

Peer Review

Review of: "The Cost of Consensus: A Simulation Study of Malignant Epistemic Herding and Adaptive Gating in Distributed Multi-Agent Search"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper studies how the **content** and the **frequency** of inter-agent messages jointly shape both task performance and the quality of a team's collective beliefs in a decentralised search problem. Four agents with partial observability (field-of-view radius $r=2$, Chebyshev) hunt a single stationary target on a 50x50 grid within $T=200$ timesteps, each maintaining a discrete belief over all 2,500 cells (Table 1). Four hand-designed protocols are compared (Table 2): no communication (C0), per-timestep broadcast of an argmax cell plus entropy (C1), per-timestep broadcast of a top-5 belief summary fused by inverse-entropy weighting (C2), and the same epistemic message emitted only when the sender's Shannon entropy has moved by at least $\theta=0.20$ nats (C3, entropy-delta gating). Crossing protocol x packet loss x latency x coordination threshold gives a full factorial of 108 conditions at 1,000 episodes each. The authors report three claims. First, richer message content buys task success: C2 reaches 0.639 success at $k=2$ against C1's 0.388 (Table 3, Success columns). Second, high-frequency epistemic communication nonetheless damages collective belief quality, because repeated inverse-entropy fusion drives agents into shared but wrong beliefs – a failure mode they name **malignant epistemic herding** (MEH), formalised in Eq. 10 as low inter-agent divergence conjoined with task failure, and argued to be invisible to ordinary coordination metrics. Third, entropy-delta gating resolves the tension: C3 cuts message volume by more than 98% (Table 4, Msgs: 19 vs 1,537 at $k=2$) while posting the highest alignment and success rates (Table 3) and the lowest MEH rate among communicating protocols (Table 4, MEH: 0.062 vs 0.990 at $k=2$). A

grid-size sweep (Table 6 / Table B.16) and a theta sweep (Table 5, Table C.17) are offered as robustness evidence.

2. Whether the conclusion holds on the data

Design arithmetic reconciles. The factorial multiplies out correctly: 4 protocols x 3 loss levels x 3 latencies x 3 coordination thresholds = 108 conditions (Section 4.6), and the appendix delivers exactly 36 rows per threshold across Tables A.7–A.9, A.10–A.12, and A.13–A.15, i.e. 108,000 episodes. The per-agent coverage figures in Section 5.5 also check out: 5 novel cells per step x T, divided by N-squared, gives 40.0%, 40.0%, 31.1%, and 25.0% for N = 25, 50, 75, 100, exactly as stated. The byte accounting is internally exact: Table A.13 gives 13,264.3 bytes over 1,894.9 messages for C1 (7.00 bytes per message) and 52,833.1 over 1,509.5 for C2 (35.00 bytes), matching the 7 and 35 bytes quoted in Section 6.1, and the alignment-per-byte column reproduces to three digits (C3, k=2, no loss, no latency: $0.704/714.7 = 985.0 \times 10^{-6}$, exactly the tabulated value). The 98% message-reduction claim survives in every cell I checked, from 98.39% in the worst case (Table A.15, C3 vs C2 at 30% loss and latency 3: 28.2 vs 1,752.4 messages) to 99.0% in the best. The C0 baseline is also physically plausible: with each agent covering at most 40% of the grid, a binomial upper bound puts $P(\text{at least 2 of 4 cover the target})$ near 0.52 and $P(\text{all 4})$ near 0.026, against observed 0.404 and 0.016 (Table 3, C0 row). The simulator is not obviously broken, and I want to say so plainly before turning to the problems.

The agents are scripted Bayesian agents, not LLMs. Section 3.2.1 specifies a log-probability vector over 2,500 cells updated by a fixed negative increment on observed-empty cells and a large positive increment on detection; Section 3.2.2 specifies greedy motion toward the belief argmax. The fusion rule is Eq. 2 with the weights of Eq. 3. Nothing in this system learns, and no language model is in the loop – the Claude usage disclosed under Use of Generative AI is authoring assistance, not agent substrate. The paper is honest about this, and I do not fault the choice. But Section 6.4 extends the claim to “networked AI agents operating over shared information environments”, and the abstract offers entropy-delta gating as a primitive “applicable to any decentralized system in which agents maintain probabilistic beliefs under uncertainty”. That conditional clause is doing enormous work. LLM-based agents do not maintain an explicit distribution over a state simplex, have no computable Shannon entropy over their latent hypothesis, and therefore admit neither the gate of Eq. 4 nor the fusion weight of Eq. 3 without a surrogate the paper does not propose. The generalisation in Section 6.4 should be narrowed to systems that literally carry an explicit belief simplex, or a surrogate uncertainty statistic should be named.

