INTERACTION DESTROYS THE DIAGNOSTICITY OF AGREEMENT WHERE TRUTH IS SCARCE

Anonymous authors
Paper under double-blind review

ABSTRACT

While multi-agent consensus is frequently used as an epistemic proxy for reliability, agreement justifies reliance only when derived from independent evidence. We find that peer interaction compromises this decision-relevant independence, triggering cascades that decouple accuracy from diagnosticity. When three heterogeneous agents deliberate on questions where their independent baseline answers disagree, the resulting two-of-three majorities are false 74.8% (86/115) of the time, versus 23.1% (31/134) for independent votes (Fisher $p = 2.1 \times 10^{-16}$). A registered blind-revision control separates visibility from revision: private revision keeps conditional falsity inside its registered bound, whereas visibility manufactures 14.2 percentage points [8.3-20.7] more false agreed answers per question on MMLU-hard (McNemar $p = 1.9 \times 10^{-5}$; replicated at $+17.5$ points), without a detected worsening of the majorities that form. We formalize this as the Diagnosticity-of-Agreement Law: agreement signals truth only to the extent agents could have agreed without interacting. Replications indicate that this collapse, which arrives predominantly at the first exchange (62.9% of the increase), tracks a panel's operative ability to verify. A weak panel collapses on GSM8K (80.0% vs. 21.2%) and MMLU (84.3% vs. 57.4%); a mid-capability panel resists on GSM8K ($+1.0$ point) but collapses on MMLU-hard (44.8% vs. 28.9%). Crucially, although deliberation is at least as accurate as self-consistency at a matched number of completions (34.7% vs. 31.0%), it severely degrades diagnosticity (79.6% vs. 46.2% conditional false agreement), indicating that rising accuracy does not by itself warrant reliance on agreement. Finally, a predicted Coverage Law ($r = -0.996$) bounds mitigation on panels that appear able to verify retrieved evidence; on a weaker panel it failed, and checkable evidence restored the bound (4.0% and 2.7% residual).

1 INTRODUCTION

Multi-agent deliberation is increasingly used to improve the reliability of large language models: several model instances exchange answers, critique one another, and converge on a joint output (Du et al., 2024; Liang et al., 2024). The design intuition comes from human committees, in which agreement among independent judges is evidence. That intuition rests on an assumption that is easy to overlook: the judges must be independent. Once agents talk to one another, their agreement can no longer be regarded as the agreement of independent witnesses. Consequently, its evidential value becomes an empirical question rather than a premise.

The question predates multi-agent systems. Philosophical work on AI distinguishes trust from reliance and treats reliance as an epistemic matter: what grounds accepting a system's output as valid (Durán and Pozzi, 2025). The human-factors tradition answers with appropriate reliance, meaning reliance that tracks a system's actual trustworthiness rather than its surface confidence (Lee and See, 2004). Both take for granted that the signal relied upon means what it appears to mean. Our question is therefore not whether interaction produces consensus, but whether the agreement it produces remains evidence that can justify reliance; we measure that for one signal, post-interaction consensus.

On 300 short-form factual-recall questions screened so that three heterogeneous weak-tier agents disagree in isolation, we compare two ways of aggregating the same agents on the same questions:

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.

a majority vote over their pre-deliberation answers, and the majority they reach after four rounds of fully connected deliberation. The first is false 23.1% of the time; the second, 74.8%. Interaction did not merely fail to improve the signal; it removed a signal that was present before the agents spoke. The remainder of the paper is evidence for, or a boundary condition on, that observation.

We report six lines of evidence: (i) the counterfactual itself, with its robustness checks and a registered blind-revision control that separates peer visibility from the opportunity to revise (Section 3.1); (ii) a mechanism, in which false consensus is manufactured where no agent independently holds the answer and adoption tracks the receiver's operative epistemic anchor (Section 3.2); (iii) intervention evidence, in which a single confident peer decides the outcome (Section 3.3); (iv) a refuted rival account, seeded incumbency, which failed all four of its pre-specified predictions (Section 3.3); (v) a quantitative law for grounded deliberation, predicted before measurement, confirmed under live retrieval, and bounded on a weak panel, with the repair program that followed (Section 3.4); and (vi) replications that fix the scope: a registered frontier replication, a matched-completion comparison in which accuracy and diagnosticity come apart, a judge-free two-substrate replication on two panels, and a round-level analysis of when the signal is lost (Sections 3.5, 3.6).

2 THREE LAWS

We name three regularities the evidence supports. Each is stated here in general terms with its falsifier and its epistemic status; the measurements behind each are in Section 3, and the formal statements, assumptions, and proofs are in Appendix C.

Definitions. An agent's *isolated answer* to a question is its answer with no access to other agents' outputs, and the question's *anchor count* k is the number of seats whose isolated answer is correct. A question is *truth-scarce* when $k = 0$: no seat holds the answer before interaction. A pool is truth-scarce when its $k = 0$ subpopulation has non-negligible mass. For a question distribution, a panel, and an aggregation rule R, let A be the agreement event, the *non-unanimous* majority (two-of-three in the three-seat studies; unanimous agreements behave differently and are reported separately), and let T be the event that the agreed answer is true. The *diagnosticity* of R is $D(R) = 2P(T \mid A) - 1$. It is a rescaled conditional probability, meaningful only against the pool's base rate: on the core pool (per-seat isolated accuracy 0.187), an independent two-of-three agreement multiplies the odds of truth by roughly 14 and a deliberated agreement by roughly 1.5 (Appendix C.8).

Diagnosticity-of-Agreement Law. On truth-scarce pools, interaction cannot create diagnosticity: $D(R_{\text{deliberated}}) \leq D(R_{\text{independent}})$, and the gap grows as independent truth-availability falls. Where truth is abundant, few items pass the divergent-start screen and there is little for the law to act on (Section 3.5). Falsifier: a protocol whose post-interaction majorities are more diagnostic than matched independent votes on a truth-scarce pool at $n \geq 150$. Status: a conjecture with a proved restricted form; the structural boundary and the rescue-and-manufacturing decomposition of deliberation's net effect are in Appendix C.8 (Proposition 4).

Anchor Principle, operative form. Adoption of a peer's claim under social pressure is governed by the receiver's *operative anchor*, its capacity to re-derive or verify the claim at run time, rather than by whether the substrate is verifiable in principle. Falsifier: adoption of confident, plausible, type-consistent falsehoods at high rates among agents whose verification accuracy has been measured and is high, at $n \geq 100$. Status: an empirical principle. Its substrate-level form was registered and confirmed on one panel and reversed on another; the operative form was adopted after that reversal, and the registered tests that have borne on it since are listed in Section 3.6.

Coverage Law. Under deliberation grounded in retrieved evidence, let p be the retrieval coverage (the probability that retrieval surfaces evidence bearing the true answer), ε the residual false-consensus rate when it does, and $\phi(k)$ the ungrounded false-consensus rate at anchor count k when it does not. Residual false consensus is then the miss rate attenuated by anchor mass, $\text{FalseRate}(p, k) = p\,\varepsilon + (1-p)\,\phi(k)$, with measured $\varepsilon \in [0.02, 0.05]$ and $\phi(0) = 1.00$, $\phi(1) = 0.69$, $\phi(2) = 0.52$. Its corollary is that residual false consensus below 10% requires coverage above $[0.918, 0.947]$ across that ε range. Status: a theorem under the assumptions stated in Appendix C, three of them measured; as a guarantee it held on the capable panels and failed on the weak one (Section 3.4), consistent with a gate on operative verification that is inferred rather than measured.

3 RESULTS

The core study uses a heterogeneous weak-tier three-agent panel (claude-haiku-4-5, gpt-5.4-mini, gemini-3.1-flash-lite), four rounds of fully connected deliberation, and $n = 300$ divergent-start SimpleQA questions (Wei et al., 2024), scored by a decomposed pairwise judge (claude-sonnet-4-6) validated against blind human labels at 95.0% agreement, $\kappa = 0.888$ [0.765-0.974], $n = 80$ (Appendix A). Intervals are 95% Wilson score intervals (Wilson, 1927), and every prediction, registered or not, appears with its verdict in Table 3.

3.1 THE COUNTERFACTUAL: INTERACTION DESTROYS DIAGNOSTICITY

After deliberation, 86 of 115 two-of-three majorities are false (74.8% [66.1-81.8]). Assembled from the same agents' pre-deliberation answers on the same questions, independent-vote majorities are false 31 of 134 times (23.1% [16.8-31.0]). The difference is the measured effect of the deliberation protocol, four rounds of peer-visible revision against none (Fisher $p = 2.1 \times 10^{-16}$); the registered EXP-14 control below separates visibility from revision. Because the two arms share questions and agents, the marginal Fisher test does not honor the pairing, so the contrast is also reported with its paired, cluster-robust complements. The gap on the paired join is $+51.1$pp (the join drops three questions without a cached equivalence judgment; Appendix A) with a question-cluster bootstrap 95% CI of [0.406, 0.616] ($p(\text{gap} \leq 0) \approx 0$), and McNemar's test on the 50 questions where both arms formed majorities is 24 to 2, exact $p = 1.05 \times 10^{-5}$. Figure 1 shows both abstention conventions.

The two arms differ in two factors: social visibility and four revision opportunities. The registered blind control (EXP-14) separated the two factors on the same pool and panel as EXP-6 by echoing each agent's own previous answer in place of the peer lines. Private revision without visibility keeps conditional falsity inside the registered bound: it moves by -0.5 points on GSM8K and $+3.6$ points on MMLU-hard (P14.1 confirmed), and on MMLU-hard the solo arm's majority formation collapses (34.9% against 68.6% independent and 79.9% deliberated). The conditional deliberated-versus-solo contrast on the divergent-start join was underpowered as registered ($n = 52$; P14.2 not confirmed); at the deployment-relevant estimand, registered by amendment, deliberation produces $+14.2$pp [8.3, 20.7] more false agreed answers per question than solo revision on MMLU-hard (McNemar 28 to 4, $p = 1.9 \times 10^{-5}$; A2.1 confirmed), and a powered three-arm replication on a fresh draw confirmed the edge ($+17.5$pp [10.2, 24.7], McNemar 35 to 6, $p = 4.9 \times 10^{-6}$; EXP-14b). Among majorities formed in both arms, no deliberated-versus-solo difference in falsity was detected (P14.2; A2.2, $n = 22$), so EXP-14 shows that visibility drives the manufacture of false agreement, not that it makes already-formed majorities falser. The loss of diagnosticity is therefore the joint effect of visibility under revision, operating on this evidence through which majorities form (Proposition 1); the full report is in Appendix A.8. Excluding majorities that contain an abstention changes nothing qualitatively: 79.6% (86/108) deliberated against 42.5% (31/73) independent, $p = 5.7 \times 10^{-7}$. Unanimous three-of-three agreement is unbiased (49.4% false, $p = 0.94$), consistent with the mechanism below: unanimity almost always means that at least one agent knew the answer.

We label the core contrast *post hoc*, because the equivalence classes among isolated answers were derived after the confirmatory results were known. Two facts anchor it regardless. Judge invariance holds (31/134 under both labelings audited), and the identical contrast was then registered and confirmed on a frontier panel (E1-P2, Section 3.5, Holm $p = 6.9 \times 10^{-5}$) and, judge-free, on two further substrates (Section 3.6).

Deliberation also changes how much agreement there is: consensus events rise from 134/300 isolated to 285/300 deliberated ($p = 2.2 \times 10^{-45}$), abstention majorities fall from 61/300 to 10/300 ($p = 4.1 \times 10^{-11}$), and mean accuracy rises from 13.7% [10.2-18.0] to 34.7% [29.5-40.2] (104/300 on the de-duplicated canonical pool, majority-of-finals convention; the answer-clustering convention gives 35.0%; the 0.3-point seam between conventions is disclosed in the appendix; 69 corrected, 5 corrupted). Both are true at once: deliberation answers more and better on average, while the agreement it produces stops tracking truth, which is what makes the signal dangerous downstream.

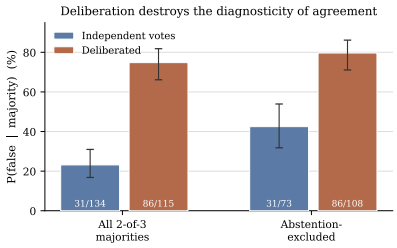

Figure 1: Conditional falsity of two-of-three majorities on the same questions: independent votes against deliberation, 23.1% versus 74.8% (Fisher $p = 2.1 \times 10^{-16}$), with the abstention-excluded convention alongside. 95% Wilson intervals; counts at bar bases.

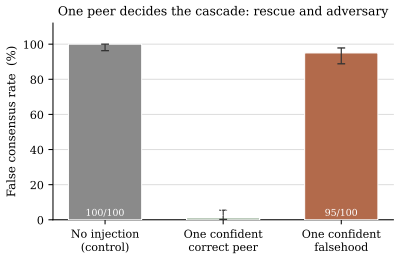

Figure 2: False consensus under single-peer injection ($n = 100$ runs per condition): one confident correct peer rescues the panel; one confident falsehood is adopted in 91.0% [83.8-95.2] of runs. 95% Wilson intervals.

3.2 Truth scarcity and the Anchor Principle

False majorities begin knowledge-poor. Deliberated false majorities start from rounds averaging 0.20 agents correct in isolation; true majorities start from 1.26. The probability that a deliberated consensus is true is 0.06, 0.71, and 0.88 when $k = 0, 1, 2$ agents were independently correct. The first anchor is nearly decisive (step odds ratio approximately 39) and the second adds little (approximately 2.9): interaction manufactures agreement on the questions to which no participant could have supplied the answer.

Confidence dynamics decouple from truth. Mean verbalized confidence rises under convergence whether the consensus is true (0.740 to 0.787) or false (0.673 to 0.704), and the highest confidence band is correct only 67% of the time. Under peer pressure, 17.0% of agent-instances [14.7-19.6] move from abstention to assertion, 86.7% of abstainers are induced to assert, and agents that abandon a correct answer move to an error in 42 of 43 cases.

The Anchor Principle predicts a substrate gradient: where agents can re-derive or verify, injected falsehoods should fail. Measured adoption of one confident injected falsehood is 12% [8-20] on GSM8K (Cobbe et al., 2021), where the answer is re-derivable by computation, 23% [17-32] on TruthfulQA (Lin et al., 2022), where partial prior knowledge is available, and 86% [81-91] on SimpleQA, which tests obscure recall ($\chi^2(2) = 94.7$, $p < 10^{-20}$). The gradient replicates across the 3×3 injector cross at 71.1% over 2,062 receiver-observations, and the pattern repeats without injection: baseline false consensus is 16.7% (GSM8K), 36.7% (TruthfulQA), and 56.5% (SimpleQA); TriviaQA and five-agent replications are in Appendix A.9.

