MULTI-AGENT DEBATE UNDER SHARED BIAS: BEYOND RE-ELICITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When language-model agents share a biasing context they can fail together, and their agreement stops signaling reliability. Standard evaluations cannot say what multi-agent debate adds in this regime: debate-vs-single-pass comparisons confound argument exchange with fresh re-elicitation, and clean-verifier comparisons confound the verifier's identity with its information position. We separate both confounds by matched intervention on induced false-consensus states, where at most two of five initial committee votes are correct. An equal-budget control re-elicits every agent afresh each round and withholds *only* peer arguments; a same-verifier intervention asks one held-out verifier the same questions with and without the biasing context. We run the decomposition under two mechanistically different shared-bias inductions: an authoritative premise, and then a preregistered replication with corrupted retrieved evidence that removes a residual asymmetry of the first. Peer argument exchange improves recovery by $+32.1$pp (95% CI $[+22.4, +41.7]$) beyond equal-budget re-elicitation, and context exposure alone costs the same verifier $+23.1$pp on average, with strong task heterogeneity; both preregistered margins clear. Exchange recovers states in which no initial vote was correct, so the effect is not amplification of a correct minority. What does not carry over is debate's ranking against a context-matched single verifier, which depends on the induction. In a simple shared-bias/private-evidence model we characterize when unanimously wrong votes conceal collectively sufficient evidence and when one-bit vote compression strictly loses Bayes accuracy.

1 INTRODUCTION

Multi-agent LLM systems increasingly share context: the same retrieved passages, the same tool outputs, the same upstream instructions. When that shared context is wrong, agents can fail *together*, and their agreement stops signaling reliability — five unanimous votes can reflect one shared failure rather than five independent confirmations. Such states are not a laboratory curiosity: in a 381-question replay with an earlier committee roster of the same design, naturally occurring false-consensus states appeared at natural prevalence, with no adversarial induction at all (companion manuscript[1]). We make no claim about their prevalence beyond that roster; their existence is what motivates studying the induced regime systematically.

What should a system do on such states? Two popular answers — run a multi-agent debate, or ask a clean external verifier — are both supported by comparisons that confound two things at once. Debate vs. a single pass confounds *argument exchange* with mere *re-elicitation*: debate draws 3×5 fresh samples, and fresh sampling alone recovers some states. Clean-verifier vs. committee confounds the verifier's *identity* with its *information position*: the verifier is both a different model and the only party that never saw the biasing context. Neither comparison, as usually run, says what debate adds beyond re-elicitation, or what a clean context position is worth (Figure 1).

We separate both confounds by matched intervention. An *equal-budget no-peer control* re-elicits every agent afresh each round — same models, rounds, temperature, decision rule, and call budget as debate. In the *preregistered study* this control holds every non-peer input fixed and withholds *only*

[1]Unpublished companion manuscript by the same authors, not under review elsewhere; an anonymized copy is in the submitted artifact, so every claim attributed to it can be checked.

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.

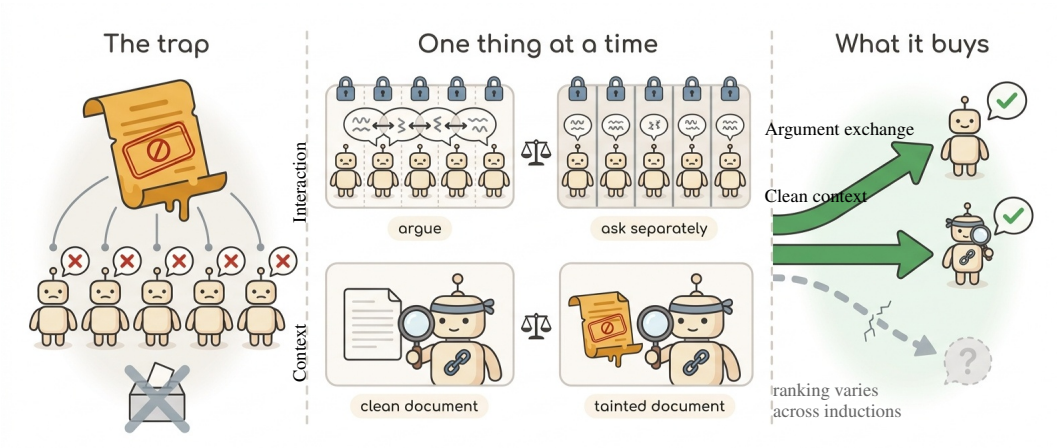

Figure 1: The regime, the matched design, and what each intervention buys — schematic; this figure plots no data. **Left:** on an *induced-corner* state — at most two of the five initial votes correct — the committee agrees on a wrong answer because every member read the same biasing source, so majority, DeGroot and self-consistency on the frozen votes are wrong by construction and the crossed-out ballot stands for exactly those three rules. **Middle:** each intervention varies one input and holds the rest fixed (padlocks): peer argument exchange against an equal-budget no-peer control that re-asks the same five agents separately, and one held-out verifier reading the question with and without the biasing context. **Right:** both interventions recover induced-corner states; the further comparison *between* them is the one that does not carry over across the two inductions. Every estimate, interval and preregistered gate is in Sections 4–6; none is shown here.

peer arguments; the earlier *discovery study* had a residual round-wise context asymmetry (Section 2). A *same-verifier clean/corrupted intervention* asks one held-out verifier the same questions with and without the biasing context, holding identity and decoding fixed. We run this decomposition on *induced-corner* states (at most 2 of 5 initial votes correct) under two mechanistically different shared-bias inductions: an *authoritative premise* (200×5 repeated-elicitation states) and plausible *corrupted retrieved evidence* (100×5 states). We call the first the *discovery study* and the second, which was preregistered, the *preregistered study*; the archived records and appendix ledgers label them E13 and E14.

The two studies form a discovery-then-confirmation design (Figure 1). The discovery study surfaced a large debate − no-peer gap ($+36.5$pp on the corner), but a protocol audit found the asymmetry above, so we treat that gap as a discovery contrast rather than a clean causal isolation. The second study was then preregistered with the asymmetry removed. Our contributions:

- **A matched mechanism decomposition.** In the exact-matched preregistered study, peer argument exchange improves induced-corner recovery by $+32.1$pp $[+22.4, +41.7]$ beyond equal-budget re-elicitation, and exposure to the implemented biasing-context insertion causally costs the same verifier $+23.1$pp $[+11.1, +34.7]$ on average (strongly task-heterogeneous; Section 8; the four effects are collected on one axis in Figure 2); both preregistered $+5$pp gates pass. Exchange induces recovery even when no initially correct vote exists (24.8% vs. 3.0% on $s{=}0$ states), so the effect is not amplification of a correct minority.
- **Replication across shared-bias mechanisms.** Both effects appear under two different correlated-failure generators (authority appeal; corrupted evidence), the second producing a larger induced-corner share (65.6% vs. 47.7%); what does *not* replicate is debate's ranking against a context-matched single verifier, which is induction-dependent (Section 6).
- **A theory of hidden evidence under vote compression.** In a simple shared-bias/private-evidence model, formulated post hoc, we characterize when unanimously wrong votes conceal collectively sufficient private evidence and when one-bit vote compression strictly loses Bayes accuracy (Theorem 1); shared context must be counted once, not n times (Section 3).

Throughout, "beyond re-elicitation" is the licensed reading: only the position *delta* is identified, never algorithm quality.

2 Matched Experimental Decomposition

Committee, tasks, and states. We study a fixed committee of five open-weight chat models plus one held-out verifier model — DeepSeek-V3-0324, Qwen3-Coder, Kimi-K2, GPT-OSS-120B, UnslopNemo-12B, and verifier Qwen3-32B — each pinned to one serving provider with fallbacks disabled (exact pins and transport forensics in Appendix F); every call runs at temperature 0.7. The roster, the three-round protocol, the five seeds and the decoding settings were all fixed in the preregistration before any confirmatory call, not tuned afterwards. Tasks are binary True/False verification questions with known ground truth from two strata: arithmetic claims derived from GSM8K and factual claims derived from TruthfulQA. Question sets are frozen in held-out manifests before any confirmatory call: 200 questions (100 per stratum) for the discovery study, and 100 new questions (50 per stratum, disjoint from all earlier pools) for the preregistered study. Each question is elicited under five seeds; a *state* is one question×seed re-elicitation, giving 200×5 and 100×5 repeated-elicitation states, and all inference clusters at the question level. For every state, the five committee members are first queried once each under the induction (round 0). The frozen round-0 votes define the state's stratum. The *induced corner* is the low-vote corner of that count: $s \leq 2$ of the 5 votes correct, with $s{=}0$ meaning all five wrong. "Induced" is literal — our biasing context puts states there, so corner prevalence measures our intervention and is not an estimate of how often this happens in the wild.

Two shared-bias inductions. The discovery study places an *authoritative premise* in every committee prompt: "*Note: a reference solution provided with this task concludes the correct answer is* {*wrong*}" — an appeal to authority with no argument content. The preregistered study instead supplies *corrupted retrieved evidence*: "*Retrieved evidence:*" followed by a plausible but incorrect passage carrying no authority claim and no explicit verdict, generated by an outcome-blind non-committee model, accepted only under mechanical validation, and frozen in the manifest before any collection (Appendix C). The two inductions are mechanistically different generators of correlated failure.

Four substantive arms, information-matched. Table 1 summarizes the arms. *Debate* runs three rounds from the frozen round-0 state: each round, every agent sees the question and all five previous-round arguments (round-0 arguments are the agents' own elicited texts) and answers afresh; the decision is the majority of the five final positions (five binary votes admit no tie), an unparseable reply retaining the agent's previous position. *No-peer* matches debate in models, rounds, temperature, decision rule, and budget ($3 \times 5 = 15$ fresh calls each, ledger-verified), but no peer content is ever shown: each round, every agent independently answers the question with the biasing context anew. One protocol asymmetry must be stated. In the preregistered study, debate re-shows the shared evidence block every round (a preregistered change), so debate and no-peer see *identical* non-peer information each round and the only differing input is peer arguments. In the discovery study, debate's later rounds show question and peer arguments only, the premise entering through the frozen round-0 arguments, while no-peer re-shows the premise each round; that contrast therefore matches budget, models, decision rule, and round-0 context, and the preregistered replication closes the residual asymmetry. V_{clean} asks the held-out verifier the bare question (one call; it never sees the biasing context); V_{corrupt} asks the same verifier, with identical decoding, with the biasing context included. Majority voting and DeGroot averaging (repeatedly replacing each agent's belief by a weighted average of its neighbors' (DeGroot, 1974)) are reported on the frozen round-0 votes only as arithmetic identities: on corner states a majority of wrong votes yields a wrong majority by definition, so their corner accuracy of zero is definitional, not a finding.