The herding result does depend on the hand-chosen update rule, and the decisive control is missing. Section 2.2 concedes that treating incoming messages as independent evidence produces the classical double-counting overconfidence, and that the inverse-entropy weighting of Eq. 3 is "an explicitly approximate method inspired by CI" with no de-correlation term. Section 6.2 then argues the mechanism is general: "any weighted fusion scheme that privileges confident senders will exhibit analogous dynamics when messages are frequent". That is an assertion, not a result. C2 and C3 differ only in the gate, so what the experiment isolates is the effect of *frequency under one specific naive fusion rule*. It cannot distinguish "high-frequency epistemic communication causes MEH" from "fusion without shared-history discounting causes MEH, and frequency merely sets how fast". The missing arm is a fourth protocol - ungated, per-timestep, epistemic content, but fused with a decorrelating rule (a discrete Covariance-Intersection surrogate, or simple message-history suppression). Until that arm exists, the paper's central mechanistic claim is under-determined by its own design, and a reader could reasonably conclude the cheaper fix is to repair the fusion rule rather than to throttle the channel. This is the single most consequential gap in the study.

The MEH metric is, for two of the three protocols, arithmetically identical to the failure rate. This is the finding that most affects how the results should be read, and it is verifiable from the paper's own numbers. In Table B.16 (and its duplicate, Table 6), the MEH rate equals 1 minus the success rate *exactly*, to three decimals, in all eight C1 and C2 rows: 25x25 C1 (success 0.472, MEH 0.528), 25x25 C2 (0.672, 0.328), 50x50 C1 (0.347, 0.653), 50x50 C2 (0.660, 0.340), 75x75 C1 (0.230, 0.770), 75x75 C2 (0.527, 0.473), 100x100 C1 (0.158, 0.842), 100x100 C2 (0.451, 0.549). The same identity holds in Tables A.13-A.15: C1 at no loss and no latency, $k=2$, shows success 0.361 (Table A.7) and MEH 0.639 (Table A.13); C2 shows 0.642 and 0.358. The reason is structural. Under Eq. 10, an episode is MEH when divergence falls below the threshold and the task fails; C1 and C2 drive final JSD to 0.0037 and 0.0060 nats respectively (Table A.10, no loss, no latency, $k=2$), which is below any plausible threshold in essentially *every* episode. So the divergence conjunct is always satisfied, and MEH collapses to "the task failed". The consequence is that the headline comparison in Table 4 ("MEH 0.990 for C2 versus 0.062 for C3 at $k=2$ ") is not an independent finding alongside the success rates of Table 3; it is very largely those same success rates re-expressed, plus the fact that C3's gate mechanically keeps JSD above the threshold (C3's final JSD is 0.0443 at $k=2$, Table A.10, an order of magnitude above C2's). Likewise, the claim in Section 5.5 that the C2-C3 MEH gap "widens systematically with grid size - from 0.249 at 25x25 to 0.514 at 100x100" is, for the C2 half of that gap, precisely a restatement of C2's success rate falling from 0.672 to 0.451, and C1's MEH rising "monotonically from

0.528 to 0.842" is exactly C1's success falling from 0.472 to 0.158. The paper needs to state what operational decision MEH supports that the failure rate does not.

The mechanism story is not evidenced, and Figure 4 argues against it. Section 5.2 and Section 6.2 describe MEH as agents locking onto "a confident but incorrect belief". But the caption of Figure 4 reports that in failed episodes *all* protocols remain near the uniform prior, with mean alignment below 0.001, approximately $1/2500 = 0.0004$. Beliefs that are near-uniform are not confident, and two near-uniform distributions have low pairwise JSD trivially. On the evidence presented, C1 and C2 failures look like agreement on ignorance rather than coordinated commitment to a wrong cell. To establish the stated mechanism, the paper must report, conditional on failure, the mean sender entropy H and the probability mass the team places on its shared argmax cell; if that mass is near $1/2500$, the word "herding" is not earned. Relatedly, Figure 3's distinction between C1/C2 collapsing below the threshold and C3 showing "increasing JSD" in failed episodes is currently visual only, and it sits uneasily with C3's overall mean final JSD of 0.0443 with SD 0.0697 (Table A.10, $k=2$, no loss, no latency): if that distribution is even roughly unimodal, well over half of all C3 episodes fall below 0.1 nats, yet C3's MEH is only 0.016 of all episodes against a 0.287 failure rate. The success-conditioned and failure-conditioned JSD distributions must be tabulated, not merely plotted.