3.3 Mechanism by intervention: one peer decides

If the failure is a content-specific, confidence-mediated cascade, a single calibrated peer should control it, and it does ($n = 100$ per condition; Figure 2). Panels heading for false consensus are rescued by one confident correct peer: false consensus falls from 100% to 1% (Fisher $p = 2.2 \times 10^{-57}$). One confident falsehood produces false consensus at 95.0% [88.8-97.8], with the group adopting the injected falsehood in 91.0% [83.8-95.2] of runs. A type-inconsistent confident assertion (a number offered for a fact question) is never adopted: 0/50, against 95% for the matched consistent falsehood ($z = 11.38$). Truth is adopted more readily than falsehood (99% against 91%, $p = 0.009$), so the cascade is content-sensitive and confidence-mediated at the level of plausible, type-consistent claims rather than content-blind; the 99/91 asymmetry is the more informative boundary, since the type-inconsistent ablation confounds inconsistency with detectability. The fully crossed injector-family design (family-robust propagation, 82.5-94.2%) is in Appendix A.9.

A refuted rival: seeded incumbency. The natural rival explanation is incumbency: receivers protect answers they already hold at round zero, so a seeded shared belief should resist injection.

We pre-specified four predictions of this account and tested them across 18 cells (1,320 runs, 1,078 confirmatory). All four failed at their pre-specified thresholds, and one reversed: seeding a false incumbency *raised* adoption of a rival falsehood from 0.371 to 0.453 ($p = 0.0019$), and 23.2% of false-seeded outcomes were novel errors matching neither seed, injection, nor gold (full table in Appendix A.9). Incumbency remains only as a descriptor of the cross, since round-zero pre-agreement marks the resistant cells; what survives the refutation is the Anchor Principle as stated in Section 2: re-derivability and knowledge, not shared priors, decide adoption.

3.4 The Coverage Law: grounding bounds what interaction can break

Because the group lacks an anchor, the mitigation must supply one before convergence: from the first round, an evidence-backed verifier asserts the retrieved answer, and answers contradicted by the evidence are excluded from the final consensus (the Verified Grounding Protocol). When retrieval is imperfect, its protection follows a quantitative law. Sweeping p_{truth} from 0.30 to 0.85 gave false-consensus rates of 68%, 45%, 39%, 30%, and 15%, a monotone decrease with residual false consensus close to $1 - p_{\text{truth}}$ ($r = -0.996$), a relationship registered as a prediction before the confirming sweep (Figure 3, left). Under live web retrieval ($n = 205$, measured $p_{\text{truth}} = 0.75$), the law's prediction of 25.0% met an observed 20.5% [15.5-26.5], with the conditional mechanism intact (5.2% on hits, 66.7% on misses) and the strata tracking $1 - p_{\text{truth}}$ at $r = 0.81$ (Figure 3, right). A registered attenuation experiment on a high-anchor pool showed the slope falling from 1.13 to 0.72 ($p = 0.010$), with the ordered anchor gradient 100%, 69%, and 52% at $k = 0, 1, 2$ (trend $p = 5.3 \times 10^{-8}$; Table 3).

The law is gated by operative verification. EXP-11 executed the registered high-coverage follow-up on the weak 3B panel with a gold-asserting verified-evidence peer from round one (150 MMLU questions per arm at $p_{\text{truth}} = 0.90$ and 0.95; Appendix A.5). All three registered predictions failed at the full draws: residual false consensus was 34.0% [26.9-41.9] (51/150) against the 15% ceiling at 0.90 and 26.7% [20.2-34.3] (40/150) against the 10% ceiling at 0.95 (P11.1-P11.3; one-sided binomial $p = 5.8 \times 10^{-9}$). The dose-response direction survives, since five additional points of coverage bought 7.3 points of rescue, so the mechanism is present but weak: the 3B panel appears unable to verify what it is shown, so a gold-asserting peer functions for it as one more conformity target. The corrected statement of the law, on this evidence, is that residual false consensus tracks the retrieval miss rate only for panels whose seats can operatively verify the evidence; coverage relocates the risk, and capability decides whether the relocation succeeds.

What repairs a weak panel. The registered repair program (EXP-11b, EXP-12, EXP-12c) probed the active ingredient with a verifiability dose-response ladder against the locked EXP-11 failure (Figure 5; design in Appendix A.5). On MMLU, a more capable verifier's checked prose verdict changed nothing: 34.0% [26.9-41.9] (51/150) at 0.90 and 28.0% [21.4-35.7] at 0.95 (P12.1-P12.4 not confirmed); an exploratory inspection localizes the null in the verifier's competence, which contradicted the correct reference on 63% of covered items (Appendix A.9). The rescue term therefore binds twice: the checker must know the answer, and the receivers must be able to check it. The same author model emitting a structured certificate with per-seat check ballots and an admission gate reduced the residual to 18.5% [13.2-25.5] at a 7.9% abstention cost (P12c.4 confirmed: a drop of 15.7pp against the locked failure, Fisher $p = 0.0016$, BH $q = 0.008$) and to 14.7% at 0.95 coverage. On GSM8K, where verification is cheap, checkable worked derivations alone restored the law in full: 4.0% [1.8-8.5] and 2.7% [1.0-6.7] at 0.90 and 0.95 (P11b.1 and P11b.2 confirmed, BH $q \leq 0.0011$), below even the ungated predictions. What repairs a weak panel is therefore a voice that is correct and checkable rather than a more capable one, and where the substrate cannot supply checkability the protocol must.

The tradeoff frontier is the risk-coverage curve of selective prediction (Geifman and El-Yaniv, 2019): treating consensus as a confidence signal is a calibration problem, and our frontier prices it. A paired baseline shows that deliberation adds no value once grounded, and EXP-8 confirms this under live retrieval (Section 3.8). The recommendation stands with its gate: retrieve, then aggregate independently; on panels too weak to verify what they retrieve, prefer abstention-default designs to consensus targets.

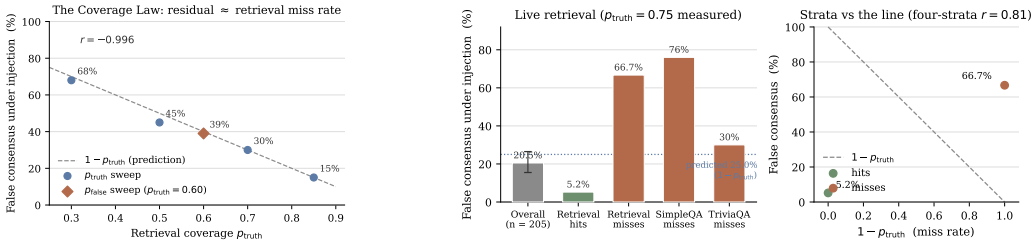

Figure 3: The Coverage Law. Left: residual false consensus tracks $1 - p_{\text{truth}}$ across the sweep ($r = -0.996$; registered before the sweep). Right: live retrieval ($p_{\text{truth}} = 0.75$), 5.2% false consensus on hits against 66.7% on misses; strata track $1 - p_{\text{truth}}$ ($r = 0.81$).

3.5 Frontier replication and the matched-completion dissociation

The registered replication E1 (OSF, anonymized view-only link in Appendix A.3) uses a generation-homogeneous current-generation panel (claude-opus-5, gpt-5.5-2026-04-23, gemini-3.1-pro-preview) over 280 items, and both principal findings replicate. Deliberated two-of-three majorities are false 66.7% [51.0-79.4] of the time (26/39) against 26.2% [18.0-36.5] (22/84) for the same agents' independent votes (one-sided Fisher, Holm-adjusted $p = 6.9 \times 10^{-5}$; paired bootstrap CI for the gap [0.237, 0.571]; P2 confirmed), and false consensus again concentrated where truth was scarce (87.5% at $k = 0$ against 33.3% at $k \geq 1$, Holm-adjusted $p = 0.0016$; P3 confirmed). The capability-shrinkage prediction failed (66.7% against 74.8%, $p = 0.2173$): no capability benefit was detected at the registered power of 39 majority events, which is not the same as capability providing no protection (P4 not evaluable at 8 events). The powered follow-up EXP-10 (registered) maps the regime boundary rather than extending the collapse: only 22 of 160 frontier items produced divergent starts and all 22 resolved to unanimity (19 true, 3 false), so the registered endpoints P10.1-P10.3 are not evaluable on this panel (Appendix A.3, Figures 6 and 7).

The matched-completion control (EXP-1, $n = 300$ per arm; registered under the name matched-compute) produces the dissociation for which the paper is named. Parity is at the level of completions, 15 per question in each arm; deliberation also re-reads the peer messages each round and is the costlier arm in tokens. Single-model self-consistency is no more accurate than deliberation (31.0% [26.0-36.4] against 34.7% [29.5-40.2]; intervals overlap) yet far more diagnostic: its conditional false-agreement rate sits at the independent-vote level (46.2% [39.0-53.7] against 79.6% deliberated). Cross-model plurality at the same completion budget is safer still but answers rarely (2.3% [1.1-4.7] unconditional false plurality; conditional 25.9% [13.2-44.7] over 27 pluralities); diversity preserves diagnosticity by declining to manufacture agreement. Accuracy rises while evidential value falls.

3.6 Generality across substrates: a registered judge-free replication

EXP-6 is the registered generality test: two new substrates (GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021)), a programmatic judge, and two panels in a new local model ecosystem. Design, amendments, stopping rules, corpora, and the judge-level repair are in Appendix A.7; both arms are reported.

The weak panel collapses on both substrates. On GSM8K, deliberated majorities are false 80.0% [67.0-88.8] of the time (40/50) against 21.2% [14.8-29.4] (25/118) for independent votes, a gap of $+58.8$ points (Fisher $p = 9.27 \times 10^{-13}$; paired bootstrap [0.454, 0.712]). On MMLU the rates are 84.3% [76.2-89.9] (91/108) against 57.4% [49.2-65.3] (81/141), a gap of $+26.8$ points (Fisher $p = 3.41 \times 10^{-6}$; bootstrap [0.167, 0.368]); McNemar tests for both substrates are in Table 3. P6.1 is confirmed. P6.2, the registered anchor-attenuation prediction, failed on this panel with the effect in the opposite direction, since the GSM8K gap is the larger one (DiD bootstrap $[-0.482, -0.152]$). This registered failure motivated the operative-anchor refinement: if a 3B model cannot reliably verify its own arithmetic at run time, the nominally re-derivable substrate offers it no operative anchor.

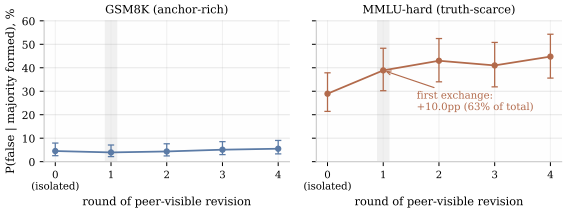

Figure 4: Conditional falsity of formed majorities by round (EXP-13, on the EXP-6 records; 95% Wilson intervals). MMLU-hard rises at the first exchange and plateaus; GSM8K is flat.

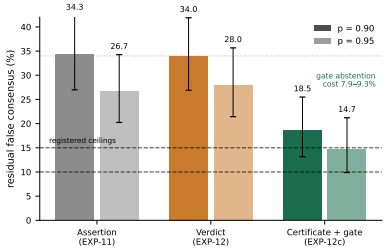

Figure 5: Repair ladder on the weak 3B panel: residual false consensus at $p = 0.90$ and 0.95 (95% Wilson intervals; dashed lines mark registered ceilings). A verifier's prose verdict matches bare assertion; the certificate-plus-gate protocol cuts failure by 15.7 points at the annotated abstention cost.

The mid-capability panel collapses only where it cannot verify. On the resumed arm the pattern reverses as the operative account predicts. On GSM8K, independent votes are false 4.5% [2.5-7.9] of the time (11/243) and deliberated majorities 5.5% [3.3-9.0] (14/254), with no meaningful gap (+1.0 point; Fisher $p = 0.384$; bootstrap $[-0.010, 0.031]$); this panel appears able to execute the verification, and interaction neither helps nor harms. On MMLU-hard, deliberated majorities are false 44.8% [35.6-54.3] of the time (47/105) against 28.9% [21.4-37.9] (33/114), a gap of +15.8 points (Fisher $p = 0.011$; paired bootstrap [0.059, 0.259], $p(\text{gap} \leq 0) = 0.0011$). P6.1 is confirmed on the registered hard stratum and P6.2 is confirmed on this panel (DiD [0.045, 0.252], $p = 0.0024$). P6.3 was not supported on either arm, since neither panel abstains at meaningful rates (Table 3). Together the two arms strongly support the operative-anchor account: the collapse tracks whether the panel can execute the verification a substrate nominally allows, not the substrate itself. Verification capability is not measured directly; the proxy is each panel's independent-arm accuracy on the same items (78.8% of the weak panel's independent GSM8K majorities are true, against 95.5% for the mid-capability panel), which is measured before any interaction and so is not derived from the collapse it explains. The operative form was adopted before the resumed arm was analyzed; the registered tests that have borne on it since (P11b.1-2 and P12c.4 confirmed, P12.1-4 failed) and its untested falsifier are in Appendix A.10. A judge-level repair of the answer-extraction pipeline is disclosed with its concordance audit ($\kappa = 0.756$) in Appendix A.7; the local arm's verdicts are invariant under it, and the resumed arm is reported on the repaired basis.

When the signal is lost: the first exchange. The records hold every agent's answer at every round, so diagnosticity can be tracked round by round (EXP-13, registered and hash-archived before its first run). On GSM8K the series is flat (4.5% at round 0, 5.5% at round 4; the paired bootstrap interval spans zero). On MMLU-hard the damage arrives at the first exchange: 28.9% of independent majorities are false, 38.9% after one round of peer visibility, and 41.0-44.8% thereafter, so the first exchange carries 62.9% of the total increase (registered share threshold 0.50; P13.2 confirmed). The registered endpoint test is borderline under its own multiplicity correction (raw one-sided bootstrap $p = 0.0206$, Benjamini-Hochberg $q = 0.041$; McNemar 5/14), so we claim the shape of the decay rather than a sharpened endpoint. For deployment, a protocol that lets agents compare notes once is not safer than full deliberation (Figure 4).

3.7 The registration ledger

Table 3 lists every prediction with its registration status, result, and verdict, including the failures and the not-evaluable outcomes, with exploratory findings labeled as such: 73 registered or pre-specified predictions (each numbered prediction counted once per arm; a verdict split across substrates counted as one confirmed and one failed), of which 42 failed, were not confirmed, or were not evaluable. Because analyses registered on already-collected records (EXP-13) and predictions added

Table 1: Conditional falsity of two-of-three majorities across regimes, independent versus deliberated, same agents and questions (95% Wilson intervals; gaps in percentage points; Δ log-odds is base-rate-free; unconditional rates in Table 2). The collapse is absent only where the panel appears able to execute the verification (mid-capability GSM8K).

