How Communication Topology Density Shapes Performance in LLM-Based Multi-Agent Systems

Anonymous authors
Paper under double-blind review

Abstract

Communication topology is a key design variable in LLM-based multi-agent reasoning. Prior work has shown that sparse communication can outperform full connectivity, but the mechanisms linking communication density and degree allocation to performance remain incompletely understood. We propose a method for analyzing how communication topology affects multi-agent reasoning performance. It first exactly decomposes per-round changes in agent-level answer information into corrective gain and misleading loss, then introduces a graph-based propagation model of evidence coverage, error dilution, and same-direction error reinforcement. Under homogeneous fixed-state assumptions, we derive an explicit optimality condition for communication density and show when it has a unique interior solution: corrective gain increases with diminishing returns, whereas misleading loss reflects a competition between error dilution and same-direction error reinforcement. Under a fixed edge budget and the stated homogeneity assumptions, regular or near-regular degree allocations maximize predicted performance among feasible graphs without isolated nodes. Experiments with two open-source models on six reasoning datasets provide qualitative support: the best accuracy among tested conditions occurs at an intermediate density in all 12 model–dataset combinations, and a fixed-budget comparison favors balanced degrees. Together, these results provide a mechanistic basis for communication-topology design.

1 Introduction

LLM-based multi-agent systems can improve reasoning by exchanging and revising answers (Du et al., 2024; Liang et al., 2024). Systems such as CAMEL, AutoGen, and MetaGPT organize collaboration through roles and workflows (Li et al., 2023; Wu et al., 2024; Hong et al., 2024). Their performance depends on agent count, independent evidence, model heterogeneity, and information allocation (Qian et al., 2025; Bertalanič & Fortuna, 2026; Yang et al., 2026; Jain et al., 2026; Lu et al., 2025), alongside reconciliation, aggregation, and interaction failures (Chen et al., 2024; Wang et al., 2025; Choi et al., 2025; Cemri et al., 2025). With the model and protocol fixed, communication topology controls which peer information each agent receives, raising a practical question: when does adding communication help, and when does it hurt?

Sparse debate can match or exceed full connectivity (Li et al., 2024), while GPTSwarm, AgentPrune, and G-Designer optimize graphs or remove redundant communication (Zhuge et al., 2024; Zhang et al., 2025a;b). Related theory studies limited attention, effective sample size, and spectral properties of consensus and robustness (Liu et al., 2026; Park & Alharthi, 2026). Collective-learning and social-influence studies likewise connect useful diffusion, premature convergence, and group judgment to interaction structure (Lazer & Friedman, 2007; Golub & Jackson, 2010; Mason & Watts, 2012; Barkoczi & Galesic, 2016; Almaatouq et al., 2020; Wang et al., 2026; Lorenz et al., 2011; Muchnik et al., 2013; Becker et al., 2017).

Mechanistic studies show that dense graphs can amplify both correct insights and misinformation (Shen et al., 2025), and distinguish correction from corruption (Reitich, 2026). Other accounts emphasize answer instability and conformity (Hao et al., 2026), or complementary knowledge and interaction noise (Qiao et al., 2026). We address a specific mechanistic question: when does the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

benefit of additional communication give way to harmful influence? Combining an exact mutual-information decomposition with a graph-based propagation model, we relate evidence coverage, error dilution, and same-direction reinforcement to conditions for an interior optimal density and balanced degree allocation under a fixed edge budget.

Our experiments compare a chain, several regular or near-regular density conditions, and full connectivity, using round-wise empirical positive transfer ($\widetilde{B}$) and negative transfer ($\widetilde{M}$) to describe state changes. These discrete topology conditions are not an exhaustive sample of all feasible graphs or a direct empirical ER ensemble. We examine density and degree-allocation trends with two models and six datasets.

The framework connects two aspects of communication design: choosing density and allocating a fixed edge budget. At a fixed information state, it characterizes the trade-off between evidence coverage and erroneous influence and favors balanced degrees under homogeneous propagation. Across rounds, state evolution and vote dependence shape how these mechanisms translate into empirical transfer and team accuracy.

Our two main theoretical contributions are an exact gain–loss decomposition paired with a propagation model that yields conditions for an interior optimal density, and a fixed-edge-budget result favoring regular or near-regular degree allocations under homogeneous fixed-state propagation. We evaluate these predictions through density sweeps, fixed-budget comparisons, and controlled interventions, and illustrate their use in validation-based density selection.

2 Problem Formulation and Performance Decomposition

We formalize LLM-based multi-agent debate as communication over a graph and decompose its performance effect into information gain and information loss. This separates the contribution of topology from the utility already present in the agents' initial responses.

2.1 Communication Graphs for LLM-Based Multi-Agent Systems

Let X denote a problem instance, $Y \in \mathcal{Y}$ the correct answer, and $V = [n]$ a set of n LLM agents.

For all mutual-information quantities, (X, Y) is drawn from a task distribution, jointly with model-generation randomness and, where specified, graph randomness. The information quantities are unconditional with respect to the full problem instance X. If Y is determined by the complete instance, then $H(Y \mid X) = 0$; conditioning all quantities on that complete instance would make the intended decomposition trivial. Thus the information objective is a population-level description, not information generated about a fixed labeled question.

The answer state of agent i after round t is Z_i^t, including both its selected answer and supporting reasoning. We represent communication by a simple undirected graph $G_\rho = (V, E_\rho)$ and define

$$\rho = \frac{2|E_\rho|}{n(n-1)}, \qquad \mathcal{N}_i(G_\rho) = \{j : (i,j) \in E_\rho\}, \qquad d_i = |\mathcal{N}_i(G_\rho)|. \tag{1}$$

Here $\rho = 0$ is the no-communication baseline: the system directly aggregates the fixed round-0 independent answers and performs no subsequent self-update. The case $\rho = 1$ is fully connected debate. During synchronous round $t+1$, agent i updates from its current state and its neighbors' states from the preceding round:

$$Z_i^{t+1} \sim P_\phi\big(\cdot \mid X, Z_i^t, \{Z_j^t : j \in \mathcal{N}_i(G_\rho)\}\big), \tag{2}$$

where P_ϕ is the stochastic update kernel induced by a fixed LLM and prompt protocol. Comparisons among communicating graphs hold the model, prompt protocol, agent count, rounds, and initial responses $Z_{1:n}^0$ fixed, allowing topology effects to be assessed under a common protocol. The round-0 baseline measures performance before communication and further updating.

2.2 Decomposing Team Performance

We characterize the answer information retained across agents by their mean mutual information with the correct answer:

$$\mathcal{I}_t(G_\rho) = \frac{1}{n}\sum_{i=1}^{n} I(Y; Z_i^t). \tag{3}$$

This agent-wise measure supports an exact gain–loss decomposition. Individual and team accuracies provide complementary measures of task performance; Appendix A.5 gives a sufficient condition under which improved individual accuracy translates into higher team accuracy. Evidence acquired by multiple agents contributes separately to their respective information terms.

Applying the mutual-information chain rule to (Z_i^t, Z_i^{t+1}) gives the exact per-round decomposition

$$\begin{aligned} I(Y; Z_i^{t+1}) - I(Y; Z_i^t) &= I(Y; Z_i^{t+1} \mid Z_i^t) - I(Y; Z_i^t \mid Z_i^{t+1}) \\ &=: B_{t,i}(G_\rho) - M_{t,i}(G_\rho). \end{aligned} \tag{4}$$

Here $B_{t,i} \geq 0$ is newly acquired information about the correct answer, which we call the corrective gain, and $M_{t,i} \geq 0$ is correct-answer information lost after the update, which we call the misleading loss. Let $B_t = n^{-1}\sum_i B_{t,i}$ and $M_t = n^{-1}\sum_i M_{t,i}$. Summing over T interaction rounds yields the core identity

$$\boxed{\mathcal{I}_T(G_\rho) = \mathcal{I}_0 + \sum_{t=0}^{T-1} [B_t(G_\rho) - M_t(G_\rho)]}. \tag{5}$$

Because all topologies reuse the same initial responses, $\mathcal{I}_0$ is independent of ρ. The cumulative gain–loss difference therefore exactly accounts for topology-dependent changes in $\mathcal{I}_T$. Taking expectations over a random graph family gives the corresponding density-level quantities.

3 How Communication Density Shapes Performance

Building on the exact decomposition in Section 2, we develop a propagation model combining evidence coverage, error dilution, and same-direction reinforcement, and examine its predictions experimentally in Section 4. Here, B_t and M_t denote modeled gains and losses that characterize how topology changes the marginal effects of communication. We test these predictions using round-wise log-score changes and final answer accuracy; Appendix A.1–A.4 gives the derivations and assumptions.

3.1 Corrective Gain as a Function of Density

We first model node-level evidence coverage and then use an Erdős–Rényi (ER) graph to obtain a closed-form density expression. Suppose U potential evidence items $R_{1:U}$ jointly determine the answer. For a fixed ordering of these items, define the following nonnegative weights for the coverage model:

$$a_{r,i}^t := I(Y; R_r \mid R_{<r}, Z_i^t), \qquad A_i^t := \sum_{r=1}^{U} a_{r,i}^t = I(Y; R_{1:U} \mid Z_i^t). \tag{6}$$

Under a homogeneous-agent approximation, let a_r^t and A_t denote the corresponding agent averages. We model useful evidence acquisition through additive weighted coverage: each item contributes its weight when first revealed, and repeated revelation adds no coverage. Each neighbor reveals an item with probability θ, while λ scales its contribution to modeled gain.

The remaining useful evidence A_t depends on the agent state. Shared round-0 responses fix A_0 across densities, but earlier communication can deplete or homogenize evidence, making later states density-dependent. We first hold A_t fixed to isolate direct coverage effects; Appendix A.7 treats $A_t(\rho)$.

For an agent of degree d_i, the probability that at least one neighbor reveals item R_r is $1-(1-\theta)^{d_i}$. Its expected weighted coverage is therefore

$$u_t(d_i) = A_t\left[1-(1-\theta)^{d_i}\right], \tag{7}$$

and the team-average gain on graph G_ρ is

$$B_t(G_\rho) = \frac{\lambda}{n}\sum_{i=1}^{n} u_t(d_i). \tag{8}$$

For an ER graph, write $N = n-1$. Because $D \sim \mathrm{Binomial}(N,\rho)$, the probability-generating function gives

$$b_t(\rho) := \mathbb{E}_{G_\rho}[B_t(G_\rho)] = \lambda A_t\left[1-(1-\rho\theta)^N\right]. \tag{9}$$

Under the fixed-state approximation, $b_t(\rho)$ increases with diminishing returns. Strict increase requires $\lambda A_t\theta > 0$ and $\rho\theta < 1$; strict concavity additionally requires $N \geq 2$. The discrete gain from an additional neighbor also decreases with degree (Appendix A.1). This describes direct evidence coverage at a fixed state: later-round empirical transfer need not be monotone because earlier communication changes the remaining useful evidence (Appendix A.7).

3.2 Misleading Loss as a Function of Density

For negative influence, assume each neighbor disrupts an evidence item with probability η. Let $K^t_{i,r}$ be the number of neighbors of agent i that disrupt item r. Disturbances can point in different incorrect directions $s \in \mathcal{S}$; multiple neighbors supporting the same direction reinforce one another. We model loss through the disturbed-neighbor fraction W, directional concentration R, and same-direction pair reinforcement P.