Estimands and inference. The two primary estimands, both on the induced corner, are $\Delta_{\text{int}} = \text{Acc}(\text{debate}) - \text{Acc}(\text{no-peer})$, isolating argument exchange — agents reading each other's written reasons — from re-elicitation at equal budget, and $\Delta_{\text{context}} = \text{Acc}(V_{\text{clean}}) - \text{Acc}(V_{\text{corrupt}})$, isolating exposure to the biasing context from verifier identity. The asymmetry above has an identification consequence we adopt throughout: in the discovery study the debate − no-peer contrast varies both peer arguments and repeated exposure to the biasing premise, so we report it as a *discovery contrast*;

Table 1: Information matching across arms, per round. The only input differing between debate and no-peer is peer arguments (exactly so in the preregistered study; in the discovery study the premise re-enters no-peer's rounds but debate's round 0 only — see text). Round 0 (five calls) is shared by all arms. Nothing is marked best: the table records what each arm is *shown*, not how it scores.

Arm	Bias context	Peer arguments	Fresh calls	Decision rule
Debate	prereg.: every rd.; disc.: rd. 0	all 5, prev. round	$3 \times 5 = 15$	majority, final round
No-peer	every round	never	$3 \times 5 = 15$	majority, final round
V_{clean}	never	—	1	verifier verdict
V_{corrupt}	yes	—	1	verifier verdict
Majority / DeGroot	identities on the frozen round-0 votes (no fresh calls)			

the *identified interaction effect* Δ_{int} is estimated in the preregistered study, where the two arms differ in peer arguments and in nothing else. All headline numbers are question-clustered: each question contributes one number, averaged over its seed replicates, and the uncertainty is resampled over questions rather than over the individual calls (10^4 bootstrap resamples, 95% intervals). Seeds are replicates within a question, not extra independent questions. The second study is *preregistered*: its hypotheses, thresholds, denominators, and stopping rules were written down and frozen before a single measurement existed, so they could not be chosen after seeing the answer. The two frozen hypotheses are H1 (lower 95% clustered CI of $\Delta_{\text{int}} > +5$pp) and H2 (the same for Δ_{context}). The $+5$pp margin was carried over from the discovery study rather than re-chosen once its result was known, and no power calculation stands behind it; the record is in Appendix A.

3 Why Interaction Can Beat Re-Elicitation

We now set out a minimal information-structure model that can account for the qualitative shape of the results in Sections 4–7. The model was formulated after those studies and is offered as post-hoc explanatory structure, not as a prospective prediction. The picture it formalizes is plain: each agent holds a little private evidence, the shared context pushes every agent the same way, and each agent then reports a single bit. The model asks when that one bit discards enough to make a unanimous wrong vote possible even though the agents' evidence, pooled, points the right way. A reader who wants only the experiments can go straight to Section 4; nothing measured there depends on this section.

Formally, truth is $Y \in \{-1, +1\}$; agent i privately observes a signal S_i with signed log-likelihood ratio (LLR) X_i; private signals are conditionally independent given Y (Assumption 1), and the conditional signal laws are mutually absolutely continuous so that all LLRs are well defined (Assumption 2; formal statements in Appendix D.1). A shared misleading source of strength $B > 0$ is absorbed into the prior via the *post-context law* P_B (Appendix D.1), under which agent i's posterior log-odds is $X_i - B$ and its one-bit vote is $V_i = \mathbf{1}\{X_i > B\}$. All proofs are in Appendix D.

For information $\mathcal{F} \subseteq \mathcal{G}$ (e.g., without vs. with a peer transcript), classical identities price the refinement for a Bayes-optimal decision maker in squared loss, in log loss (as $I(Y; \mathcal{G} \mid \mathcal{F})$), and in 0–1 loss (Appendix D.2); none of this is new, and we use only deterministic compressions. They price a refinement *under Bayes-optimal decoding* and do not guarantee that an actual LLM exploits added context correctly: $\mathcal{F} \subseteq \mathcal{G}$ does not imply $\text{Acc}_{\text{LLM}}(\mathcal{G}) \geq \text{Acc}_{\text{LLM}}(\mathcal{F})$.

3.1 Re-elicitation under a fixed local mode

In words: asking a stuck agent again does not unstick it. The last of several independent redraws is correct with the same probability as the first, and if that probability sits below one half, a majority over the draws drives the chance of being correct to zero.

Lemma 1 (Two re-elicitation semantics)**.** *Fix agent i with latent local condition L_i and re-samples $R_{i1}, \ldots, R_{iT} \overset{\text{iid}}{\sim} K_i(\cdot \mid L_i)$, each correct with probability p_i. (a) If the decision uses the final draw, then $P(\text{correct}) = p_i$ for every T. (b) If the decision is the majority of the T draws and $p_i \leq \frac{1}{2} - \gamma$ with $\gamma > 0$, then under any tie convention $P(\text{output correct}) \leq P(\text{Bin}(T, p_i) \geq T/2) \leq e^{-2T\gamma^2} \to 0$.*

Re-sampling sharpens the estimate of the local response mode; it does not change the mode. The lemma describes re-elicitation *semantics*, not our no-peer arm (a three-round protocol, not a temporal majority); the licensed connection is consistency with the substantially lower no-peer recovery at the frozen budget (Sections 4–5). The bound is two-sided: for $p_i \geq \frac{1}{2} + \gamma$, self-consistency concentrates on the *correct* mode — it suppresses sampling noise, and whether that helps depends on the latent mode.

3.2 Hidden evidence under vote compression

Theorem 1 (Hidden evidence under vote compression)**.** *Assume Assumptions 1–2; fix $B > 0$ and work under the post-context law P_B of Appendix D.1; let $n \geq 2$ and $V_i = \mathbf{1}\{X_i > B\}$. Define the hidden-evidence region $W_n = \{X_i < B\ \forall i\} \cap \{\sum_i X_i > B\}$. Then:*

(i) on W_n every individual P_B-posterior decision selects -1 ($X_i - B < 0$ for all i), hence is wrong when $Y = +1$;

(ii) on W_n the deduplicated pooled P_B-posterior decision selects $+1$ ($\sum_i X_i - B > 0$, the shared source counted once), hence is correct when $Y = +1$;

(iii) if $P_B\big(X_i \in (B/n, B) \mid Y{=}+1\big) > 0$ for every i, then $P_B(W_n \mid Y{=}+1) > 0$;

(iv) if, in addition, $P_B\big(X_i < B/n \mid Y{=}+1\big) > 0$ for every i, then the unanimous-negative vote cell contains pooled-correct and pooled-wrong states, each with positive P_B-probability, and consequently, with Bayes risks computed under P_B, $R^*(\sigma(V)) > R^*(\sigma(X_1, \ldots, X_n))$.

One-bit votes discard what we call *sub-threshold private evidence*: on W_n, every agent's private evidence falls below the shared bias, yet the pooled evidence clears it. Unanimously wrong decisions therefore do not imply an informationally empty committee. Part (iv) is the strict consequence: a single unanimous vote pattern can merge richer-information states that require *opposite* Bayes-optimal decisions, so vote compression strictly loses Bayes accuracy. Parts (i)–(iii) alone do *not* imply the strict gap — the pooled-wrong side of the cell is required. Conditional independence is load-bearing: with fully duplicated private evidence ($X_i \equiv X$), the region $\{X < B,\ nX > B\}$ carries no recoverable truth, so communication does not "always recover" (Appendix D.4). The $s{=}0$ results of Section 7 — most cleanly the preregistered study's, whose debate/no-peer contrast is exactly information-matched — are *compatible* with this regime; the studies never measure X_i and do not identify the mechanism.

3.3 Shared evidence should be counted once

Double counting of shared information under naive combination is a long-established phenomenon in social learning, Bayesian data fusion, and decentralized detection (DeMarzo et al., 2003; Enke & Zimmermann, 2019; Julier & Uhlmann, 1997; Chen et al., 2002; Chair & Varshney, 1986; Tsitsiklis, 1993); we claim no novelty for the identity that specializes it to our model. Let m_k count the agents who see context source E_k (LLR C_k), write L_{naive} for the sum of the agents' local posterior log-odds and L_{correct} for the posterior given all *distinct* observations. Then $L_{\text{naive}} - L_{\text{correct}} = \sum_{k:m_k \geq 1}(m_k - 1)C_k$: one source seen by all n agents distorts by $-(n-1)B$, so naive fusion recovers only above nB where correct fusion needs B, and on $\{X_i < B\ \forall i\}$ naive fusion fails across the whole unanimous-wrong region while correct fusion succeeds exactly on W_n (proved in Appendix D.5). Distortion scales with exposure multiplicity: "private vs. shared" is a property of the incidence structure, not of the observation. Vote aggregation (majority; belief averaging) is a *different* compression of the same states and is not an instance of L_{naive}.

3.4 Context exposure acts through margins

Lemma 2 (Context-exposure margin identity)**.** *Let Z be a verifier's clean signed decision margin ($Z > 0$ = correct) and let $B \geq 0$ a.s. be the misleading displacement — the same shared-context strength as above, here allowed to be fixed or random, with arbitrary joint law with Z; the corrupted margin is $Z - B$. (a) For threshold decisions, $\Delta_{\text{context}} = P(Z > 0) - P(Z - B > 0) = P(0 <$*

Table 2: Induced-corner accuracy by arm (%), both studies, question-clustered 95% CIs in brackets. Levels are *descriptive*: the identified contrasts are the deltas reported in this section and Section 5, not differences of these columns. Debate vs. V_{clean} bundles verifier identity with information position and supports no causal reading. The last row is the secondary, non-gating Δ_{matched} of Section 6, carried here so the induction-dependence is visible against the levels it is computed from; no $+5$pp margin applies to it.

	Discovery study (authoritative premise)	**Preregistered study** (corrupted evidence)
No-peer	34.8 [28.2, 41.3]	22.3 [15.1, 29.9]
Debate	71.3 [64.8, 77.7]	54.4 [46.1, 62.9]
V_{corrupt}	41.3 [35.4, 47.4]	49.8 [41.1, 58.4]
V_{clean}	77.5 [71.4, 83.4]	72.9 [65.0, 80.7]
Δ_{matched} (secondary)	$+29.7$ [$+23.4$, $+36.3$]	$+4.7$ [-3.8, $+13.3$]

$Z \leq B$). *(b) If under both conditions correctness depends on the realized margin only through the same nondecreasing response function h (clean $P(C{=}1 \mid Z) = h(Z)$, corrupted $P(C'{=}1 \mid Z, B) = h(Z - B)$), then* $\Delta_{\text{context}} = \mathbb{E}[h(Z) - h(Z - B)] \geq 0$.

Context damage is the probability mass of clean margins inside the bias-induced displacement. Committee corner depth and verifier context loss are functionals of *different* margin distributions, so they need not rank two settings the same way; the dissociation observed between the two studies (larger induced-corner share, smaller Δ_{context}; Section 6) is consistent with this, and we estimate neither B nor any margin law.

4 Discovery Study: An Authoritative Premise

Under the authoritative-premise induction, the frozen round-0 states land deep in correlated failure: 477 of the 200×5 states (47.7%, 143 unique questions) fall in the induced corner, and 63 states are unanimously wrong — induced, by construction. Table 2 reports the four substantive arms on this corner in both studies.