Effect sizes are absent, and with 1,000 episodes per arm, significance is close to free. Section 4.6 is the paper's entire statistical apparatus: two-sided Mann-Whitney U tests for continuous metrics, two-proportion z-tests for binary rates, Benjamini-Hochberg correction, and the verdict that "all protocol comparisons across primary conditions are statistically significant at $FDR < 0.05$ ". No test statistic, no p-value, no confidence interval, and no standardised effect size appear anywhere in the paper, including the appendix. That is not adequate at this sample size. A worked example from the authors' own appendix: C2 versus C3 on final alignment at $k=3$, no loss, no latency (Table A.11) is 0.662 (SD 0.472) against 0.720 (SD 0.441). The standard error of that difference is 0.0204, so $z = 2.84$, and the result is indeed significant - but Cohen's d is $0.058 / 0.457 = 0.13$, a small effect by any convention. The paper reports the verdict and suppresses the magnitude. Note also what those SDs reveal: at roughly 0.44-0.48 against means of 0.66-0.72, final alignment is strongly bimodal, because an episode either locates the target (mass near 1) or does not (mass near 0.0004). Mean alignment is therefore nearly collinear with the success rate - compare Table 3 at $k=2$, where C2 records alignment 0.633 against success 0.639 and C3 records 0.700 against 0.734. Between this collinearity and the MEH identity established above, Tables 3 and 4 are largely three presentations of one underlying quantity, $P(\text{target found})$. Every comparison needs a

reported effect size (rank-biserial correlation for the Mann-Whitney tests, risk difference with a 95% CI for the proportions), and the number of tests entering the BH correction must be stated – the paper never says whether the family is the 3 pairwise comparisons at the primary condition, the 6 protocol pairs across 27 network-by-threshold cells, or something larger.

Multiple-comparison correction is claimed but not auditable. Section 4.6 and the Appendix A preamble both assert BH correction "across all tests", with two named exceptions (C0 versus C1 success at high latency; C2 versus C3 final alignment at maximum packet loss and latency). Since neither the family size nor any raw p-value is given, a reader cannot check either the correction or the exceptions. A supplementary table of every comparison with raw p, BH-adjusted q, effect size, and CI is the minimum here, and at 108 conditions, it is not optional.

The generative model partly bakes in the C3 advantage – though not on the headline result. Eq. 5 makes packet loss congestion-dependent, $P(\text{drop}) = 1 - (1 - p_{\text{base}})^n$ in the number of simultaneous senders. At $p_{\text{base}} = 0.3$ with all four agents transmitting, this is $1 - 0.7^4 = 0.760$, whereas a lone C3 sender faces 0.300; at $p_{\text{base}} = 0.1$, the figures are 0.344 and 0.100. The sparse protocol is thus handed a delivery-probability advantage of roughly 2.5x by the noise model itself, over and above any property of gating. Section 3.4 acknowledges this in one sentence, but no control with congestion-independent loss is run. In fairness, the MEH results the paper leads with are reported at zero packet loss and zero latency (Figure 3 caption, Table 5, Table B.16), where Eq. 5 is inert, so the central claim does not rest on this. Two further points should be made, however. First, the functional form of Eq. 5 is asserted without citation to any medium-access model, and it reuses the symbol n , defined as the number of agents ($n = 4$) in Section 3, for the number of simultaneous senders. Second – and this is an interesting result the authors have not noticed – packet loss itself suppresses herding. In Table A.13 at $k=2$, C1's MEH falls from 0.639 at no loss and no latency to 0.516 at 30% loss and latency 3, a 19% relative reduction, and the trend is visible across all three tables. If random message loss is itself a partial remedy for MEH, then entropy-delta gating is in one sense engineered packet loss, and the paper should say what gating adds beyond stochastic dropping – a testable and, I suspect, favourable comparison, since gating drops the *uninformative* messages selectively.

The C1-versus-C2 contrast confounds three factors at once. Table 2 shows C1 and C2 differ in message content (argmax cell versus top-5 distribution), in payload (7 versus 35 bytes, Section 6.1), and in fusion rule ("Argmax replacement" versus "Inv-entropy weighted"). Section 6.1 acknowledges only the first two and offers alignment-per-byte as the remedy. That does not license the paper's framing that "message

content ... determines how well the team coordinates" (Section 6.1), because the fusion rule varies simultaneously and is never isolated. The C2-versus-C3 contrast, by contrast, is clean – identical content, identical fusion, gate only – and that is the comparison the paper should lead with.