Let $c^t_{r,i} \geq 0$ weight the disturbance of item r, and let μ scale the receiver's response to erroneous influence. Conditional on degree d_i, assume

$$K^t_{i,r} \mid d_i \sim \mathrm{Binomial}(d_i,\eta), \qquad K^t_{i,r,s} \mid d_i \sim \mathrm{Binomial}(d_i,\eta\pi_s), \tag{10}$$

where π_s is the conditional distribution of error directions and $\sum_s \pi_s = 1$. Define

$$W^t_{i,r} := \frac{K^t_{i,r}}{d_i}, \qquad R^t_{i,r} := \sum_{s\in\mathcal{S}}\left(\frac{K^t_{i,r,s}}{d_i}\right)^2, \qquad P^t_{i,r} := \sum_{s\in\mathcal{S}}\binom{K^t_{i,r,s}}{2}. \tag{11}$$

For $d_i = 0$, all three quantities are defined as zero. The term W measures the fraction of disruptive neighbors. The normalized concentration R captures how a small number of aligned errors can appear salient at low degree but become diluted as more diverse information arrives. The pair count P captures mutual reinforcement between neighbors supporting the same wrong direction.

Let nonnegative weights $\omega_1,\omega_2,\omega_3$ control these mechanisms. Define

$$\begin{aligned} q^t_{i,r} &:= \omega_1 W^t_{i,r} + \omega_2 R^t_{i,r} + \omega_3 P^t_{i,r}, \\ M_{t,i}(G_\rho) &:= \mu\sum_{r=1}^{U} c^t_{r,i} q^t_{i,r}, \qquad M_t(G_\rho) := \frac{1}{n}\sum_{i=1}^{n} M_{t,i}(G_\rho). \end{aligned} \tag{12}$$

Let $S_\pi = \sum_s \pi_s^2$ and $C_t = n^{-1} \sum_i \sum_r c_{r,i}^t$. We use the mean-degree approximation $m_t(\rho) = \mu C_t \psi_t(N\rho)$, where $\psi_t(d)$ is the expected local disturbance multiplier derived in Appendix A.2. With density-independent state parameters, this approximation is analyzed for $N\rho \geq 1$; no communication is treated separately. The dilution component decreases with degree, whereas same-direction pair reinforcement increases. Their competition determines whether additional connectivity lowers or raises modeled loss, while $m_t''(\rho) \geq 0$ under this closure. Appendix A.2 gives the derivatives, and Appendix A.6 states the stability conditions for extending the conclusions to the exact binomial degree distribution.

3.3 Overall Density–Performance Relationship

The cumulative model score combines these competing effects:

$$F_T(\rho) := \mathcal{I}_0 + \sum_{t=0}^{T-1} \left[b_t(\rho) - m_t(\rho) \right]. \tag{13}$$

For nondegenerate parameters, $b_t''(\rho) \leq 0$ and $m_t''(\rho) \geq 0$, so

$$F_T''(\rho) = \sum_{t=0}^{T-1} \left[b_t''(\rho) - m_t''(\rho) \right] < 0. \tag{14}$$

On a feasible interval $[\rho_{\min}, 1]$ with $N\rho_{\min} \geq 1$, strict concavity gives a unique interior maximizer when $F_T'(\rho_{\min}) > 0$ and $F_T'(1) < 0$; otherwise the optimum occurs at an endpoint. This provides a principled criterion for communication design: additional edges are useful only while their marginal corrective benefit exceeds their marginal misleading loss. The criterion concerns the fixed-state performance proxy; Appendix A.7 discusses evolving states across rounds.

Appendix A.3 derives the optimal-density equation and comparative statics. Holding other effective parameters fixed, greater correct-evidence strength relative to error reinforcement, a higher relative dilution weight, or greater error-direction dispersion shifts an interior optimum toward higher density. The effect of revelation probability θ changes sign at $\theta N\rho^* = 1$: it raises the optimum below this threshold and lowers it above the threshold. Under the stated positivity and interior-optimum conditions, these results identify when additional connectivity is valuable and when sufficient evidence coverage can be achieved with fewer edges.

These conclusions motivate treating density as a task-specific validation choice instead of defaulting to full connectivity. Our application study operationalizes the marginal-benefit principle using empirical correct-option log-score transfers to select density on validation data. On RACE-High and MathQA, the selected graphs use 80% and 40% fewer edges than the complete graph while achieving slightly higher test accuracy. The theory thus supplies an interpretable basis for communication-budget selection, complemented by held-out empirical evaluation.

3.4 Degree Allocation under a Fixed Edge Budget

Density does not uniquely determine performance. Holding n and $|E|$ fixed, $u_t(d) = A_t[1-(1-\theta)^d]$ is concave in d, so Jensen's inequality gives

$$B_t(G) \leq \lambda A_t \left[1 - (1-\theta)^{\bar{d}} \right], \qquad \bar{d} = \frac{2|E|}{n}. \tag{15}$$

At fixed node and edge counts, concentrated degrees create redundant evidence at hubs while leaving other agents underserved. Under homogeneous fixed-state propagation, concave coverage favors balanced degrees. The corresponding local loss is convex for degrees at least one, so balancing also reduces modeled loss among feasible graphs without isolated nodes. A regular graph attains the bound when feasible; otherwise a near-regular allocation is optimal under these assumptions (Appendix A.4).

4 EXPERIMENTS

We organize the experiments around three theoretical predictions: how density affects team performance and round-wise transfers, how degree allocation matters under a fixed edge budget, and whether the proposed local error mechanisms appear under controlled intervention. All topology conditions reuse the same round-0 independent responses.

4.1 EXPERIMENTAL SETUP

Models and debate protocol. We use Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Qwen Team, 2025). Eleven agents answer each question and participate in three communication rounds. Agents answer independently at round 0, then receive only their graph neighbors' preceding-round answers and explanations before updating. The final team prediction is the majority vote.

Tasks. The six datasets cover commonsense reasoning, scientific question answering, medical knowledge, and mathematical reasoning: HellaSwag (Zellers et al., 2019), ARC-Challenge (Clark et al., 2018), MMLU hard balanced and Professional Medicine (Hendrycks et al., 2021), OpenBookQA (Mihaylov et al., 2018), and AQuA-RAT (Ling et al., 2017).

Metrics. We report final team accuracy and mean individual accuracy. Let $p_{i,q}^t$ be the correct-option probability from agent i's normalized self-reported confidence distribution on question q, not an API token log probability, and set $s_{i,q}^t = \log \max(p_{i,q}^t, \epsilon)$ with $\epsilon = 10^{-6}$. For $\Delta s_{i,q}^t = s_{i,q}^{t+1} - s_{i,q}^t$ and $[x]_+ = \max(x, 0)$, define empirical positive and negative transfer as

$$\begin{aligned} \widetilde{B}_{t\to t+1}(G_\rho) &= \frac{1}{n|\mathcal{D}|} \sum_{q\in\mathcal{D}} \sum_{i=1}^{n} [\Delta s_{i,q}^t]_+, \\ \widetilde{M}_{t\to t+1}(G_\rho) &= \frac{1}{n|\mathcal{D}|} \sum_{q\in\mathcal{D}} \sum_{i=1}^{n} [-\Delta s_{i,q}^t]_+. \end{aligned} \tag{16}$$

These statistics describe beneficial and harmful correct-answer confidence changes; they are not numerical estimators of conditional mutual information. Section 4.2 and Appendix B report the first transition, while Appendix C reports later transitions. Final team accuracy is measured after all three rounds.

4.2 TESTING THE THEORETICAL PREDICTIONS

Density sweep. With 11 agents and three communication rounds fixed, we compare the round-0 no-communication baseline, Chain, and regular or near-regular graphs at densities $0.2, 0.4, 0.6, 0.8$, and 1.0. All conditions reuse the same initial responses. Figure 1 reports final team accuracy on AQuA-RAT, HellaSwag, OpenBookQA, and Professional Medicine; Figure 2 reports first-round transfers. Appendix B adds ARC-Challenge and Balanced MMLU. These representative graphs test qualitative density trends rather than an empirical ER ensemble.

Final accuracy. Across all 12 model–dataset combinations, the highest accuracy among tested conditions occurs at a density between 0.4 and 0.8; full connectivity is never best. The optimum depends on both model and task. For example, both models peak at 0.6 on HellaSwag, whereas AQuA-RAT peaks at 0.4 for Llama and 0.8 for Qwen. Descriptive results on final-accuracy variability are reported in Appendix E.

First-round transfers. All eight positive-transfer curves in Figure 2 end above their Chain values, supporting improved evidence coverage, although not every curve is strictly concave. Negative transfer is task-dependent: both HellaSwag curves reach their minimum at 0.6 and both OpenBookQA curves at 0.8 before rising, whereas Qwen on AQuA-RAT decreases throughout. The model permits monotone loss when one mechanism dominates; a U-shape is not required. Other curvature departures may reflect differences between the homogeneous, independent-revelation model and finite-sample confidence proxies on representative graphs rather than the analytical ER average. These are plausible sources of mismatch, not separately identified causes.

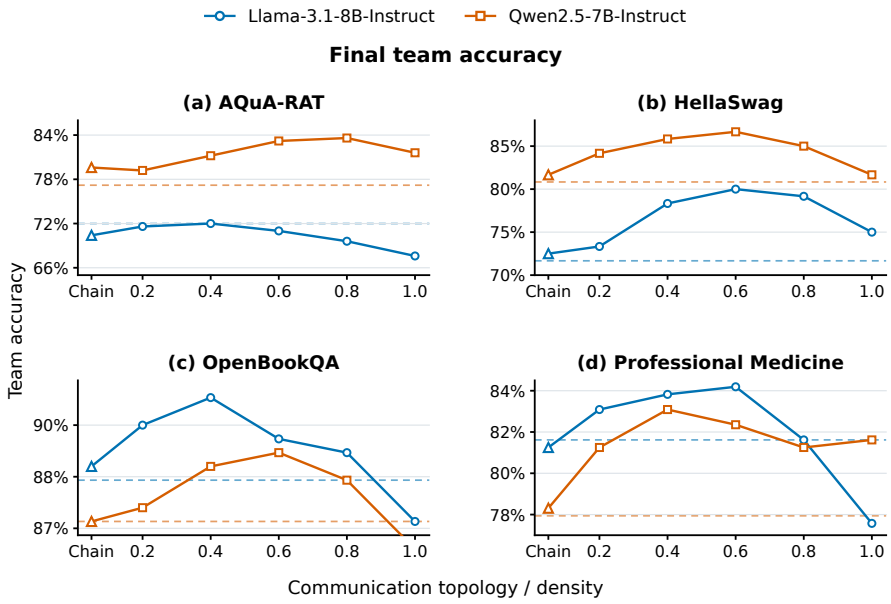


Figure 1: Final team accuracy after three rounds on the four main-text datasets. Blue circles: Llama; orange squares: Qwen; triangles: Chain; dashed lines: no-communication baselines.

Across rounds. First-round transfers do not alone determine the final optimum. On AQuA-RAT with Qwen, positive transfer rises and negative transfer falls in the first round, yet final accuracy declines beyond 0.8. Subsequent communication changes the remaining useful evidence and error composition, while correlated answers affect majority voting; final accuracy therefore need not inherit the curvature of the fixed-state information proxy. These considerations motivate the state-dependent analysis and later-round measurements (Appendices A.5, A.7, and C).

Degree allocation under a fixed edge budget. With Qwen2.5-7B-Instruct on Balanced MMLU, we hold node count, edge count, and rounds fixed while varying degree heterogeneity; a smaller degree standard deviation means a more uniform allocation.

Figure 3(a)–(b) shows that accuracy falls from 73.0% to 71.1% and positive transfer from 0.3725 to 0.3229 as degree standard deviation increases from 0 to 2.954. Balanced degrees achieve the highest accuracy and positive transfer among the tested allocations. Negative transfer follows a non-monotone pattern, reaching its minimum at a degree standard deviation of 1.044.