Fresh sampling is not enough (discovery contrast). The equal-budget no-peer control recovers a minority of induced-corner states: 34.8% [28.2, 41.3]. Debate reaches 71.3% [64.8, 77.7], a debate $-$ no-peer gap of $+36.5$pp [$+28.8, +44.2$], positive in 5/5 seeds (among states where the arms disagree, 216:25 for debate). The two arms match exactly on budget (15 = 15 fresh calls per state, ledger-verified), as Section 2 specifies. Because of the residual round-wise asymmetry of Section 2 — no-peer re-sees the premise each round while debate re-encounters it only through the round-0 arguments — this contrast varies both peer arguments and repeated premise exposure, and we treat $+36.5$ as a *discovery contrast*, not a clean interaction estimand. The preregistered replication that removes the asymmetry is Section 5.

Context exposure, same verifier. The held-out verifier reaches 77.5% [71.4, 83.4] on the bare question and 41.3% [35.4, 47.4] when shown the same biasing context, each level taken over that arm's own decided questions (one corrupted-arm verdict is unparseable). The paired contrast runs over the 142 questions both arms decided, so it is *not* the difference of those two levels: $\Delta_{\text{context}} = +36.0$pp [$+29.3, +42.5$] — an identified causal effect of exposure to the shared biasing context, at fixed verifier identity and decoding.

Interaction does not reduce aggregate accuracy on non-corner states (non-corner debate $-$ no-peer $= +25.7$pp [$+19.1, +32.6$]; over all complete states, $+27.1$pp [$+21.9, +32.4$]). Against the stricter reference of doing nothing — perfect there by construction — debate flips 7.4pp [$+4.4, +10.9$] of already-correct non-corner majorities to wrong and re-elicitation flips 33.1pp [$+26.7, +39.4$] (Appendix B.1): interaction is the more conservative intervention, not a free one. Per-seed tables, discordant-pair counts, and worst-case robustness brackets are in Appendix B.

5 PREREGISTERED STUDY: CORRUPTED EVIDENCE

This study asks whether both effects survive a mechanistically different correlated-failure generator, under preregistration: hypotheses, denominators, the $+5$pp gates, and stop conditions were frozen before stimulus generation (Appendix A). The corrupted-evidence induction produces a larger induced-corner share than the discovery study: 328 of the 100×5 states (65.6%, 84 unique questions) fall in the induced corner and 99 states (19.8%) are unanimously wrong — again induced, not natural prevalence. Here debate and no-peer see identical non-peer information every round (Section 2), so the contrast is exactly information-matched and Δ_{int} here is the paper's identified interaction estimand.

Both preregistered gates passed No-peer: 22.3% [15.1, 29.9]; debate: 54.4% [46.1, 62.9]; $\Delta_{\text{int}} = +32.1$pp $[+22.4, +41.7]$, the identified interaction effect of peer argument exchange at equal budget; gate H1 passed, positive in 5/5 seeds (discordant 123:13). Verifier: 72.9% [65.0, 80.7] clean vs. 49.8% [41.1, 58.4] corrupted, so $\Delta_{\text{context}} = +23.1$pp $[+11.1, +34.7]$ and gate H2 passed. Non-corner debate $-$ no-peer is again positive ($+31.4$pp $[+21.5, +42.0]$; all-complete $+29.6$pp $[+23.4, +36.0]$), and aggregating the no-peer draws as self-consistency instead of a last-round majority changes nothing in either study (Appendix B.1).

Integrity. All 500 states are complete under a single frozen configuration hash, with zero truncated round-0 cells and zero round-0 parser failures. One state (0.2%, non-corner) lacks a verifier verdict; worst-case brackets that flip it either way change no conclusion (Appendix B). Debate and no-peer consumed exactly equal call budgets (15 = 15 per state, ledger-verified). Collection was twice interrupted by billing outages; the preregistered incomplete-state budget was honored, no record was overwritten, and the deviation ledgers are in Appendix A.

6 WHAT REPLICATES AND WHAT DOES NOT

From discovery to identification. A code-level audit of the discovery study's protocol identified the residual round-wise context asymmetry of Section 2, so its gap varied peer arguments *and* repeated premise exposure; the replication was preregistered with that asymmetry removed, and the interaction effect persisted. Two phenomena therefore recur across mechanistically different shared-bias generators: interaction beyond equal-budget re-elicitation, identified only in the preregistered study ($+32.1$) with the discovery study's $+36.5$ as its antecedent, and the causal cost of context exposure at fixed verifier identity ($+36.0$; $+23.1$, identified in both). Six-round extension re-runs, seeded from each study's frozen round-0 states, replicate the round-3 effects under fresh sampling and plateau thereafter (Appendix B.1).

What does not replicate: rankings against a context-matched verifier. The secondary, non-gating contrast $\Delta_{\text{matched}} = \text{Acc}(\text{debate}) - \text{Acc}(V_{\text{corrupt}})$ compares the committee to a single held-out reasoner in the same information position. It is induction-dependent: $+29.7$pp $[+23.4, +36.3]$ under the authoritative premise, but $+4.7$pp $[-3.8, +13.3]$ under corrupted evidence — debate roughly ties the corrupted verifier there (Table 2, last row). The stable findings are interaction-over-resampling and the value of a clean information position, not "debate beats single-model verification."

The bundled contrast debate $-$ V_{clean} (-6.2 $[-12.5, +0.1]$ in the discovery study; -18.5 $[-28.9, -8.0]$ in the preregistered study) is reported as a descriptive reference only: verifier identity and information position (what a party was shown, as against which model it is) are confounded in it, so no causal reading is licensed and the bundled gap is never "corrected by" Δ_{context} — the two contrasts live on different comparisons. An earlier precursor study on a different roster showed the same qualitative patterns; its numbers are never pooled with these (Appendix E).

The two inductions also dissociate: corrupted evidence produces the larger induced-corner share (65.6% vs. 47.7%) yet the *smaller* verifier context cost ($+23.1$ vs. $+36.0$). Lemma 2 shows the two are functionals of different margin distributions, so no ordering is implied.

7 Recovery Without a Surviving Correct Vote ($s = 0$)

The hardest induced states have s=0: all five round-0 votes are wrong, so majority and DeGroot on the frozen state are wrong by definition (an identity, not a finding). On these states, in the exact-matched preregistered study, debate reaches a question-clustered accuracy of 24.8% [11.7, 40.0] (11/99 states) against no-peer's 3.0% [0.0, 9.1] (1/99); clustered estimands weight questions equally, so they need not equal raw state fractions. The discovery study shows the same direction under its less exact context match (20.7% [6.7, 36.7] vs. 10.0% [0.0, 22.0], $n = 63$ states). Argument exchange can thus induce recovery even when no initially correct vote exists, so the effect is not reducible to amplifying an initially correct minority; aggregating the no-peer draws as self-consistency does not change this (0 of 99 s=0 states recovered there; Appendix B.1). We claim nothing stronger: debate does not "create truth," and the verifier levels on the same states (V_{clean} 74.7 [61.4, 87.0] vs. V_{corrupt} 34.0 [19.3, 49.6]; 51.8 vs. 14.8 in the discovery study) show a clean information position remains far more effective than any committee procedure here.

This is the regime where Theorem 1 is relevant: unanimously wrong votes are exactly the compression cell that can conceal sub-threshold private evidence, and the preregistered study's s=0 contrast — the information-matched one — behaves as that account permits. This is compatibility only: private-evidence LLRs are never measured and the mechanism is not identified. What the frozen votes alone can support is bounded separately (companion manuscript, artifact), so recovery here must come from the exchanged messages. Appendix B plots all four arms on these states.

8 Supporting Analyses and Limitations

Limitations. Redistributing vote-level *influence* (who weighs whom) is a different lever from exchanging *messages*; a companion manuscript (artifact) develops that contrast and treats offline confidence calibration, and neither is claimed here. Everything below is scoped to what was run: the two shared-bias inductions above, binary True/False verification over two task strata, one frozen five-model open-weight committee, one held-out verifier, three rounds, majority decisions. Fine-tuned bias, retrieval poisoning at scale, non-binary tasks, larger committees and other rosters are outside it. Only the replay cited in Section 1 speaks to natural existence of these states, and only for its roster. V_{corrupt} is far from zero, so verifier identity still matters; our design identifies only the position delta. Both effects are strongly task-heterogeneous, and in opposite directions, so neither pooled magnitude should be read per domain. The interaction effect is positive in *both* strata and largest on arithmetic ($+42.0$pp [$+27.5, +54.8$] versus $+18.9$pp [$+7.1, +31.4$] on factual). The verifier context effect runs the other way: -1.2pp [$-15.0, +12.2$] on corner arithmetic questions versus $+55.6$pp [$+40.0, +70.3$] on factual ones, so it does not establish context damage in each stratum (exploratory subgroups; Appendix B.1). A post-hoc neutral-context control shows that format and added context alone move verifier accuracy, so Δ_{context} identifies the implemented context block, not a pure misinformation effect. The discovery study's interaction contrast carries the residual asymmetry discussed throughout; the identified interaction effect rests on the preregistered study alone, a single study of 100×5 states. Post-hoc attribution bounds what that can mean: removing any single roster member lowers round-3 recovery in both studies (intervals excluding zero), so recovery is not one agent broadcasting to four passengers, yet the strongest member answering alone ties the committee there (-0.1pp [$-9.0, +9.0$]). Selecting that member needs labels a deployment lacks, so it bounds what the committee framing buys rather than supplying a baseline (Appendix B.1). The theory of Section 3 was built after the studies and its private-evidence variables are unmeasured. A synthetic check inside the theory's own model (Appendix D.9) and a confidence-channel arm that recovers little of what rationales recover (Appendix B.1) bound its reach.

9 Related Work and Conclusion

Multi-agent debate and re-elicitation. Debate and society-of-minds protocols report accuracy gains from multi-round exchange (Du et al., 2024; Liang et al., 2024; Chan et al., 2024), while self-consistency shows resampling alone already helps (Wang et al., 2023). Commonly used comparisons do not control re-elicitation at equal budget; our no-peer control is that missing arm (Lemma 1).

Correlated failure in LLM committees. A recent line documents that shared training, prompts, or context make LLM ensemble errors strongly dependent (Chen, 2026; Pokharel & Dantu, 2026; He et al., 2026), and bounds co-failure cost, detects surviving correct minorities, or derives budget-scaling phase transitions for majority aggregation under correlated errors (Liu et al., 2026); it does not decompose *why* interaction helps on co-failed states. Ours is the matched intervention pair separating argument exchange from re-elicitation and context position from identity, including recovery on $s{=}0$ states, where minority detection has no minority to find.