Regime	Substrate	Judge	Panel	Independent	Deliberated	Gap	Δ log-odds
Core (post hoc)	SimpleQA	LLM, human-validated	weak API tier	23.1% [16.8-31.0]	74.8% [66.1-81.8]	+51.7	-2.29
E1 (registered)	SimpleQA-family	LLM	frontier API	26.2% [18.0-36.5]	66.7% [51.0-79.4]	+40.5	-1.73
EXP-6 (registered)	GSM8K	programmatic	local weak 3B	21.2% [14.8-29.4]	80.0% [67.0-88.8]	+58.8	-2.70
EXP-6 (registered)	MMLU	programmatic	local weak 3B	57.4% [49.2-65.3]	84.3% [76.2-89.9]	+26.8	-1.38
EXP-6 (registered)	MMLU-hard	programmatic	mid-capability 30B	28.9% [21.4-37.9]	44.8% [35.6-54.3]	+15.8	-0.69
EXP-6 (registered)	GSM8K	programmatic	mid-capability 30B	4.5% [2.5-7.9]	5.5% [3.3-9.0]	+1.0	-0.21

by amendment (EXP-14 A2) are tabulated alongside the rest, selective reporting cannot silently inflate the confirmations; the tabulated failures are what make the confirmations credible.

3.8 Boundary experiments: a summary

Five registered follow-ups have completed (Appendix A.8). EXP-8 re-ran the grounded pairing under live Wikipedia retrieval, addressing the oracle-window caveat under a weak retrieval channel: grounded deliberation produced false consensus at 42.4% [37.1-47.9] against 21.8% [17.6-26.7] for grounded independent voting on byte-identical evidence (McNemar $p = 2.5 \times 10^{-8}$). EXP-9 extended the collapse to five seats (84.4% against 55.5%); EXP-7, EXP-7b, and EXP-7c scoped the poisoned channel (adoption 4.8-100% across configurations; no effect of evidence realism); and the H1 human pilot is reported with its $n = 1$ provenance.

4 Discussion: the calibration boundary

For a practitioner, the Coverage Law is a calibration boundary rather than a protocol: if a deployment reads agreement among agents as evidence, the value of that signal is set by the panel's independent-vote diagnosticity on the question distribution and by the measured coverage of the retrieval channel. Four rules follow. Collect votes before agents interact. Harvest uncertainty from isolated responses, because deliberation converts abstention into confident assertion. Treat same-family pairs as preformed consensus at round zero. Ground the deliberation from the first round, and expect residual false consensus close to the retrieval miss rate only on panels able to verify the evidence (EXP-11's gate); on weaker tiers, prefer abstention-default aggregation, and treat grounding sources as single points of attack (Section 3.8). At every measured coverage level, deliberation added risk rather than value.

The accuracy-diagnosticity dissociation is therefore also a dissociation between performance and epistemic warrant: higher accuracy does not by itself give stronger grounds for relying on an agreed answer. Trust should rest on properties that indicate trustworthiness rather than on cues that merely raise the willingness to rely (Jacovi et al., 2021), and reliance should be calibrated to what a system can actually do (Lee and See, 2004; Durán and Pozzi, 2025). Post-deliberation agreement is such a cue: after interaction it is far less diagnostic of correctness yet no less salient.

Two inversions of the field's default engineering moves follow. First, invest in anchor probability rather than in seats: the first anchored agent is worth more than several additional unanchored ones (step odds ratio approximately 39), and panel size only bounds an adversary. Second, one verifier seat outperforms additional debater seats: a single confident correct peer reduced false consensus from 100% to 1%, and the grounded pairing shows that the verifier's value lies in its evidence.

The claims are scoped to fully connected, unstructured deliberation. Structured protocols with roles, critique, or hierarchical aggregation (Appendix B) are untested here; since Proposition 2 binds any protocol that manufactures agreement on the truth-scarce stratum, they escape it only by declining to manufacture agreement or by grounding, a prediction rather than a result. The findings also change how positive multi-agent results should be read: deliberation helps when at least some agents can derive or verify the answer, and manufactures agreement when they cannot. In the terms of the reliance literature, agreement among agents justifies reliance only insofar as the agreeing judgments are decision-relevantly independent, and interaction removes that independence (Appendix B). Post-

interaction agreement thus keeps the appearance of corroboration while losing much of its evidential content. Our experiments do not address longitudinal effects on humans, but recurrent passive reliance on human-AI collaboration can become habituated and weaken independent scrutiny (Herrera, 2026); whether repeated exposure to interaction-induced consensus reinforces reliance on degraded agreement signals is a question for future work.

5 Related work

The conformity result has a long lineage: Asch (1951) showed that unanimous confederate majorities move human judgments, and informational cascade theory (Bikhchandani et al., 1992) explains why public agreement can carry no private information. Multi-agent debate is widely reported to raise accuracy (Du et al., 2024; Liang et al., 2024), and self-consistency shows that sampling plus voting is a strong single-model baseline (Wang et al., 2023); our accuracy results are consistent with both (34.7% after deliberation; 31.0% for SC-15), and the divergence concerns what agreement means. The closest concurrent work compares debate against voting on accuracy (Choi et al., 2025), audits false-consensus risk correlationally in a medical pipeline (Zhu et al., 2025), or documents diversity collapse and manufactured overconfidence at the trajectory level (Dai, 2026); none measures the conditional evidential value of the agreement event, matched at the question level, with the interaction protocol as the manipulated variable. The human literature already contains the phenomenon (Lorenz et al., 2011; Becker et al., 2017; Mavrodiev et al., 2012; Tian et al., 2022; Jayles et al., 2017). An adjacent literature asks when AI outputs give appropriate grounds for reliance: Durán and Pozzi (2025) distinguish reliance from trust and characterize it in epistemic terms, and Lee and See (2004) make appropriate reliance, rather than trust maximization, the objective of calibrated human-automation interaction. Our results identify a specific multi-agent failure of that calibration: interaction can make agreement a substantially weaker indicator of correctness than the same agents' independent agreement. Appendix B engages these works, and the recent failure literature on multi-agent debate, individually.

6 Limitations

Appendix A.10 states each limitation with the reason the core claim survives it. In brief: the counterfactual is confirmed on three substrate classes, with contested, non-English, and multi-hop or expert-level questions open; the judge is human-validated and the result replicates judge-free; the frontier contrast is disclosed at registered power; EXP-14's conditional contrast is underpowered on the join; verbalized confidence is used only ordinally; the protocol is one fully connected scaffold, window, and language, and structured topologies are untested; verification capability is inferred from independent-arm accuracy, not measured; the human pilot has $n = 1$ and is corroboration only; benchmark exposure cancels in the within-item contrast; frequency claims are scoped to panels that converge; the coverage guarantee carries its operative-verification gate; and the follow-ups have a thin live channel, small strata, synthetic evidence by design, and infeasible strict certificate ballots.

7 Conclusion

Agreement among agents is evidence of truth only to the extent that the agents could have agreed without interacting. Where truth is scarce, fully connected deliberation manufactures agreement and removes much of its evidential value. The collapse replicates under a pre-registered, judge-free design on two panels, is absent where a panel appears able to execute the verification, persists at the frontier where pools remain truth-scarce, and, where it occurs, arrives predominantly at the first exchange. It is bounded by a coverage law whose guarantee held only on panels that appear able to verify the evidence they receive. Independent voting is safe because it abstains when it does not know, and deliberation removes that abstention: it remains a good way to obtain an answer and a poor way to learn whether an answer is true.

More broadly, the results separate performance from epistemic warrant: a system can become a better answerer while consensus becomes a worse reason for believing the answer. For multi-agent systems, the question is not how many agents agree, but what independent or verifiable basis remains behind their agreement.

REPRODUCIBILITY STATEMENT

The supplementary material contains the anonymized reproducibility package: the collection, judging, and analysis code for every experiment, the pre-registrations (Tier-0, Tier-2, E1, EXP-6 through EXP-14b) with their amendments, the deviation ledger, and the scripts that regenerate every figure (`figures/make_figures.py` and `figures/gen_fig8_repair_ladder.py`), under a `SHA256SUMS` manifest. The original core-study harness ships unmodified in `original_harness/` (collection runner, canonical tally script with by-question de-duplication, and both judge-validation scripts); the cross-platform `harness/` package used by the registered replications is a protocol-identical port of it. Run transcripts, judge calls and caches, question pools, verdict outputs, and the auditor memos are deposited in the anonymized OSF project at https://osf.io/a5t3v/overview?view_only=be06a28afaf74a3f8cbc02d33cd337da. A post-registration harness fix in E1 is logged as a deviation entry with its hash transition, diff, and date (Appendix A).

ETHICS STATEMENT

Human input enters this work in two disclosed places. Judge validation used blind human labels produced by the authors (80 for the core judge, 60 per additional audit, and two further sets of $n = 180$). The H1 pilot is a single-respondent experiment: one volunteer expert answered 40 benchmark questions under two counterbalanced conditions, one of which exposed the respondent to two fabricated confident answers attributed to fictitious peers. The respondent was told in advance that the study measured the influence of peer answers on accuracy, gave informed consent, and was debriefed on the fabricated-peer manipulation and its purpose. The protocol is minimal-risk; its pre-registration, released with the deposit, records the design, counterbalancing, consent, and debriefing. The response sheet is released under an anonymized respondent code, and no personal data was collected beyond it.

The failure mode documented here, manufactured consensus that reads as reliability, bears directly on high-stakes deployments of multi-agent pipelines, and the poisoned-verifier variant shows that it is adversary-accessible through the evidence channel. We disclose the injection methodology so that developers can audit their own systems. The adversary needs only one confident, type-consistent peer, a capability already present in any heterogeneous multi-agent deployment; no capability for harm is introduced beyond what existing frameworks expose.

USE OF GENERATIVE AI

Generative AI systems are the object of study in this work: all agent, injector, verifier, and judge behaviour reported in the Results is experimental data produced under the pre-specified protocols and is unedited.

The study itself was designed, driven, and judged by the authors. The research questions, hypotheses, experimental designs, pre-registered predictions and their failure branches, the registration and deviation-ledger discipline that governs the program, and every scientific judgment were formulated and made by the authors, who also directed, reviewed, and stands behind every line of the work. Within that framework, LLM-based assistants were used as research tools under our direction: to implement parts of the experimental harness and analysis scripts, to critique designs and review draft claims adversarially, to recompute reported statistics independently as a check on our own analysis code, to search and format references, and to draft and edit manuscript text. No assistant designed an experiment, wrote a prediction, chose a verdict, or decided what the paper claims. Every AI-assisted analysis was re-derived from raw data by the released analysis code; every AI-suggested reference was verified against its primary source; every substantive correction arising from these checks is recorded in the deviation ledger released with the deposit. We take full responsibility for all content.

REFERENCES

S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (ed.), *Groups, Leadership and Men*, pp. 177-190. Carnegie Press, 1951.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992-1026, 1992.

H. K. Choi, X. Zhu, and Y. Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Y. Cui, H. Fu, H. Zhang, and L. Wang et al. Free-MAD: Consensus-free multi-agent debate. *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

J. M. Durán and G. Pozzi. Trust and trustworthiness in AI. *Philosophy & Technology*, 38(1):16, 2025.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations (ICLR)*, 2021.

J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50-80, 2004.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9459-9474, 2020.

T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi. Encouraging divergent thinking in large language models through multi-agent debate. *Proceedings of EMNLP*, 2024.

J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020-9025, 2011.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Y. Okawa. Emergence of biased consensus in multi-agent large language model debates. *arXiv*, 2026.

P. Pitre, N. Ramakrishnan, and X. Wang. CONSENSAGENT. *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations (ICLR)*, 2023.

J. Wei, N. Karina, H. W. Chung, Y. Xin Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models with SimpleQA. *arXiv:2411.04368*, 2024.

E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209-212, 1927.

Y. Zhu, L. Gu, Z. Wang, H. Sang, D. Sui, and W. Tang et al. Auditing medical multi-agent AI reveals risks of false consensus. *arXiv:2510.10185*, 2025.

X. Zhu, C. Zhang, Y. Chi, and T. Stafford et al. Demystifying multi-agent debate: The role of confidence and diversity. *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026.

C. Dai. When consensus is not correctness: Diversity collapse and manufactured overconfidence in multi-agent LLM debate. *KDD 2026 Workshop*, 2026.

A. Wynn, A. Satija, and D. Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv:2509.05396*, 2025.

M. Cemri, M. Z. Pan, S. Yang, L. A. Agrawal, B. Chopra, R. Tiwari, K. Keutzer, A. Parameswaran, D. Klein, K. Ramchandran, M. Zaharia, J. E. Gonzalez, and I. Stoica. Why do multi-agent LLM systems fail? *arXiv:2503.13657*, 2025.

H. Zhang et al. Stop overvaluing multi-agent debate: We must rethink evaluation and embrace model heterogeneity. *arXiv:2502.08788*, 2025.

J. Oh, S. Jeong, J. Ko, and S. Yun et al. From belief entrenchment to robust reasoning in LLM agents (DReaMAD). *arXiv:2503.16814*, 2025.

Y. Pan, A. Ai, D. Simchi-Levi, M. Tambe, and A. Xu et al. Beyond majority voting: LLM aggregation by leveraging higher-order information. *arXiv:2510.01499*, 2025.

Y. Yang et al. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv:2505.22960*, 2025.

H. Wu, Z. Li, and L. Li. Can LLM agents really debate? A controlled study of multi-agent debate in logical reasoning. *arXiv:2511.07784*, 2025.

C. He et al. Minority sentinel: When to overturn majority voting in multi-agent LLM debates. *arXiv:2606.29270*, 2026.

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

P. Mavrodiev, C. J. Tessone, and F. Schweitzer. Effects of social influence on the wisdom of crowds. *Proceedings of the Conference on Collective Intelligence (CI-2012)*, 2012. *arXiv:1204.3463*.

Y. Tian, L. Wang, and F. Bullo. How social influence affects the wisdom of crowds in influence networks. *arXiv:2204.13610*, 2022. SIAM Journal on Control and Information Systems, 2023.

B. Jayles, H. Kim, R. Escobedo, et al. How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47):12620-12625, 2017.

Y. Geifman and R. El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624-635, 2021. doi:10.1145/3442188.3445923.

M. J. A. N. de Caritat, Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

C. List and R. E. Goodin. Epistemic democracy: Generalizing the Condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277-306, 2001.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *International Conference on Learning Representations (ICLR)*, 2024.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv:2308.08155*, 2023.

M. Minsky. *The Society of Mind*. Simon & Schuster, New York, 1986.

F. Herrera. Habituation and the human will: How passive reliance on human-AI collaboration erodes ethical agency. *Science and Engineering Ethics*, 2026.