Alignment per byte is a degenerate objective. Section 4.5 defines APB as final mean alignment divided by bytes transmitted, then excludes C0 because the metric is "undefined" at zero bytes. That exclusion is the diagnosis, not the fix: the metric diverges to infinity as transmission goes to zero for any protocol with non-zero alignment, and C0 achieves a mean alignment of 0.299 (recomputed across the nine C0 rows of Table A.10, $k=2$) at exactly zero bytes. A ratio whose supremum is attained by sending nothing cannot serve as an efficiency criterion. Report instead alignment against bytes as a Pareto frontier (which is what Figure 2 already shows), or the bytes required to reach a fixed alignment target.

Threshold selection is tuned on the evaluation set, and the robustness claim is contradicted by Table 5. Section 3.3 states that "performance is robust in theta within [0.15, 0.35]". Table 5 shows success rising monotonically across that interval from 0.658 at theta = 0.15 to 0.755 at theta = 0.35 – 9.7 percentage points, a 14.7% relative change in the primary task metric – and MEH more than doubling from 0.089 at theta = 0.20 to 0.190 at theta = 0.35. That is not robustness. The claim in Section 5.4 that message volume varies by fewer than three messages over theta in [0.20, 0.35] does check out (22.9, 22.4, 21.3, 20.8 in Table 5, a range of 2.1), but bandwidth stability is not performance stability. Worse, theta = 0.20 is selected because it minimises MEH – the paper's own novel metric – while theta = 0.35 in the same Table 5 delivers better success (0.755 versus 0.695), better alignment (0.790 versus 0.711), better APB (1,095.8 versus 893.5), and fewer messages (20.8 versus 22.9). Selecting the operating point on the metric one is introducing, against the conventional metric, needs explicit defence. And the selection is made on the same 1,000-episode sample used for the headline results; a held-out seed block is required.

The Appendix C generalisation of that selection is false at $k=2$. The Appendix C preamble states that "the selected operating point theta = 0.20 minimizes MEH within the low-bandwidth plateau across all coordination thresholds". In Table C.17 at $k = 2$, theta = 0.25 gives MEH 0.010 against theta = 0.20's 0.022 – less than half. The claim holds at $k = 3$ (0.025, minimum) and $k = 4$ (0.034, minimum) but not at $k = 2$, and it fails on either MEH denominator.

External validity. Beyond the LLM point above, three structural limits bound the result more tightly than Section 6.5 concedes. Team size is fixed at $n = 4$ throughout and is never varied, although MEH is advanced as a *team-level* phenomenon whose severity should depend on how many correlated senders feed each fusion step – the scaling analysis varies the state space but not the team. Sensing is noiseless

(Section 6.5, Sensor model), so there is no false-positive channel through which a genuinely confident wrong belief could form, which may be exactly why failed episodes sit at the uniform prior in Figure 4. And the environment holds a single stationary target, so the "wrong hypothesis" the team is said to converge on has no competing attractor to converge onto. These are acknowledged as limitations, but their joint effect is that the paper may be observing coordinated *indecision* rather than coordinated *error*, and the generalisation in Section 6.4 should be conditioned accordingly.

3. Reproducibility

The Code Availability statement names a public repository (github.com/davidthfarr/agent2agent) said to contain all four protocols, the environment, and "deterministic random seeds for all 108 experimental conditions", and Section 4.6 states that "all experiments are fully deterministic given seed values". This is above the median for simulation papers and should be credited. It is nonetheless incomplete in five specific respects.

First, no commit hash, release tag, or archival DOI pins the repository state to the manuscript, and no language, library, or RNG-implementation versions are given; a paper whose reproducibility claim rests on bitwise determinism must pin the artefact. Second, the seed guarantee covers only the 108 main conditions, but the study actually reports roughly 158,000 episodes: 108,000 in the factorial, 9,000 in Table 5, 27,000 in Table C.17 (9 theta values x 3 thresholds x 1,000), and 14,000 in Table B.16 (12 conditions at 1,000 plus 4 at 500). The additional 50,000 episodes – which carry the theta selection and the entire scaling argument – fall outside the stated seed provision. Third, it is not stated whether the same target-placement seeds are reused across protocols. This matters twice over: a paired design would make the Mann-Whitney tests of Section 4.6 substantially more powerful, and if seeds *are* shared, the comparisons are dependent and the BH assumptions in Section 4.6 need restating. Fourth, Section 4.6 promises that "results are reported as means with standard deviations across seeds unless otherwise noted", but no dispersion whatsoever accompanies Table 3, Table 4, or Table 5, and the appendix supplies SDs only for time-to-success (Tables A.7-A.9) and for JSD and alignment (Tables A.10-A.12) – never for success rate, MEH rate, message count, or APB. Fifth, the parameter table is otherwise good (Table 1 fixes N , $|S|$, n , r , T , k , $k_{\text{msg}} = 5$, $w_{\text{max}} = 0.8$, $\theta = 0.20$, p_{base} , and latency grids), but two quantities that materially drive the dynamics are described only in prose and never given numerically: the magnitude of the negative log-probability decrement on observed-empty cells and of the positive update on detection (Section 3.2.1), and the size of the "fixed probability boost" $C1$ applies to a reported argmax cell, together with how