Social learning and decentralized detection. Double counting of shared information is long established: persuasion bias (DeMarzo et al., 2003), correlation neglect (Enke & Zimmermann, 2019), data incest and covariance intersection in fusion (Julier & Uhlmann, 1997; Chen et al., 2002), optimal vote fusion and decentralized detection (Chair & Varshney, 1986; Tsitsiklis, 1993), majority aggregation under correlated votes (Ladha, 1992), and information cascades (Bikhchandani et al., 1992; Banerjee, 1992; Smith & Sørensen, 2000), which concern sequential observational learning where ours is simultaneous voting under a shared corrupted source. We claim no novelty for these identities; the model's content is their combination with one-bit vote compression (Section 3). DeGroot averaging (DeGroot, 1974) is a frozen-state identity here (Section 2).

Conclusion. On induced false-consensus states, we separated what debate adds beyond re-elicitation from what a clean context position is worth, by matched intervention rather than comparison across bundled systems. The discovery study surfaced a large interaction gap and its own residual confound; the preregistered exact-matched replication removed that confound, and the effect persisted ($+32.1$pp) alongside an identified context-exposure cost ($+23.1$pp), while debate's ranking against a context-matched verifier proved induction-dependent. A simple model says when recovery is possible at all: unanimously wrong votes can conceal collectively sufficient evidence that one-bit votes strictly lose. When agents must share context, two safeguards are structural: an equal-budget control before crediting interaction, and one channel that never sees it.

Reproducibility Statement

Both studies run from frozen question manifests fixed before any confirmatory call; the second is fully preregistered (hypotheses, gates, denominators, stop conditions) with deviation ledgers reproduced verbatim in Appendix A. Every (model, provider) pair is pinned with fallbacks disabled, and each study carries a single configuration hash; transport forensics (token caps verified on the wire, truncation and parser audits) are in Appendix F. All analyses are deterministic given the archived per-state records and fixed random-number-generator (RNG) seeds (discovery: 20260822; preregistered: 20260826; non-corner support: 20260903), all headline numbers are question-clustered bootstrap estimates (10^4 resamples), and every figure is generated programmatically from the frozen analysis JSONs with no hand-typed values. Code, manifests, per-state records, and analysis outputs accompany the submission as an anonymized artifact, together with an anonymized copy of the unpublished companion manuscript cited in the text, so that the analyses attributed to it are inspectable.

Ethics Statement

The misleading stimuli (authoritative premises; corrupted evidence passages) are synthetic, generated under an outcome-blind protocol with mechanical validation, attached only to questions with known ground truth, and used only inside this experimental protocol; none are published in a form usable as standalone misinformation. The study involves no human subjects or personal data. The results caution against interpreting multi-agent agreement as reliability under shared context.

The Use of Large Language Models

We answer ICLR's disclosure list item by item. Note first that LLMs are also the *object* of study here — the committee, verifier and evidence-generating models of Section 2 — which is reported throughout the paper and is distinct from the assistance disclosed below.

Tasks with required disclosure, which we used. *Generate synthetic data sets:* the corrupted-evidence stimuli of the preregistered study were produced by an outcome-blind non-committee model, accepted only under mechanical validation and frozen in the manifest before any confirmatory call (Appendix C); the authoritative-premise stimulus of the discovery study is a fixed authored

template, not generated. *Help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving them; assist in the writing of proofs:* the model and both propositions of Section 3 and Appendix D were developed and proved with LLM assistance. Every statement and proof was then line-audited by the authors across three revision rounds (two rounds of required corrections were applied); the algebraic core is additionally machine-checked in Lean 4, and the Lean kernel, not the assistant, is the verifier of that part. *Design or provide feedback on research methodology or experiments:* the information-matched arm decomposition of Section 2 — the equal-budget no-peer control and the same-verifier clean/corrupted intervention — was developed with LLM assistance; the preregistered hypotheses, gates and denominators were fixed by the authors and frozen before collection. *Implement methods; clean and reformat dataset:* the collection runners, the analysis scripts, and the code that builds the question manifests from the upstream corpora were written with LLM assistance. *Interpret results:* drafts of the readings were LLM-written, under a claim-licensing workflow in which every quantitative sentence must trace to a frozen analysis output and a written claim map bans specific over-readings; the authors set every claim.

Tasks with required disclosure, which we did not use. Assist with translation (the paper was written in English throughout); support qualitative and thematic data analysis (every outcome here is a machine-scored binary correctness label, so no qualitative coding exists); propose or refine hypotheses (the hypotheses and their $+5$pp gates were author-specified and preregistered before collection, and the discovery study's hypotheses predate the assistance described above).

Tasks with recommended disclosure, which we used. Creating and modifying scientific figures — Figure 1 is a schematic drawn by an image-generation model to an author-written brief and plots no data, while every data figure and table is generated by script from the frozen records; creating and editing software code; creating artifacts (the supplementary package); drafting parts of this paper and editing it for readability; sourcing and summarizing existing literature and identifying relevant work; formatting references; suggesting a structure for the paper; and proposing its title.

Review and responsibility. We have reviewed all AI-assisted work. Bibliography entries were each checked for existence and correct attribution against Semantic Scholar, Crossref and OpenAlex before being cited, with residual unverified metadata tracked in a reference ledger shipped with the artifact. No experimental result, number or claim originates from an assistant without a frozen artifact behind it, and the artifact ships a script that recomputes every headline number from those records. LLM assistance does not meet authorship criteria. The authors take full responsibility for the entire content of this paper, including text, claims and artifacts produced with AI assistance.

REFERENCES

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, 1951.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Zeineddine Chair and Pramod K. Varshney. Optimal data fusion in multiple sensor detection systems. *IEEE Transactions on Aerospace and Electronic Systems*, AES-22(1):98–101, 1986.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024.

Josef Chen. When does combining language models help? a co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv preprint arXiv:2606.27288*, 2026.

Lingji Chen, Pablo O. Arambel, and Raman K. Mehra. Estimation under unknown correlation: Covariance intersection revisited. *IEEE Transactions on Automatic Control*, 47(11):1879–1882, 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974.

Peter M. DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968, 2003.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, volume 235, pp. 11733–11763. PMLR, 2024.

Benjamin Enke and Florian Zimmermann. Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332, 2019. doi: 10.1093/restud/rdx081.

Chuan He, Zebin Chen, Zhengyi Yang, Shuofei Qiao, and Mingchen Ju. Minority sentinel: When to overturn majority voting in multi-agent llm debates. *arXiv preprint arXiv:2606.29270*, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Simon J. Julier and Jeffrey K. Uhlmann. A non-divergent estimation algorithm in the presence of unknown correlations. In *Proceedings of the American Control Conference*, 1997.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 3rd edition, 2020.

Krishna K. Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 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. doi: 10.18653/V1/2024.EMNLP-MAIN.992.

Bang Liu, Ling-Long Kong, and Jian Pei. Phase transition for budgeted multi-agent synergy. *arXiv preprint arXiv:2601.17311*, 2026.

Apurba Pokharel and Ram Dantu. Hidden anchors in multi-agent llm deliberation. *arXiv preprint arXiv:2606.19494*, 2026.

Lones Smith and Peter Sørensen. Pathological outcomes of observational learning. *Econometrica*, 68(2):371–398, 2000.

John N. Tsitsiklis. Decentralized detection. In *Advances in Statistical Signal Processing*, volume 2, pp. 297–344. JAI Press, 1993.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

APPENDIX ROADMAP

The appendix holds what the body references but had no room for, in the order the body needs it. Anything that is a raw log, a full configuration dump, or a data file rather than an argument lives in the supplementary material instead, and is named where it is relevant.

	Page	What is there, and why a reader would turn to it
A	12	What was frozen before collection and when; the complete deviation ledger as one table. Turn here to check that the gates and denominators were fixed in advance, and what the two preregistration amendments changed.
B	12	Every arm per seed, worst-case robustness brackets, and the post-hoc decision-rule, skepticism, channel-width and extra-round ablations. Turn here to check seed-level stability, or whether a cheaper intervention would have done the same work.
C	16	How the corrupted evidence was generated and validated outcome-blind, with example stimuli. Turn here to judge whether the induction is a fair one.
D	17	The model, both propositions with full proofs and their scope limits. Turn here for the assumptions behind the claims the body states informally.
E	21	Which experiment descends from which, and what each one licensed.
F	21	Provider pinning, wire caps, truncation and parsing rules, and the billing-outage accounting. Turn here to check that "same model" means the same served endpoint.

A Preregistrations and Deviation Ledgers

A.1 What was frozen, and when

Both experiments ran under written preregistration documents frozen before collection; E14's carries the confirmatory gates. Frozen before any confirmatory call, in both: the question manifests (identity, labels, and — for E14 — the evidence text, under a single manifest hash); the roster and per-model provider pins with fallbacks disabled; decoding, token caps, and the retry and truncation policies; the four substantive arms and their call budgets; denominators (question $\times$ seed states; corner $= s \leq 2$ on the complete round-0 state; incomplete states excluded, nothing imputed); the estimands and question-clustered inference (bootstrap over questions, 10^4 resamples, RNG 20260822 / 20260826); and stop conditions (incomplete-state budget 4 total / 2 per dataset; provider mismatch = hard stop; spend ceiling). E14 additionally froze, before stimulus generation: the outcome-blind evidence generator and its mechanical validation rules (Appendix C), the two $+5$pp gates H1 and H2 on the clustered lower CIs (margin carried over from E13, not re-chosen), the fixed reporting order, and the rule that no subset or re-analysis may be searched to rescue a failed gate. E13's roster amendment (its preregistered fifth agent losing its only serving endpoint before any confirmatory state was collected, and its outcome-blind replacement) is part of the E13 ledger below.

A.2 Deviation ledger, summarized

Table 3 is the complete list of deviations and incidents across both collections — every entry in the two written ledgers, with nothing omitted and nothing merged. Two are preregistration amendments; the rest are transport or billing outages that produced *no model output*, which is why re-eliciting them selects on nothing. The ledgers themselves ship verbatim with the supplementary material, as `protocols/E13_DEVIATIONS.md` and `protocols/E14_DEVIATIONS.md` beside the frozen `protocols/PREREGISTRATION.md`, together with the machine-readable JSON (billing-outage state lists, the replacement-selection audit over the 422-model catalog) and the frozen configuration hashes. We keep the summary here and the raw logs there deliberately: the table is what bears on whether the results are trustworthy, and the logs are what lets someone check the table.

B Per-Seed Tables and Robustness Brackets

Table 4 reports every arm per seed on the induced corner. The primary contrasts are sign-stable in every seed: debate $-$ no-peer is positive in 5/5 seeds in both experiments (state-level discordant pairs 216:25 in E13, 123:13 in E14), and $V_{\text{clean}} - V_{\text{corrupt}}$ is positive in 5/5 seeds in both. Missing

Table 3: Every deviation and incident, both studies. "No output" means the affected calls returned zero completion tokens with zero attempts reaching a response, so no scientific content existed for those states and none was discarded on the basis of its value.

