation remained near zero at every fan-in level, because additional weak verdicts never justified abandoning a strong correct belief. Actual capitulation rose from 0.08 at fan-in 2 to 0.42 at fan-in 10, and the cell-level correlation between fan-in and excess capitulation was $r = 0.91$. Because the normative posterior was computed from the verdicts the agents actually received, the growing excess

identifies a response to the number of contrary verdicts rather than to their confidence-weighted content. This is the behavioral signature of vote counting, and it is not entailed by the content manipulation.

Section S6.7. Dispersion of terminal collective belief and the collective-Brier decomposition

For truth-normalized pooled probability p , mean collective Brier error decomposes exactly as $E[(1 - p)^2] = (1 - E[p])^2 + \text{Var}(p)$, separating a mean-location component from an across-trial variance component. The identity held to within 10^{-12} in every cell. Population variance is used throughout. These analyses can be reproduced with `analyze_dispersion.py`.

Misleading-majority, all-channel, logarithmic pool (Table S9).

Table S9. Terminal collective belief and Brier decomposition under the logarithmic pool.

Family	Condition	N	mean p	SD	Brier	$(1 - E[p])^2$	Var(p)	wrong-side	mean p if wrong
gpt-4.1-mini	no communication	100	0.632	0.015	0.1355	0.1353	0.0002	0.00	—
	evidence, terminal	100	0.805	0.225	0.0886	0.0382	0.0504	0.10	0.197
	conclusion, terminal	100	0.643	0.381	0.2723	0.1275	0.1448	0.36	0.165
claude-sonnet-4-6	no communication	50	0.634	0.015	0.1344	0.1342	0.0002	0.00	—
	evidence, terminal	50	0.898	0.158	0.0354	0.0103	0.0250	0.04	0.185
	conclusion, terminal	50	0.659	0.393	0.2708	0.1162	0.1546	0.38	0.171

Decomposition of the change, by contrast and pooling rule (Table S10).

Table S10. Decomposition of the change in Brier error, by contrast and pooling rule.

Contrast	Pool	Δ Brier	mean component	variance component
gpt: conclusion terminal – own round 0	log	+0.1368	−0.0078 (−6%)	+0.1446 (106%)
gpt: conclusion terminal – own round 0	arithmetic	+0.0342	−0.0308 (−90%)	+0.0650 (190%)
claude: conclusion terminal – own round 0	log	+0.1366	−0.0178 (−13%)	+0.1544 (113%)
claude: conclusion terminal – own round 0	arithmetic	+0.0003	−0.0616	+0.0619
gpt: conclusion – evidence, terminal, paired	log	+0.1837	+0.0893 (49%)	+0.0944 (51%)
gpt: conclusion – evidence, terminal, paired	arithmetic	+0.1353	+0.0895 (66%)	+0.0458 (34%)
claude: conclusion – evidence, terminal, paired	log	+0.2355	+0.1059 (45%)	+0.1296 (55%)
claude: conclusion – evidence, terminal, paired	arithmetic	+0.1356	+0.0878 (65%)	+0.0478 (35%)

Relative to each condition's own pre-communication baseline, the variance component more than accounts for the increase in Brier error under conclusion sharing, because the mean component improves slightly. This holds under both pooling rules, although the *net* baseline increase is small under arithmetic pooling in gpt-4.1-mini and approximately zero in claude-sonnet-4-6, where a variance term of +0.0619 is offset by a mean-component improvement of −0.0616. The baseline decomposition should therefore be read as a property of the prespecified logarithmic pool. Relative to evidence sharing, roughly half the terminal gap under the log pool and two thirds under the arithmetic pool is attributable to the mean component, because evidence sharing raises mean p while conclusion sharing does not.

Direction of movement relative to each trial's own round 0. Movement toward the misleading majority is a decrease in pooled $P(\text{true state})$ (Table S11).

Table S11. Direction of movement relative to each trial's own round 0.

Family	Condition	toward majority	mean Δ	away	mean Δ	net mean Δ	crossed to wrong side
gpt-4.1-mini	evidence	11%	-0.399	89%	+0.243	+0.173	10%
gpt-4.1-mini	conclusion	39%	-0.440	61%	+0.299	+0.011	36%
claude-sonnet-4-6	evidence	4%	-0.445	96%	+0.294	+0.265	4%
claude-sonnet-4-6	conclusion	38%	-0.461	62%	+0.323	+0.025	38%

Conclusion sharing increases the frequency of majority-ward movement without materially changing its depth, and most trials still move away from the misleading majority. A literal vote-counting rule would be incorrect on 100% of misleading-majority trials by construction.

Terminal distribution shape. Counts of terminal pooled P(true state), gpt-4.1-mini, misleading-majority, all-channel, logarithmic pool (Table S12).

Table S12. Distribution of terminal pooled P(true state) in gpt-4.1-mini.

Bin	0-.05	.05-.1	.1-.2	.2-.3	.3-.4	.4-.5	.5-.6	.6-.7	.7-.8	.8-.9	.9-.95	.95-1
evidence	1	2	3	1	2	1	1	4	15	27	25	18
conclusion	10	6	7	6	3	4	3	3	3	10	9	36

The evidence-sharing distribution is unimodal and concentrated at high probabilities (quantiles $p_{05} = 0.164$, $p_{50} = 0.881$, $p_{95} = 0.990$). The conclusion-sharing distribution accumulates mass near both certainty bounds with a sparse interior ($p_{05} = 0.016$, $p_{50} = 0.859$, $p_{95} = 1.000$). We describe this shape descriptively and do not fit a modality test.