4.3 Controlled Validation of Local Propagation Mechanisms

We fix a focal agent whose initial answer is correct and control its neighbors' opinions. The horizontal axis denotes the total number of neighbors, k. Condition D uses one incorrect and $k-1$ correct neighbors; R uses k neighbors supporting the same incorrect answer; N uses k correct neighbors. For question q under condition c, we measure local disturbance as $L_{q,c} = [p_{q,c}^{\text{before}} - p_{q,c}^{\text{after}}]_+$, where p denotes the focal agent's normalized self-reported probability for the correct option.

Figure 3(c)–(d) shows that D falls from about 33% to 14% on Balanced MMLU and from 19% to 9% on Professional Medicine as total neighborhood size increases from one to five, while R rises to about 40% and 24%. The low, relatively flat N curve indicates that neighborhood composition matters beyond message count. Together, the interventions support error dilution and the harmful effect of repeated support for the same wrong answer.

4.4 Application-Oriented Density Selection

We evaluate validation-based density selection on RACE-High (Lai et al., 2017) and MathQA (Amini et al., 2019), using Qwen2.5-7B-Instruct with 11 agents and three rounds. Each task uses 100



Figure 2: First-round positive transfer $\widetilde{B}_{0\to1}$ (upper) and negative transfer $\widetilde{M}_{0\to1}$ (lower) on four main-text datasets. Blue/left axes: Llama; orange/right axes: Qwen; units: nats; triangles: Chain.

Table 1: Validation-selected densities: test accuracy and edge savings relative to full connectivity.

Dataset	ρ_{sel} (edges)	Selected acc.	Oracle ρ (acc.)	Full acc.	Edge savings
RACE-High	0.2 (11)	84.00%	0.2 (84.00%)	81.33%	80%
MathQA	0.6 (33)	90.67%	0.8 (91.33%)	90.00%	40%

validation and 150 disjoint test questions, and all densities share round-0 responses. Scanning densities from 0.2 to 1.0, we select the value before the first non-positive marginal change in cumulative net correct-option log-score transfer; Chain is an external baseline. Appendix D gives the full protocol.

Validation selects $\rho = 0.2$ on RACE-High and $\rho = 0.6$ on MathQA (Table 1), yielding 84.00% and 90.67% test accuracy versus 81.33% and 90.00% under full connectivity while using 80% and 40% fewer edges. The selected density matches the retrospective test oracle on RACE-High; on MathQA, its test accuracy is 0.67 percentage points lower. The oracle is used only for retrospective comparison. These tasks demonstrate communication-efficient operating points, but selection should be revalidated when the model, protocol, or task changes.

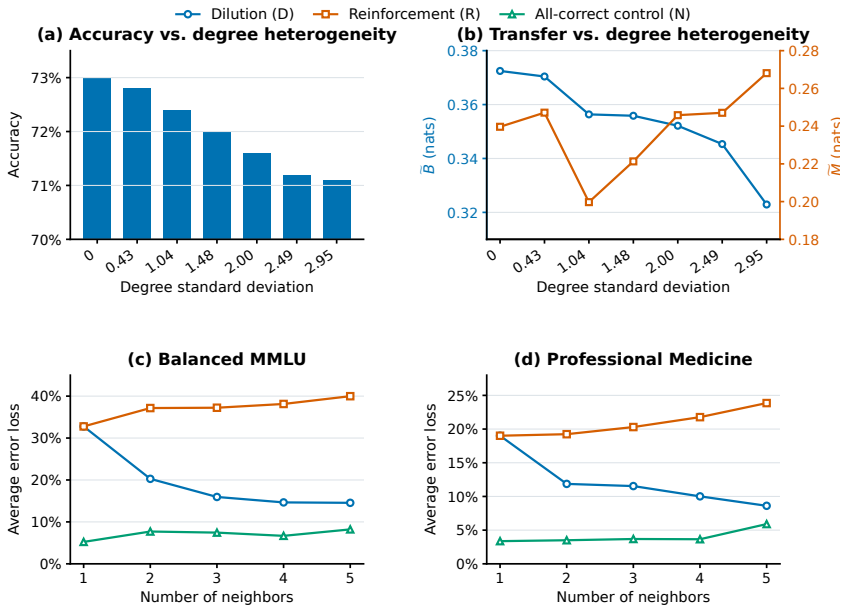


Figure 3: Fixed-budget degree allocation (a)–(b; 11 nodes, 22 edges) and controlled mechanisms (c)–(d): correct-neighbor dilution (D), same-error reinforcement (R), and all-correct control (N). Panels (c)–(d) report mean probability drops, expressed in percentage points.

5 Conclusion

We developed an information-theoretic decomposition and a graph-based propagation model to explain how communication density shapes LLM-based multi-agent reasoning. The framework connects evidence coverage, error dilution, and same-direction reinforcement to density-dependent gains and losses. Under the fixed-state model, denser graphs increase evidence coverage with diminishing marginal returns. Misleading loss reflects a competition between error dilution and same-direction reinforcement, and may therefore vary monotonically or non-monotonically with density. Their difference yields a strictly concave predicted performance curve under the stated nondegeneracy conditions. When the edge budget is fixed, a more uniform degree allocation reduces both evidence redundancy and local error concentration.

Experiments with two models and six datasets support these qualitative predictions. Across all 12 model–dataset combinations, the highest observed accuracy occurs at an intermediate tested density. First-round transfers show broad gains in correct-answer confidence and task-dependent losses, while controlled interventions support dilution and the harmful effect of repeated same-error support. Later-round curves also motivate the density-dependent state extension in Appendix A.7: high density can improve early coverage while depleting or homogenizing useful evidence sooner.

The theory relies on idealizations including homogeneous agents, independent evidence revelation, and simplified error directions. Future work should study heterogeneous models, directed and adaptive communication, and state-dependent topology selection using complete reasoning trajectories. The framework thus explains when additional connectivity helps and guides validation-based density selection, with the application study demonstrating fewer communication edges alongside slightly higher test accuracy.

Reproducibility statement

The paper specifies the communication protocol, model families, number of agents and rounds, topology conditions, evaluation metrics, and the theoretical assumptions used in each result. Full derivations appear in Appendix A, and all round-wise empirical transfer curves appear in Appendices B and C. Additional implementation details and supporting code will be made available subject to the venue's anonymity and artifact-release requirements.

AI use statement

Generative AI tools assisted with literature retrieval and discovery, research ideation, drafting selected sections, language translation and polishing, document formatting, and developing and checking mathematical proofs. The authors reviewed and revised the AI-assisted content and take full responsibility for the scientific claims, proofs, citations, and reported results.

References

Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, Peter M. Krafft, Mehdi Moussaid, and Alex Pentland. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21):11379–11386, 2020.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2357–2367, 2019.

Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, 7:13109, 2016.

Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26):E5070–E5076, 2017.

Blaž Bertalanič and Carolina Fortuna. The ringelmann effect in multi-agent LLM systems: A scaling law for effective team size. *arXiv preprint arXiv:2606.02646*, 2026.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7066–7085, 2024.

Hyun Kyung Choi, Xuan Zhu, and Sheng Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, 2024.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

X. Hao, Z. Wu, Y.-X. Qiu, C. Xiao, R. Xu, S. Zheng, and J. Qin. Not all flips are conformity: Decomposing stance convergence in multi-agent LLM debate. *arXiv preprint arXiv:2606.00820*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Juergen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

A. Jain, V. Krishnamurthy, and Y. Zhang. Collaborative QA using interacting LLMs: Impact of network structure, node capability and distributed data. *Transactions on Machine Learning Research*, 2026.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017.

David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 158–167, 2017.

K. Liu, G. Xiong, W. Zhang, and S. Tang. Social networks of LLM agents. *arXiv preprint arXiv:2607.03695*, 2026.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

D. Lu, J. Zhang, C. Yuan, J. Shao, C. Zhang, and X. Li. The law of multi-model collaboration: Scaling limits of model ensembling for large language models. *arXiv preprint arXiv:2512.23340*, 2025.

Winter Mason and Duncan J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Lev Muchnik, Sinan Aral, and Sean J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

E. D. J. Park and D. Alharthi. Predictive maps of multi-agent reasoning: A successor-representation spectrum for LLM communication topologies. *arXiv preprint arXiv:2605.11453*, 2026.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Ke Zhu, Hao Xia, Yufan Dang, Zekun Du, Weize Chen, Yusheng Cheng, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration. In *International Conference on Learning Representations*, 2025.

D. Qiao, B. Chen, F. Cai, J. Chen, W. Li, F. Jiang, Z. Chen, H. Zha, T. Zhang, and B. Wang. Epistemic gain, aleatoric cost: Uncertainty decomposition in multi-agent debate for math reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Fernando Reitich. Correction and corruption: A two-rate view of error flow in LLM protocols. *arXiv preprint arXiv:2604.18245*, 2026.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12347–12361, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, 2025.

Y. Wang, A. Civilini, A. Sheng, X. Chen, L. Wang, and V. Latora. Optimal network structure for collective performance with strategic information sharing. *arXiv preprint arXiv:2605.00758*, 2026.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed H. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Conference on Language Modeling*, 2024.

Y. Yang, C. Qu, M. Wen, L. Shi, Y. Wen, W. Zhang, A. Wierman, and S. Gu. Understanding agent scaling in LLM-based multi-agent systems via diversity. In *ICLR Workshop on AI for Mechanism Design and Strategic Decision Making*, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

G. Zhang, Y. Yue, Z. Li, S. Yun, G. Wan, K. Wang, D. Cheng, J. X. Yu, and T. Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2025a.

G. Zhang, Y. Yue, X. Sun, G. Wan, M. Yu, J. Fang, K. Wang, T. Chen, and D. Cheng. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 76678–76692, 2025b.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Juergen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 62743–62767, 2024.

A Theoretical Derivations

Unless stated otherwise, let $N = n - 1$, $0 \le \rho \le 1$, and $0 \le \theta, \eta \le 1$. We retain the independent-revelation, independent-disturbance, and homogeneous-agent assumptions from the main text. Strict inequalities additionally require the corresponding information strength, absorption rate, and mechanism weight to be positive.

A.1 Corrective Gain and Density

Proposition A.1 (expected weighted coverage). Conditional on degree d_i, the expected weighted evidence coverage is

$$u_t(d_i) = A_t[1 - (1-\theta)^{d_i}]. \tag{A.1}$$

Proof. Let J_r indicate whether item r is revealed by at least one neighbor. Its probability of revelation is $\mathbb{E}[J_r \mid d_i] = 1 - (1-\theta)^{d_i}$. By the additive coverage definition and linearity of expectation,

$$u_t(d_i) = \mathbb{E}\left[\sum_{r=1}^{U} a_r^t J_r \mid d_i\right] = \sum_{r=1}^{U} a_r^t \mathbb{E}[J_r \mid d_i] = A_t[1 - (1-\theta)^{d_i}]. \tag{A.2}$$

Because the receiver absorbs a fraction λ, the team-average corrective gain conditional on G is

$$B_t(G) = \frac{\lambda}{n} \sum_{i=1}^{n} u_t(d_i). \qquad \square \tag{A.3}$$

Proposition A.2 (expected gain on a random graph). In $G_\rho \sim G(n, \rho)$, let D be the degree of any node. Then $D \sim \mathrm{Binomial}(N, \rho)$ and

$$b_t(\rho) := \mathbb{E}_{G_\rho}[B_t(G_\rho)] = \lambda A_t[1 - (1 - \rho\theta)^N]. \tag{A.4}$$

Proof. Exchangeability implies that the expected team average equals the expected gain of any node:

$$b_t(\rho) = \lambda A_t[1 - \mathbb{E}(1-\theta)^D]. \tag{A.5}$$

The binomial probability-generating function is $\mathbb{E}[z^D] = (1 - \rho + \rho z)^N$. Setting $z = 1 - \theta$ yields

$$\mathbb{E}(1-\theta)^D = [1 - \rho + \rho(1-\theta)]^N = (1 - \rho\theta)^N, \tag{A.6}$$

which proves the result. $\square$

Differentiating,

$$b_t'(\rho) = \lambda A_t N \theta (1 - \rho\theta)^{N-1}, \tag{A.7}$$

$$b_t''(\rho) = -\lambda A_t N(N-1)\theta^2 (1 - \rho\theta)^{N-2} \le 0. \tag{A.8}$$

If $\lambda A_t \theta > 0$ and $\rho\theta < 1$, then $b_t'(\rho) > 0$; if additionally $N \ge 2$, the function is strictly concave. Equivalently, the discrete increment is

$$\begin{aligned} u_t(d+1) - u_t(d) &= A_t[(1-\theta)^d - (1-\theta)^{d+1}] \\ &= A_t \theta (1-\theta)^d, \end{aligned} \tag{A.9}$$

which is non-increasing in d and strictly decreasing when $0 < \theta < 1$ and $A_t > 0$.