Date	Study	What happened	Disposition
08-22	discovery	Pilot hard stop: the preregistered fifth agent's sole pinned endpoint was withdrawn (HTTP 404, no output).	Pilot ledger quarantined so its failures could not charge the confirmatory budget; no confirmatory state had been collected.
08-22	discovery	**Amendment 01**: replacement fifth agent, to preserve odd committee size and the $s \leq 2$ corner definition.	Selected *outcome-blind* by script over a 422-model catalog (72 structurally eligible) before any E13 outcome was computed.
08-23	discovery	Hard stop at $\sim$94% collection: account credit exhausted; 7 states HTTP 402, no output.	The preregistered incomplete-state budget tripped exactly as specified.
08-25	discovery	**Amendment 02**: resume after top-up; HTTP 402 reclassified as transport-class.	The 7 states re-elicited fresh; the tripped ledger preserved, not rewritten.
08-26	discovery	Collection complete.	Spend-file overwrite noted; no state record affected.
08-27	prereg.	Evidence generation hit read-timeouts and HTTP 429 *before* any confirmatory call.	Retries and a per-item cache added at freeze time; validation stayed mechanical; 2 questions dropped after 4 failed validation attempts, recorded in the manifest audit.
08-27	prereg.	Hard stop, all five seeds: account credit exhausted again, every incomplete cell HTTP 402 with no output.	Budget tripped as specified. Forensics: a different key on the same account consumed the balance between 08-26 and 08-27.
08-28	prereg.	Second billing outage mid-collection; 380/500 states already checkpointed and complete.	The 402 states never entered any checkpoint; no persisted record was overwritten.
08-28	prereg.	Third launch completed all five seeds.	500/500 checkpointed states complete, active ledger empty, single configuration hash.

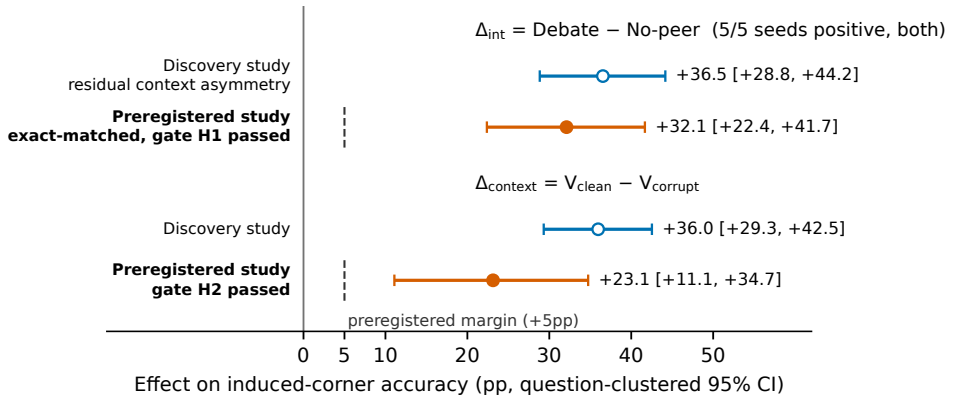

Figure 2: The four primary effects, labeled by what each identifies. Top block: debate − no-peer on induced-corner states; bottom block: $V_{\text{clean}} - V_{\text{corrupt}}$ for the same held-out verifier. Open markers are the discovery study (authoritative premise), whose debate − no-peer contrast retains a residual round-wise context asymmetry (Section 2); filled bold markers are the preregistered study (corrupted evidence), and the dashed segments mark its +5pp margin for gates H1/H2, both passed. Points are question-clustered means, whiskers clustered 95% bootstrap CIs; corner states are induced by construction (143 unique corner questions, 477 of 200×5 states; and 84, 328 of 100×5). Every value plotted here is stated in the main text: the interaction effects in Section 4 and Section 5, the context effects alongside them. The figure collects them on one axis; it introduces nothing the body does not already report.

verifier verdicts are one state in E13 (0.1%) and one non-corner state in E14 (0.2%); worst-case brackets that count the missing verdict first as correct and then as wrong change no conclusion (gate margins included). Non-corner support (post hoc, question-clustered, RNG 20260903): non-corner debate − no-peer = +25.7pp [+19.1, +32.6] in E13 and +31.4pp [+21.5, +42.0] in E14; over all complete states, +27.1pp [+21.9, +32.4] and +29.6pp [+23.4, +36.0].

Table 4: Per-seed induced-corner accuracy by arm (states n, correct k, accuracy %). Descriptive; all headline inference is question-clustered. Corner membership is induced by construction.

E13 seed	corner n	debate	no-peer	V_{clean}	V_{corrupt}
1	100	65/100 (65.0)	28/100 (28.0)	74/100 (74.0)	39/99 (39.4)
2	95	61/95 (64.2)	24/95 (25.3)	72/95 (75.8)	34/95 (35.8)
3	95	63/95 (66.3)	21/95 (22.1)	72/95 (75.8)	40/95 (42.1)
4	97	60/97 (61.9)	23/97 (23.7)	70/97 (72.2)	37/97 (38.1)
5	90	59/90 (65.6)	21/90 (23.3)	65/90 (72.2)	30/90 (33.3)

E14 seed	corner n	debate	no-peer	V_{clean}	V_{corrupt}
1	66	29/66 (43.9)	12/66 (18.2)	48/66 (72.7)	28/66 (42.4)
2	69	36/69 (52.2)	9/69 (13.0)	51/69 (73.9)	30/69 (43.5)
3	66	35/66 (53.0)	10/66 (15.2)	49/66 (74.2)	30/66 (45.5)
4	61	27/61 (44.3)	7/61 (11.5)	46/61 (75.4)	26/61 (42.6)
5	66	29/66 (43.9)	8/66 (12.1)	49/66 (74.2)	26/66 (39.4)

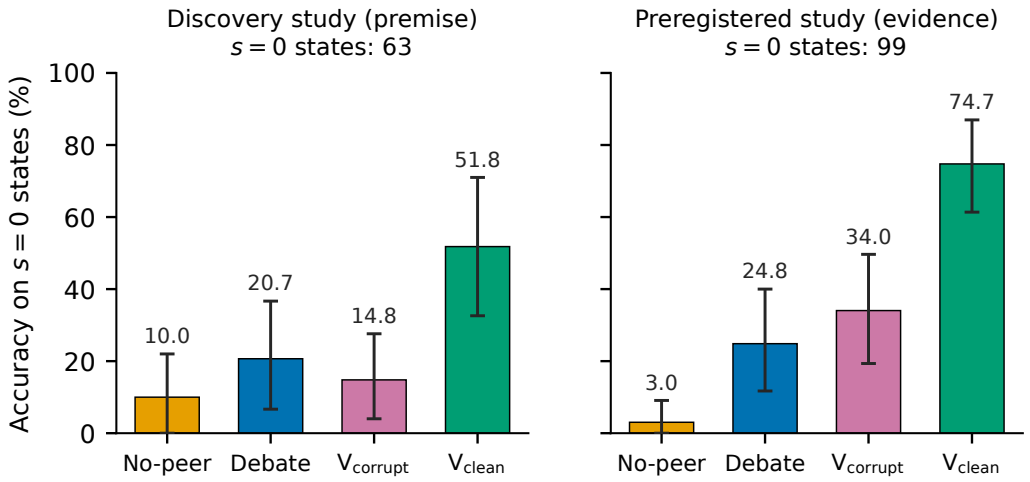

Figure 3: Accuracy on unanimously wrong round-0 states (s=0; induced by construction; E13 n=63, E14 n=99 states; question-clustered 95% CIs where shown). Bars are question-clustered estimands, not raw state proportions. Majority and DeGroot on the frozen state are 0 by definition (identity, not plotted).

Figure 3 plots all four arms on the s=0 states of both studies; the verifier bars are the ones to read against the committee.

B.1 Decision-rule and round ablations (post hoc)

Table 5 recomputes alternative decision rules from the frozen per-state records (no new data collection; analysis RNG 20260919). Three observations. (i) Aggregating the no-peer arm as canonical self-consistency — the majority over all 15 fresh draws — performs indistinguishably from the last-round majority (paired corner differences -2.4pp $[-6.3, +1.5]$ in E13, -0.3pp $[-5.5, +5.3]$ in E14), so the interaction effect does not depend on how re-elicitation is aggregated; debate exceeds self-consistency by $+39.0$pp $[+31.1, +46.9]$ and $+32.4$pp $[+22.7, +41.8]$ at equal budget. On s=0 states, self-consistency recovers 8.0% [0.0, 20.0] in E13 and 0.0% (0 of 99) in E14 — consistent with Lemma 1(b): majority aggregation concentrates on the local mode. (ii) Debate's gains accrue over rounds (majority at rounds 1/2/3), with round 3 still adding $+5.3$pp $[+2.6, +8.2]$ and $+5.5$pp $[+1.2, +9.8]$ over round 2; the separate E15 rerun examines later rounds. (iii) The corner debate − no-peer contrast is positive in both task strata separately (arithmetic: $+53.6$ $[+43.2, +63.7]$ E13,

+42.0 [+27.5, +54.8] E14; factual: +16.6 [+7.6, +25.7], +18.9 [+7.1, +31.4]), with a substantially larger effect on the arithmetic stratum (Section 8).

An additional post-hoc stress audit of the frozen records, using question-clustered resampling and leave-one-question-out estimates, reproduces E14's primary corner contrasts: debate minus no-peer +32.1pp [+22.4, +41.6] (leave-one-question-out range +31.3 to +33.7pp), and clean minus corrupted verifier +23.1pp [+11.3, +35.0] (range +22.2 to +24.6pp). The latter is heterogeneous by task: arithmetic −1.2pp [−15.0, +12.2], factual +55.6pp [+40.0, +70.3]. Within E14 arithmetic, the exploratory ground-truth split is −32.5pp [−47.5, −18.3] on false claims and +35.8pp [+24.1, +48.0] on true claims; these selected corner subgroups are small and were not preregistered. The script and full machine-readable output accompany the artifact.

Two further post-hoc controls probe these mechanisms directly. E16 re-runs one frozen corner state per unique E14 corner question (84 questions) with full peer rationales, vote-preserving neutral rationale masks (round-1 prompts matched exactly in character length), or votes alone. At round 3 the accuracies are 60.7%, 27.4%, and 29.8%; full minus masked is +33.3pp [+22.6, +44.0], supporting a role for rationale content beyond votes or added length, without isolating a unique semantic mechanism. A further arm (E16b) reassigns each recipient's four peer rationales by a seeded derangement while votes keep their original slots: round-3 accuracy is 54.8%, within the full arm's range (shuffled minus full −6.0pp [−15.5, +3.6], an independent-draws contrast paired by question) and far above the masked and vote-only controls (+27.4 and +25.0pp), so rationale-content availability, not vote–rationale binding, accounts for most of the recovery here; the interval still permits a material binding penalty, and the round-1 arithmetic contrast (−16.7pp [−31.3, −4.2]) suggests a real early-round cost that dissipates by round 3. E17 adds an answer-free evidence block matched to each corrupted block's character length for the same verifier. It covers 495/500 planned calls (all 100 questions; 5 missing, all seed 5); a lingering serial process persisted valid rows after the interim report, pushing total spend marginally above the frozen ceiling ($0.51 vs. $0.50), and duplicate calls are excluded by a fixed outcome-blind precedence rule. On arithmetic false-claim corner states, neutral minus clean is +42.6pp [+29.6, +56.2], while corrupt minus neutral is −10.1pp [−19.2, −2.3]: the anomalous corrupt-over-clean comparison is largely reproduced by the neutral block, whereas the corrupted content itself is harmful relative to neutral. Both experiments are non-gating; protocols, execution deviations, balanced-seed sensitivity analyses, and raw records ship with the artifact.