Within-trial both-extremes configurations. To determine whether increased within-trial diversity reflected diffuse disagreement or agents occupying opposing probability extremes, we constructed a post hoc trial-level indicator equal to one when at least one agent reported a terminal probability below 0.20 and at least one reported a probability above 0.80. Because the threshold was selected after inspection of the distribution, we repeated the analysis at 0.10/0.90 and 0.25/0.75. Inference used paired seed bootstrap intervals.

At the 0.20/0.80 threshold, prevalence was 0.00 at round 0 in both families. In gpt-4.1-mini it was 0.37 under evidence sharing and 0.54 under conclusion sharing, a conclusion-minus-evidence difference of +0.17 [+0.03, +0.31]. In claude-sonnet-4-6 it was 0.18 and 0.52, respectively, a difference of +0.34 [+0.18, +0.50]. The conclusion-minus-evidence difference remained positive at both alternative thresholds: +0.27 [+0.15, +0.39] and +0.16 [+0.02, +0.30] in gpt-4.1-mini at 0.10/0.90 and 0.25/0.75, and +0.36 [+0.20, +0.52] and +0.30 [+0.14, +0.46] in claude-sonnet-4-6.

The configuration is therefore induced by communication and is more frequent under conclusion sharing, but it is not specific to the failing condition. Evidence sharing produces both-extremes configurations on 37% of gpt-4.1-mini trials at the primary descriptive threshold. These results support describing the terminal beliefs as sometimes bipolar or extremized, but they do not identify polarization or bounded-confidence updating as the mechanism.

Among conclusion-sharing trials, the 0.20/0.80 configuration occurred on 0.47 of collectively correct and 0.67 of collectively incorrect gpt-4.1-mini trials, and on 0.39 of correct and 0.74 of incorrect claude-sonnet-4-6 trials. These are descriptive post-treatment associations and are not causal estimates.

Pooling sensitivity of wrong-side depth. Under the logarithmic pool, the wrong-side rate was 0.36 in gpt-4.1-mini and 0.38 in claude-sonnet-4-6, with mean P(true) among wrong trials of 0.165 and 0.171.

Under arithmetic pooling, wrong-side rates were 0.31 and 0.26, with corresponding means of 0.316 and 0.334. Wrong-side outcomes therefore persist under both rules, but the characterization of those outcomes as highly confident is substantially stronger under the prespecified logarithmic pool. These quantities can be reproduced with `adversarial_followup_analysis.py`.

Diversity across the degree ladder. Belief diversity D and mean individual error $\bar{e}$, concordant evidence sharing at $n = 11$, the condition in which the high-fan-in failure appears without any misleading majority (Table S13).

Table S13. Belief diversity and mean individual error across the degree ladder.

Degree	Topology	N	$\bar{e}$ round 0	$\bar{e}$ terminal	D round 0	D terminal	ΔD	Accuracy
2	circle	50	0.076	0.005	0.007	0.002	-0.004	1.00
2.55	ring + shortcuts	50	0.076	0.010	0.007	0.008	+0.001	1.00
4	lattice_k4	30	0.076	0.021	0.006	0.016	+0.010	1.00
6	lattice_k6	30	0.077	0.093	0.007	0.071	+0.064	1.00
8	lattice_k8	30	0.076	0.314	0.006	0.128	+0.122	0.90
10	all-channel	50	0.077	0.434	0.007	0.091	+0.084	0.68

Mean individual error rose over the same interval, from 0.210 at round 0 to 0.288 in gpt-4.1-mini and 0.255 in claude-sonnet-4-6 under conclusion sharing, compared with 0.120 and 0.052 respectively under evidence sharing.

Diversity expands as accuracy declines. At low degree, evidence sharing both improves accuracy and reduces diversity; the failure regimes are accompanied by increased diversity. Under conclusion sharing in the misleading-majority regime the same direction holds ($\Delta D = +0.051$, $+0.046$ and $+0.072$ at degrees 2, 2.55 and 10).

Section S6.8. Aggregation-rule comparison

Six rules applied to identical terminal beliefs: the logarithmic pool (primary), the arithmetic mean, an unweighted majority head-count, the median belief, a logarithmic pool with beliefs clipped to $[0.05, 0.95]$, and a logarithmic pool omitting each trial's largest-magnitude logit. Reproducible with `analyze_pools.py`.

A linear confidence-weighted vote, assigning weight $|2p - 1|$ to the vote $\text{sign}(p - 0.5)$, sums to $2n(\text{mean}(p) - 0.5)$ and therefore has the same sign as the arithmetic pool. It is not a distinct classification rule, so the median pool is reported instead as the robust reference point.

Content contrast by model family (misleading-majority, all-channel, terminal) (Table S14).

Table S14. Terminal accuracy under six aggregation rules, by model family.