A.2 Misleading Loss and Density

Fix agent i, evidence item r, and round t, and suppress these indices. Given $d \ge 1$, each neighbor produces a disturbance in direction s with probability $\eta\pi_s$ and no disturbance with probability $1 - \eta$. The vector $(K_s)_{s \in \mathcal{S}}$ and the undisturbed count are multinomial, with marginals

$$K_s \mid d \sim \mathrm{Binomial}(d, \eta\pi_s), \qquad K := \sum_{s\in\mathcal{S}} K_s \mid d \sim \mathrm{Binomial}(d, \eta). \tag{A.10}$$

Define the error-direction concentration

$$S_\pi := \sum_{s\in\mathcal{S}} \pi_s^2. \tag{A.11}$$

The three local quantities are

$$W = \frac{K}{d}, \qquad R = \sum_{s\in\mathcal{S}} \left(\frac{K_s}{d}\right)^2, \qquad P = \sum_{s\in\mathcal{S}} \binom{K_s}{2}. \tag{A.12}$$

Because $\mathbb{E}[K \mid d] = d\eta$,

$$\mathbb{E}[W \mid d] = \eta. \tag{A.13}$$

For $p_s = \eta\pi_s$,

$$\mathbb{E}[K_s^2 \mid d] = dp_s(1-p_s) + d^2p_s^2 = dp_s + d(d-1)p_s^2. \tag{A.14}$$

Hence

$$\begin{aligned} \mathbb{E}[R \mid d] &= \frac{1}{d^2}\sum_s \mathbb{E}[K_s^2 \mid d] \\ &= \frac{\eta}{d} + \frac{d-1}{d}\eta^2 S_\pi \\ &= \eta^2 S_\pi + \frac{\eta - \eta^2 S_\pi}{d}. \end{aligned} \tag{A.15}$$

The binomial coefficient $\binom{K_s}{2}$ counts unordered pairs that produce the same directional disturbance. Therefore

$$\mathbb{E}\left[\binom{K_s}{2}\middle| d\right] = \binom{d}{2}p_s^2, \tag{A.16}$$

and

$$\mathbb{E}[P \mid d] = \binom{d}{2}\eta^2 S_\pi. \tag{A.17}$$

Let

$$a := \eta - \eta^2 S_\pi \geq 0, \qquad b := \eta^2 S_\pi \geq 0. \tag{A.18}$$

For fixed $d \geq 1$, the expected local disturbance multiplier is

$$\begin{aligned} \psi_t(d) &:= \mathbb{E}[q_{i,r}^t \mid d_i = d] \\ &= \omega_1\eta + \omega_2\left(b + \frac{a}{d}\right) + \frac{\omega_3 b}{2}d(d-1). \end{aligned} \tag{A.19}$$

Set $\psi_t(0) = 0$. Under the homogeneous approximation, let

$$C_t := \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{U} c_{r,i}^{t}. \tag{A.20}$$

Treating C_t as independent of graph sampling and local disturbance counts, the exact ER expectation is

$$\mathbb{E}_{G_\rho}[M_t(G_\rho)] = \mu C_t\, \mathbb{E}_D[\psi_t(D)], \qquad D \sim \mathrm{Binomial}(N,\rho), \tag{A.21}$$

with

$$\mathbb{E}_D[\psi_t(D)] = \sum_{d=0}^{N} \psi_t(d)\binom{N}{d}\rho^d(1-\rho)^{N-d}. \tag{A.22}$$

The $1/d$ term and isolated-node boundary make a compact curvature criterion inconvenient. The main text therefore uses $\bar{d} = \mathbb{E}[D] = N\rho$ and the approximation $\mathbb{E}[\psi_t(D)] \approx \psi_t(\bar{d})$, yielding

$$\begin{aligned} m_t(\rho) &:= \mu C_t \psi_t(\bar{d}) \\ &= \mu C_t \left[\omega_1\eta + \omega_2\left(b + \frac{a}{\bar{d}}\right) + \frac{\omega_3 b}{2}\bar{d}(\bar{d}-1)\right], \qquad \bar{d} = N\rho \geq 1. \end{aligned} \tag{A.23}$$

The degree variance and relative variance are

$$\mathrm{Var}(D) = N\rho(1-\rho), \qquad \frac{\mathrm{Var}(D)}{\bar{d}^2} = \frac{1-\rho}{N\rho}, \tag{A.24}$$

and $\Pr(D=0) = (1-\rho)^N$. When expected degree is not small, isolation is rare and degrees concentrate near $\bar{d}$. A second-order expansion gives

$$\mathbb{E}[\psi_t(D)] = \psi_t(\bar{d}) + \frac{1}{2}\psi_t''(\bar{d})\mathrm{Var}(D) + \text{higher-order terms}. \tag{A.25}$$

Thus the approximation is appropriate away from extremely sparse or highly heterogeneous regimes. We use it only for $N\rho \geq 1$ and treat $\rho = 0$ separately with $M_t = 0$. Differentiation gives

$$\begin{aligned} m_t'(\rho) &= \mu C_t N\left[-\frac{\omega_2 a}{\bar{d}^2} + \frac{\omega_3 b}{2}(2\bar{d}-1)\right], \\ m_t''(\rho) &= \mu C_t N^2\left[\frac{2\omega_2 a}{\bar{d}^3} + \omega_3 b\right] \geq 0. \end{aligned} \tag{A.26}$$

The W expectation is degree-independent. The $a/\bar{d}$ component of R decreases with density, while the pair-reinforcement term $\bar{d}(\bar{d}-1)/2$ accelerates upward.

A.3 Overall Performance and the Optimal Density

Define the cumulative model score

$$F_T(\rho) := \mathcal{I}_0 + \sum_{t=0}^{T-1}[b_t(\rho) - m_t(\rho)]. \tag{A.27}$$

Then

$$F_T''(\rho) = \sum_{t=0}^{T-1} [b_t''(\rho) - m_t''(\rho)] \leq 0. \tag{A.28}$$

If at least one round has $b_t''(\rho) < 0$ or $m_t''(\rho) > 0$, then $F_T''(\rho) < 0$, so F_T is strictly concave and F_T' strictly decreases. Nondegeneracy means that at least one curvature-producing mechanism is active, for example $\lambda A_t > 0$, $0 < \theta < 1$, and $N \geq 2$, or $\mu C_t > 0$ with $\omega_2 a > 0$ or $\omega_3 b > 0$.

On $[\rho_{\min}, 1]$, if

$$F_T'(\rho_{\min}) > 0, \qquad F_T'(1) < 0, \tag{A.29}$$

continuity and strict monotonicity of the derivative imply a unique $\rho^* \in (\rho_{\min}, 1)$ satisfying $F_T'(\rho^*) = 0$. Strict concavity makes it the unique global maximizer. If $F_T'(\rho_{\min}) \leq 0$, the left endpoint is optimal; if $F_T'(1) \geq 0$, the right endpoint is optimal.

Let

$$\mathcal{A} := \lambda \sum_{t=0}^{T-1} A_t, \qquad \mathcal{C} := \mu \sum_{t=0}^{T-1} C_t, \tag{A.30}$$

assume $\mathcal{C}\omega_3\eta^2 S_\pi > 0$, and define

$$\kappa := \frac{\mathcal{A}}{\mathcal{C}\omega_3\eta^2 S_\pi}, \qquad \tau := \frac{\omega_2}{\omega_3}, \qquad \zeta := \frac{1}{\eta S_\pi} - 1. \tag{A.31}$$

Because $\eta - \eta^2 S_\pi = \eta^2 S_\pi \zeta$, dividing $F_T'(\rho^*) = 0$ by the positive quantity $N\mathcal{C}\omega_3\eta^2 S_\pi$ gives

$$h(\rho; \kappa, \theta, \tau, \zeta) = \kappa\theta(1 - \rho\theta)^{N-1} + \frac{\tau\zeta}{\bar{d}^2} - \left(\bar{d} - \frac{1}{2}\right) = 0, \qquad \bar{d} = N\rho. \tag{A.32}$$

Moreover,

$$\partial_\rho h = -\kappa\theta^2(N-1)(1-\rho\theta)^{N-2} - \frac{2\tau\zeta N}{\bar{d}^3} - N < 0. \tag{A.33}$$

The implicit-function theorem applies near an interior optimum:

$$\frac{\partial\rho^*}{\partial\xi} = -\frac{\partial_\xi h}{\partial_\rho h}. \tag{A.34}$$

The relevant partial derivatives are

$$\partial_\kappa h = \theta(1-\rho\theta)^{N-1} \geq 0, \tag{A.35}$$

$$\partial_\tau h = \frac{\zeta}{\bar{d}^2} \geq 0, \qquad \partial_\zeta h = \frac{\tau}{\bar{d}^2} \geq 0. \tag{A.36}$$

Consequently,

$$\frac{\partial\rho^*}{\partial\kappa} \geq 0, \qquad \frac{\partial\rho^*}{\partial\tau} \geq 0, \qquad \frac{\partial\rho^*}{\partial\zeta} \geq 0, \tag{A.37}$$

with strict inequalities when the corresponding parameters are positive. For θ,

$$\begin{aligned}\partial_\theta h &= \kappa(1-\rho\theta)^{N-2}[1-N\rho\theta] \\ &= \kappa(1-\rho\theta)^{N-2}(1-\theta\bar{d}),\end{aligned} \tag{A.38}$$

so

$$\operatorname{sign}\left(\frac{\partial\rho^*}{\partial\theta}\right) = \operatorname{sign}(1-\theta\bar{d}^*). \tag{A.39}$$

This proves the comparative-statics statements in the main text.

A.4 Degree Allocation under a Fixed Edge Budget

Fix n and $|E|$, so every feasible graph has

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{2|E|}{n}. \tag{A.40}$$

Corrective gain. For $0 < \theta < 1$, extending $u_t(d) = A_t[1-(1-\theta)^d]$ to real $d \geq 0$ gives

$$u_t''(d) = -A_t(1-\theta)^d[\log(1-\theta)]^2 \leq 0. \tag{A.41}$$

Jensen's inequality therefore yields

$$\begin{aligned}B_t(G) &= \frac{\lambda}{n}\sum_{i=1}^{n} u_t(d_i) \\ &\leq \lambda u_t\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right) \\ &= \lambda A_t[1-(1-\theta)^{\bar{d}}].\end{aligned} \tag{A.42}$$

Under strict concavity, equality requires equal degrees. If $\bar{d}$ is nonintegral, take two nodes with $d_i \geq d_j + 2$. The marginal gain $\Delta u_t(d) = A_t\theta(1-\theta)^d$ decreases with d, so shifting one degree unit from i to j changes total gain by

$$u_t(d_i-1) + u_t(d_j+1) - u_t(d_i) - u_t(d_j) = \Delta u_t(d_j) - \Delta u_t(d_i-1) \geq 0. \tag{A.43}$$

In a simple graph, such balancing can be implemented by an edge reallocation: some node k is adjacent to i but not j; replace (i,k) with (j,k). Repetition yields a near-regular graph whose degrees differ by at most one.