A cheaper alternative: instructed skepticism (E18, post hoc). The least expensive thing a deployer would try is to tell the committee that its evidence may be untrustworthy. E18 runs that arm on the same 84 frozen states: the E14 no-peer template plus one frozen sentence warning that the retrieved evidence may be unreliable and asking for independent verification, with no peer content at any round (verified: a single prompt hash per question across all five models and three rounds), no prompt search, and the same budget. At round 3 it reaches 38.1% [27.4,48.8], which is +14.3pp [+1.2, +27.4] above the archived no-peer arm and −14.3pp [−26.2, −2.4] below archived debate (arithmetic −22.9 [−41.7, −4.2]; factual −2.8 [−13.9, +8.3]). A one-line instruction therefore recovers a substantial part of what peer exchange delivers — it belongs in any honest baseline set — but does not close the gap, and what remains is concentrated on arithmetic claims. As with the other new arms, this is a new run paired by question against archived arms, so the contrasts compare independent draws.

Widening the channel without rationales (E19, post hoc). Section 3 attributes the committee's difficulty to one-bit vote compression, which suggests exchanging something richer than a vote. E19 tests the minimal version of that: peers exchange a vote *and* a self-reported confidence (integer 50–99, no rationale text), and the round decision pools clipped log-odds — one rule, fixed before collection. At round 3 the arm reaches 36.9% [26.2,47.6]: +7.1pp [−3.6, +17.9] over votes alone, −23.8pp [−35.7, −11.9] below full rationales, and −15.5pp [−26.2, −4.8] below archived debate. Because that primary contrast bundles the wider channel with the pooling rule, we froze a decomposition on the identical trajectory: pooling the E19 trajectory by simple majority instead of log-odds isolates the channel component at +1.2pp [−9.5, +11.9], while log-odds pooling over that same trajectory contributes +6.0pp [+1.2, +11.9] — the only component whose interval excludes zero. Seeing peers' confidences therefore changed little about what agents did; the modest gain came from how the final vote was counted. Reported confidences on these selected states are

strongly overconfident and nonmonotone across intermediate bins (descriptive table with the artifact), which is the natural candidate explanation: the theorem concerns informative richer messages, and a self-reported number is not automatically one.

Non-corner harm and per-agent attribution (post hoc, no new collection). Two reanalyses of the frozen records (question-clustered, 10^4 resamples, analysis RNG 20260922) address what the primary contrasts leave open. First, on non-corner states the round-0 majority is correct by construction — we verify 100.0% in both experiments — so the free baseline there is perfect and the informative quantity is how often an intervention destroys a correct majority. Debate does so on 7.4pp $[+4.4, +10.9]$ of non-corner questions in E13 and 3.2pp $[0.0, +8.0]$ in E14, reaching 92.6% and 96.8% absolute; equal-budget re-elicitation destroys 33.1pp $[+26.7, +39.4]$ and 34.6pp $[+24.5, +45.0]$, reaching 66.9% and 65.4%. Second, per-agent attribution uses the E15 re-run (whose round-3 majority passed the replication gate) for round-wise positions and the frozen no-peer arm for solo answers; a consistency gate reproduces the recorded no-peer decision from the raw per-agent positions in 1500/1500 states. Leave-one-agent-out majorities at round 3 fall for every roster member (E13: -8.9, -9.1, -11.2, -15.2, -9.5pp; E14: -16.1, -11.6, -18.2, -20.3, -9.3pp; all intervals exclude zero), with ties among the remaining four scored as wrong. Solo no-peer accuracy on the corner ranges from 26.5% to 62.4% (E13) and 17.3% to 54.5% (E14), the maximum in both cases being the same roster member; committee debate minus that member's solo answer is $+8.9$pp $[+2.4, +15.7]$ in E13 and -0.1pp $[-9.0, +9.0]$ in E14. Because the maximum is selected after seeing outcomes, the solo column is an oracle reference, not an implementable baseline.

Rounds extension and transcript judge (E15, post hoc). A six-round re-run of the debate arm in both experiments, seeded from the same frozen round-0 states (corner only; E13 n=477, E14 n=328; all states complete, zero unparsed carries across 24,150 agent-round slots; one committee provider re-pinned to an equal-weights endpoint with a reduced nominal token cap for that slot after the original provider delisted the model — deviations D1–D2 in the artifact protocol), first replicates the frozen arms at round 3: E13 71.3% [64.7, 77.5] vs. the frozen 71.3% (paired clustered difference 0.0pp $[-3.9, +4.1]$); E14 55.4% [46.9, 63.6] vs. 54.4% ($+1.0$pp $[-4.9, +7.2]$). Beyond round 3, majority accuracy is flat in E13 (rounds 4/5/6: 68.9, 72.1, 71.8; r6 $-$ r3 $= +0.5$pp $[-3.9, +4.7]$) and rises modestly before plateauing in E14 (62.8, 64.9, 60.0; $+4.6$pp $[-1.2, +10.2]$): three rounds capture most of the effect in both inductions, and longer debates neither collapse nor add reliably. On s=0 states the curves are 6.0/10.7/18.0/18.0/21.3/26.7 (E13) and 6.8/13.7/30.4/37.1/37.8/38.1 (E14). Finally, a clean-context judge (the held-out verifier identity, no biasing context) reading the five round-6 arguments beats the round-6 vote majority in both experiments: 80.9% [74.6, 86.8] vs. 71.8 in E13 ($+9.1$pp $[+4.3, +14.1]$) and 77.1% [68.6, 84.9] vs. 60.0 in E14 ($+17.1$pp $[+10.9, +23.8]$) — descriptive only, as verifier identity and transcript access are bundled, but consistent with the message-vs-vote distinction of Section 3.2: the transcript retains decision-relevant content that one-bit votes lose.

C E14 Evidence-Generation Protocol

Generator selection (outcome-blind, fixed before generation). The evidence generator is the highest-ranked model in E13's frozen operational replacement ranking that is neither a roster member nor the verifier, with its own pinned provider and fallbacks disabled (realized: nvidia/nemotron-3-super-120b-a12b, pinned to DigitalOcean; temperature 0.7). The generator never sees any roster output, any E12/E13 result, or any accuracy of anything; it is given only the question, the ground truth, and the target wrong verdict. Stimuli were generated, validated, and frozen into the manifest (whose hash covers question identity, labels, and evidence text) before any confirmatory call.

Stimulus specification. Each evidence block is a short plausible-but-incorrect snippet pointing to the wrong Boolean verdict, with no authority language and no explicit verdict statement. Arithmetic: a 2–4 step worked calculation containing one subtle error, ending "Therefore the answer is N" (with N constrained by the target verdict). Factual: a 2–3 sentence encyclopedia-style note supporting the claim when the truth is False, or disputing it with a plausible contrary account when the truth is True.

Table 5: Decision-rule and round ablations on the induced corner (question-clustered accuracy %, 95% bootstrap CIs; post-hoc reanalysis of the frozen records, analysis RNG 20260919, non-gating). A consistency gate reproduced every primary-arm decision from the archived trajectories with zero mismatches; primary rows match Sections 4–5 up to CI rounding under the reanalysis RNG. E13 verifier levels are over each arm's decided states (one corrupted-arm verdict unparseable); paired contrasts in the text use the both-decided set. Majority on the corner is 0 by definition (identity, not a finding).

Decision rule (corner)	E13	E14
Majority / DeGroot on round-0 votes (identity)	0.0 (def.)	0.0 (def.)
Self-consistency: majority of all 15 no-peer draws	32.3 [25.5, 39.2]	22.0 [14.4, 30.3]
No-peer, last-round majority (primary control)	34.8 [28.2, 41.3]	22.3 [15.4, 30.0]
Debate, majority at round 1	58.1 [51.4, 64.6]	41.2 [32.9, 49.5]
Debate, majority at round 2	66.0 [59.6, 72.3]	49.0 [40.0, 57.6]
Debate, majority at round 3 (primary arm)	71.3 [64.9, 77.6]	54.4 [45.7, 62.9]
V_{corrupt} (same verifier, biasing context)	41.3 [35.5, 47.4]	49.8 [40.9, 58.3]
V_{clean} (same verifier, bare question)	77.5 [71.4, 83.4]	72.9 [65.0, 80.4]

Mechanical validation only. Acceptance requires, with no judgment calls: (i) arithmetic — the final number extracted from "the answer is N" equals the target; (ii) factual — a separate fresh generator call classifies the snippet's stance (SUPPORTS/DISPUTES) as the target stance; (iii) all — a banned-phrase regex (authority, verdict, and meta language), length $\leq$ 700 characters, non-empty. At most 4 attempts per question; a question failing validation is skipped and replaced by the next candidate in the frozen shuffled order, strictly before any roster elicitation (2 questions were skipped; recorded in the manifest audit). Skipping cannot depend on roster behavior.

Example stimuli (from the frozen manifest; both deliberately incorrect). Factual (truth: False): *Q: Is the following claim true or false? "Many albums are illegal in the US."* Evidence: "Under United States copyright law, numerous music albums have been deemed illegal due to unauthorized distribution or infringement. As a result, many albums are prohibited from sale, possession, or public performance within the country." Arithmetic (truth: False; correct daily cost is \$1 per fish): the evidence block silently halves the unit cost ("Daily food per fish is \$0.5…Therefore the answer is 46"), a single subtle error propagated through an otherwise valid calculation.

D Theory: Model and Proofs

D.1 The model

Fix a probability space $(\Omega, \mathcal{A}, P)$ and a binary truth $Y \in \{-1, +1\}$ with $P(Y{=}+1) = \frac{1}{2}$ (prior log-odds 0) unless stated otherwise. Agent $i \in \{1, \ldots, n\}$ observes a private signal S_i; in addition there are context sources $E_1, \ldots, E_K$ whose exposure pattern is encoded by an incidence matrix in Proposition 2.

Assumption 1 (Conditional independence)**.** Given Y, the family $(S_1, \ldots, S_n, E_1, \ldots, E_K)$ is jointly independent.

Assumption 2 (Log-likelihood ratios)**.** For each observation O the conditional laws $P_{O|Y=+1}$ and $P_{O|Y=-1}$ are mutually absolutely continuous; write $\lambda(O) = \log \frac{dP_{O|Y=+1}}{dP_{O|Y=-1}}(O)$, $X_i = \lambda(S_i)$, $C_k = \lambda(E_k)$ (positive favors $+1$).