A METHODS

A.1 CORE STUDY

The panel is claude-haiku-4-5, gpt-5.4-mini, and gemini-3.1-flash-lite (a heterogeneous weak tier), with four rounds of fully connected message passing and a fixed seat order rotated per question. The pool is 300 SimpleQA-style questions (Wei et al., 2024) screened for divergent starts: excluded are starts where the three isolated answers form any unanimous cluster (converging true, false, or abstaining together) and starts where all three are independently correct; partial two-of-three agreements are admitted by design, which is why the independent-vote counterfactual contains 134 two-of-three majorities. The screen rule is stated here as the deposit code implements it, and the count matches the design. Three of the 134 independent-majority questions lack a cached equivalence judgment and drop out of the pairing join used for the bootstrap and McNemar complements; the independent rate moves from 23.1% to 23.7%, inside every interval. The frontier replication carries the same complements (gap $+40.5$pp, bootstrap CI [0.237, 0.571]). The effect sizes throughout (headline $h = 1.09$; the registered judge-free replication adds $+15.8$pp on MMLU-hard, bootstrap [0.059, 0.259], with the anchor-rich GSM8K arm flat as predicted) make a qualitative reversal under a plausible design effect unlikely. Aggregation is by two-of-three majority with abstention as a first-class outcome; the independent-vote counterfactual applies the same rule to pre-deliberation answers on the same items. Denominators: the core-study collection contains 301 records because of one resume-time duplicate run record; the canonical basis de-duplicates by question (first occurrence kept), and every core-study number in the paper is reported on this 300-question basis.

A.2 JUDGE VALIDITY AND THE MEASUREMENT CORRECTION

Correctness and answer equivalence were first scored by a holistic LLM judge. That judge failed validation against blind human labels (16.7% agreement, $\kappa = 0.000$) and inflated convergence roughly 2.5-fold; it was discarded, and every number in this paper comes from its replacement, a decomposed pairwise judge (claude-sonnet-4-6) validated at 95.0% agreement with blind human labels, $\kappa = 0.888$ [0.765-0.974], $n = 80$. Two further judge families agree with it at 99.4% and 98.9% ($n = 180$), and a re-judging audit found 5 disagreements in 900 records, all in the conservative direction. Confidence bands (0.25/0.50/0.75/0.95) are ordinal and are used only for dynamics.

A.3 FRONTIER EXPERIMENTS: TWO EXPERIMENTS, DESCRIBED SEPARATELY

Generation-mixed exploratory run. The panel was claude-opus-4-8, gpt-4.1, and gemini-2.5-pro, with its own earlier registration record (OSF, anonymized view). This run mixed model generations and served to scope the frontier regime. It is exploratory, and no registered verdicts are drawn from it.

E1, the registered replication. The panel was claude-opus-5, gpt-5.5-2026-04-23, and gemini-3.1-pro-preview (generation-homogeneous), registered at https://osf.io/a5t3v/overview?view_only=be06a28afaf74a3f8cbc02d33cd337da; 280 items with 50 injection items; vendor-default decoding (the harness sent no temperature or top-p parameter; completions were

capped at 1,024 tokens), an asymmetry against the weak-tier settings that is disclosed because decoding affects both answer diversity and cascade dynamics. Vendor defaults are not versioned, so the model strings and collection dates in the deposit define the setting, and a re-collection should record the defaults then in force. Predictions P1-P4 and their verdicts are in Table 3.

Deviation entry (E1 audit trail). The registered artifact hashes and the executed artifact hashes differ: `eseries_harness.py 17708706...` → `193f2d96...`; `e1_frontier_rerun.py 12483b3c...` → `db60938d...`. The cause was a post-registration harness fix (Google batch shim, new file `google_batch_shim.py`, `1170fb75...`); the diff and the date are logged in the deposit. Per the registration's own rule, this entry is recorded here and in the audit trail.

EXP-10, the powered follow-up. EXP-10 (registered) has executed on the frontier panel, and it maps a boundary rather than extending the collapse. Only 22 of 160 items produced divergent starts, because capable agents rarely disagree. All 22 resolved to unanimity under deliberation, 19 true and 3 false, so the registered two-of-three endpoints are not evaluable on this panel: the minimum-events rule fires on an empty cell (zero non-unanimous deliberated majorities), and P10.1-P10.3 are reported as not evaluable rather than rescued. Descriptively, where truth-availability is high there is little divergence to exploit, and interaction manufactures mostly true consensus; the collapse requires the scarcity precondition. Read against E1, whose SimpleQA-family items remain truth-scarce even at the frontier and where deliberated falsity was 66.7%, EXP-10 is the regime boundary the Law's scope clause names: the effect is gated by independent truth-availability, a property of the pool and the panel together, rather than by the act of deliberation itself. Capability does not remove the risk; it relocates it to the frontier of model knowledge, which is where multi-agent systems are deployed to help.

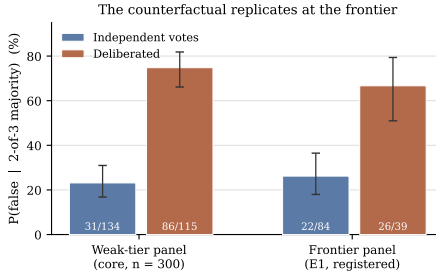

Figure 6: The diagnosticity counterfactual at weak and frontier tiers. Independent two-of-three agreement is false 23.1-26.2% of the time; after deliberation, 74.8% (weak) and 66.7% [51.0-79.4] (frontier, registered). No capability benefit was detected at registered power. Error bars are 95% Wilson intervals.

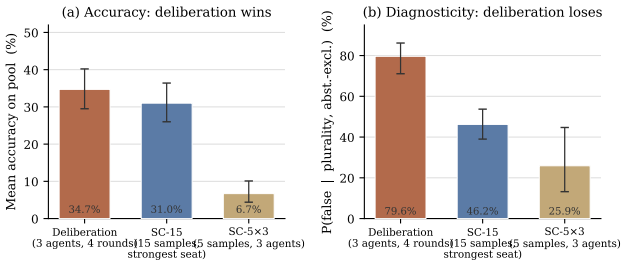

Figure 7: The matched-completion dissociation (EXP-1). Accuracy: deliberation 34.7%, self-consistency 31.0%, cross-model plurality 6.7%. Conditional false agreement: 79.6%, 46.2%, and 25.9%. Accuracy and diagnosticity come apart at an equal number of completions.

A.4 Injection, cross, and incumbency designs

Injection used $n = 100$ runs per condition, with the injected peer calibrated for confidence and type consistency. The cross used 3×3 injector families with 103 questions per cell; the anomalous same-family cell was independently replicated (0.27, then 0.30; pooled 0.286 [0.245-0.332]). Seeded incumbency used 18 cells, 1,320 runs, and 1,078 confirmatory runs on 60 common questions; P5.1-P5.4 were pre-specified, with P5.3's threshold amended in the registration before data collection.

A.5 Coverage-law measurements and the repair program

Simulated-coverage sweeps vary p_{truth} at $\{0.30, 0.50, 0.70, 0.85\}$ (four points); the fifth sweep point plotted in Figure 3, at $p_{\text{truth}} = 0.60$, comes from the companion p_{false} sweep, so the figure's x-axis is auditable as stated in the caption. Live retrieval used web search with no oracle

window; coverage was measured post hoc at $p_{\text{truth}} = 0.75$ on $n = 205$ screened questions. The paired grounded comparison ($n = 167$; 80 covered, 87 uncovered) used an evidence cache and channel that differ from the deposit's (deviations #7 and #8); the canonical re-pairing is pending, and the comparison is labeled oracle-windowed in the text. The poisoned-verifier observation was exploratory ($n = 48$ truth-scarce; high-anchor contrast at 57%), and its confirmatory program became EXP-7, EXP-7b, and EXP-7c (Appendix A.8).

The high-coverage program ran on the weak 3B panel with the programmatic judge. EXP-11 collected 150 valid MMLU records per arm at $p_{\text{truth}} = 0.90$ and 0.95 with a gold-asserting verified-evidence peer from round one; its registered ceilings were 15% and 10%. EXP-12 kept the design and replaced the assertion with a 7B verifier's checked prose verdict (the sweep and live-retrieval panels are capable API models that can engage with the content of evidence, which the 3B panel cannot); EXP-12c replaced it with a structured certificate (atomic sub-claims, per-seat check ballots, and an admission gate under which consensus counts only if two seats verify every load-bearing claim, with abstention otherwise); EXP-11b ran on GSM8K with checkable worked derivations and no ballot or gate. All four were registered before collection (`PREREGISTRATION_EXP11.md`, `PREREGISTRATION_EXP11B_EXP12.md` with amendments A1-A3), and the EXP-12 predictions were tested against EXP-11's locked interim comparator (49/143, 34.3%); the final EXP-11 draw is 51/150 (34.0%). A frontier-verifier arm (EXP-12d) is registered as a signed-off draft and has not been collected; it appears in Table 3 as pending.

A.6 EXP-1 MATCHED-COMPLETION ARMS

The arms used the Tier-2 canonical panel with $n = 300$ per arm and 15 completions per question in each arm; deliberation's token cost is higher because each round re-reads the peer messages. SC-15 draws 15 samples from the strongest seat, defined in the registration as the seat with the highest isolated accuracy on the released core-study answers, ties broken by vendor order: gemini-3.1-flash-lite, correct on 109/300 canonical questions in isolation against 49/300 for gpt-5.4-mini and 10/300 for claude-haiku-4-5 (ledger entry 97). Sampling in both arms used temperature 1.0, a registered and disclosed departure from the core-study sampling, and aggregation is by plurality; SC-5×3 draws 5 samples from each of the three agents. SC-15's plurality accuracy (31.0%) lies below the seat's single-sample isolated accuracy on the same pool (36.3%), which is possible when the truth is often not the seat's modal answer (the premise of Lemma 2); P1.3 was not confirmed. In both arms the plurality is formed by clustering the samples with an incremental variant of the validated pairwise judge and matching the plurality cluster to gold. The variant was checked against the full pairwise clustering on a 10-record sample before adoption (10/10 plurality-label agreement, 9/10 cluster-structure agreement; ledger entry 10); that check is small and is disclosed as such. Thresholds P1.1-P1.3 were pre-registered with inconclusive bands, and the binomial test for P1.1 was one-sided against 0.50.

A.7 EXP-6 EXECUTION RECORD

EXP-6 tests the counterfactual on two new substrates, GSM8K (multi-step arithmetic; numeric match against gold) and MMLU (broad-knowledge multiple choice; letter match), with a pre-registered stopping rule of $n \geq 150$ divergent-start questions per substrate or draw exhaustion. Correctness is judged programmatically, so the LLM judge is out of the measurement loop. The registration carries three amendments, all filed before the collections they govern: A1 restored three-family heterogeneity when a third local family became available; A2 extended the draw (the full GSM8K test split of 1,319 items and 1,500 MMLU questions, same shuffle seed, prefix-preserving) and defined the MMLU-hard subset (at most one of three isolated seats correct) after early screening showed that the divergent-start yield would otherwise exhaust below the $n \geq 150$ stop, with side-by-side reporting of the full and hard sets; A3 moved the primary arm to a local weak-tier panel after the mid-size host became unreachable, quarantining the 327 mid-size records as a suspended arm.

EXP-6 ran as two arms, and both are reported; the panel composition of every record is verified from the deposit rather than from run logs. The registered primary arm (amendment A3) is a local weak-tier three-seat panel (`qwen2.5:3b-instruct`, `gemma3:4b`, `llama3.2:3b`), four rounds at temperature 0.7, protocol identical to the core study, with 310 GSM8K records (159 divergent-start, yield 51%) and 277 MMLU records (180 divergent-start, yield 65%); it closed at the registered

stops on 2026-08-20. The second arm is the resumed mid-capability panel of the original design (amendment A1: `qwen3.6:35b`, `qwen3-coder:30.5B`, `gemma3:27b`), quarantined when the host was lost and resumed on its return, with 455 GSM8K records (260 divergent-start, yield 57%) and 311 MMLU-hard records (150 divergent-start, yield 48%); 100% of its judged records carry the A1 seats. The raw collection files hold 490 and 375 records; the excess over the analyzed corpus is exactly 35 and 64 error records (failed-generation tombstones carrying an `error` field, excluded from analysis and disclosed in the ledger), so the analyzed corpus is the full valid corpus. The claims' scope is these sub-populations.

Inference followed the Tier-2 common policies: dual conditioning, Wilson intervals beside 10,000-resample question-cluster bootstrap intervals (bootstrap primary on discordance), Fisher exact contrasts, paired McNemar tests on both-majority intersections, and Benjamini-Hochberg correction within the EXP-6 family. Twenty-three ledger entries cover the program's deviations, among them a reasoning-model truncation bug caught by a six-question pilot (the pilot records are quarantined) and four interim-analysis inspections, each disclosed when it occurred; the definitive analyses ran at the registered stops and supersede the interim outputs.

Table 2: Unconditional false-agreement rates, $\mathrm{P}(\text{false agreed answer})$ over all questions, for the regimes of Table 1; gaps in percentage points.

Regime	Substrate	Judge	Panel	Independent	Deliberated	Gap
Core (post hoc)	SimpleQA	LLM	weak API tier	10.3%	56.7%	+46.4
EXP-6 (registered)	GSM8K	programmatic	local weak 3B	15.7%	25.2%	+9.4
EXP-6 (registered)	MMLU	programmatic	local weak 3B	45.0%	50.6%	+5.6
EXP-6 (registered)	MMLU-hard	programmatic	mid-capability 30B	22.0%	31.3%	+9.3
EXP-6 (registered)	GSM8K	programmatic	mid-capability 30B	4.2%	5.4%	+1.2

The judge repair. A judge-level audit found and fixed a parse failure class in the pipeline's answer extraction: template-scaffold echoes (unfilled `ANSWER`/`CONFIDENCE` structure text) leaking into recorded long-form answers, which programmatic matching then scored as false. The repaired pipeline was validated against a frontier judge on a 100-item sample of the resumed arm's corpus (95.0% agreement, $\kappa = 0.756$; the registered $\kappa \geq 0.7$ criterion met; five residual discordants, all scaffold-echo records, disclosed). Two integrity statements follow. First, the local arm is invariant under the repair: identical headline counts before and after, and zero verdict changes, so its collapse is behavior rather than artefact. Second, the resumed arm is reported on the repaired basis throughout; its first analysis on the pre-repair basis had shown a far larger GSM8K collapse, and the repair reversed the P6.2 verdict on that arm from failed to confirmed. Both arms' pre-repair artefacts are retained in the deposit, and the repair is a ledger entry.

Formation margin. The two arms behave differently at the formation margin, and both are disclosed. On the local weak panel, deliberation reduced majority formation on GSM8K (74.2% to 31.4%, partly format decay in long finals) while manufacturing the program's largest conditional falsity; on the resumed arm, formation moved from 93.5% to 97.7% (GSM8K) and from 76.0% to 70.0% (MMLU-hard). The Law conditions on formed majorities and is unaffected; the frequency claims are scoped to panels that converge, and the divergence is itself reported.

A.8 Boundary experiments: full reports

Five registered experiments, each aimed at a disclosed soft spot, have completed; verdicts are reported as registered, thresholds included (`PREREGISTRATION_EXP7_EXP8.md`, amendments A1-A2; `PREREGISTRATION_EXP7B.md`; `PREREGISTRATION_EXP7C.md`).