Misleading loss. Restrict attention to feasible graphs without isolated nodes and retain the homogeneous approximation. For $d \geq 1$,

$$\psi_t(d) = \omega_1\eta + \omega_2\left(b + \frac{a}{d}\right) + \frac{\omega_3 b}{2}d(d-1), \tag{A.44}$$

with

$$\psi_t''(d) = \frac{2\omega_2 a}{d^3} + \omega_3 b \geq 0. \tag{A.45}$$

Thus

$$\frac{1}{n}\sum_{i=1}^{n}\psi_t(d_i) \geq \psi_t(\bar{d}), \tag{A.46}$$

and the mean misleading loss is minimized by uniform degrees. The discrete increment $\Delta\psi_t(d) = \psi_t(d+1) - \psi_t(d)$ is non-decreasing, so the same balancing operation changes loss by

$$\psi_t(d_i - 1) + \psi_t(d_j + 1) - \psi_t(d_i) - \psi_t(d_j) = \Delta\psi_t(d_j) - \Delta\psi_t(d_i - 1) \leq 0. \tag{A.47}$$

Therefore a regular or near-regular graph maximizes predicted performance under the stated fixed-budget assumptions. This conclusion should not be extrapolated unconditionally to extremely sparse graphs with many isolated nodes, because $\psi_t(0) = 0$ is a separate boundary case, or to settings with strongly heterogeneous node-specific loss strength.

A.5 Individual Accuracy and Team Aggregation

To clarify how individual performance translates into team decisions, let $a_i(\rho)$ denote agent i's final-round answer and define

$$\chi_i(\rho) = \mathbf{1}\{a_i(\rho) = Y\}, \qquad p(\rho) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\chi_i(\rho)]. \tag{A.48}$$

Here, $p(\rho)$ is the mean individual accuracy. Under an independent, homogeneous-vote baseline in which each agent is correct with probability p, the probability of a strict correct majority is

$$B_n(p) = \sum_{k=q}^{n}\binom{n}{k}p^k(1-p)^{n-k}, \qquad q = \lfloor n/2 \rfloor + 1. \tag{A.49}$$

For multiclass plurality voting, a strict correct majority is sufficient for a correct team decision.

For $0 < p < 1$, $B_n(p)$ is strictly increasing because

$$B_n'(p) = n\binom{n-1}{q-1}p^{q-1}(1-p)^{n-q} > 0. \tag{A.50}$$

Communication introduces dependence, common errors, and conformity. Write actual team accuracy as

$$A_{\text{team}}(\rho) = B_n(p(\rho)) + \Delta(\rho), \tag{A.51}$$

where $\Delta(\rho)$ captures departures from this baseline due to vote dependence, heterogeneous accuracies, and the aggregation rule. For densities ρ_1 and ρ_2, the independent-baseline direction is preserved whenever

$$B_n(p(\rho_2)) - B_n(p(\rho_1)) > |\Delta(\rho_2)| + |\Delta(\rho_1)|. \tag{A.52}$$

Indeed, $|\Delta(\rho_2) - \Delta(\rho_1)| \leq |\Delta(\rho_2)| + |\Delta(\rho_1)|$, so the condition implies

$$A_{\text{team}}(\rho_2) - A_{\text{team}}(\rho_1) > 0. \tag{A.53}$$

It therefore provides a sufficient bridge from improved mean individual accuracy to higher team accuracy. We evaluate final team accuracy separately to capture the combined effects of individual performance, answer dependence, heterogeneous accuracies, and aggregation.

A.6 Scope and Stability of the Mean-Degree Approximation

The main text replaces the exact expectation over random degrees with the loss function evaluated at mean degree. This is not a numerical identity. The qualitative conclusions remain exact on the experimental interval whenever the discrete curvature of the exact misleading loss is nonnegative. The approximation primarily supplies an interpretable first-order condition and comparative statics; the fixed-edge degree-allocation result does not depend on it.

A.6.1 Curvature under the exact binomial expectation

Let

$$D \sim \mathrm{Binomial}(N, \rho), \qquad \bar{d} = N\rho, \tag{A.54}$$

and write

$$\psi(0) = 0, \qquad \psi(d) = c_0 + \frac{c_1}{d} + \frac{c_2}{2} d(d-1), \quad d \geq 1, \tag{A.55}$$

where

$$c_0 = \omega_1 \eta + \omega_2 \eta^2 S_\pi, \quad c_1 = \omega_2(\eta - \eta^2 S_\pi), \quad c_2 = \omega_3 \eta^2 S_\pi. \tag{A.56}$$

The exact and mean-field losses are

$$m_{t,\mathrm{exact}}(\rho) = \mu C_t \mathbb{E}[\psi(D)], \qquad m_{t,\mathrm{MF}}(\rho) = \mu C_t \psi(\bar{d}). \tag{A.57}$$

Binomial finite-difference identities give

$$\frac{d}{d\rho}\mathbb{E}[\psi(D)] = N\mathbb{E}[\Delta\psi(K)], \qquad K \sim \mathrm{Binomial}(N-1, \rho), \tag{A.58}$$

$$\frac{d^2}{d\rho^2}\mathbb{E}[\psi(D)] = N(N-1)\mathbb{E}[\Delta^2\psi(L)], \qquad L \sim \mathrm{Binomial}(N-2, \rho). \tag{A.59}$$

At the isolated-node boundary,

$$\Delta\psi(0) = c_0 + c_1, \qquad \Delta^2\psi(0) = c_2 - c_0 - \frac{3c_1}{2}. \tag{A.60}$$

For $k \geq 1$,

$$\Delta\psi(k) = -\frac{c_1}{k(k+1)} + c_2 k, \tag{A.61}$$

$$\Delta^2\psi(k) = c_2 + \frac{2c_1}{k(k+1)(k+2)} \geq 0. \tag{A.62}$$

Thus discrete curvature is always nonnegative above degree zero; only the transition from isolation to the first edge can violate global convexity. On the experimental density interval $\mathcal{R}$, define

$$\Gamma_t(\rho) := \mathbb{E}[\Delta^2\psi(L)]. \tag{A.63}$$

If $\Gamma_t(\rho) \geq 0$ throughout $\mathcal{R}$, the exact misleading loss is convex. Since corrective gain is exactly concave, the exact total performance is strictly concave whenever at least one component has strict curvature:

$$F''_{T,\mathrm{exact}}(\rho) = \sum_t [b''_t(\rho) - m''_{t,\mathrm{exact}}(\rho)] < 0. \tag{A.64}$$

Likewise define

$$\Lambda_t(\rho) := \mathbb{E}[\Delta\psi(K)]. \tag{A.65}$$

If $\Lambda_t(\rho_{\min}) < 0 < \Lambda_t(1)$ and exact discrete curvature is nonnegative, then Λ_t is non-decreasing and exact misleading loss first falls and then rises, with at most one interior minimum. The no-communication point remains a separate boundary and should not be extrapolated from the continuous mean-degree expression.

A.6.2 Stability of the optimal density

Let

$$F_{T,\mathrm{exact}}(\rho) = F_{\mathrm{MF}}(\rho) + R(\rho), \qquad \varepsilon := \sup_{\rho\in\mathcal{R}} |R'(\rho)|, \tag{A.66}$$

and suppose $-F''_{T,\mathrm{exact}}(\rho) \geq c > 0$. If both optima are interior, then

$$|\rho^*_{\mathrm{exact}} - \rho^*_{\mathrm{MF}}| \leq \frac{\varepsilon}{c}. \tag{A.67}$$

The bound quantifies the sensitivity of the continuous optimum to the mean-degree approximation. Comparing ε/c with the experimental density step assesses this sensitivity at the tested resolution; discrete candidate choices can then be checked using the exact finite-sum scores.

A.6.3 Fixed-edge degree allocation

For $0 < \theta < 1$ and degrees at least one, this result compares allocations with the same total degree and therefore requires no mean-degree substitution:

$$u''_t(d) = -A_t[\log(1-\theta)]^2(1-\theta)^d \leq 0, \qquad \psi''(d) = \frac{2c_1}{d^3} + c_2 \geq 0. \tag{A.68}$$

Consequently,

$$\frac{1}{n}\sum_i u_t(d_i) \leq u_t(\bar{d}), \qquad \frac{1}{n}\sum_i \psi(d_i) \geq \psi(\bar{d}). \tag{A.69}$$

For a fixed edge budget and degrees at least one, a more uniform degree allocation increases mean corrective gain and decreases mean misleading loss. If isolated nodes are allowed, the $\psi(0) = 0$ boundary must be treated separately.

A.7 Multi-Round Dynamic Extension with Density-Dependent States

The main text treats the remaining evidence strength and absorption rate in round t as given parameters to obtain interpretable closed forms. In a multi-round interaction, the round-t state is produced by earlier communication, so remaining evidence, effective absorption, and error structure can all depend on density.

A.7.1 Density-dependent effective positive evidence

In the ER baseline, define the coverage probability

$$q(\rho) = 1 - (1 - \rho\theta)^N. \tag{A.70}$$

Combine remaining evidence and absorption into $E_t(\rho) = \lambda_t(\rho)A_t(\rho)$. Dynamic corrective gain becomes

$$b_t(\rho) = E_t(\rho)q(\rho). \tag{A.71}$$

All densities share the same round-0 responses, but $E_t(\rho)$ need not be constant for $t \geq 1$. This representation does not claim that $\widetilde{B}$ alone separates remaining-evidence depletion from a lower absorption rate; it treats them as a joint effective state.

A.7.2 Direct coverage and state evolution

Differentiating,

$$b_t'(\rho) = E_t'(\rho)q(\rho) + E_t(\rho)q'(\rho), \tag{A.72}$$

$$b_t''(\rho) = E_t''(\rho)q(\rho) + 2E_t'(\rho)q'(\rho) + E_t(\rho)q''(\rho). \tag{A.73}$$

The term $E_t(\rho)q'(\rho)$ is the direct coverage effect: more edges raise the chance of receiving useful evidence. The term $E_t'(\rho)q(\rho)$ captures state evolution caused by earlier communication. If high density causes evidence to be absorbed, repeated, or homogenized earlier, the latter term can be negative. When this state effect outweighs direct coverage, the modeled later-round gain b_t decreases with density, providing a mechanism consistent with non-monotone empirical transfers.

A.7.3 A minimal cross-round recursion

Let γ_t be the fraction of remaining evidence consumed or homogenized in one communication round and let $g_t(\rho)$ denote useful evidence newly generated or reorganized during updating:

$$A_{t+1}(\rho) = A_t(\rho)[1 - \gamma_t q(\rho)] + g_t(\rho). \tag{A.74}$$

With no newly generated evidence and constant γ and λ,

$$b_t(\rho) = \lambda A_0 q(\rho)[1 - \gamma q(\rho)]^t, \tag{A.75}$$

and

$$\frac{\partial b_t}{\partial q} = \lambda A_0[1 - \gamma q]^{t-1}[1 - (t+1)\gamma q]. \tag{A.76}$$

Later-round gain can therefore fall once coverage passes a round- and depletion-dependent threshold. For otherwise equal parameters, that threshold is lower in later rounds. The decline does not require the direct effect of an added edge to become harmful.