it is "scaled by the sender's reported entropy" (Section 3.3). Neither is recoverable from the text, and the C1-versus-C2 comparison cannot be assessed without the latter.

The definition of the epistemic-alignment metric is not stable across the paper, and this affects reported numbers. Three distinct problems, all checkable. (a) The MEH threshold takes two different values. Eq. 10 sets $\epsilon = 0.2$ nats, and Section 4.4 justifies it as "approximately 14% of the maximum possible JSD ($\log 2 = 0.693$ nats)" - but $0.2/0.693 = 28.9\%$, not 14%; it is 0.1 nats that is 14.4% of $\log 2$. Consistently with that, the Figure 3 caption uses " $\epsilon = 0.1$ nats (dashed)" and the captions of Tables A.13, A.14, and A.15 all define MEH at " $\text{JSD} < 0.1$ nats". Eq. 10 appears to be a typographical error, but as written, the paper defines its headline construct at two thresholds that differ by a factor of two. (b) The MEH *denominator* also differs between tables, silently. Table 5 and Table 4 report MEH among failed episodes; Tables A.13–A.15, Table B.16, and Table C.17 report it over all episodes. The two reconcile once the convention is flipped - at $k=3$, $\theta = 0.20$, Table 5's 0.089 among failures with success 0.695 implies 0.027 of all episodes, against Table C.17's 0.025; at $\theta = 0$, Table 5's 1.000 with success 0.629 implies 0.371, against Table C.17's 0.354 - so the underlying data are consistent and only the labelling is wrong. But the captions of Tables A.13–A.15, which read "MEH = malignant epistemic herding rate among failed episodes", are demonstrably incorrect: C1 at $k=2$, no loss, no latency has success 0.361 (Table A.7) and tabulated MEH 0.639, which is the whole-sample rate. As it stands, a reader comparing Table 5 with Table C.17 will conclude the two runs disagree by a factor of 3.6 at $\theta = 0.20$. (c) The timestep at which the metrics are evaluated is inconsistent. Section 4.2 says JSD and alignment are "recorded at every timestep throughout each episode" and reports the mean across timesteps, whereas Tables A.10–A.12 define both "at episode end", and Eq. 10 never says at which timestep (or over which aggregation) A_{div} is compared against ϵ - final, time-averaged, or minimum over the episode. Since the minimum over 200 timesteps will fall below any threshold far more often than the final value, this choice alone can move every MEH number in the paper. All three definitions must be fixed to a single convention and applied uniformly.

4. What specifically to fix

1. **Add the decorrelating-fusion control arm.** Introduce a protocol with C2's content and per-timestep frequency but a fusion rule that discounts shared history (a discrete Covariance-Intersection surrogate, or suppression of information already received). Without it, Section 6.2's claim that the mechanism generalises to "any weighted fusion scheme that privileges confident senders" is

untested, and the study cannot separate frequency from the naive fusion of Eq. 2-3. This is the change that most affects the paper's standing.