Cell	N	log	arithmetic	majority	median	clip .05	drop-max
gpt-4.1-mini, evidence	100	0.90	0.95	0.96	0.96	0.96	0.94
gpt-4.1-mini, conclusion	100	0.64	0.69	0.68	0.68	0.70	0.66
gpt-4.1-mini, both	100	0.81	0.84	0.83	0.83	0.89	0.76
claude-sonnet-4-6, evidence	50	0.96	0.98	0.98	0.98	0.98	0.98
claude-sonnet-4-6, conclusion	50	0.62	0.74	0.78	0.78	0.76	0.72
claude-sonnet-4-6, both	50	0.88	0.96	0.96	0.96	0.98	0.96
gemini-2.5-flash, evidence	30	0.97	1.00	1.00	1.00	1.00	1.00
gemini-2.5-flash, conclusion	30	0.93	0.93	0.90	0.90	0.93	0.97

Evidence minus conclusion, by rule (Table S15).

Table S15. Evidence-minus-conclusion accuracy difference, by aggregation rule.

Family	log	arithmetic	majority	median	clip .05	drop-max
gpt-4.1-mini	+0.26	+0.26	+0.28	+0.28	+0.26	+0.28
claude-sonnet-4-6	+0.34	+0.24	+0.20	+0.20	+0.22	+0.26
gemini-2.5-flash	+0.03	+0.07	+0.10	+0.10	+0.07	+0.03

The ordering is preserved under every rule. In gpt-4.1-mini the magnitude is nearly rule-invariant. In claude-sonnet-4-6 the logarithmic pool is the most favorable of the six rules for the content contrast, and we report that asymmetry explicitly.

Degree ladder, concordant evidence sharing (the high-fan-in failure) (Table S16).

Table S16. Degree ladder under concordant evidence sharing, by aggregation rule.

Degree	Topology	N	log	arithmetic	majority	median	clip .05	drop-max
2	circle	50	1.00	1.00	1.00	1.00	1.00	1.00
2.55	ring + shortcuts	50	1.00	1.00	1.00	1.00	1.00	1.00
4	lattice_k4	30	1.00	1.00	1.00	1.00	1.00	1.00
6	lattice_k6	30	1.00	1.00	1.00	1.00	1.00	1.00
8	lattice_k8	30	0.90	0.67	0.63	0.63	0.67	0.87
10	all-channel	50	0.68	0.50	0.50	0.50	0.50	0.64

Accuracy remains 1.00 through degree 6 under all six rules. At degree 10 it is 0.68 under the logarithmic pool and between 0.50 and 0.64 under the five alternatives, so the high-fan-in failure is larger under the alternative rules than under the primary one and is not produced by the primary rule.

Degree ladder, misleading-majority conditions (Table S17).

Table S17. Degree ladder under misleading-majority conditions, by aggregation rule.

Degree	Topology	N	Content	log	arithmetic	majority	median	clip .05	drop-max
2	circle	50	conclusion	1.00	1.00	1.00	1.00	1.00	1.00
2.55	ring + shortcuts	50	conclusion	0.94	1.00	1.00	1.00	1.00	0.98
10	all-channel	50	conclusion	0.58	0.58	0.56	0.56	0.58	0.64
2	circle	50	evidence	1.00	1.00	1.00	1.00	1.00	1.00
2.55	ring + shortcuts	50	evidence	1.00	1.00	1.00	1.00	1.00	1.00
10	all-channel	50	evidence	0.38	0.48	0.52	0.52	0.48	0.38

In the misleading-majority evidence cell at degree 10 the logarithmic pool is the *least* favorable of the rules (0.38 against 0.52 under majority and median). The primary rule is therefore not uniformly favorable to the reported claims.

Section S6.9. Neighbor fan-in

The within-graph analysis in Section S6.5 shows that an agent's own fan-in is associated with capitulation. We tested whether a neighbor's fan-in contributes beyond that, using the heterogeneous-degree topologies (wheel, hierarchy, chain) among 600 strong-minority agents in 300 trials under conclusion sharing. A neighbor is an agent whose message was received in round 1 (Table S18). Reproducible with `analyze_load.py`.

Table S18. Correlations between capitulation and own or neighbor fan-in.

Quantity	r
Own fan-in and capitulation	+0.357
Neighbor mean fan-in and capitulation	−0.248
Neighbor fan-in and capitulation, controlling for own fan-in	+0.007
Own fan-in and capitulation, controlling for neighbor fan-in	+0.265
Own and neighbor fan-in	−0.709

Across 600 strong-minority agent-trials in 300 trial-graphs, own fan-in was positively associated with capitulation ($r = +0.357$, cluster-bootstrap 95% CI $[+0.257, +0.450]$), while mean neighbor fan-in had an unconditional negative association ($r = -0.248$ $[-0.319, -0.168]$) because own and neighbor fan-in are negatively correlated in the heterogeneous-degree graphs. Conditional on an agent's own fan-in, neighbor mean fan-in had partial $r = +0.007$ $[-0.056, +0.075]$. Conversely, the focal agent's own fan-in remained positively associated with capitulation after controlling for neighbor mean fan-in (partial $r = +0.265$ $[+0.172, +0.356]$). Bootstrap uncertainty was clustered on trial.

The neighbor interval is sufficiently narrow to exclude more than a small second-order association in these tested graphs. The result therefore supports localization of the measurable association to focal own fan-in rather than a detectable neighbor-fan-in term, without establishing a strictly local causal mechanism.