A.7.4 Negative states and overall performance

Let the effective error information, absorption rate, and local structural term all evolve with density:

$$m_t(\rho) = \mu_t(\rho)C_t(\rho)\Phi_t(\rho). \tag{A.77}$$

Dilution and same-direction reinforcement still describe the local structural effect conditional on the current state, while later-round differences can also arise from changing error strength, absorption, and directional distribution. The dynamic model score is

$$F_T(\rho) = \mathcal{I}_0 + \sum_{t=0}^{T-1}[b_t(\rho) - m_t(\rho)]. \tag{A.78}$$

If every dynamic positive-gain term remains concave and every negative-loss term remains convex, the main-text concavity result survives. Otherwise, single-round or cumulative curves may plateau, locally decline, or take more complex shapes. The static model is the special case in which effective positive and negative states are density-independent.

B Density Sweeps on the Remaining Datasets

This appendix reports ARC-Challenge and Balanced MMLU. Professional Medicine is included in the four-panel main-text density sweep. Figure 4 shows final team accuracy after three rounds; Figures 5 and 6 show first-round empirical positive and negative transfer. The node count, topology conditions, and models match Section 4.2.

For all six benchmark sweeps, reported confidence percentages are converted to probabilities where needed, negative entries are clipped to zero, and values are renormalized across options; all required probabilities were parsed. Each non-complete density uses five independently sampled graphs with three independent generation runs per graph, while the complete graph has one realization with three runs. Density-sweep values are averaged equally over runs and graph realizations.

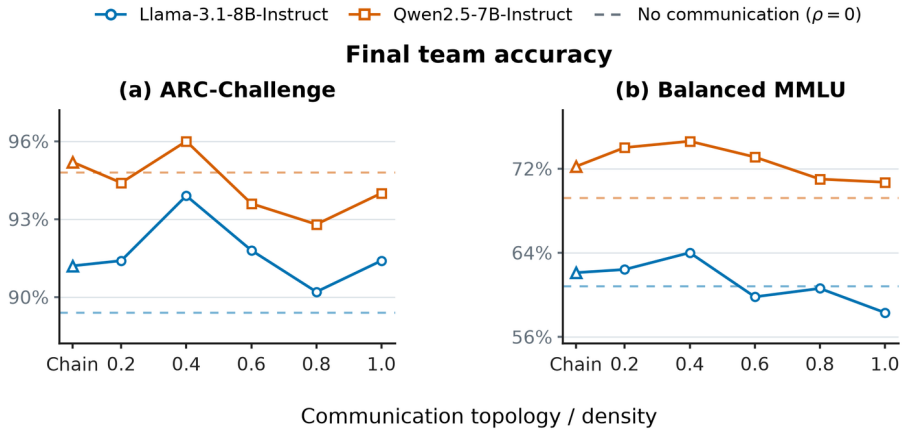


Figure 4: Final team accuracy on the remaining datasets. (a) ARC-Challenge; (b) Balanced MMLU. Conventions match Figure 1.

Both models reach their highest accuracy among the tested topologies at $\rho = 0.4$ on ARC-Challenge and Balanced MMLU. All four optima occur at an intermediate density, and the complete graph is never optimal.

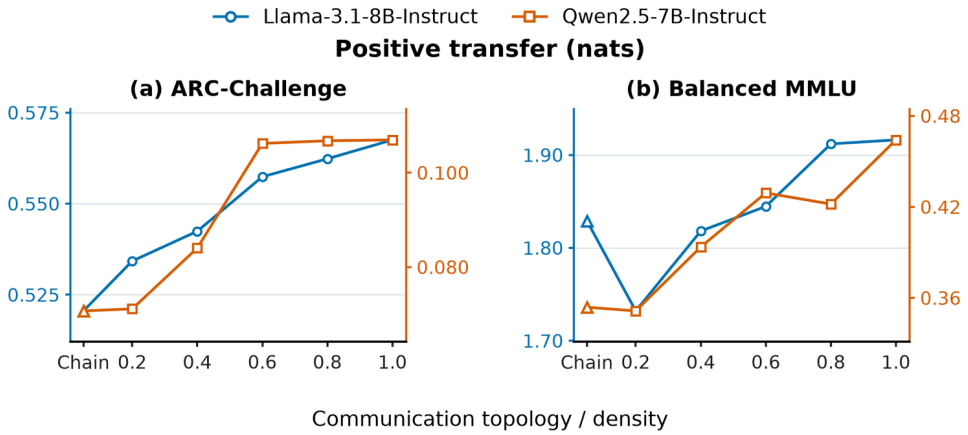


Figure 5: First-round empirical positive transfer $\widetilde{B}_{0\to1}$ on the remaining datasets. (a) ARC-Challenge; (b) Balanced MMLU. Conventions match Figure 2; values are in nats.

On ARC-Challenge, $\widetilde{B}_{0\to1}$ rises over the sweep for both models, although local increments vary. On Balanced MMLU, both models decrease slightly from the chain to $\rho = 0.2$ and then generally rise; Qwen also dips from $\rho = 0.6$ to 0.8. Across all 12 model–dataset conditions, the complete-graph endpoint remains above the chain endpoint, while the regular or near-regular portion is non-decreasing in 11 conditions.

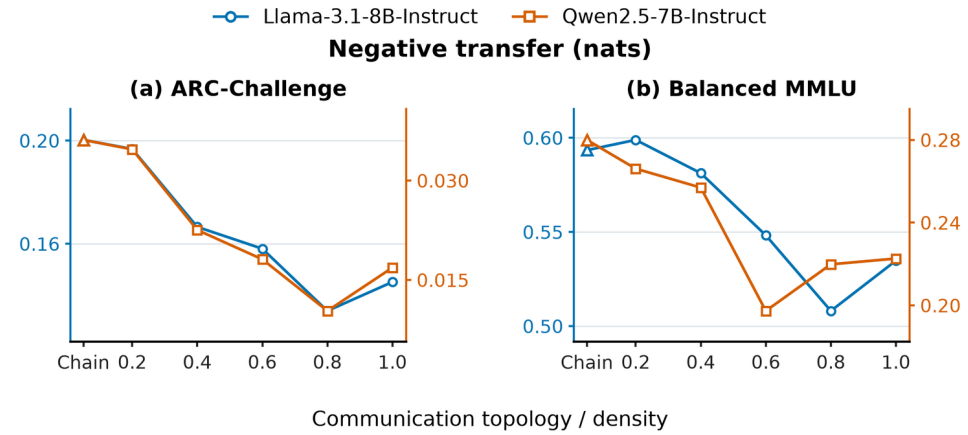


Figure 6: First-round empirical negative transfer $\widetilde{M}_{0\to1}$ on the remaining datasets. (a) ARC-Challenge; (b) Balanced MMLU. Conventions match Figure 2; values are in nats.

Both ARC-Challenge curves are lowest at $\rho = 0.8$ and rise under full connectivity. On Balanced MMLU, Llama and Qwen are lowest at $\rho = 0.8$ and 0.6, respectively; Llama rises slightly from the chain to $\rho = 0.2$. These results support a conditional density response for misleading loss, not a universal strictly U-shaped curve. The first-round extrema also need not coincide with final-accuracy optima; for example, ARC-Challenge $\widetilde{M}_{0\to1}$ is lowest at 0.8 while accuracy peaks at 0.4.

C Empirical Transfers in Rounds 2 and 3

This appendix reports empirical positive and negative transfer for round $1 \to 2$ and round $2 \to 3$. Axis, color, chain-marker, and dual-axis conventions follow Figure 2. Density zero has no communication and therefore zero transfer by definition, so it is omitted from the curves.

C.1 Round-2 Empirical Positive Transfer

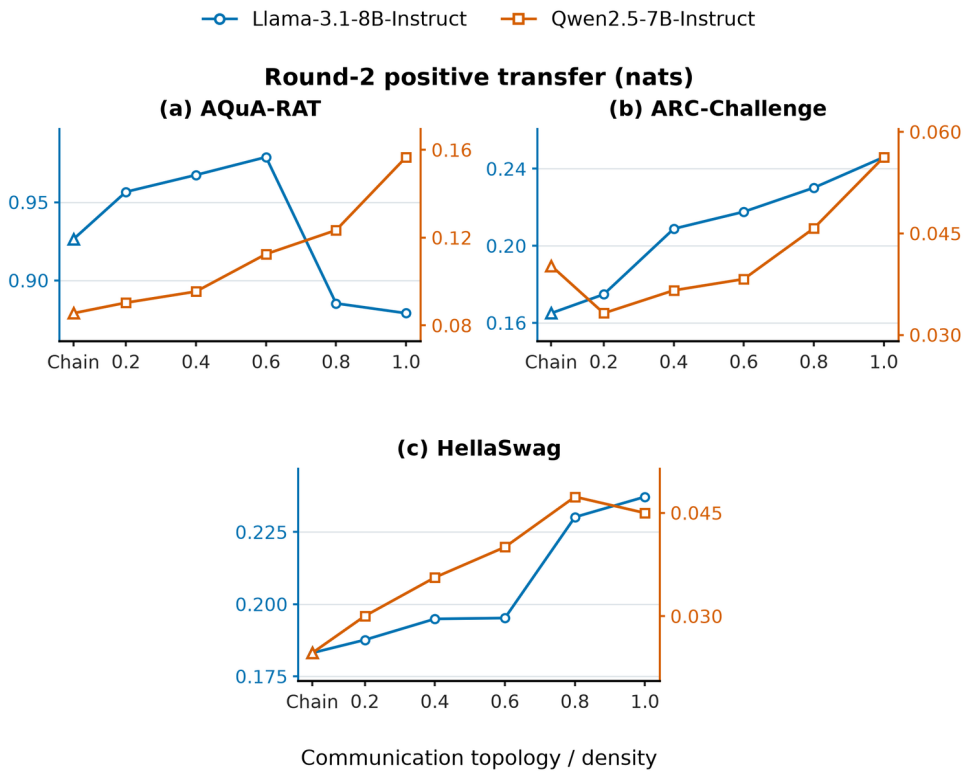


Figure 7: Round-2 empirical positive transfer $\widetilde{B}_{1\to 2}$. (a) AQuA-RAT; (b) ARC-Challenge; (c) HellaSwag. Blue/left and orange/right axes correspond to Llama and Qwen; values are in nats.

For AQuA-RAT, Llama rises through $\rho = 0.6$ and then drops, whereas Qwen increases across the sweep. Both ARC-Challenge curves rise overall after a small low-density variation, and both HellaSwag curves also increase. Thus direct coverage remains visible in several conditions, but AQuA-RAT already shows that the round-2 response depends on the state created in round 1.

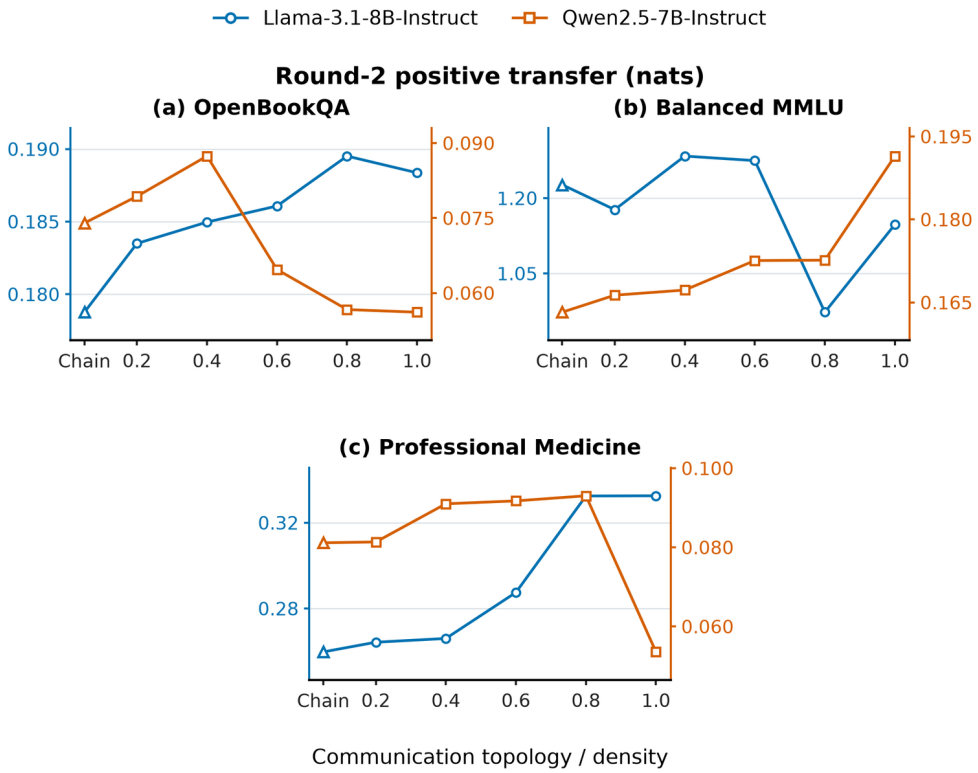


Figure 8: Round-2 empirical positive transfer $\widetilde{B}_{1\to 2}$. (a) OpenBookQA; (b) Balanced MMLU; (c) Professional Medicine. Conventions match the preceding figure.