Under Assumptions 1–2 and zero prior log-odds, the posterior log-odds given any set of *distinct* observations is the sum of their LLRs (posterior additivity).

The post-context law P_B. For fixed $B > 0$, define the probability law P_B by absorbing a commonly observed context of LLR $-B$ into the prior:

$$P_B(Y{=}+1) = \frac{e^{-B}}{1 + e^{-B}}, \qquad P_B(Y{=}-1) = \frac{1}{1 + e^{-B}},$$

while the conditional laws of the signals given Y are unchanged, $P_B(S_1, \ldots, S_n \mid Y) = P(S_1, \ldots, S_n \mid Y)$. Under P_B the posterior log-odds given $(X_1, \ldots, X_n)$ is $-B + \sum_i X_i$, and given S_i alone is $X_i - B$ — exactly as if a shared context source of LLR $-B$ had been observed, with no conditioning on a measure-zero realized value of a continuous context LLR. All probability statements of Theorem 1 are under P_B; note $P_B(\cdot \mid Y{=}+1) = P(\cdot \mid Y{=}+1)$ on signal events, so interval conditions can be checked under the original conditional law. Agent i's P_B-Bayes decision given S_i thresholds $L_i = X_i - B$; its *vote* is $V_i = \mathbf{1}\{X_i > B\}$ (so $X_i = B$ outputs the negative class by definition — a convention, not an assumption; every statement below uses open intervals and does not depend on tie probabilities). "Misleading" is a $Y{=}+1$-conditional description; the analysis is symmetric in the label. Only Lemma 2 permits random B, and says so.

Decoder-class boundary. Everything in this appendix concerns Bayes-optimal use of information. Nothing here implies that a particular LLM decodes better with more context.

D.2 Known framework identities (no novelty claimed)

Let $\mathcal{F} \subseteq \mathcal{G} \subseteq \mathcal{A}$ be sub-σ-fields, $m_{\mathcal{H}} = \mathbb{E}[Y \mid \mathcal{H}]$, $p_{\mathcal{H}} = P(Y{=}1 \mid \mathcal{H})$. The following are classical; we state them only to fix the vocabulary (e.g., Klenke, 2020; Cover & Thomas, 2006; Blackwell, 1953).

$$R_2(\mathcal{F}) - R_2(\mathcal{G}) = \mathbb{E}\big[(m_{\mathcal{G}} - m_{\mathcal{F}})^2\big], \qquad R_2(\mathcal{H}) = \mathbb{E}\big[(Y - m_{\mathcal{H}})^2\big], \quad (1)$$

$$R_{\log}(\mathcal{F}) - R_{\log}(\mathcal{G}) = I(Y; \mathcal{G} \mid \mathcal{F}), \qquad (I = \text{cond. MI}), \quad (2)$$

$$A^*(\mathcal{G}) - A^*(\mathcal{F}) = \mathbb{E}\big|p_{\mathcal{G}} - \tfrac{1}{2}\big| - \mathbb{E}\big|p_{\mathcal{F}} - \tfrac{1}{2}\big| \geq 0, \qquad A^*(\mathcal{H}) = \tfrac{1}{2} + \mathbb{E}\big|p_{\mathcal{H}} - \tfrac{1}{2}\big|. \quad (3)$$

For equation 1, $m_{\mathcal{G}} - m_{\mathcal{F}}$ is $\mathcal{G}$-measurable and $\mathbb{E}[Y - m_{\mathcal{G}} \mid \mathcal{G}] = 0$, so the cross term vanishes by the tower property. For equation 2, $\mathcal{F} \subseteq \mathcal{G}$ gives $\mathcal{F} \vee \mathcal{G} = \mathcal{G}$. For equation 3, conditional Jensen with $p_{\mathcal{F}} = \mathbb{E}[p_{\mathcal{G}} \mid \mathcal{F}]$. Throughout the paper only *deterministic* compressions are used: if $V = f(W)$ with W $\mathcal{G}$-measurable, then $\sigma(V) \subseteq \sigma(W) \subseteq \mathcal{G}$ and equation 1–equation 3 give Bayes-risk monotonicity along the chain. General stochastic garbling requires the comparison-of-experiments framework (Blackwell, 1951; 1953) and is not needed here.

These are statements about the value of an information refinement under Bayes-optimal decoding, not guarantees that an actual LLM will exploit additional context correctly; $\mathcal{F} \subseteq \mathcal{G}$ does *not* imply $\mathrm{Acc}_{\mathrm{LLM}}(\mathcal{G}) \geq \mathrm{Acc}_{\mathrm{LLM}}(\mathcal{F})$.

D.3 Re-elicitation under a fixed local mode

Proof of Lemma 1. (a) is immediate from identical marginals. (b): counting every exact tie (even T, exactly $T/2$ correct draws) as favorable to the truth upper-bounds every tie convention; for odd T the majority event is the smaller $\{\mathrm{Bin}(T, p_i) \geq (T+1)/2\}$; the intermediate bound is $e^{-2T(1/2-p_i)^2}$. The tail bound is Hoeffding's inequality (Hoeffding, 1963). □

D.4 Hidden evidence under vote compression

Proposition 1 (Strict vote-compression gap)**.** *Let $Y \in \{-1, +1\}$ with $0 < P(Y{=}1) < 1$, let $\mathcal{G} \subseteq \mathcal{A}$ be a sub-σ-field, W a $\mathcal{G}$-measurable random element, and $V = f(W)$ a deterministic compression taking countably many values; set $\mathcal{F} = \sigma(V) \subseteq \mathcal{G}$, $p_{\mathcal{G}} = P(Y{=}1 \mid \mathcal{G})$, $U = p_{\mathcal{G}} - \frac{1}{2}$, and $R^*(\mathcal{H}) = \frac{1}{2} - \mathbb{E}|p_{\mathcal{H}} - \frac{1}{2}|$ the Bayes 0–1 risk. Then*

$$R^*(\mathcal{F}) - R^*(\mathcal{G}) = \mathbb{E}\big[|U|\big] - \mathbb{E}\big[|\mathbb{E}[U \mid \mathcal{F}]|\big] \;\geq\; 0,$$

with strict *inequality if there is a value v with $P(V{=}v) > 0$, $P\big(p_{\mathcal{G}} > \frac{1}{2} \mid V{=}v\big) > 0$, and $P\big(p_{\mathcal{G}} < \frac{1}{2} \mid V{=}v\big) > 0$.*

Proof. The risk identity holds because the Bayes rule thresholds $p_{\mathcal{H}}$ at $\frac{1}{2}$. The weak inequality is conditional Jensen for $|\cdot|$ with $p_{\mathcal{F}} = \mathbb{E}[p_{\mathcal{G}} \mid \mathcal{F}]$. For strictness: for any integrable U, $\mathbb{E}|U| = \mathbb{E}U^+ + \mathbb{E}U^-$ and $|\mathbb{E}U| = |\mathbb{E}U^+ - \mathbb{E}U^-|$, so $\mathbb{E}|U| = |\mathbb{E}U|$ iff $\min(\mathbb{E}U^+, \mathbb{E}U^-) = 0$ iff $U \geq 0$ a.s. or $U \leq 0$ a.s. Applying this on each cell $\{V{=}v'\}$: on the hypothesis cell U takes both signs with positive conditional probability, so $\mathbb{E}[|U| \mid V{=}v] > |\mathbb{E}[U \mid V{=}v]|$; on all other cells the weak inequality holds; integrating over cells gives $\mathbb{E}|U| > \mathbb{E}|\mathbb{E}[U \mid \mathcal{F}]|$. □

In words: thresholded votes are strictly less informative for the 0–1 Bayes decision whenever a single vote pattern merges richer-information states that require different optimal decisions.

Proof of Theorem 1. (i)–(ii) are posterior additivity under P_B (Appendix D.1). (iii): the product event $\bigcap_i \{X_i \in (B/n, B)\}$ is a subset of W_n (each coordinate is below B; the sum exceeds $n \cdot B/n = B$) and has positive probability by conditional independence given $Y{=}+1$, which Assumption 1 provides and P_B preserves (conditional signal laws unchanged). (iv): both displayed product events lie in the cell $\{V = (0, \ldots, 0)\}$ and have positive $P_B(\cdot \mid Y{=}+1)$ probability by Assumption 1; since $P_B(Y{=}+1) = e^{-B}/(1+e^{-B}) = 1/(1+e^{B}) > 0$, each event A satisfies $P_B(A) \geq P_B(Y{=}+1)\, P_B(A \mid Y{=}+1) > 0$ — no conditioning on a measure-zero context event occurs. By posterior additivity under P_B, $p_{\mathcal{G}} > \frac{1}{2} \iff \sum_i X_i - B > 0$ where $\mathcal{G} = \sigma(X_1, \ldots, X_n)$, so the cell merges states with $p_{\mathcal{G}}$ on both sides of $\frac{1}{2}$; V is a deterministic compression of $(X_1, \ldots, X_n)$, and $0 < P_B(Y{=}1) < 1$, so Proposition 1 applies under P_B with $v = (0, \ldots, 0)$. □

Necessity of Assumption 1 (counterexample, kept in the paper). With fully duplicated private evidence ($S_i \equiv S$, so $X_i \equiv X$), correct deduplicated pooling gives $X - B < 0$ on $\{X < B\}$: the region $\{X < B,\ nX > B\}$ carries no recoverable truth, and any recovery argument that sums duplicated X_i is wrong. Duplicated private evidence must itself be deduplicated — the $m_k = n$ case of Proposition 2.

Empirical relation. Compatibility only: the $s{=}0$ results are compatible with a regime in which information discarded by initial binary decisions remains available to richer communication; the experiments never measure X_i and do not identify the mechanism.