2. **Demonstrate that MEH carries information beyond the failure rate.** In Table B.16 the MEH rate equals 1 minus the success rate exactly, to three decimals, in all eight C1 and C2 rows (25x25 C1: 0.472/0.528; 25x25 C2: 0.672/0.328; 50x50 C1: 0.347/0.653; 50x50 C2: 0.660/0.340; 75x75 C1: 0.230/0.770; 75x75 C2: 0.527/0.473; 100x100 C1: 0.158/0.842; 100x100 C2: 0.451/0.549), and the same identity holds throughout Tables A.13-A.15. Either report MEH as a fraction of failures alongside the failure rate itself so the two are visibly distinct, or add a discriminating statistic - for instance, the fraction of failed episodes in which the team's shared argmax is a single specific wrong cell.
3. **Evidence the "confident but incorrect" mechanism, or restate it.** The Figure 4 caption reports mean alignment below 0.001 (approximately 1/2500) in failed episodes for all protocols, which is the uniform prior and is inconsistent with the confident-wrong-belief account of Section 5.2 and Section 6.2. Add, conditional on failure, the mean sender entropy H and the probability mass on the team's modal cell.
4. **Reconcile the MEH threshold.** Eq. 10 states $\epsilon = 0.2$ nats while Section 4.4 describes it as "approximately 14% of the maximum possible JSD" - $0.2/\log 2 = 28.9\%$, whereas $0.1/\log 2 = 14.4\%$. The Figure 3 caption and the captions of Tables A.13-A.15 all use 0.1 nats. Fix Eq. 10 and confirm which threshold generated every reported MEH figure.
5. **Correct the MEH denominator captions.** Tables A.13, A.14, and A.15 state "MEH = ... rate among failed episodes", but their values are whole-sample rates (C1, $k=2$, no loss, no latency: success 0.361 in Table A.7, MEH 0.639 in Table A.13). Table 4 and Table 5 use the among-failures convention, Table C.17 and Table B.16 the whole-sample one. State the convention in every caption, or adopt one throughout - this alone makes Table 5 and Table C.17 appear to disagree by a factor of 3.6 at $\theta = 0.20$ (0.089 versus 0.025).
6. **Specify the timestep at which A_{div} in Eq. 10 is evaluated.** Section 4.2 describes per-timestep recording with a mean across timesteps; Tables A.10-A.12 specify "at episode end". Whether MEH uses the final, mean, or minimum divergence changes every MEH value reported.
7. **Report effect sizes and confidence intervals for every comparison asserted in Section 4.6.** No p -value, test statistic, or effect size appears anywhere in the paper. Worked example: C2 versus C3 final alignment at $k=3$, no loss, no latency (Table A.11) is 0.662 (SD 0.472) against 0.720 (SD 0.441), giving $z = 2.84$ but Cohen's $d = 0.13$ - significant and small. Add rank-biserial correlation for the Mann-Whitney tests and risk difference with 95% CI for the proportion tests.

8. **State the family size entering the Benjamini-Hochberg correction** (Section 4.6 and the Appendix A preamble) and supply a supplementary table of all comparisons with raw p, adjusted q, effect size, and CI. At 108 conditions, the assertion "all ... significant at $FDR < 0.05$ " is not auditable as written, and the two named exceptions cannot be verified.
9. **Quote the Monte Carlo noise floor.** The nine C0 rows of Table A.7 span 0.370 to 0.407 in success at $k=2$ across network conditions that cannot affect C0 by construction (Section 4.6). That 0.037 spread is pure sampling noise at 1,000 episodes and should be stated as the resolution limit below which differences in Tables 3, A.7-A.9, and B.16 are not interpretable.
10. **Reconcile Table 3 with Appendix A, and state its aggregation rule.** Table 3 gives no aggregation. Recomputing the nine-condition means from Tables A.7 and A.10 at $k=2$ yields success 0.390/0.387/0.629/0.713 and alignment 0.299/0.343/0.628/0.677 for C0-C3, against Table 3's 0.404/0.388/0.639/0.734 and 0.303/0.345/0.633/0.700. The discrepancy reaches 0.021 in success and 0.023 in alignment for C3, and every discrepancy favours the proposed protocol. The no-loss/no-latency cells do not match either (C3 success 0.713 in Table A.7 versus 0.734 in Table 3), so Table 3 appears to come from a separate run.
11. **Correct the "over 80x" APB claim in Section 6.1.** The claim that "C3 dominates both C1 and C2 on this metric by over 80x" holds for C2 only: from Table 4, C3's cross-threshold mean is 939.3×10^{-6} and C2's is 11.4×10^{-6} (82.6x), but C1's is 26.7×10^{-6} , giving 35.1x. Section 5.3 states the C2 comparison correctly; Section 6.1 over-extends it to C1.
12. **Recheck the 80x figure against the appendix.** Averaging the APB columns of Tables A.13-A.15 gives $C3 = 874.9 \times 10^{-6}$ and $C2 = 11.27 \times 10^{-6}$, a ratio of 777x, which does not reach the "over 80x" of Section 5.3. Table 4's C3 APB values (1091.4, 887.5, 838.9) each exceed the corresponding appendix means (1007.7, 828.2, 788.9) by 6-8%; the same tables show a smaller discrepancy in C2's message counts (Table 4: 1537, 1622, 1715; appendix means: 1544.7, 1626.0, 1720.7). Identify which set is authoritative.
13. **Withdraw or qualify the Appendix C claim that $\theta = 0.20$ minimises MEH "across all coordination thresholds".** In Table C.17 at $k=2$, $\theta = 0.25$ gives 0.010 against $\theta = 0.20$'s 0.022. The claim holds at $k=3$ and $k=4$ only.
14. **Withdraw the robustness claim in Section 3.3.** "Performance is robust in θ within [0.15, 0.35]" is contradicted by Table 5, where success rises monotonically from 0.658 to 0.755 across that interval (9.7 points, 14.7% relative) and MEH rises from 0.089 at $\theta = 0.20$ to 0.190 at $\theta = 0.35$. Bandwidth is flat over that range (22.9 to 20.8 messages, Table 5); performance is not.