The second group is more mixed. OpenBookQA rises to an intermediate or high density before Qwen declines; Balanced MMLU contains a pronounced Llama dip at $\rho = 0.8$; and Professional Medicine combines a high-density Llama plateau with a Qwen drop at full connectivity. By the second update, empirical positive transfer is therefore no longer determined by edge count alone; remaining evidence from the first update also matters.

C.2 Round-2 Empirical Negative Transfer

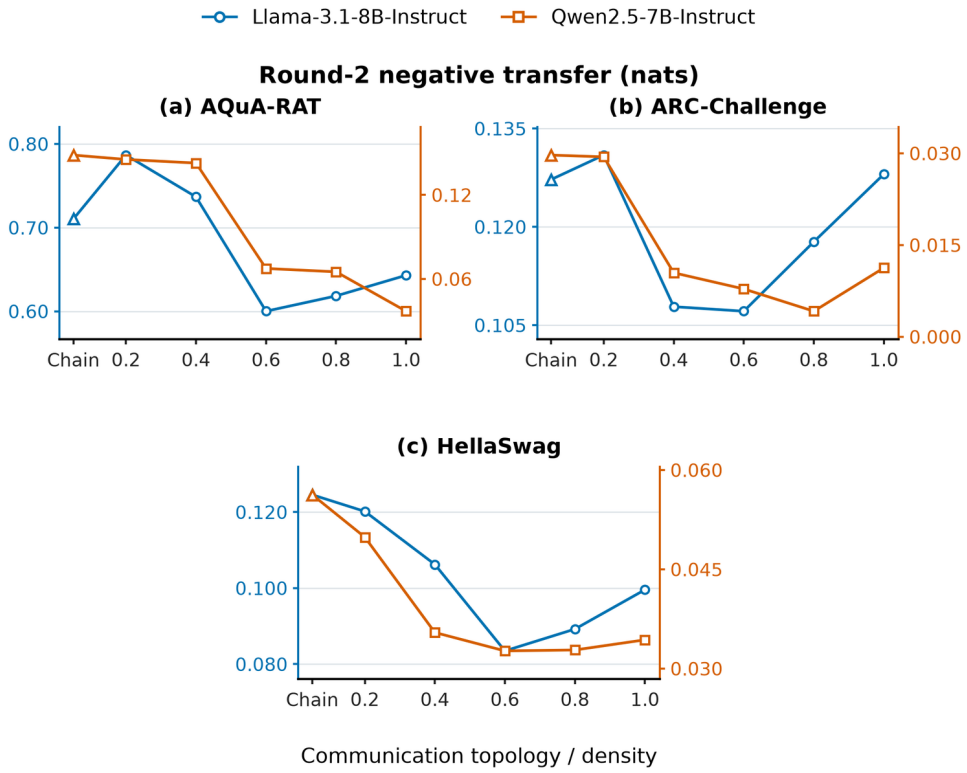


Figure 9: Round-2 empirical negative transfer $\widetilde{M}_{1\to 2}$. (a) AQuA-RAT; (b) ARC-Challenge; (c) HellaSwag. Conventions match the round-2 positive-transfer figures.

Across AQuA-RAT, ARC-Challenge, and HellaSwag, empirical negative transfer generally falls toward an intermediate density, although the precise minimum and the full-connectivity rebound differ by model and dataset. The curves support a conditional trade-off between error dilution and aligned-error reinforcement rather than a shared universal minimum.

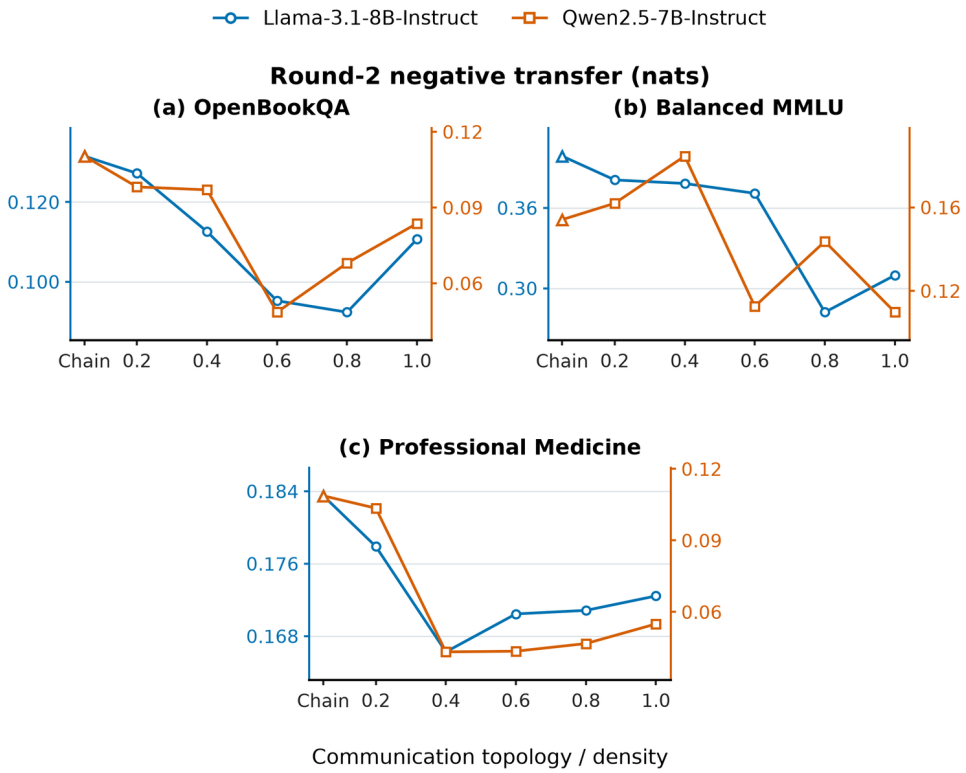


Figure 10: Round-2 empirical negative transfer $\widetilde{M}_{1\to 2}$. (a) OpenBookQA; (b) Balanced MMLU; (c) Professional Medicine.

OpenBookQA reaches its lowest region around $\rho = 0.6$ before rising, while Professional Medicine drops sharply by $\rho = 0.4$ and then changes only modestly. Balanced MMLU is more irregular across models. These differences are consistent with density-dependent error states shaped by earlier communication, as captured by the dynamic extension in Appendix A.7.

C.3 Round-3 Empirical Positive Transfer

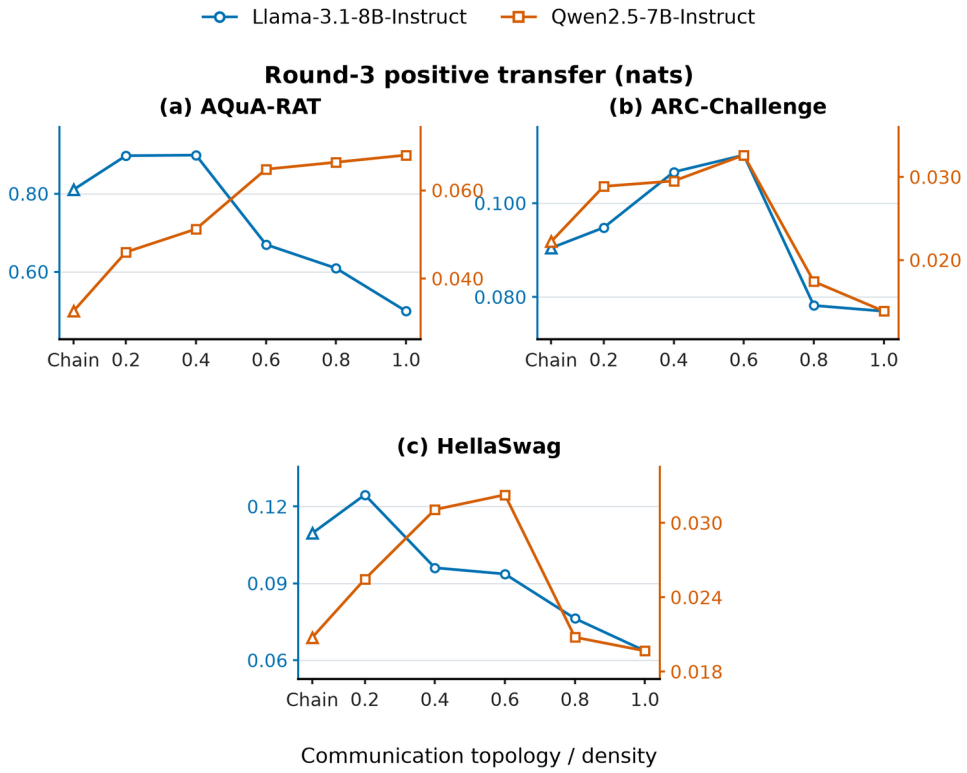


Figure 11: Round-3 empirical positive transfer $\widetilde{B}_{2\to 3}$. (a) AQuA-RAT; (b) ARC-Challenge; (c) HellaSwag.

High-density declines are more visible in round 3. Both ARC-Challenge models peak at an intermediate density and then decline, and HellaSwag shows a similar late decrease. Qwen on AQuA-RAT continues to rise while Llama falls after its low-density plateau, showing that later-round positive transfer varies across models and tasks.

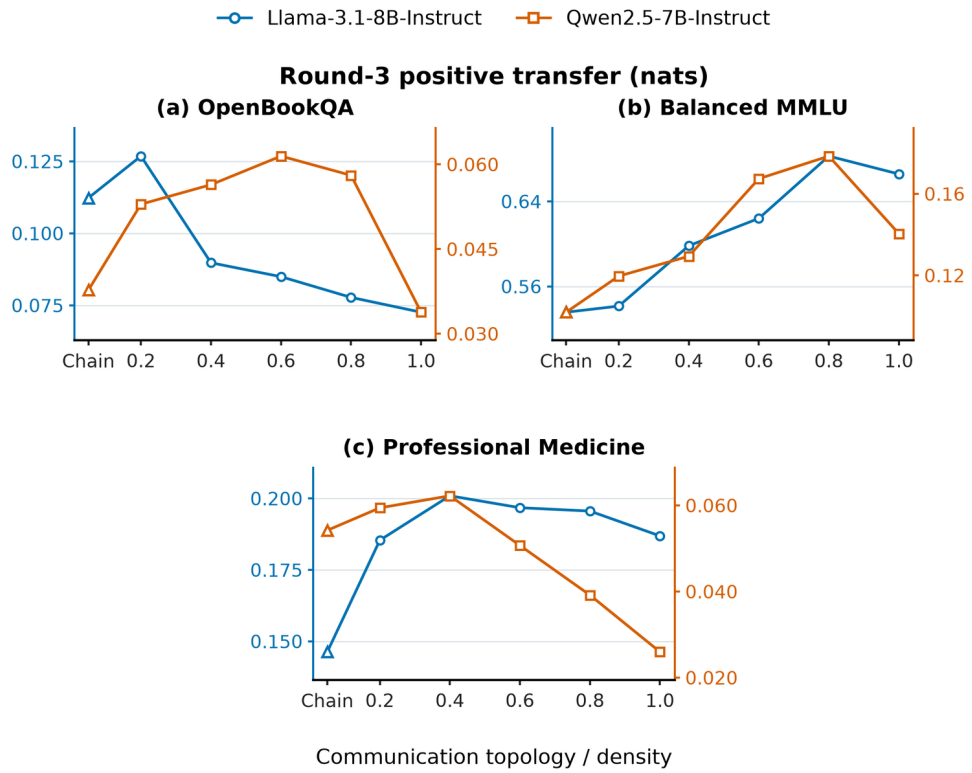


Figure 12: Round-3 empirical positive transfer $\widetilde{B}_{2\to3}$. (a) OpenBookQA; (b) Balanced MMLU; (c) Professional Medicine.