EXP-7 and EXP-7b: the poisoned channel has a measured boundary. The exploratory corollary (a poisoned verifier adopted 48/48 on truth-scarce SimpleQA) fails to reproduce with thin template evidence on the local weak-tier panel: truth-scarce adoption was 4.8% [0.8-22.7] on GSM8K and 24.3% [13.4-40.1] on MMLU against a 0.80 threshold, and all three pre-registered predictions failed. With realistic 27B-authored prose, adoption rises but stays bounded (41-56% on small truth-scarce strata), and the GSM8K rescue arm confirms (false consensus 3.1% [1.1-8.7]). Read with

the content ablation (Section 3.3), the evidence points to the mechanism: what propagates is a plausible, type-consistent claim, and the 48/48 rate belongs to its original configuration. A powered item-matched test (EXP-7c, MMLU) sharpened the result: the realism of the supporting evidence is irrelevant, since thin template and 27B-authored realistic prose produced identical truth-scarce adoption (47.4% against 47.4%, 19 matched items, McNemar 5/5, one-sided $p = 0.623$); agents respond to the claim, not to its justification. The deployment warning is scoped rather than withdrawn: measured adoption of evidence-labelled falsehoods ranges from 4.8% to 100% across configurations, with a powered truth-scarce estimate of 43.1% [31.2-55.9] on the local panel (full verdicts in Appendix A.9).

EXP-8: the practical claim survives live retrieval. The paired grounded comparison was re-run with live Wikipedia retrieval on MMLU (byte-identical evidence across arms, $n = 316$; the coverage channel was weak, $p_{\text{hit}} = 0.025$, and 119 records were transport-failed and excluded from strata). Grounded deliberation produced false consensus at 42.4% [37.1-47.9] against 21.8% [17.6-26.7] for grounded independent voting on the same evidence: gap $+20.6$pp, paired bootstrap [0.136, 0.272], McNemar 101 to 36, $p = 2.5 \times 10^{-8}$ (P8.1 confirmed). The oracle-window caveat is addressed under a live but weak channel (with $p_{\text{hit}} = 0.025$, nearly all items were retrieval misses): retrieve, then aggregate independently, is the recommendation under real retrieval and not only under clean evidence, although stronger retrievers remain untested. P8.3 is reclassified as not evaluable, because its registered test (the arm-I rate within 10 points of the coverage-rule value) could not fail: the dropped hit term is bounded by the measured coverage (2.5 points against a 10-point tolerance), so the test was a tautology, and the ledger says so. The hit-stratum concentration test (P8.2) was underpowered (8 hits) and is reported as not confirmed.

H1: a human baseline (single-expert pilot). The objection that the collapse is an artifact of language models is answered by the human social-influence literature (Appendix B); a single-respondent pilot adds a check in the same direction and carries no evidential weight of its own: one AI-specialist respondent, 40 MMLU-hard items, counterbalanced isolation against exposure to two confident wrong peer answers. Accuracy fell from 55.0% (11/20) in isolation to 37.5% (6/16) under exposure, and the respondent adopted the majority's wrong option on 18.8% of answered items, the same direction of accuracy-truth decoupling as every LLM panel in this paper. One behavior diverged: abstention rose from 0% to 20% under exposure, whereas the LLM panels abandon abstention under pressure. The option of declining to answer was preserved by the human and not by the models. With $n = 1$, this single-expert pilot is qualitative corroboration only; a powered scale-up is registered as future work.

EXP-14 and EXP-14b: visibility versus revision. The matched-rounds blind control (EXP-14, registered) ran on the same pool and panel as EXP-6, with the same agents and budget and with each agent's own previous answer echoed in place of the peer lines. Four rounds of private revision without visibility keep conditional falsity inside the registered bound: it moves by -0.5 points [-0.014, 0.000] on GSM8K and by $+3.6$ points [0.000, 0.089] on MMLU-hard (P14.1 confirmed on both substrates). On MMLU-hard the solo arm's majority formation collapses (34.9% against 68.6% independent and 79.9% deliberated; P14.3 failed in the stronger direction), so the manufacture of agreement requires an audience. The deliberated-versus-solo conditional contrast on the divergent-start join was underpowered as registered (-1.9pp [-9.6, $+3.8$], $n = 52$; P14.2 not confirmed). The registered amendment (A2, filed when the conditional join proved limited to $n = 52$ and before the analysis it adds was run) answered the question at the deployment-relevant estimand: unconditionally, deliberation produces $+14.2$pp [8.3, 20.7] more false agreed answers per question than solo revision on MMLU-hard (McNemar 28 to 4, $p = 1.9 \times 10^{-5}$; A2.1 confirmed) and $+2.1$pp [0.4, 4.1] on GSM8K, while solo revision stays within $+10$pp of independent voting unconditionally (GSM8K -1.2pp, MMLU-hard -3.0pp; A2.3 confirmed). Where both screens agree that an item is truth-scarce, no difference between arms was detected among formed majorities (39%, 58%, and 47%; A2.2 not confirmed, MMLU-hard -4.5pp [-18.2, $+9.1$], $n = 22$), so the evidence locates visibility's damage in the manufacture of the agreement event; a worsening of the majorities that form was not detected, which at this sample size does not show that none exists. A powered three-arm replication on a fresh draw (EXP-14b, 166 divergent-start items) confirmed the edge independently: P14b.1 (revision alone, inside its registered $+10$pp bound) $+2.8$pp [-2.8, $+8.5$]; P14b.2

unconditional +17.5pp [10.2, 24.7], McNemar 35 to 6, $p = 4.9 \times 10^{-6}$; and the registered ordering holds ($0.759 > 0.747 > 0.446$).

EXP-9: the collapse holds at five seats. On a five-seat local panel, deliberated-majority falsity reaches 84.4% [74.7-90.9] against 55.5% [46.8-63.8] independent (Fisher $p = 1.2 \times 10^{-5}$; paired bootstrap [0.179, 0.399]; P9.1 confirmed), with non-unanimous-majority falsity of 82.9% [72.4-89.9] (P9.2 confirmed), within two points of the five-agent SimpleQA signature in a different ecosystem.

A.9 Extended experimental detail

TriviaQA and five-agent replications of the gradient. On TriviaQA ($n = 81$ screened), false consensus is 19.8% [12.5-29.7], and false majorities again begin knowledge-poorer than true ones (0.50 against 1.72 agents correct). A five-agent replication on fresh SimpleQA items ($n = 100$) reproduces the pattern at 80.8% [62.1-91.5] non-unanimous falsity (21/26), again governed by initial knowledge rather than majority strength (97.8% false unanimous consensus at zero independent anchors).

Five-agent replication (fresh SimpleQA, $n = 100$). A majority formed on 97 of 100 questions. Non-unanimous majorities were false at 80.8% (21/26) [62.1-91.5], flat across majority strength: 76.9% at three-of-five, 84.6% at four-of-five, and 71.8% at five-of-five. The governing variable is initial knowledge rather than majority size: unanimous consensus was false at 97.8% (44/45) when no agent knew the answer in isolation, 30.0% with one anchored agent, and 16.7% with two (trend $p = 9 \times 10^{-10}$). The five-agent screen admits a different knowledge mix than the three-agent screen, so cross-size comparisons are made only knowledge-conditioned.

Injector cross, family level. A fully crossed 3×3 injector-family design (103 questions per cell, the anomalous cell independently replicated) shows that propagation is family-robust: every injecting family drives adoption to 82.5-94.2%. The residual heterogeneity sits in round-zero pre-agreement and in receiver-family update propensity (adoption 0.075 against 0.562 across families, extreme-pair Fisher $p = 1.4 \times 10^{-32}$). Panel size bounds the attack: one adversary's falsehood falls to 11% adoption at four or more honest agents and saturates there.

Injector cross, cell-level detail. The one resistant configuration, a same-family receiver pair facing the weakest injector, adopted at 0.27, replicated at 0.30, pooled 0.286 [0.245-0.332]. In about two thirds of the resistant runs both receivers already agreed on the same wrong answer at round zero (48/70 and 43/68 in the original and the replication), which is why incumbency survives as a descriptor of the cross and nothing more. Conditional on a peer adopting, receiver adoption is uniform at 0.82-0.89 across families. Two adversaries at panel size seven held adoption at 18.0% ($p = 0.16$, not significant).

Seeded incumbency: the full prediction table. Four predictions were pre-specified across 18 cells (1,320 runs; 1,078 confirmatory on the 60 questions common to all cells). All four failed at their registered thresholds.

Prediction	Registered content	Outcome	Verdict
P5.1	Seeded false incumbency lowers adoption of a rival falsehood	Adoption *rose* 0.371 → 0.453 ($p = 0.0019$); most resistant family 0.075 → 0.227	Failed, reversed
P5.2	Incumbency × injection interaction	$p = 0.070$	Not significant
P5.3	Protection effect ≥ 20pp	8.4pp	Failed threshold
P5.4	Effect range across cells ≤ 10pp	0.652-0.835 (18.3pp)	Failed threshold

What the refutation leaves standing: receiver-family update propensity is the dominant stable moderator (adoption 0.075, 0.475, and 0.562 across families, extreme-pair Fisher $p = 1.4 \times 10^{-32}$); seeded truth survives at 0.821 against 0.121 for seeded falsehood ($p = 5.7 \times 10^{-58}$); 23.2% of false-seeded outcomes were novel errors matching neither the seed, the injection, nor gold; and double-seeded truth still converged false in 49.2% of panels.

EXP-7, EXP-7b, and EXP-7c: full verdicts. EXP-7 (thin synthetic poison evidence, local weak panel): truth-scarce adoption 4.8% [0.8-22.7] on GSM8K and 24.3% [13.4-40.1] on MMLU against the registered 0.80 threshold; rescue leaked at 16.5% and 27.0%; P7.1-P7.3 all failed. The 48/48 exploratory rate does not travel to this configuration. EXP-7b (27B-authored realistic poison prose): adoption 55.6% [26.7-81.1] on GSM8K ($n = 9$ truth-scarce) and 41.2% [21.6-64.0] on MMLU; the GSM8K rescue arm confirmed at 3.1% [1.1-8.7]; P7.3 confirmed on GSM8K only. EXP-7c (powered item-matched follow-up, MMLU, 19 matched items): thin and realistic evidence produced identical truth-scarce adoption, 47.4% against 47.4% (McNemar 5/5, one-sided $p = 0.623$); the powered truth-scarce adoption estimate on the local panel is 43.1% [31.2-55.9]. The mechanism sentence this program supports is that agents respond to the claim rather than to its justification.

EXP-12 verdict arm: exploratory localization of the null. Post-hoc inspection of the prose-verdict arm, labeled exploratory, localizes its null result in the verifier's competence. The 7B verifier contradicted the correct reference on 63% of covered items (88/139), and the panel amplified whichever stance it took: false consensus ran 9% (4/46) behind affirmations of the reference against 43% (38/88) behind rejections. A frontier replay of the identical prompt on the identical items affirmed the reference on 83% (113/136) of items, against 33% for the 7B verifier. The verdict arm's null is therefore a verifier-competence effect on a hard substrate rather than evidence that verdict voices cannot work.

Table 3: Registration ledger: every prediction, its status, and its outcome. Intervals are 95% Wilson intervals. Verdicts use the registered or pre-specified thresholds where they exist.

#	Prediction	Registered?	Result	Verdict
Core-P1	2-of-3 false consensus > chance (50%)	Pre-registered (Tier-0)	74.8% [66.1-81.8], $p = 9.8 \times 10^{-8}$	**Confirmed**
Core-P2	Deliberation improves mean accuracy on divergent-start pool	Pre-registered	13.7% → 34.7% (canonical de-duplicated pool), $p < 10^{-11}$	**Confirmed**
Core-P3	Truth-scarcity gradient (false majorities begin knowledge-poorer)	Pre-registered	0.20 vs 1.26 agents correct	**Confirmed**
Core-P4	Confidence rises with convergence regardless of truth	Pre-registered	0.673→0.704 (false), 0.740→0.787 (true)	**Confirmed**
§3.1 counterfactual	Deliberated agreement less diagnostic than independent	Post hoc (labeled)	74.8% vs 23.1%, Fisher $p = 2.1 \times 10^{-16}$	Exploratory; replicated as E1-P2
Inj-F4	One confident truth collapses false consensus	Pre-registered	100% → 1%, $p = 2.2 \times 10^{-57}$	**Confirmed**
Inj-F5	One confident falsehood propagates	Pre-registered	91.0% adoption [83.8-95.2]	**Confirmed**
Ablation	Type-inconsistent assertion rejected	Pre-registered	0/50 vs 95%, $z = 11.38$	**Confirmed**
Cross	Propagation family-robust	Pre-registered	82.5-94.2%, all families	**Confirmed**
Incumbency P5.1	Shared false incumbent lowers adoption ≥30pp	Pre-specified	Reversed: 0.371 → 0.453, $p = 0.0019$	**Failed**
Incumbency P5.2	Protection rank-ordered by update propensity	Pre-specified	Interaction $p = 0.070$, n.s.	**Failed**
Incumbency P5.3	B-adoption asymmetry ≥20pp (amended)	Pre-specified (amended)	8.4pp	**Failed**
Incumbency P5.4	Composition-invariant conditional adoption	Pre-specified	Range 0.652-0.835 (18.3pp)	**Failed**
Coverage law	FalseRate $\approx 1 - p_{\text{truth}}$	Predicted before sweep	$r = -0.996$; live 20.5% vs 25.0% predicted	**Confirmed**
Attenuation	Slope falls on high-anchor pool	Registered	1.13 → 0.72, $p = 0.010$	**Confirmed**
Attenuation (logistic interaction)	Registered test statistic	Registered	Undefined (complete separation, 57/57)	Not evaluable (as registered)
Attenuation (quadratic ϕ)	Registered functional form	Registered	Overpredicts decline	**Rejected**
E1-P1	Frontier false consensus < weak tier	Registered (OSF)	66.7% vs 74.8%, $p = 0.22$	**Failed**
E1-P2	Deliberated > independent falsity (frontier)	Registered	66.7% vs 26.2%, Holm $p = 6.9 \times 10^{-5}$	**Confirmed**
E1-P3	Scarcity gradient (frontier)	Registered	87.5% vs 33.3%, Holm $p = 0.0016$	**Confirmed**
E1-P4	Anchored-stratum adoption contrast	Registered	2/8 vs 1/42 (events < 10)	Not evaluable
EXP1-P1.1	SC-15 ≤ 50% (cond. excl.)	Pre-registered (Tier-2)	46.2% [39.0-53.7], binomial $p = 0.181$	Not confirmed (descriptively supportive)
EXP1-P1.2	SC-5×3 ≤ SC-15 + 5pp	Pre-registered	2.3% vs 26.7% unconditional	**Confirmed**
EXP1-P1.3	SC-15 accuracy ≥ 33.2%	Pre-registered	31.0% [26.0-36.4]	Not confirmed
EXP-5b paired	Grounded deliberation worse than grounded vote	Pre-specified decision rule	12.6% vs 4.2%, $p = 0.0092$	**Confirmed** (oracle-windowed pairing)
EXP-6 P6.1 (local 3B arm)	Deliberated-majority falsity > independent	Pre-registered (amendments A1-A3)	MMLU 84.3% [76.2-89.9] vs 57.4% [49.2-65.3], Fisher $p = 3.41 \times 10^{-6}$; GSM8K 80.0% [67.0-88.8] vs 21.2% [14.8-29.4], Fisher $p = 9.27 \times 10^{-13}$	**Confirmed** (both substrates)
EXP-6 P6.2 (local 3B arm)	Gap smaller on GSM8K than MMLU (anchor attenuation)	Pre-registered	Opposite direction: +58.8pp vs +26.8pp, DiD boot [−0.482, −0.152]	**Failed** (operative-anchor refinement)
EXP-6 P6.1 (resumed 30B arm)	Same, MMLU-hard stratum	Pre-registered	44.8% [35.6-54.3] vs 28.9% [21.4-37.9], Fisher $p = 0.011$, paired boot [0.059, 0.259]	**Confirmed**
EXP-6 P6.2 (resumed 30B arm)	Same attenuation	Pre-registered	+1.0pp vs +15.8pp, DiD boot [0.045, 0.252], $p = 0.0024$	**Confirmed**
EXP-6 P6.3	Deliberation suppresses abstention, both substrates	Pre-registered	Neither panel abstains at meaningful rates; MMLU-hard deliberated abstention outcomes rose (1/150 → 11/150)	Not supported (panel-dependent)
EXP-6 judge repair	Template-scaffold parse leaks in answer extraction	Deviation entry	Concordance $\kappa = 0.756$ on the resumed arm's corpus, criterion met; local arm invariant under repair (identical counts, zero verdict changes)	Measurement correction, disclosed (§3.6)
EXP-10 P10.1-P10.3	Frontier shrinkage / persistence / scarcity	Registered	0 non-unanimous deliberated majorities (all 22 divergent starts resolved unanimous); min-events rule fires	Not evaluable (empty cell)

continued on next page

Registration ledger (continued)