15. **Justify selecting theta on MEH rather than on task success, and select it out of sample.** In Table 5, $\theta = 0.35$ beats the chosen $\theta = 0.20$ on success (0.755 versus 0.695), alignment (0.790 versus 0.711), APB (1,095.8 versus 893.5), and message count (20.8 versus 22.9). The selection is also made on the same 1,000-episode sample that carries the headline results; use a held-out seed block.
16. **Add a congestion-independent packet-loss control.** Eq. 5 gives four simultaneous senders a drop probability of $1 - 0.7^4 = 0.760$ at $p_{\text{base}} = 0.3$ against 0.300 for a lone C3 sender (and 0.344 versus 0.100 at $p_{\text{base}} = 0.1$), handing the sparse protocol a roughly 2.5x delivery advantage from the noise model alone. Section 3.4 notes this but runs no control. Also cite a medium-access model justifying the functional form, and rename the sender count in Eq. 5, which currently collides with $n = 4$ agents from Section 3.
17. **Report and discuss the finding that packet loss itself suppresses MEH.** In Table A.13 at $k=2$, C1's MEH falls from 0.639 at no loss and no latency to 0.516 at 30% loss and latency 3. If stochastic dropping is already a partial remedy, the paper should benchmark entropy-delta gating against random dropping at matched message volume – the natural null for a gating mechanism, and one C3 should win.
18. **Correct the message-count range in the Figure 2 caption.** It states that C1 and C2 "transmit continuously at 1,300-1,600 messages per episode", but Table 4 gives C1 at 1,856, 1,891, and 1,947, and Tables A.13-A.15 put C1 between 1,788.5 and 1,996.3 and C2 between 1,498.7 and 1,753.6 – both partly or wholly above the stated range.
19. **Correct the episode count and the budget-scaling description in the scaling analysis.** Section 5.5 says "running 1,000 episodes" for all four grids, while the captions of Table 6 and Table B.16 say "1,000 episodes (500 for 100x100)". Separately, Appendix B states that budgets "scale proportionally with grid side length", but T/N is 2.0, 4.0, 4.67, and 5.0 for $N = 25, 50, 75, 100$ – manifestly not proportional, as the paper's own coverage ratios (40.0%, 40.0%, 31.1%, 25.0%) confirm.
20. **Reconcile the 50x50 block of Table B.16 with the main factorial.** That block should be identical to the $k=3$, no-loss, no-latency condition already reported, but gives success 0.705, alignment 0.723, and JSD 0.0431 for C3 against 0.698 (Table A.8), 0.720, and 0.0435 (Table A.11), and success 0.660 for C2 against 0.629 (Table A.8). Either reuse the factorial run or state that the scaling sweep is an independent replication and report both.
21. **Supply the missing update magnitudes.** Section 3.2.1 describes the empty-cell decrement and the detection increment only qualitatively, and Section 3.3 gives neither the magnitude of C1's "fixed

probability boost" nor its entropy scaling function. Add both to Table 1; the C1-versus-C2 comparison is not assessable without the latter.