The wheel cannot contribute to the conditional own-versus-neighbor comparison because a hub's neighbors are exactly the spokes and a spoke's only neighbor is the hub, making own and neighbor fan-in perfectly collinear. The conditional estimate is therefore identified by the hierarchy and chain. Separately, the wheel provides a strong direct positional contrast: hub-minus-spoke capitulation was $+0.439$ $[+0.288, +0.590]$.

These uncertainty estimates can be reproduced with `adversarial_followup_analysis.py`.

Section S6.10. Channel-purity audit

Both communication channels were audited by automated text classification of every transmitted message. Reproducible with `check_leakage.py`.

In the evidence-only channel we scanned for a stated preference for a state, which the condition prohibits; likelihood phrasing describing a reading under each state is permitted by design. Across all 32,520 evidence-only messages, no message disclosed a verdict.

In the conclusion-only channel we scanned for disclosure of evidence: the identity of a reading, its probability under each state, or a likelihood ratio. The verdict and its confidence value are permitted by design and were removed before scanning. Two categories are distinguished. A message that merely refers to possessing a private reading, without disclosing anything about it, transmits no evidence and is recorded as a referential mention. A message that states a reading's identity or diagnosticity is recorded as substantive disclosure. Referential mentions were tracked separately and were not counted as substantive disclosure. Across all 40,335 conclusion-only messages, 145 (0.36%) were substantive disclosures.

Substantive disclosure was concentrated in one model family (Table S19).

Table S19. Substantive evidentiary disclosure in conclusion-only messages, by dataset.

Dataset	Messages	Substantive	Rate
gpt-4.1-mini	15,000	1	0.01%
gpt-4.1-mini, n = 11 / degree / signal sweep	22,050	0	0.00%
claude-sonnet-4-6	1,500	141	9.40%
claude-sonnet-4-6, extended reasoning	900	0	0.00%
gemini-2.5-flash	885	3	0.34%

In the focal misleading-majority all-channel cell, no gpt-4.1-mini conclusion message disclosed evidence, across all 1,500 messages in 100 trials. In claude-sonnet-4-6, 141 of 750 messages (18.8%) did so, a mean of 2.82 messages per 15-message trial. Disclosure was concentrated in the first communication round: 55.6% of round-1 messages, 0.8% of round-2 messages, and none at round 0.

Trials containing more disclosure ended more accurate (Table S20).

Table S20. Trial accuracy by number of disclosing messages.

Disclosing messages in trial	0	1	2	3	4	5
Trials	2	9	10	10	13	6
Terminal accuracy	1.00	0.33	0.50	0.60	0.77	0.83
Mean P(true state)	0.901	0.454	0.584	0.610	0.768	0.858

Because disclosure enriched the nominal conclusion-only channel with evidentiary information, it would be expected to reduce rather than create the measured content contrast. This association is observational: the number of disclosing messages is a property of model behavior on that trial and may covary with trial features, and only two trials in this cell are free of disclosure, so the relationship does not identify the causal magnitude of any attenuation. The manipulation was implemented exactly as designed in gpt-4.1-mini, the family carrying the largest sample.

Section S6.11. Incoming message load by degree and content mode

Per-node degree changes several quantities simultaneously, including the number of senders, the volume of incoming text, and the demand of tracking dependence among messages. This section characterizes incoming message volume so that message volume can be assessed separately from the number of

senders. The analysis is descriptive: message length is chosen by the agents and is therefore endogenous to their behavior (Table S21). Values are per agent-round, at $n = 11$, and can be reproduced with `analyze_load.py`.

Table S21. Incoming message load by degree and content mode.

Topology	Per-node degree	Content	Senders	Incoming characters	Characters per sender
circle	2	conclusion	2.0	141	71
circle	2	evidence	2.0	241	120
ring + shortcuts	2.55	conclusion	2.5	171	67
ring + shortcuts	2.55	evidence	2.5	305	120
all-channel	10	conclusion	10.0	653	65
all-channel	10	evidence	10.0	1,188	119

Two observations follow. First, characters per sender are nearly constant within a content condition across the degree ladder, so total incoming volume scales almost linearly with the number of senders. Because within-cell variation in incoming volume is small (coefficient of variation 0.02 to 0.25), these data cannot separate the number of senders from the volume of incoming text as drivers of high-fan-in failure. Second, evidence messages carried approximately 1.8 times as many characters per sender as conclusion messages at every degree tested. At low and moderate fan-in, evidence sharing was nevertheless more accurate, so incoming text volume is not sufficient to explain the focal content effect. At degree 10, the sender count is identical at 10 per receiving agent, while evidence messages generate approximately 1,188 and 1,185 incoming characters per agent-round in the misleading-majority and concordant regimes compared with 653 and 650 under conclusion sharing, a ratio of 1.82. The condition carrying more text is also less accurate at this highest fan-in. This is consistent with incoming volume or complexity contributing to the upper-degree reversal, but content, message length, evidentiary complexity, provenance, and dependency-tracking burden are bundled by the design and cannot be separated causally.

Section S7. Capability screening

Candidate models were admitted only if they could integrate multiple signals with unequal diagnosticity rather than rely on signal counts. In an independent screening task, a single agent observed all five readings and reported its belief without communicating with other agents. Each model completed 12 trials per regime (Table S22).

Table S22. Capability-screening results and admission decisions.