#	Prediction	Registered?	Result	Verdict
EXP-10 boundary	Divergent-start yield at the frontier	Registered	22/160 divergent; deliberated unanimity 19 true / 3 false (descriptive)	Boundary mapped (scarcity gate)
EXP-13 P13.1	Endpoint decay $F_4 > F_0$, both substrates	Analysis registered	MMLU-hard $+11.1$pp [0.012, 0.210], raw $p = 0.0206$, BH $q = 0.041$; GSM8K flat ($+0.4$pp)	Not confirmed at the registered correction (borderline)
EXP-13 P13.2	First-exchange share of total decay ≥ 0.50	Analysis registered	MMLU-hard 0.629; GSM8K: no decay to decompose	**Confirmed** (MMLU-hard)
EXP-13 P13.3	Round-series monotonicity (descriptive)	Analysis registered	Falsity jumps at round 1 then plateaus (MMLU-hard); flat (GSM8K)	Descriptive; direction erratum logged
EXP-14 P14.1	Revision alone does not destroy diagnosticity (bound +10pp)	Registered	GSM8K -0.5pp [-0.014, 0.000]; MMLU-hard $+3.6$pp [0.000, 0.089]	**Confirmed** (both substrates)
EXP-14 P14.2	Deliberated $>$ solo falsity on MMLU-hard	Registered	-1.9pp [-9.6, $+3.8$], $n = 52$	Not confirmed (joint-effect reading registered)
EXP-14 P14.3	Solo formation within 10pp of independent	Registered	GSM8K $+3.3$pp (in band); MMLU-hard -33.7pp (formation collapses without visibility)	**Failed** on MMLU-hard, in the stronger direction
EXP-14 A2.1	Unconditional falsity, deliberated $>$ solo, MMLU-hard	Registered amendment (2026-09-07)	$+14.2$pp [8.3, 20.7], McNemar 28/4, $p = 1.9 \times 10^{-5}$; GSM8K $+2.1$pp [0.4, 4.1]	**Confirmed**
EXP-14 A2.2	Strict both-screens subset: $F_{\mathrm{del}} > F_{\mathrm{solo}}$	Registered amendment	MMLU-hard -4.5pp [-18.2, $+9.1$], $n = 22$; GSM8K $+2.4$pp [0.0, 6.0], $n = 84$	Not confirmed (formation channel carries the damage)
EXP-14 A2.3	Unconditional solo $-$ ind $\leq$ +10pp	Registered amendment	GSM8K -1.2pp; MMLU-hard -3.0pp	**Confirmed**
EXP-14b	Powered three-arm replication of the visibility edge	Registered	166 divergent-start on the fresh draw: P14b.1 revision alone $+2.8$pp [-2.8, $+8.5$], inside the $+10$pp bound, confirmed; P14b.2 unconditional $+17.5$pp [10.2, 24.7], McNemar 35/6, $p = 4.9 \times 10^{-6}$; P14b.3 ordering holds ($0.759 > 0.747 > 0.446$)	**Confirmed**
EXP-7 P7.1	Thin-poison adoption ≥ 0.80 (0-correct stratum)	Pre-registered	4.8% [0.8-22.7] GSM8K; 24.3% [13.4-40.1] MMLU	**Failed** (both)
EXP-7 P7.2	Anchor protection gradient under thin poison	Pre-registered	Reversed ordering (adoption rises with anchors)	**Failed**
EXP-7 P7.3	Rescue false consensus ≤ 0.10 (thin template)	Pre-registered	16.5% [12.0-22.3] GSM8K; 27.0% [21.3-33.5] MMLU	**Failed** (both)
EXP-7b P7.1	Realistic-poison adoption ≥ 0.80 (0-correct)	Pre-registered	55.6% [26.7-81.1] GSM8K ($n = 9$); 41.2% [21.6-64.0] MMLU	**Failed** (both)
EXP-7b P7.2	Anchor protection under realistic poison	Pre-registered	GSM8K not evaluable ($n < 10$); MMLU n.s. ($p = 0.232$)	Not confirmed
EXP-7b P7.3	Rescue false consensus ≤ 0.10 (realistic)	Pre-registered	GSM8K 3.1% [1.1-8.7], $p = 0.0099$; MMLU 35.9% [26.1-47.0]	**Confirmed** (GSM8K) / **Failed** (MMLU)
EXP-7b P7b.3	Realistic $>$ thin adoption by $\geq$ 30pp, item-matched	Pre-registered	GSM8K $+37.5$pp ($n = 8$, McNemar $p = 0.25$); MMLU $+8.3$pp ($n = 12$, $p = 1$)	Not confirmed (underpowered)
EXP-8 P8.1	Grounded deliberation $\geq$ grounded vote $+5$pp, live retrieval	Pre-registered	42.4% [37.1-47.9] vs 21.8% [17.6-26.7], paired boot [0.136, 0.272], McNemar $p = 2.5 \times 10^{-8}$	**Confirmed**
EXP-8 P8.2	Gap concentrates in retrieval-miss stratum	Pre-registered	8 hits; DiD boot [-0.290, 0.706]	Not confirmed (underpowered)
EXP-8 P8.3	Arm-I rate tracks coverage rule within 10pp	Pre-registered	Not evaluable: the identity holds by construction (hit term bounded by $p_{\mathrm{hit}} = 0.025$)	**Not evaluable** (reclassified)
EXP-9 P9.1	Five-seat collapse: deliberated $>$ independent	Pre-registered	84.4% [74.7-90.9] vs 55.5% [46.8-63.8], Fisher $p = 1.2 \times 10^{-5}$	**Confirmed**
EXP-9 P9.2	Non-unanimous falsity ≥ 0.60 at five seats	Pre-registered	82.9% [72.4-89.9], binomial $p = 3.4 \times 10^{-5}$	**Confirmed**
EXP-7c P7c.1	Realistic $>$ thin evidence adoption ($\geq$20pp, item-matched)	Pre-registered	47.4% vs 47.4% (19 items, McNemar 5/5, $p = 0.623$)	**Failed** (claim, not evidence, is the vector)
EXP-7c P7c.2	Realistic-evidence 0-correct adoption ≥ 0.60	Pre-registered	50.0% [31.4-68.6] ($n = 24$)	**Failed**
EXP-7c P7c.3	Realistic rescue false consensus ≤ 0.10 (MMLU)	Pre-registered	31.2% [23.7-39.8] ($n = 125$)	**Failed**
EXP-11 P11.1	Residual false consensus ≤ 0.15 at $p_{\mathrm{truth}} = 0.90$ (weak panel)	Pre-registered (OSF memo 2026-08-28)	34.0% [26.9-41.9] (51/150), binomial $p = 5.8 \times 10^{-9}$	**Failed**: capability gate on the rescue term
EXP-11 P11.2	Residual false consensus ≤ 0.10 at $p_{\mathrm{truth}} = 0.95$ (weak panel)	Pre-registered	26.7% [20.2-34.3] (40/150), $p = 5.8 \times 10^{-9}$	**Failed**

continued on next page

Registration ledger (continued)

#	Prediction	Registered?	Result	Verdict
EXP-11 P11.3	Both points within 5pp of the law's prediction	Pre-registered	$\|\text{diff}\| = 0.21$ and 0.18 vs tolerance 0.05	**Failed** (law refined: operative-coverage gate)
EXP-11b P11b.1/2	Derivation-evidence rescue ≤ 0.15 / ≤ 0.10 (GSM8K, weak panel)	Pre-registered (v2, amend. A1-A3)	4.0% [1.8-8.5] at 0.90; 2.7% [1.0-6.7] at 0.95	**Confirmed** (both; BH $q \leq 0.0011$)
EXP-11b P11b.3	Within 5pp of the ungated prediction	Pre-registered	$\|\text{diff}\| = 0.091, 0.057$	**Failed**: observed *better* than predicted
EXP-12 P12.1-P12.4	Prose-verdict rescue ≤ 0.15/≤ 0.10; drop $\geq$ 10pp vs EXP-11	Pre-registered	34.0% [26.9-41.9] at 0.90; 28.0% [21.4-35.7] at 0.95; drop 0.3pp	**Failed** (all): capability alone does not rescue
EXP-12c P12c.1/2	Certificate-arm residual ≤ 0.15 / ≤ 0.10	Pre-registered	18.5% [13.2-25.5] at 0.90 ($p = 0.135$); 14.7% [9.9-21.2] at 0.95 ($p = 0.044$); abstention 7.9-9.3%	Not confirmed (marginal; descriptively supportive)
EXP-12c P12c.4	Certificate drop $\geq$ 10pp vs EXP-11's locked 0.343	Pre-registered	Drop 15.7pp, Fisher $p = 0.0016$	**Confirmed** (BH $q = 0.008$)
EXP-12c P12c.5	Ballot feasibility $\geq 60\%$	Pre-registered	0/139 complete (strict rule: all seats $\times$ all claims)	**Failed** as operationalized; gate engaged correctly (ledger #72)
EXP-12d P12d.1-P12d.4	Frontier-verifier arm of the repair ladder (delta registration against EXP-11b/12)	Registered as a signed-off draft (2026-09-03)	No collection has been run	Pending (registered, not executed)

A.10 LIMITATIONS: FULL TABLE

Limitation	Why the core claim survives
Substrate scope	The counterfactual is registered and confirmed on three substrate classes (SimpleQA, GSM8K, MMLU), with the anchor gradient and the TriviaQA replication bounding the rest. Open: contested or subjective questions, non-English pools, and multi-hop or expert-level reasoning (HotpotQA, GPQA, MATH), where a panel might construct an anchor collaboratively; the operative form predicts that such construction protects only the seats able to check the constructed chain.
LLM-mediated measurement	Human-validated judge ($\kappa = 0.888$ [0.765-0.974], $n = 80$), cross-family checks ($> 98.9\%$), and EXP-6's judge-free replication on two panels with the same qualitative result.
Frontier sample size	The 39-event E1 contrast is disclosed at registered power; P2 is Holm-significant regardless; the powered follow-up (EXP-10) has executed and maps the boundary, with 22/160 divergent starts and all 22 resolving to unanimity (19 true), so its two-of-three endpoints are not evaluable (Appendix A.3).
Interaction versus revision	The headline arm differs from the independent arm in social visibility and in four revision opportunities. The registered blind control (EXP-14) has executed: revision without visibility keeps conditional falsity inside the registered bound and scatters consensus; the deliberated-versus-solo falsity contrast is underpowered on the join ($n = 52$), so the loss of diagnosticity is claimed as the joint effect, and visibility is credited with manufacturing false agreement rather than with worsening the majorities that form (§3.1).
Verbalized confidence	Ordinal bands are used only for dynamics, never as interval-scale measurements.
Single scaffold, one window, English	The protocol is fully connected and unstructured; structured topologies (roles, critique, hierarchical aggregation) are untested. The injector cross is family-robust (82.5-94.2% across nine pairs). Proposition 2 binds any protocol that manufactures agreement on the truth-scarce stratum, so a structured protocol escapes it only by declining to manufacture agreement or by grounding (Appendix B).
Operative verification is inferred	Verification capability is not measured directly. The proxy is the independent arm's accuracy on the same items, measured before any interaction, so it is not derived from the collapse it explains. The operative form was adopted at the local arm's closure (ledger entry 22), before the resumed arm was analyzed, and the registered tests since then are P11b.1-2 and P12c.4 (confirmed) and P12.1-4 (failed, as the form expects for evidence the panel cannot check). Its falsifier, per-seat verification accuracy measured outside deliberation, is untested.
Human baseline	The single-respondent pilot (H1) is corroboration only; the artifact objection is answered by the human social-influence literature (Appendix B), not by H1.
Benchmark exposure	The divergent-start screen removes recalled items, and the counterfactual is internal to each item, so exposure cancels in the contrast.
Panel-dependent dynamics	The two EXP-6 panels disagree at the formation margin (the local 3B panel's GSM8K deliberation scatters, 74.2% $\rightarrow$ 31.4%) and neither abstains at meaningful rates; the Law conditions on formed majorities and is unaffected, and frequency claims are scoped to panels that converge.
Coverage guarantee's scope	The sweep and live-retrieval confirmation ran on capable API panels; EXP-11 measured the bound on the weak tier and it fails there (34.0% and 26.7% against 15% and 10% ceilings), so the law is restated with the operative-verification gate, a measured boundary rather than an extrapolation.
Follow-up constraints	EXP-8's live channel was thin ($p_{\text{hit}} = 0.025$; 119/316 transport failures excluded); EXP-7b's strata are small; EXP-7's evidence was synthetic by design; EXP-12c's certificate ballots were infeasible under the strict rule (0/139 complete), so its gate engaged on partial ballots.