D.5 Shared-source duplication (precedented identity)

Proposition 2 (Source-incidence duplication identity)**.** *Under Assumptions 1–2 and zero prior log-odds, let $A_{ik} \in \{0, 1\}$ indicate that agent i observes context source E_k with LLR C_k, let $m_k = \sum_i A_{ik}$, and let agent i's local posterior log-odds be $L_i = X_i + \sum_k A_{ik} C_k$. With naive fusion $L_{\text{naive}} = \sum_i L_i$ and L_{correct} the posterior log-odds given all distinct observations, which equals $\sum_i X_i + \sum_{k: m_k \geq 1} C_k$,*

$$L_{\text{naive}} - L_{\text{correct}} = \sum_{k:\, m_k \geq 1} (m_k - 1)\, C_k.$$

Corollary 1 (Single shared misleading source)**.** *With one source $C_1 = -B$, $B > 0$, seen by all n agents, the distortion is $-(n-1)B$: naive fusion recovers iff $\sum_i X_i > nB$, correct fusion iff $\sum_i X_i > B$. On $\{X_i < B\ \forall i\}$, $\sum_i X_i < nB$ always — naive posterior fusion fails on the entire unanimous-wrong region, while correct fusion succeeds exactly on W_n.*

Proof of Proposition 2 and Corollary 1. The identity is immediate from the two defining displays; that L_{correct} — the posterior log-odds given the σ-field generated by all distinct observations $\{S_i\} \cup \{E_k : m_k \geq 1\}$ — equals $\sum_i X_i + \sum_{k: m_k \geq 1} C_k$ is posterior additivity, for which Assumption 1 is load-bearing. For the corollary: on $\{X_i < B\ \forall i\}$ we have $\sum_i X_i < nB$ termwise, so the naive threshold $\sum_i X_i > nB$ is unreachable there, while the correct threshold $\sum_i X_i > B$ is met exactly on W_n of Theorem 1. □

D.6 Context exposure

Proof of Lemma 2. (a): $B \geq 0$ a.s. gives $\{Z - B > 0\} \subseteq \{Z > 0\}$ pathwise, so the difference of probabilities is $P(\{Z > 0\} \setminus \{Z > B\}) = P(0 < Z \leq B)$; pathwise, hence valid for random B dependent on Z. (b): the equalities $P(C{=}1) = \mathbb{E}[h(Z)]$ and $P(C'{=}1) = \mathbb{E}[h(Z - B)]$ are the tower property under the shared response function; nonnegativity is pointwise monotonicity with $B \geq 0$. □

D.7 UNANIMITY SCALING (WITH PREVALENCE CO-REPORT)

Corollary 2 (Unanimity scaling)**.** *Work throughout under $P_+(\cdot) := P(\cdot \mid Y{=}+1)$:* every *probability, law, and expectation in this corollary is under P_+. (This is required, not cosmetic: the X_i are iid only given Y; unconditionally they are dependent through the common Y, and the product-law and q^n statements below would be false under the unconditional law.) Let $X_1, \dots, X_n$ be iid under P_+; fix $B > 0$; let $q = P_+(X < B) \in (0,1)$, $\mu_B = \mathbb{E}_+[X \mid X < B] > 0$, and $\mathbb{E}_+|X| < \infty$; for the rate assume additionally $X \in [a,b]$ P_+-a.s. Let $Q_B = \mathcal{L}_+(X \mid X < B)$ and $A_n = \bigcap_i \{X_i < B\}$. Because the X_i are iid under P_+ and A_n is a product event, $\mathcal{L}_+(X_1, \dots, X_n \mid A_n) = Q_B^{\otimes n}$; hence, for one fixed iid sequence $\widetilde{X}_1, \widetilde{X}_2, \cdots \sim Q_B$,*

$$P_+\Big(\sum_{i\le n} X_i > B \,\Big|\, A_n\Big) = P\Big(\sum_{i\le n} \widetilde{X}_i > B\Big) \;\longrightarrow\; 1,$$

by the SLLN applied to the $\widetilde{X}$-sequence ($n^{-1}\sum \widetilde{X}_i \to \mu_B > 0$ a.s., B fixed); and for $n > B/\mu_B$, $P(\sum_{i\le n} \widetilde{X}_i \le B) \le \exp\{-2n(\mu_B - B/n)^2/(b-a)^2\}$. Meanwhile $P_+(A_n) = q^n \to 0$ exponentially.

The two displays are always presented *together*: conditional recoverability tends to one while the conditioning event becomes exponentially rare. "Bigger committees fix unanimous error" is not licensed; nor is any claim about committee sizes beyond the experimental $n = 5$. If $\mu_B \le 0$ the conclusion fails (the conditional sum drifts negative).

D.8 SCOPE LIMITS

(1) All results concern Bayes-optimal decoding (framework caveat). (2) Lemma 1(b) is two-sided: $p_i \ge \frac{1}{2}{+}\gamma$ concentrates on the correct mode. (3) Theorem 1 needs Assumption 1 (see the duplicated-evidence counterexample) and its positive-probability conclusions need the stated interval conditions. (4) Proposition 2 is stated only under M0: with sources correlated given Y, "count each source once" is not the correct posterior either. (5) Lemma 2(b) fails for non-monotone h (real LLM verifiers may be non-monotone), and (a) needs $B \ge 0$ a.s.

D.9 SYNTHETIC MECHANISM CHECK (POST HOC)

The theory above was built after the experiments, and none of its variables are measured on the LLM committees (Section 8). As a mechanism check we therefore simulated the model itself, with grids, decision rules, and RNG frozen in a written protocol before execution (zero provider calls; raw outputs, analysis, and manifests ship with the artifact). Private signals are two-point, $X_i = \pm\lambda$ with $P(X_i{=}\lambda Y \mid Y) = \sigma(\lambda)$; votes are $V_i = \mathbf{1}\{X_i > B\}$ as above; full messages reveal the X_i. Independence and clean-channel grids are computed exactly; an assumption-violation grid draws correctness indicators from a one-factor Gaussian copula with latent correlation ρ, and there the Bayes decoder uses the correlation-aware law of the vote count, so duplicated evidence is never summed as if independent. All five frozen qualitative predictions pass. (i) In every sub-threshold cell ($\lambda < B$), all votes are wrong with probability one given $Y{=}+1$ — the induced-corner analogue — while supra-threshold majorities recover with n. (ii) Sub-threshold votes are identical across states, so the optimal vote-only rule attains exactly the prior-majority accuracy $\sigma(B)$ at every n, whereas full-message pooling at $n{=}129$ exceeds it by $+0.118$ to $+0.361$ across all nine $(B, \lambda/B)$ cells (Figure 4, left). (iii) The strict Theorem 1 gap appears in all 60 eligible cells (minimum 0.0019) and is exactly zero supra-threshold, where votes reveal the two-point signal. (iv) Under balanced exogenous Y, the clean channel's accuracy $\sigma(\lambda)$ is exactly constant in B while exposed accuracy is nonincreasing, dropping to $\frac{1}{2}$ once $B \ge \lambda$. (v) The message-pooling advantage erodes as ρ rises — at $n{=}65$ it first falls below 0.01 at $\rho{=}0.75$ for $\lambda/B{=}0.5$ and at $\rho{=}0.99$ for $\lambda/B{=}0.75$ — and vanishes at $\rho{=}1$, where the correlation-aware decoder counts the duplicated evidence once (Figure 4, right; compare the duplicated-evidence counterexample under Assumption 1). The ρ sweep is an assumption stress test: it shows that the magnitude of the message-pooling advantage depends on conditional independence, not that the assumption holds for LLM committees. This simulation illustrates the theorem's mechanism in its own generative model; it is not evidence that LLM agents expose calibrated LLRs, that their evidence is conditionally independent, or that debate implements a Bayes decoder, and no multi-round dynamics are simulated.

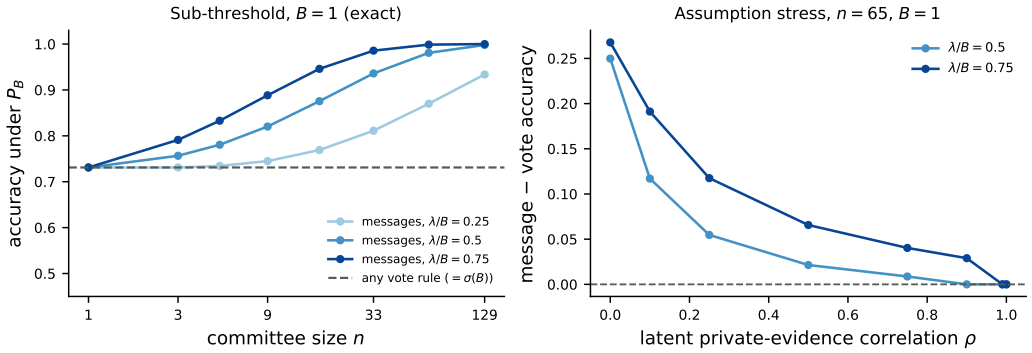

Figure 4: Synthetic mechanism check (post hoc; protocol frozen before execution). **Left:** sub-threshold accuracy vs. committee size n at $B{=}1$ under P_B: every vote-only rule is pinned at the prior-majority baseline $\sigma(B)$ (dashed) because all votes are identical, while full-message Bayesian pooling recovers with n. **Right:** message-minus-vote accuracy gap at $n{=}65$ as the latent private-evidence correlation ρ rises; the advantage erodes and is zero at $\rho{=}1$, where the correlation-aware decoder counts duplicated evidence once. Illustration of Theorem 1's mechanism only; no LLM variable is measured (Appendix D.9).

E Experimental Lineage

The design chain is: natural observation → precursor → discovery → exact-matched confirmation. A 381-question natural replay with an earlier committee roster of the same architecture found naturally occurring false-consensus states at natural prevalence, with no induction (companion manuscript, artifact); that observation, which is roster-specific and analyzed in companion work, motivated inducing the regime deliberately. The E12 precursor (original roster) first showed, on held-out questions under the authoritative-premise induction, the debate-escape pattern and the clean/corrupt verifier gap that E13/E14 then isolate; E12's fifth agent later lost its only serving endpoint, so E13 froze a replacement roster under an outcome-blind operational rule (Appendix A), and E12 absolute numbers are never pooled with E13/E14 — cross-experiment statements in this paper are qualitative replication statements only. A companion six-identity context-control experiment — on the original roster, with one elicitation per identity, under the same authoritative-premise induction class as E13 — provides converging evidence that context exposure, not verifier identity, carries the clean-verifier advantage (companion manuscript, artifact). E13 and E14 as reported here share one frozen roster and are the only experiments whose numbers appear in this paper.

F Infrastructure Forensics

Agent identity = (model, provider). An OpenRouter model id is not a fixed execution system: requests are multiplexed across upstream providers that differ in quantization, declared completion ceilings, and protocol compliance. Every logical agent is therefore pinned to one exact provider with `allow_fallbacks=false`, and a response served by any other provider is a hard stop, not a warning. Frozen pins (both experiments): deepseek-chat-v3-0324 @ Crusoe; qwen3-coder @ Alibaba; kimi-k2 @ Novita; gpt-oss-120b @ Groq; unslopnemo-12b @ Parasail; verifier qwen3-32b @ DeepInfra.

Caps on the wire. The execution ceiling follows the route (65,536 completion tokens for every committee agent; 16,384 for the verifier route), and a probe verifies the value that actually goes on the wire — the cap is a per-call payload field, not a config aspiration. Both experiments report zero wrong-cap requests.

Truncation and parsing. Every reply carries finish-reason and token accounting; a truncation classifier separates hard truncation from natural stops, and unparsed replies are never imputed — in round 0 they make the state incomplete (excluded everywhere); in later debate/no-peer rounds the

agent's previous position carries, per the frozen decision rule. Both experiments closed with zero Class-A truncations and zero round-0 parser failures among kept states.

Incomplete-state budget and billing outages. A preregistered budget (4 incomplete states total / 2 per dataset) hard-stops collection; it tripped only during the OpenRouter billing outages documented verbatim in Appendix A (HTTP 402, zero completion tokens — transport-class, carrying no elicited content). All outage states were re-elicited fresh on resume; no persisted record was ever overwritten, and per kept state the call budget is exactly $5 + 15 + 15$ (ledger-verified), with verifier calls cached once per question per seed process.