Model	Concordant accuracy	Misleading-majority accuracy	Decision
claude-sonnet-4-6	1.00	0.83	Admit
gpt-4.1-mini	1.00	0.83	Admit
gemini-2.5-flash	0.92	1.00	Admit with thinking budget of 8,192 tokens
gemini-2.5-pro	1.00	1.00	Admit with extended thinking
claude-haiku-4.5	0.75	0.50	Exclude
gpt-4o-mini	0.83	0.50	Exclude

Results are stored in `code/data/adequacy/adequacy.csv`. Admitted models achieved misleading-majority accuracy of at least 0.83, whereas the two excluded models achieved 0.50 and showed a

tendency to count signals rather than weight their likelihoods. The screening criterion was specified before the focal experiments. The screening used a separate task, prompts, and seeds, and no model was excluded after its focal outcomes were inspected. This screen reduces the likelihood that the focal effects reflect baseline inability to integrate unequal evidence, but it does not establish general reasoning competence outside the tested task family.

Section S8. Planted-conclusion intervention

In the misleading-majority, all-channel, conclusion-sharing condition, we planted a confident first-mover conclusion that was visible to all agents from round 0. Every agent's private evidence was held fixed. On matched seeds, only the direction of the planted conclusion varied, with conditions favoring the true state, favoring the false state, and containing no plant (Table S23).

The average causal effect of plant direction was defined as

$$\text{ACE} = E[P(\text{truth}) \mid \text{do}(\text{correct plant})] - E[P(\text{truth}) \mid \text{do}(\text{incorrect plant})].$$

Table S23. Planted-conclusion intervention: terminal P(true state) by plant direction.

Model	Matched seeds	Correct plant	No plant	Incorrect plant	ACE
claude-sonnet-4-6	15	0.99	0.76	0.36	+0.64
gpt-4.1-mini	40	0.85	0.59	0.34	+0.51

The ACE was calculated from unrounded trial-level probabilities. For claude-sonnet-4-6, the unrounded means were 0.994 under the correct plant and 0.355 under the incorrect plant, yielding 0.639, which rounds to 0.64. Results are stored in `code/data/replay-<model>/`.

Because private evidence was identical across intervention arms, the difference identifies a causal effect of conclusion direction on terminal collective belief. This intervention complements the factorial manipulation of communication content, but does not by itself estimate the full causal effect of conclusion sharing relative to evidence sharing.

Section S9. Reasoning-effort manipulation

We conducted a within-family manipulation using claude-sonnet-4-6. The task, topology, and communication condition were held fixed at misleading-majority, all-channel, and conclusion sharing. The only manipulated factor was whether extended private reasoning was enabled for each agent using an Anthropic thinking budget of 2,048 tokens (Table S24).

Table S24. Reasoning-effort manipulation: accuracy and individual-error trajectory.

Private reasoning	N	Collective accuracy	Mean individual error trajectory
Off	50	0.62 [0.48–0.74]	$\bar{e}$ increased
On	30	1.00 [0.89–1.00]	$\bar{e}$ declined to 0.004

The difference yielded $z = 3.87$ and $p = 1.1 \times 10^{-4}$. Enabling extended reasoning eliminated the conclusion-sharing accuracy loss in this condition. Agents assigned high probability to the true state rather than following the misleading numerical majority.

Extended reasoning increased output-token use by approximately 4.5-fold, from 5,707 to 25,917 tokens per trial, and increased expenditure per trial by approximately 3.3-fold. These results show that additional private computation can eliminate the conclusion-sharing accuracy loss in one model family and one experimental condition. Reasoning effort is therefore a demonstrated moderator in a single within-family manipulation, not an established general mechanism, and these data do not establish a general reasoning-

effort effect across models, tasks, or communication structures. Data are stored in `code/data/claude-sonnet-4-6_think/`.

Section S10. Signal-structure robustness

We varied the diagnosticity gap and composition of the misleading majority while holding the all-channel topology, group size, and trial count fixed at $n = 5$ and $N = 50$ per cell (Table S25).

Table S25. Collective accuracy by signal structure and communication condition.

Signal structure	No communication	Evidence sharing	Conclusion sharing
Narrow gap	1.00 [0.93–1.00]	0.82 [0.69–0.90]	0.54 [0.40–0.67]
Wide gap	1.00 [0.93–1.00]	0.84 [0.71–0.92]	0.42 [0.29–0.56]
Heavy majority	1.00 [0.93–1.00]	0.82 [0.69–0.90]	0.26 [0.16–0.40]

Table S25 reports the Wilson 95% intervals for the corresponding estimates.

The qualitative content effect persisted across all three signal structures. Accuracy before communication remained 1.00, and evidence-sharing accuracy remained between 0.82 and 0.84. Conclusion-sharing accuracy declined from 0.54 in the narrow-gap condition to 0.42 in the wide-gap condition and 0.26 under the heavy majority. This graded pattern indicates that the conclusion-sharing accuracy loss became more severe as the misleading-majority structure strengthened. Data are stored in `code/data/gpt-4.1-mini_sigsweep/`.

Section S11. Naturalistic hidden-profile probe

We conducted a preliminary probe using a fictional materials-comparison hidden-profile task. Each analyst privately received one prose finding. The correct material was determined by the combined weight of all findings, while the numerical majority of individual snippets was engineered to favor the wrong material. We applied the same evidence-versus-conclusion and topology manipulations using `gpt-4.1-mini`, with $N = 50$ trials per cell.