A.11 THREATS TO VALIDITY AND THEIR HANDLING

Threat	Hostile reading	Handling (and where)
Paired structure of the headline test	Fisher's test on the 2×2 table ignores same-question pairing	Bootstrap CI [0.406, 0.616] and McNemar 24-to-2 ($p = 1.05 \times 10^{-5}$) printed beside Fisher's p (§3.1); E1 complements alongside (§3.5)
Post-hoc provenance of the core contrast	The central result rests on an exploratory re-analysis	Registered frontier replication (E1-P2, Holm $p = 6.9 \times 10^{-5}$) and registered EXP-6 on two substrates and two panels (§3.6)
Visibility versus revision	The headline arm also adds four revision opportunities	Registered blind control EXP-14: private revision keeps conditional falsity inside the registered bound; visibility adds $+14.2$pp false agreed answers per question, replicated at $+17.5$pp in EXP-14b; no effect on the falsity of formed majorities is claimed (§3.1)
Small-n claims	Frontier P2 rests on 39 events; the EXP-10 boundary on 22	The claims the paper rests on are adequately powered (EXP-6, EXP-8, EXP-9, EXP-14b); the 39-event null is reported as absence of detected benefit, not absence of benefit; the 22-item EXP-10 stratum is a boundary observation, labeled as such and not load-bearing
LLM-judge circularity	The measurement model belongs to a panel family	Human-validated decomposed judge ($\kappa = 0.888$ [0.765-0.974]), two other judge families at $> 98.9\%$, and EXP-6's programmatic judge with no LLM in the loop and the same qualitative result
Training contamination (GSM8K/MMLU)	Models may have seen benchmark items	The divergent-start screen removes unanimous-correct items, and the counterfactual is internal (same items, same agents, both arms), so exposure cancels in the contrast
Oracle-evidence practical claim	The recommendation rests on clean evidence	Addressed: EXP-8 replicates under live Wikipedia retrieval with a weak channel ($p_{\text{hit}} = 0.025$), McNemar $p = 2.5 \times 10^{-8}$ (§3.8)
Coverage guarantee's scope	The sweep and live-retrieval confirmation ran on capable API panels	Measured on the weak tier by EXP-11: the bound fails there (34.0%/26.7% against 15%/10% ceilings), the law is restated with the operative-verification gate, and the repair ladder (EXP-11b/12/12c) identifies checkable evidence as the ingredient that restores it (§3.4)
Poisoned-channel extrapolation	48/48 reads as universal vulnerability	Scoped by EXP-7/7b/7c: measured adoption 4.8-100% across configurations; content plausibility and panel capability are the measured moderators
Protocol singularity	One scaffold, one window, one language	Bounded by the injector cross (family-robust), two ecosystems, four panel tiers, two network sizes; disclosed in §6
Open items	P8.2, the EXP-13 endpoint, EXP-12d	P8.2 is underpowered at the snippet-quality channel and awaits a stronger retriever; the EXP-13 endpoint test is borderline at its registered correction and only the decay shape is claimed; EXP-12d is registered and uncollected. No claim in the paper depends on them

B EXTENDED RELATED WORK

The conformity result has a long lineage. Asch (1951) showed that unanimous confederate majorities move human judgments, and informational cascade theory (Bikhchandani et al., 1992) explains why public agreement can carry no private information. What this paper adds is a controlled identification inside deployed LLM practice: the same agents, the same questions, the same judge, with the interaction protocol as the manipulated variable and the revision-versus-visibility decomposition supplied by a registered control (EXP-14).

Multi-agent debate and related deliberation schemes are widely reported to raise accuracy (Du et al., 2024; Liang et al., 2024), and self-consistency shows that sampling plus voting is a strong single-model baseline (Wang et al., 2023).

Our accuracy results are consistent with both lines (34.7% after deliberation; 31.0% for SC-15 at a matched number of completions); the divergence concerns what agreement means, not whether debate helps accuracy.

The closest concurrent work compares debate against voting directly. Choi et al. (2025) ask which decision rule yields better accuracy across tasks; they measure the accuracy of decision rules, whereas we measure the conditional diagnosticity of agreement itself, matched at the question level, with a mechanism (injection, anchor gradient) and boundary conditions (the Coverage Law). Zhu et al. (2025) audit a medical multi-agent pipeline and document false-consensus risk in that domain; their audit is correlational, whereas the present design is controlled (an independent-vote counterfactual on the same footprint) and ends in a quantitative law that was predicted and then confirmed. The KDD 2026 workshop paper *When Consensus Is Not Correctness* (Dai, 2026) documents diversity collapse and manufactured overconfidence at the trajectory level; it offers diagnostics without an intervention, and its overconfidence finding is consistent with the confidence dynamics in Section 3.2. On mechanism, Zhu et al. (2026) decompose multi-agent debate into confidence and diversity effects; their diversity emphasis matches the SC-5×3 arm, in which cross-model diversity preserves diagnosticity by declining to agree. On design, Cui et al. (2026) (Free-MAD) remove the consensus

requirement from debate and Pitre et al. (2025) (CONSENSAGENT) shape consensus behavior directly; both are engineering responses to the phenomenon this paper measures. Okawa (2026) report the emergence of biased consensus in debates, which we read as a trajectory-level correlate of the cascade measured here under a controlled design.

Structured orchestration frameworks are the natural next test of the topology scope. Tree of Thoughts (Yao et al., 2023) and Reflexion (Shinn et al., 2023) structure a single agent's search and self-critique; ChatEval (Chan et al., 2024) assigns evaluator roles across agents; AutoGen (Wu et al., 2023) composes role-conditioned conversations; Free-MAD (Cui et al., 2026) removes the consensus requirement. None was run here. The theory says where to look: Proposition 2 conditions on manufactured agreement, not on topology, so a structured protocol preserves diagnosticity to the extent that its roles keep agreement from forming on items where no seat holds or can verify the answer, and the Coverage Law's gate applies unchanged to any verifier role whose output the seats cannot check. The society-of-mind view (Minsky, 1986) poses the same question in older terms: what a committee of agents can know depends on what its members can independently establish.

Grounding work on retrieval-augmented generation (Lewis et al., 2020) supplies the tool the Coverage Law prices; the law's contribution is to make measured coverage, rather than the presence of a retriever, the quantity that bounds residual risk.

The failure literature of the last twelve months is dense and is engaged directly, one sentence per work. Wynn et al. (2025) document debate-driven accuracy loss with conformity as a factor; Cemri et al. (2025) build the MAST failure taxonomy from annotated traces; Zhang et al. (2025) show that MAD often loses to self-agent baselines and name heterogeneity as the remedy; Oh et al. (2025) decompose belief entrenchment and repair it with prompt-level diversity; Pan et al. (2025) build higher-order aggregation rules that outperform majority voting; Yang et al. (2025) reframe debate as conditional test-time scaling; Wu et al. (2025) find that majority pressure suppresses independent correction; and He et al. (2026) train a meta-classifier to overturn majority votes from debate logs. All eight are correlational, system-level, or engineering responses; none measures the conditional evidential value of the agreement event, matched at the question level, with the interaction protocol as the manipulated variable.

The human literature already contains the phenomenon, which bears on the question of whether the result is an artifact of language models. Lorenz et al. (2011) showed that social influence narrows the diversity of human estimates and raises confidence without improving collective error; Becker et al. (2017) showed network conditions under which influence improves accuracy even as beliefs homogenize, which is consonant with our accuracy-diagnosticity dissociation, since they measure accuracy and we measure what agreement is worth; Mavrodiev et al. (2012) and Tian et al. (2022) formalized the observation that whether influence helps or harms is ill-defined without the population's starting configuration, a thesis our anchor-mass account instantiates with every term measured in a new setting; and Jayles et al. (2017) documented the improvement direction in human estimation, making the sign of the social-influence effect configuration-dependent across species and substrates. The reliance literature frames the normative side: Lee and See (2004) define appropriate reliance in automation as trust calibrated to actual trustworthiness, and Durán and Pozzi (2025) separate trust, trustworthiness, and reliance, arguing that reliance is not entailed by accuracy. Both are normative or behavioral accounts of single-system settings; the present work supplies the measurement for a multi-agent signal, the evidential value of agreement itself, under a registered control.

The epistemic status of consensus. In the terms of the reliance literature, the central measurement concerns reliance rather than trust: we have not asked whether anyone should trust a multi-agent system (Durán and Pozzi, 2025), but whether one signal, agreement among agents, justifies reliance. It does so only insofar as the agreeing judgments are decision-relevantly independent, and interaction removes that independence, so that after four rounds the number of agents behind an answer no longer counts independent sources. The matched-completion dissociation (Section 3.5) is the case Lee and See (2004) describe, a trust signal and system performance coming apart, measured in a new setting: deliberation made the panel more accurate and its agreement less evidentiary at the same time. The evidential weight of a consensus therefore depends on how the agreeing judgments were produced, and on whether they could have been produced independently, not on the consensus itself.

C FORMAL STATEMENTS AND PROOFS

This appendix gives the formal statement of each law, a proof of the Diagnosticity-of-Agreement inequality in a stylized model, and the derivation of the engineering corollary. The companion theory file (`appendix_theory.tex`, in the deposit) is included in full.

Epistemic labeling. Every numbered statement is a Theorem, Lemma, Proposition, or Corollary proved in full from stated assumptions; a Conjecture, which is unproved and falsifiable; or an Empirical Regularity, which is measured and not derived. Measured quantities carry 95% Wilson intervals. Assumptions are stated before they are used.

C.1 SETUP

A question pool is a distribution Q over questions q; each question has a ground-truth answer $y^*(q)$ in a large answer space $\mathcal{Y}$. A panel consists of agents $i = 1, \dots, n$ ($n = 3$ in the core study, $n = 5$ in the replication).

Definition 1 (isolated answer, correctness, anchor count)**.** Agent i's *isolated answer* $X_i(q)$ is produced with no access to other agents' outputs. The correctness indicator is $C_i(q) = \mathbf{1}\{X_i(q) = y^*(q)\}$, and the *anchor count* is $K(q) = \sum_{i=1}^{n} C_i(q)$. A question with $K(q) = 0$ is *truth-scarce* (unanchored); a question with $K(q) \geq 1$ is *anchored*.

Definition 2 (aggregation protocols)**.** *Independent-vote aggregation* is $A_{\mathrm{ind}}(q) = \mathrm{plurality}(X_1(q), \dots, X_n(q))$, with $A_{\mathrm{ind}}(q) = \perp$ when the votes tie or all agents abstain. *Deliberated aggregation* exchanges messages for r rounds ($r = 4$, fully connected, in the core study), producing post-interaction answers $\tilde{X}_i(q)$, and outputs $A_{\mathrm{del}}(q) = \mathrm{plurality}(\tilde{X}_1(q), \dots, \tilde{X}_n(q))$ under the same abstention convention. A protocol Π is any mechanism of this form.

Definition 3 (agreement event, conditional falsity, diagnosticity)**.** Let $s_\Pi(q)$ be the size of the modal cluster under Π (the plurality's support). The *agreement event* under Π is the *non-unanimous* majority event $E_\Pi = \{\lceil (n+1)/2 \rceil \leq s_\Pi(q) < n\}$: the rule returns a majority on a split panel. This is the estimand of every headline number in the paper. Unanimity, $U_\Pi = \{s_\Pi(q) = n\}$, is a separate event, always reported separately because it behaves differently (Lemma 2; empirically unbiased under interaction, Section 3.1). A modal cluster may be abstention-valued; the primary convention counts such agreements as formed and non-false, the secondary excludes them, and both are printed wherever they differ. The *conditional falsity* and the *diagnosticity of agreement* are

$$F_\Pi = P\big(A_\Pi(q) \neq y^*(q) \,\big|\, E_\Pi\big), \qquad D_\Pi = P\big(A_\Pi(q) = y^*(q) \,\big|\, E_\Pi\big) - P\big(A_\Pi(q) \neq y^*(q) \,\big|\, E_\Pi\big).$$

Remark 1. Conditional on agreement, the agreed answer is the truth or is not; reading non-truth as false (the primary convention counts abstention-valued agreements as non-false), the outcome is binary and $D_\Pi = 1 - 2F_\Pi$. We use F for estimation (it is a proportion; Wilson intervals apply) and D for the law statements (it is signed evidence). The central measurement of the paper, in this notation, is $F_{\mathrm{ind}} = 23.1\%$ $[16.8, 31.0]$ $(31/134)$ against $F_{\mathrm{del}} = 74.8\%$ $[66.1, 81.8]$ $(86/115)$ for non-unanimous three-agent majorities (Fisher $p = 2.1 \times 10^{-16}$), replicated on a frontier panel at 26.2% $[18.0, 36.5]$ against 66.7% $[51.0, 79.4]$ (Holm-adjusted one-sided $p = 6.9 \times 10^{-5}$). The estimand is the same everywhere in this paper: E_Π in the law statements and in every empirical headline is the *non-unanimous* majority event (two-of-three in the three-seat studies). Unanimous agreements are always reported separately because they behave differently: interaction does not inflate their falsity (49.4% false, $p = 0.94$ against the independent counterpart), consistent with Lemma 2.

Remark 2 (information relative to what baseline). D_Π is a rescaled conditional probability, so "agreement carries information" is made operational against the pool's own base rate. On the core pool the per-seat isolated accuracy is 0.187 (deposit-verified anchor masses on the canonical basis $\pi = (0.577, 0.287, 0.137)$; Proposition 4). Against that prior, an independent two-of-three agreement multiplies the odds of truth by ≈ 14 (posterior $P(T \mid A_{\mathrm{ind}}) = 0.769$), while a deliberated agreement multiplies them by ≈ 1.5 (posterior 0.252). Interaction leaves the agreement signal barely above the base rate it started from; the D gap is the same statement on the signed scale.

C.2 THE INDEPENDENCE BENCHMARK

The two lemmas below are standard: Condorcet's jury theorem (Condorcet, 1785) and the plurality direction of List and Goodin (2001). They are restated because the paper's results depart from their premises, not from their conclusions.

Lemma 1 (Condorcet majority). *Let $C_1, \dots, C_n$ be independent* Bernoulli(p) *with $p > \frac{1}{2}$, n odd, and $M_n = \mathbf{1}\{\sum_i C_i > n/2\}$. Then $P(M_n = 1) > p$, the probability is strictly increasing in n over odd n, and $P(M_n = 0) \le \exp(-2n(p - \frac{1}{2})^2)$.*

Proof. The bound is Hoeffding's inequality applied to $\sum_i C_i$, a sum of n independent $[0, 1]$-valued variables with mean np. For monotonicity, write $n = 2m + 1$ and pass from n to $n + 2$. The majority outcome changes only when the $2m + 1$ existing voters split as evenly as possible. A split of m correct and $m + 1$ wrong (probability $\binom{2m+1}{m} p^m (1-p)^{m+1}$) followed by two correct votes (probability p^2) flips the majority from wrong to right; a split of $m + 1$ correct and m wrong (probability $\binom{2m+1}{m} p^{m+1}(1-p)^m$) followed by two wrong votes (probability $(1-p)^2$) flips it from right to wrong. Hence

$$P(M_{2m+3} = 1) - P(M_{2m+1} = 1) = \binom{2m+1}{m} \big[p(1-p)\big]^{m+1} (2p - 1) > 0,$$

Condorcet's recursion. Since $P(M_1 = 1) = p$ and each increment is positive, $P(M_n = 1) > p$ for all odd $n \ge 3$. □

Lemma 2 (unanimity with dispersed errors; List–Goodin direction). *Let $X_1, \dots, X_n$ be mutually independent and identically distributed answers with $P(X_i = y^*) = p$ and $q_a := P(X_i = a)$ for $a \ne y^*$, the same for every i, and suppose the truth is the modal answer: $\bar{q} := \max_{a \ne y^*} q_a \le p$. Then, conditional on unanimous agreement,*

$$P\big(\textit{agreed answer} = y^* \,\big|\, \textit{all } n \textit{ agree}\big) \;\ge\; \frac{p^n}{p^n + \bar{q}^{\,n-1}(1-p)}.$$

If $\bar{q} < p$, this probability tends to 1 exponentially fast in n, at rate $\log(p/\bar{q})$.