OpenBookQA and Professional Medicine contain medium- or high-density declines, while Balanced MMLU generally rises until a high density and then declines slightly. Together with the preceding figure, these curves support the view that dense communication can expand coverage quickly in early rounds while consuming or homogenizing useful evidence more rapidly.

C.4 Round-3 Empirical Negative Transfer

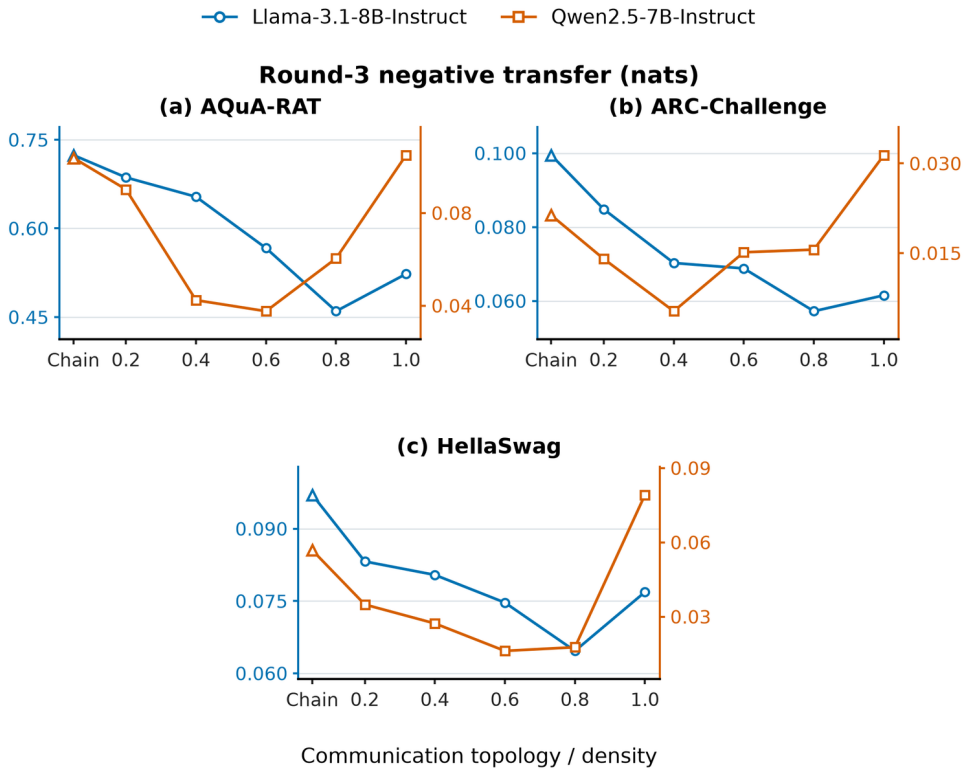


Figure 13: Round-3 empirical negative transfer $\widetilde{M}_{2\to 3}$. (a) AQuA-RAT; (b) ARC-Challenge; (c) HellaSwag.

For AQuA-RAT, ARC-Challenge, and HellaSwag, empirical negative transfer generally decreases from the chain toward an intermediate or high density, followed by a rebound under full connectivity in several curves. The location and size of that rebound differ across models, so the harmful-transfer response remains state dependent.

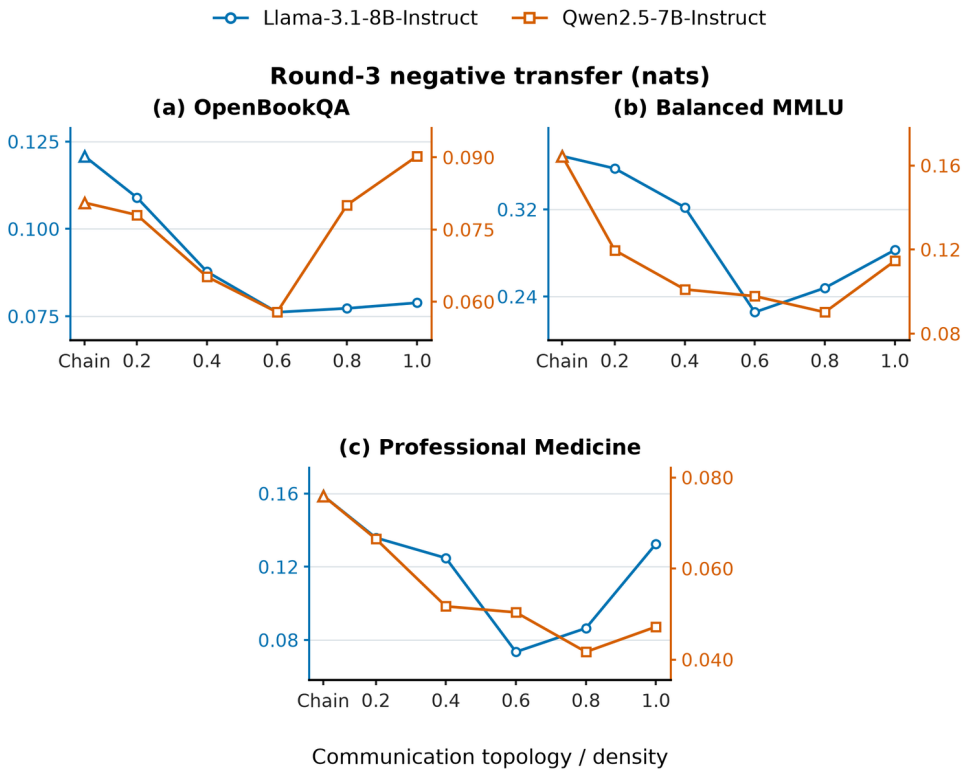


Figure 14: Round-3 empirical negative transfer $\widetilde{M}_{2\to3}$. (a) OpenBookQA; (b) Balanced MMLU; (c) Professional Medicine.

OpenBookQA and Balanced MMLU reach their lowest regions at intermediate or high densities and then rise, while Professional Medicine shows a smaller rebound after its minimum. Across both round-3 figures, dilution and reinforcement operate on a state distribution created by earlier rounds rather than on a density-independent population; the current round's coverage gain therefore cannot be interpreted without the preceding communication history.

D Application-Oriented Density Selection Protocol

We use Qwen2.5-7B-Instruct, 11 agents, three synchronous communication rounds, and the topology family described in Section 4.2. A fixed set of 250 questions is split into 100 validation and 150 disjoint test questions; the same split is used across topology conditions. Each non-complete density uses five independently sampled graphs with three independent generation runs per graph; the complete graph has one realization with three runs. Topology conditions share the same round-0 responses within each run. Decoding settings are held fixed across the compared topologies. Final answers use majority vote, with ties broken by aggregate self-reported confidence and then option order.

Validation evaluates $\rho \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ in ascending order; Chain is retained only as an external baseline. After each round, the model reports a confidence distribution over candidate options. We convert percentages to probabilities, clip negative entries to zero, and renormalize across options. The resulting correct-option probability is self-reported, not a native token log probability. All required probabilities were successfully parsed. Using $\epsilon = 10^{-6}$, we compute the round-wise positive and negative correct-option log-score transfers as in Section 4.1, with the validation set as $\mathcal{D}$. Let r index graph realizations and j index generation runs. We sum the three round-wise transfers within each run and average equally across runs and graphs at each density:

$$
\begin{aligned}
B_T^{(r,j)}(\rho) &= \sum_{t=0}^{2} \widetilde{B}_{t\to t+1}^{(r,j)}(G_\rho), \qquad \overline{B}_T(\rho) = \frac{1}{3R_\rho} \sum_{r=1}^{R_\rho} \sum_{j=1}^{3} B_T^{(r,j)}(\rho), \\
M_T^{(r,j)}(\rho) &= \sum_{t=0}^{2} \widetilde{M}_{t\to t+1}^{(r,j)}(G_\rho), \quad \overline{M}_T(\rho) = \frac{1}{3R_\rho} \sum_{r=1}^{R_\rho} \sum_{j=1}^{3} M_T^{(r,j)}(\rho),
\end{aligned}
\tag{D.1}
$$

where $R_\rho = 5$ for non-complete densities and $R_1 = 1$. With all required probabilities successfully parsed, cumulative net transfer equals the change in mean correct-option log score over the three rounds. The selection rule therefore compares the additional validation log-score improvement obtained by increasing density. These empirical transfers are not numerical estimates of the conditional mutual information in Section 2. For adjacent densities, define

$$
g_k = [\overline{B}_T(\rho_k) - \overline{B}_T(\rho_{k-1})] - [\overline{M}_T(\rho_k) - \overline{M}_T(\rho_{k-1})]. \tag{D.2}
$$

At the first $g_k \le 0$, select ρ_{k-1}; if every increment is positive, select $\rho = 1.0$. The test set is used only after selection. A full test-set scan defines the oracle solely for retrospective evaluation.

The rule selects $\rho = 0.2$ on RACE-High and $\rho = 0.6$ on MathQA. Chain reaches 83.33% on RACE-High and 88.00% on MathQA. The selected graphs contain 11 and 33 edges, compared with 55 in the complete graph, giving edge savings of 80% and 40%. Selection remains task- and setting-dependent.

E Accuracy Variability and Paired Confidence Intervals

Table E.1 reports standard deviations of final team accuracy and question-level paired 95% confidence intervals for the observed best-performing tested density versus full connectivity. All values are in percentage points.

Table E.1: Final-accuracy SDs at the observed best-performing tested density $\rho^\star$ and full connectivity, with paired 95% CIs for $\rho^\star$ minus full connectivity (percentage points).

Dataset	Model	$\rho^\star$	SD at $\rho^\star$	SD at 1.0	Paired 95% CI
HellaSwag	Llama	0.6	0.99	1.67	[3.55, 6.79]
	Qwen	0.6	0.68	0.83	[3.46, 6.54]
ARC-Challenge	Llama	0.4	0.80	1.29	[1.64, 3.21]
	Qwen	0.4	0.34	1.20	[1.13, 2.87]
Balanced MMLU	Llama	0.4	0.54	0.46	[4.81, 6.65]
	Qwen	0.4	0.65	1.29	[3.15, 4.74]
Professional Medicine	Llama	0.6	0.31	0.37	[5.19, 7.31]
	Qwen	0.4	0.68	0.37	[0.94, 2.00]
OpenBookQA	Llama	0.4	0.69	1.06	[2.59, 4.61]
	Qwen	0.6	0.45	1.06	[1.66, 3.14]
AQuA-RAT	Llama	0.4	0.92	1.20	[3.51, 5.29]
	Qwen	0.8	0.39	0.69	[1.34, 2.66]