Task construction and prompts. Each trial drew a hidden correct material, then generated five prose findings from distinct engineering tests. In the misleading-majority regime, three analysts received marginal findings favoring the wrong material and two received decisive findings favoring the correct one, so a vote over individual findings favors the wrong material while the combined weight of the evidence favors the correct one. The latent weight of each finding was never shown to agents. Trials regenerate deterministically from `generate_hidden_trial(seed, regime, n_agents)` in `code/tasks_hidden.py`, and the run pipeline and prompts are in `code/run_hidden.py`.

The system prompt was used for both belief elicitation and message generation:

You are one of several analysts evaluating which of two materials, Material Zeta or Material Theta, is the better choice for an engineering application. Each analyst has privately reviewed a DIFFERENT engineering test, and no single test is decisive: the correct choice is the one favored by the weight of ALL the evidence combined. Weigh a decisive, large-margin finding far more heavily than a marginal, small-margin one. Combine independent findings; do not double-count a finding that reaches you again after being relayed by someone else.

Beliefs were elicited with:

Your private evidence: you reviewed {finding}. {messages received this round} Combine your own finding with each DISTINCT finding reported by others, weighting decisive findings far more than

marginal ones. Work through it in 1-3 short sentences, then end with exactly this line: FINAL=<your probability that Material Zeta is the better choice, between 0 and 1>

Messages used the same system prompt and the same private-evidence and received-message preamble, followed by one of three content instructions:

evidence_only. Write a short message sharing ONLY your finding: which test you reviewed and what it showed (which material it favored, and whether the margin was decisive or marginal). Do NOT say which material you overall think is better. One or two sentences.

conclusion_only. Write a short message sharing ONLY your current best guess of which material is the better choice (Zeta or Theta) and how confident you are. Do NOT describe your test or finding. One sentence.

both. Write a short message sharing your finding (test + which material it favored, decisive or marginal) AND your current best guess of the better material with your confidence. Two or three sentences.

Received messages were presented as Messages you received this round: followed by one – [from analyst s] message line per sender. Belief elicitation used temperature 0 and message generation temperature 0.7, matching the focal task. Probabilities were elicited for Material Zeta and mapped to the P(A) convention used throughout, with Zeta as A and Theta as B.

Example findings. The five findings below are the materials for the worked trial reported later in this section, regenerated from code/tasks_hidden.py at seed 0 of the misleading-majority regime. Latent weights are shown here for the reader and were not visible to agents. The correct material is Theta, favored by a net combined weight of 3.21, while three of the five individual findings favor Zeta (Table S26).

Table S26. Example hidden-profile findings for the worked trial.

Analyst	Finding as shown to that analyst	Favors	Latent weight
0	a marginal finding from the load-cycling evaluation: Material Zeta slightly edged out Material Theta, by a small margin within normal sample-to-sample variation	Zeta	0.15
1	a marginal finding from the thermal-cycling test: Material Zeta slightly edged out Material Theta, by a small margin within normal sample-to-sample variation	Zeta	0.25
2	a decisive finding from the fatigue-life assessment: Material Theta clearly and consistently outperformed Material Zeta by a large margin across all samples	Theta	1.98
3	a decisive finding from the vibration-endurance run: Material Theta clearly and consistently outperformed Material Zeta by a large margin across all samples	Theta	1.98
4	a marginal finding from the corrosion-resistance trial: Material Zeta slightly edged out Material Theta, by a small margin within normal sample-to-sample variation	Zeta	0.34

Collective accuracy was 1.00 in every one of the ten communication cells tested, across both signal regimes and both topologies. The no-communication controls were 1.00 in the misleading-majority regime and 0.92 in the concordant regime (Table S27).

Table S27. Hidden-profile results by regime, topology, and content condition.

Regime	Topology	Content	N	Accuracy	$\bar{e}$	D	P(true state)
Concordant	all-channel	evidence	50	1.00	0.006	0.001	0.969
Concordant	all-channel	conclusion	50	1.00	0.015	0.001	0.941
Concordant	all-channel	both	50	1.00	0.010	0.001	0.963
Concordant	circle	evidence	50	1.00	0.023	0.008	0.946
Concordant	circle	conclusion	50	1.00	0.031	0.008	0.909
Misleading majority	all-channel	evidence	50	1.00	0.013	0.003	0.943
Misleading majority	all-channel	conclusion	50	1.00	0.007	0.002	0.958
Misleading majority	all-channel	both	50	1.00	0.004	0.001	0.970
Misleading majority	circle	evidence	50	1.00	0.081	0.040	0.897
Misleading majority	circle	conclusion	50	1.00	0.047	0.020	0.896

Mean individual error at round 0 was between 0.128 and 0.208 across cells, and every communication condition terminated at an accuracy of 1.00. Mean individual error declined under conclusion sharing in the misleading-majority all-channel cell from 0.128 to 0.007. Inspection of the outputs indicated that agents could often reconstruct task-relevant evidence from peers' prose conclusions, despite the instruction to transmit conclusions rather than findings.

Worked trial. The trial below is the lowest-numbered seed, seed 0, in the misleading-majority, all-channel, conclusion-sharing cell. It was selected by that rule rather than for its outcome. Its five findings are the example given above. A vote over the individual findings favors Zeta, the wrong material, while the combined weight of the evidence favors Theta (Table S28).