22. **Disentangle or reframe the C1-versus-C2 comparison.** Table 2 shows the two differ in content, payload (7 versus 35 bytes, Section 6.1), and fusion rule simultaneously. Either add a top-k-content protocol using argmax-replacement fusion, or restate Section 6.1's conclusion as being about protocol families rather than about "message content" as an isolated factor.
23. **Replace alignment per byte with a Pareto presentation.** Section 4.5 excludes C0 because APB is undefined at zero bytes, but C0 attains mean alignment 0.299 (recomputed from Table A.10, $k=2$) at zero cost, so the metric's supremum is attained by silence. Report bytes-to-reach-target-alignment, or present the frontier of Figure 2 directly.
24. **Report dispersion in the main tables.** Section 4.6 promises means with standard deviations, but Tables 3, 4, and 5 carry none, and no table anywhere reports dispersion for success rate, MEH rate, message count, or APB. Binomial confidence intervals on the rate columns would also make item 9's noise floor visible in place.
25. **Vary team size.** $n = 4$ is fixed throughout (Table 1) while MEH is advanced as a team-level phenomenon; the severity of double counting under Eq. 2-3 should grow with the number of correlated senders. A sweep over n in the manner of Section 5.5's grid sweep would materially strengthen the generality claim, and is more informative here than further grid sizes.
26. **Narrow the generalisation in Section 6.4 and the abstract.** Entropy-delta gating as specified requires an explicit distribution over a finite state space (Eq. 4 and Eq. 3). Either restrict the claim to systems carrying such a belief, or name the surrogate uncertainty statistic that would carry the mechanism to agents - LLM-based ones in particular - that do not.
27. **Pin the artefact.** Add a commit hash or archival DOI to the Code Availability statement, together with language and library versions; state whether target-placement seeds are shared across protocols (which bears on both statistical power and the independence assumption behind the BH correction in Section 4.6); and extend the seed provision to the roughly 50,000 additional episodes behind Table 5, Table B.16, and Table C.17, which currently fall outside the "all 108 experimental conditions" guarantee.
28. **Minor corrections.** Section 5.1 asserts that "all three communication protocols improve over the no-communication baseline (C0) in both alignment and success rate", but Table 3 itself shows C1 at $k=2$ with success 0.388 against C0's 0.404 - lower, not higher. The abstract's claim of the "lowest MEH rate across all coordination thresholds" should read "lowest among communicating protocols", since

C0 records 0.000 throughout Tables A.13-A.15 against C3's 0.062, 0.093, and 0.168 (Table 4). Section 5.3 quotes only C3's $k=2$ MEH of 0.062; the $k=4$ figure of 0.168 in the same Table 4 is nearly three times higher, and the degradation with coordination threshold deserves a sentence. Table 6 and Table B.16 are identical, and one should be removed. Reference [1] gives the year as 2002, but the volume and pages (14(2):220-240) correspond to the 1998 ALLIANCE paper. Finally, several sentences are comma-spliced where a semicolon or full stop is needed – for example, "more confident senders contribute more to the update than uncertain ones" in Section 3.2.1 and "lower-entropy senders produce a larger boost" in Section 3.3.

5. Recommendation

Major revisions.

I want to be clear about why this is not a rejection. The question the paper asks is a good one and is not well covered in the literature: the field has studied overconfidence in individual belief updates far more carefully than it has studied what happens to a *team's* belief diversity under frequent fusion, and framing that as a distinct, named failure mode is a useful contribution. The simulator appears sound – the factorial multiplies out correctly at 108 conditions, the coverage arithmetic in Section 5.5 checks out, the byte accounting in Tables A.13-A.15 is internally exact, and the C0 success rates in Table 3 sit where a coverage argument says they should. The C2-versus-C3 comparison is a genuinely clean single-factor manipulation, the 98% message reduction is robust in every cell I checked, and the code is public. The limitations section is unusually candid.

But three things must be repaired before the central claims can be accepted. First, the headline metric is currently near-degenerate: MEH equals one minus the success rate exactly in all eight C1 and C2 rows of Table B.16 and throughout Tables A.13-A.15, so the paper's principal novelty must be shown to carry information the success rate does not (items 2 and 3). Second, the mechanism is under-determined by the design – with no decorrelating-fusion arm, the study cannot distinguish frequency from naive fusion as the cause (item 1), and Figure 4's own finding that failed episodes sit at the uniform prior argues the observed phenomenon may be coordinated indecision rather than coordinated error. Third, the definitional and numerical inconsistencies are too numerous to leave standing: two values of epsilon (Eq. 10 versus Figure 3 and Tables A.13-A.15), two MEH denominators across Tables 4, 5, A.13-A.15, B.16, and C.17, a Table 3 that reconciles with neither the appendix means nor the zero-loss cells, an "over 80x" claim that is 35x for C1 and 77.7x when recomputed from the appendix, a robustness interval contradicted

by Table 5, and a scaling section whose episode counts and budget-scaling description contradict their own captions. None of these is individually fatal, and several are plainly typographical, but together they mean a careful reader cannot at present verify the paper's arithmetic against itself.

Items 1, 2, 3, 5, 7, 10, and 14 are the ones I would consider blocking. The remainder are corrections and additions that would substantially strengthen a paper I would be glad to see published once the mechanism is properly isolated and the numbers reconcile.

Declarations

Potential competing interests: No potential competing interests to declare.