Table S28. Hidden-profile worked trial: round-by-round beliefs and pooled probabilities.

Round	Analyst beliefs P(Zeta)	Pooled P(Zeta)	Pooled P(true material)
0	0.55, 0.30, 0.05, 0.05, 0.30	0.186	0.814
1	0.30, 0.30, 0.10, 0.10, 0.30	0.200	0.800
2	0.20, 0.20, 0.05, 0.10, 0.10	0.117	0.883
3	0.10, 0.10, 0.00, 0.05, 0.10	0.009	0.991

The trial illustrates the ceiling rather than a content effect. The round-0 pooled belief already favors the correct material at 0.814, before any communication, and three rounds of conclusion sharing move it further toward the truth, ending at 0.991. This is representative of the cell rather than exceptional, because every trial in every communication cell terminated correctly.

The preliminary hidden-profile runs retained the task materials and round-level belief trajectories but did not persist the generated message text: `code/run_hidden.py` passes each round's messages to the next round but writes only the per-round belief vectors to `code/data/hidden-gpt-4.1-mini/full/transcripts.jsonl`. The worked example above therefore reports the archived findings, beliefs, and pooled trajectory rather than reconstructed messages. The message-generation prompts given earlier in this section specify exactly what agents were asked to write, and Section S13 provides verbatim message examples from the focal binary-inference experiments.

Because collective accuracy is 1.00 in every communication cell, the probe has no sensitivity to any content effect: there is no headroom within which a difference between the evidence and conclusion conditions could have appeared. The misleading-majority no-communication control is likewise at 1.00; the concordant control, at 0.92, is the only condition below ceiling. The improvement in mean individual

error and the qualitative observation of evidence reconstruction are therefore suggestive only, and are consistent with a task that is easy for the collective regardless of what is communicated.

The absence of a content effect shows that the focal result does not transfer automatically to this prose-based hidden-profile task. It does not establish why, and because the communication conditions are at ceiling it provides no positive evidence about generalization beyond the focal task. Because collective accuracy is at ceiling in every communication condition, the probe cannot distinguish recoverability of evidence from a conclusion message, task difficulty, and ceiling performance, and it cannot be used as evidence that natural-language conclusions preserve their evidentiary grounds. A design that independently manipulates the recoverability and salience of findings would be required to distinguish these explanations. We therefore treat this result as a preliminary boundary probe rather than evidence of generalization. Data are stored in `code/data/hidden-gpt-4.1-mini/`.

Section S12. Statistics, reproducibility, and data audit

Statistical analysis. The primary collective belief was calculated using a logarithmic opinion pool, defined as the mean log odds across agents. The arithmetic-mean pool was used as a robustness measure. Individual error was measured using the Brier score. Mean individual error is denoted $\bar{e}$, and diversity D is the variance of individual beliefs.

For the arithmetic-mean pool, collective Brier error satisfies the identity

collective error = $\bar{e} - D$.

We use this identity as a diagnostic decomposition of the distribution of individual beliefs. It is not a decomposition of the logarithmic opinion pool.

Accuracy is reported with Wilson 95% confidence intervals. Paired content contrasts use McNemar tests on shared seeds. Logistic models use cluster-robust inference with clustering by task seed. The content-by-topology interaction was assessed with a likelihood-ratio test. The $n = 11$ content $\times$ fan-in interaction used 10,000 within-seed content-label permutations across the three paired degree levels (2, 2.55, and 10), with $RD(\text{degree } 10) - \text{mean}[RD(\text{degree } 2), RD(\text{degree } 2.55)]$ as the statistic. Degree-specific content and matched-topology risk differences used seed bootstrap intervals. Agent-level positional analyses used trial-clustered bootstrap intervals. The both-extremes diagnostic used paired seed bootstrap intervals and was repeated at three probability thresholds. Differences in $P(\text{truth})$ use a 10,000-resample bootstrap. The degree trend uses seed-clustered logistic regression of collective accuracy on per-node degree. Because degree and diameter covary across the tested graph family, the fan-in claim relies on matched contrasts rather than independent coefficients from a joint model.

Effect sizes, aggregation-rule robustness, and the error-diversity analysis can be reproduced with `analyze_stats3.py`. Assignment-based and belief-based role-transition analyses can be reproduced with `analyze_stats9.py` and `analyze_stats8.py`, respectively. The between-network degree analysis, within-network fan-in analysis, positional balance tests, and graph-fixed-effects analysis can be reproduced with `analyze_stats10.py`, `analyze_stats11.py`, and `analyze_stats12.py`. Core contrasts can be reproduced with `analyze_stats.py` and `analyze_stats2.py`.

Models. The experiments used OpenAI gpt-4.1-mini, Anthropic claude-sonnet-4-6, and Google gemini-2.5-flash/pro. The capability screen also included claude-haiku-4.5 and gpt-4o-mini. Models were queried through public application programming interfaces without fine-tuning. All manipulations were implemented through prompts. Exact model snapshot strings, sampling parameters, and query dates are recorded in each run's `run_meta.json` file under `code/data/`.

Data and code availability. All study materials are publicly available on the Open Science Framework (DOI 10.17605/OSF.IO/AU2GS, <https://doi.org/10.17605/OSF.IO/AU2GS>) and at

<https://github.com/ericcgldstone-maker/m2m-collective-belief-dynamics>. The study materials include complete transcripts for the focal binary-inference experiments, containing each signal, message, source, and belief for every round, along with cell-level CSV files, run metadata, analysis scripts, figure-generation scripts, and the audit script. The principal analysis file is `analyze_stats.py`, figure generation is implemented in `make_figures.py`, and validation is implemented in `audit.py`, which incorporates the revision-era checks in `audit_revision.py`. The additional analyses introduced in this revision are implemented in `analyze_dispersion.py`, `analyze_pools.py`, `analyze_load.py`, `analyze_benchmark.py`, `analyze_recipient.py`, `check_leakage.py`, `extract_worked_trial.py`, and `adversarial_followup_analysis.py`.

Re-running `audit.py` re-derives the reported quantities from the archived data. The original audit and the revision-era audit are integrated under `code/audit.py`, which is the single entry point for verifying this manuscript and its Supplementary Materials: it performs 82 structural and original-submission checks and then delegates to `audit_revision.py`, which re-derives 428 quantities introduced or changed in this revision. Every audited quantity is recomputed from the raw or archived processed data; reported values are used only as assertion targets and never as inputs. The audit is deterministic, with all resampling under a fixed seed of 20260909 and 10,000 draws, and repeated runs produce identical output. Results are written to `code/audit_results.json`, which records for each assertion its identifier, claim family, source data file, reproduced value, reported value, tolerance, and status, together with 23 informational records that are reported for transparency and carry no assertion. All 82 checks and all 428 assertions currently pass. Coverage includes the reported cell accuracies, the nonemptiness of approximately 100,000 logged messages, the error-diversity identity, the causal-effect estimates and capability-screening scores including the gemini gate, the absence of verdict leakage from evidence-only messages, the confidence-weighted benchmark and normative-recipient counterfactual (Section S6.6), the absence of duplicate trial keys, the dispersion and Brier decomposition (Section S6.7), the six-rule aggregation comparison (Section S6.8), the neighbor-fan-in analysis (Section S6.9), the channel-purity audit (Section S6.10), the message-load characterization (Section S6.11), the content $\times$ fan-in randomization test and paired risk differences, the both-extremes threshold-sensitivity analysis, the path-length contrast intervals, the cluster-aware neighbor-fan-in intervals, the wheel contrast interval, the degree-10 message-volume comparison, the planted-conclusion intervention, the reasoning-effort manipulation, the signal-structure sweep, and the hidden-profile probe including the worked example reported in Section S11.

Section S13. Worked example trial

A single complete trial, reproducible with `extract_worked_trial.py`. Seed 2, misleading-majority regime, all-channel network, $n = 5$, gpt-4.1-mini. The true state is A. The full-information Bayesian posterior is $P(A) = 0.945$. A head-count of round-0 verdicts favors B, the false state (Table S29).

Table S29. Worked example trial: agent readings, likelihood ratios, and round-0 beliefs.

Agent	Reading	P(reading A)	P(reading B)	Log-likelihood ratio	Normative P(A)	Elicited round-0 belief
0	r3	0.88	0.12	+2.04	0.885	0.880
1	r0	0.39	0.61	-0.46	0.387	0.390
2	r4	0.89	0.11	+2.07	0.888	0.890
3	r1	0.41	0.59	-0.37	0.409	0.410
4	r2	0.39	0.61	-0.43	0.394	0.390

Three agents hold weak readings favoring the false state B and two hold strong readings favoring the true state A. The round-0 pooled belief is $P(A) = 0.638$, on the correct side, while a vote over the same five

verdicts favors B.

Evidence sharing. Round-1 messages, verbatim:

Agent 0: *"My private reading is 'r3,' which occurs with probability 0.88 if the true state is A and 0.12 if the true state is B."*

Agent 1: *"My private reading is 'r0,' which occurs with probability 0.39 if the true state is A and 0.61 if the true state is B."*

Agent 2: *"My private reading is 'r4,' which occurs with probability 0.89 if the true state is A and 0.11 if the true state is B."*

Agent 3: *"My private reading is 'r1,' which occurs with probability 0.41 if the true state is A and 0.59 if the true state is B."*

Agent 4: *"My private reading is 'r2,' which occurs with probability 0.39 if the true state is A and 0.61 if the true state is B."*

Terminal beliefs: 0.941, 0.939, 0.168, 0.937, 0.939. Terminal pooled $P(A) = 0.866$. The collective is correct.

Conclusion sharing. The same trial, same private evidence, round-1 messages verbatim:

Agent 0: *"I believe the hidden state is A with about 94% confidence."*

Agent 1: *"I believe the hidden state is B with about 75% confidence."*

Agent 2: *"I believe the hidden state is A with about 90% confidence."*

Agent 3: *"I believe the hidden state is B with about 92% confidence."*

Agent 4: *"I believe the hidden state is B with about 67% confidence."*

Terminal beliefs: 0.000, 0.600, 0.515, 0.030, 0.020. Terminal pooled $P(A) = 0.016$. The collective is incorrect, and confidently so rather than marginally. Individual beliefs remain widely dispersed at the terminal round, spanning 0.000 to 0.600, so the failure occurs without convergence on a shared belief.