Proof. Write $q_a = P(X_i = a)$ for $a \ne y^*$. Independence gives $P(\text{all agree on } a) = q_a^n$ and $P(\text{all agree on } y^*) = p^n$. The conditional probability is

$$\frac{p^n}{p^n + \sum_{a \ne y^*} q_a^n} \;\ge\; \frac{p^n}{p^n + \bar{q}^{\,n-1} \sum_{a \ne y^*} q_a} = \frac{p^n}{p^n + \bar{q}^{\,n-1}(1-p)},$$

using $q_a^n \le \bar{q}^{\,n-1} q_a$ and $\sum_{a \ne y^*} q_a = 1 - p$. If $\bar{q} < p$, the ratio of the second denominator term to p^n is $(\bar{q}/p)^{n-1} \cdot (1-p)/p \to 0$ geometrically. □

Remark 3 (heterogeneous agents). The identical-distribution premise is a convenience of the stylized benchmark, not a requirement. With agent-specific $p_i = P(X_i = y^*)$ and $\bar{q}_i = \max_{a \ne y^*} P(X_i = a)$, independence gives $P(\text{all agree on } y^*) = \prod_i p_i$ and $P(\text{all agree on } a) = \prod_i P(X_i = a) \le P(X_1 = a) \prod_{i \ge 2} \bar{q}_i$, so summing over $a \ne y^*$,

$$P\big(\text{agreed answer} = y^* \,\big|\, \text{all } n \text{ agree}\big) \;\ge\; \frac{\prod_i p_i}{\prod_i p_i + (1 - p_1) \prod_{i \ge 2} \bar{q}_i},$$

which reduces to the lemma when the agents are identical and tends to 1 geometrically whenever $\bar{q}_i < p_i$ for every i. The premise that fails on the truth-scarce support is the same in both forms: that the truth is each agent's modal answer.

Remark 4 (what the benchmark needs). Both lemmas rest on (i) mutual independence of the votes and (ii) competence in the sense that the truth is each agent's modal answer. Premise (ii) fails exactly on the truth-scarce support: when $K(q) = 0$, no agent holds the truth, and any agreement that forms must form on some $a \ne y^*$. The measured anchor gradient is the empirical fingerprint of premise (ii): $P(\text{true consensus}) = 0.06, 0.71, 0.88$ at $K = 0, 1, 2$ (Empirical Regularity 1).

C.3 CONDITIONAL-FALSITY INFLATION UNDER INTERACTION

Deliberation changes the agreement event, not only the answers. Measured on the core pool: consensus events rose from $134/300$ under isolated voting to $285/300$ after deliberation (Fisher $p = 2.2 \times 10^{-45}$), while abstention majorities fell from $61/300$ to $10/300$ ($p = 4.1 \times 10^{-11}$).

Definition 4 (inherited and manufactured agreements)**.** Partition the deliberated agreement event question-wise into *inherited* agreements $E_{\text{inh}} = E_{\text{ind}} \cap E_{\text{del}}$ (an independent plurality existed and deliberation kept one) and *manufactured* agreements $E_{\text{new}} = E_{\text{del}} \cap E_{\text{ind}}^{c}$ (independent voting had no plurality and deliberation produced one). Write

$$w = P(E_{\text{inh}} \mid E_{\text{del}}), \qquad f_{\text{inh}} = P(A_{\text{del}} \neq y^* \mid E_{\text{inh}}), \qquad f_{\text{new}} = P(A_{\text{del}} \neq y^* \mid E_{\text{new}}).$$

The empirically rare reverse flow (independent agreement destroyed by interaction) is neglected; the measured net flow is overwhelmingly in the manufacturing direction.

Proposition 1 (conditional-falsity inflation)**.** *Assume* $0 < w < 1$. *Then:*

(i) $F_{\text{del}} = w f_{\text{inh}} + (1 - w) f_{\text{new}}$.

(ii) $F_{\text{del}} > F_{\text{ind}}$ *if and only if* $(1 - w)\big(f_{\text{new}} - F_{\text{ind}}\big) > w\big(F_{\text{ind}} - f_{\text{inh}}\big)$.

(iii) Sufficient condition: if $f_{\text{inh}} \geq F_{\text{ind}}$ *(deliberation does not improve the truth content of pre-existing agreements on average) and* $f_{\text{new}} > F_{\text{ind}}$ *(manufactured agreements are falser than the independent-agreement average), then* $F_{\text{del}} > F_{\text{ind}}$*, strictly.*

On the status of this statement. This is a decomposition whose force comes from the measured assumptions (T1)–(T3), not a theorem about deliberation dynamics: given the measurements, the inequality follows by total probability. It is labeled a proposition accordingly. The nearest published theory is the belief-trajectory martingale of Choi et al. (2025) (debate alone does not improve expected correctness); that result is compatible with, and orthogonal to, the present one: a martingale on beliefs constrains expected correctness, while the object here is the diagnosticity of the agreement *event*, a different functional, whose degradation is driven by the changing composition of the agreement set rather than by the mean of the belief process.

Proof. Part (i) is the law of total probability over the partition $\{E_{\text{inh}}, E_{\text{new}}\}$ of E_{del}. Part (ii) rearranges $w f_{\text{inh}} + (1 - w) f_{\text{new}} > F_{\text{ind}}$ by collecting terms in $f_{\text{new}} - F_{\text{ind}}$ and $F_{\text{ind}} - f_{\text{inh}}$. For (iii), if $f_{\text{inh}} \geq F_{\text{ind}}$ then

$$F_{\text{del}} \;\geq\; w F_{\text{ind}} + (1 - w) f_{\text{new}} \;>\; w F_{\text{ind}} + (1 - w) F_{\text{ind}} = F_{\text{ind}},$$

using $w < 1$ and $f_{\text{new}} > F_{\text{ind}}$ for the strict step. □

Remark 5 (the data meet the condition of (iii)). (a) *Manufactured agreements are falser than the independent average.* Independent-agreement falsity is $F_{\text{ind}} = 23.1\%$ $[16.8, 31.0]$ (weak tier) and 26.2% $[18.0, 36.5]$ (frontier). Manufactured agreements concentrate on the truth-scarce support: false majorities began with a mean of 0.20 agents correct in isolation, against 1.26 for true majorities. On the $K = 0$ stratum the measured shared-fabrication rates are 81.5% (three-agent), 90.6% (five-agent), $19/19 = 100\%$ $[83.2, 100]$ (TriviaQA, three-agent), and $44/45 = 97.8\%$ $[88.4, 99.6]$ for five-agent unanimities in which no agent knew the answer in isolation. Every interval lies far above the independent-agreement interval. (b) *Inherited agreements are not improved on average.* Directly measured, $f_{\text{inh}} = 0.706$ (36/51) against $F_{\text{ind}} = 0.231$: on the questions where an independent majority already existed, deliberation made the surviving majority markedly falser, which is the same movement the paired McNemar test records (24 to 2, Section 3.1). At the agent level, of the instances that changed away from a correct answer, $42/43$ moved to error; 17.0% $[14.7, 19.6]$ of agent-instances moved from abstention to error; 86.7% of abstainers were induced to assert. On the paired grounded comparison, the uncovered stratum went from 4.6% false under independent voting to 17.2% under deliberation ($p = 0.013$). (c) *The pooled conclusion is the measured headline.* $F_{\text{del}} = 74.8\%$ $[66.1, 81.8]$ against 23.1% $[16.8, 31.0]$ (Fisher $p = 2.1 \times 10^{-16}$); abstention-excluded, 79.6% $(86/108)$ against 42.5% $(31/73)$, $p = 5.7 \times 10^{-7}$; frontier replication 66.7% $[51.0, 79.4]$ against 26.2% $[18.0, 36.5]$, Holm-adjusted $p = 6.9 \times 10^{-5}$ (pre-registered prediction P2, confirmed).

Remark 6 (the decomposition, measured). The three quantities are measured directly on the canonical pool rather than inferred from marginals (`tier2/close_items.py decomp`). Of the 115 two-of-three majorities that form after deliberation, 51 are inherited ($w = 0.443$) and 64 are manufactured, with conditional falsities $f_{\text{inh}} = 0.706$ (36/51) and $f_{\text{new}} = 0.781$ (50/64). They reproduce the headline through part (i): $wf_{\text{inh}} + (1-w)f_{\text{new}} = 0.748 = F_{\text{del}}$. Both terms of the sufficient condition in part (iii) are therefore measured, not assumed: $f_{\text{inh}} = 0.706 \geq F_{\text{ind}} = 0.231$ and $f_{\text{new}} = 0.781 > F_{\text{ind}}$. Manufactured majorities fall on the unanchored stratum in 45 of 64 cases, and their falsity is 0.933 there against 0.421 on the anchored stratum, which is the two-regime model's central asymmetry in its most direct form.

C.4 The Diagnosticity-of-Agreement Law

The law is stated as a conjecture; the restricted case that the two-regime model supports is proved. Nothing beyond what the model gives is claimed.

Definition 5 (two-regime model)**.** Partition the pool by truth-availability: the *unanchored* stratum $U = \{q : K(q) = 0\}$ with mass $1-\alpha$, and the *anchored* stratum $A = \{q : K(q) \geq 1\}$ with mass α.

Assumption 1. (T1) *Truth-scarcity bound.* Any agreement on U, however produced, is true with probability at most τ, where $\tau < 1 - F_{\text{ind}}$. On the studied pool, $P(\text{true consensus} \mid K = 0) = 0.06$, so $\tau \approx 0.06$ is the measured value.

(T2) *Manufacture.* The protocol creates agreements on U where independent voting does not agree: $P(E_{\text{del}} \cap U \cap E_{\text{ind}}^c) > 0$. Measured: consensus events $134/300 \to 285/300$ (canonical basis).

(T3) *No average improvement of inherited agreements.* $f_{\text{inh}} \geq F_{\text{ind}}$ (Remark 5(b); relaxed in Remark 8).

(T4) *Unanchored share of manufacture.* Let $\mu = P(U \mid E_{\text{new}})$ be the fraction of manufactured agreements that fall on the unanchored stratum. Then $\mu > F_{\text{ind}}/(1-\tau)$. Nothing is assumed about the falsity of manufactured agreements on the anchored stratum A; the bound below holds even if every one of them is true. On the studied pool the threshold is $0.231/0.94 \approx 0.25$ and the measured value is $\mu = 0.703$ (Remark 7).

Proposition 2 (restricted Diagnosticity-of-Agreement inequality)**.** *Under the two-regime model with* (T1)–(T4)*,*

$$D_{\text{del}} \;\leq\; D_{\text{ind}},$$

*with equality if and only if $w = 1$ and $f_{\text{inh}} = F_{\text{ind}}$; that is, if and only if the protocol's agreement events coincide almost surely with the independent agreement events and preserve their truth content (*preservation of decision-relevant independence*). Whenever the protocol manufactures agreement on the truth-scarce stratum, the inequality is strict.*

Proof. Since $D_{\Pi} = 1 - 2F_{\Pi}$, the claim $D_{\text{del}} \leq D_{\text{ind}}$ is equivalent to $F_{\text{del}} \geq F_{\text{ind}}$. By Proposition 1(ii),

$$F_{\text{del}} - F_{\text{ind}} = w\big(f_{\text{inh}} - F_{\text{ind}}\big) + (1-w)\big(f_{\text{new}} - F_{\text{ind}}\big).$$

The first term is nonnegative by (T3). By (T2), manufactured agreements exist ($w < 1$). Decompose their falsity over the two strata: $f_{\text{new}} = \mu f_{\text{new}}^U + (1-\mu) f_{\text{new}}^A$, where f_{new}^U and f_{new}^A are the falsity rates of manufactured agreements on U and on A. By (T1), $f_{\text{new}}^U \geq 1 - \tau$, and $f_{\text{new}}^A \geq 0$ trivially, so $f_{\text{new}} \geq \mu(1-\tau) > F_{\text{ind}}$ by (T4); the second term is therefore strictly positive. Hence $F_{\text{del}} > F_{\text{ind}}$ whenever $w < 1$. If $w = 1$, then $F_{\text{del}} = f_{\text{inh}} \geq F_{\text{ind}}$ by (T3), with equality if and only if $f_{\text{inh}} = F_{\text{ind}}$. □

Remark 7 (measured support for (T4)). Measured on the canonical pool, manufactured two-of-three majorities fall on the unanchored stratum in 45 of 64 cases, so $\mu = 0.703$, nearly three times the threshold $F_{\text{ind}}/(1-\tau) = 0.246$ that (T4) requires. Their falsity on that stratum is 0.933, against the bound $1 - \tau = 0.94$ that (T1) supplies.

Remark 8 (how much anchored-stratum improvement could offset manufacture). Relax (T3). The inequality $D_{\text{del}} \leq D_{\text{ind}}$ survives any improvement of inherited agreements down to

$$f_{\text{inh}} \;\geq\; F_{\text{ind}} - \frac{1-w}{w}\big(f_{\text{new}} - F_{\text{ind}}\big).$$

Instantiating the measured core-study magnitudes ($w = 0.443$, $F_{\text{ind}} = 0.231$, $f_{\text{new}} = 0.781$; Remark 6), the bound is $0.231 - 1.26 \times 0.550 \approx -0.46$: overturning the inequality would require inherited agreements to be *improved* by deliberation down to a negative falsity, which is impossible. On these measurements the inequality is not close. This also clarifies the relationship to pool accuracy: deliberation raised mean pool accuracy from 13.7% $[10.2, 18.0]$ to 34.7% $[29.5, 40.2]$ ($104/300$ on the de-duplicated canonical pool, majority-of-finals convention; 35.0% under the cluster-label convention; 69 corrections against 5 corruptions), mostly on the anchored stratum, and still destroyed diagnosticity, because diagnosticity prices the agreements that deliberation *adds*, not only the ones it repairs. Accuracy and diagnosticity of agreement are different functionals of the same transcript.

Conjecture 1 (Diagnosticity-of-Agreement Law, general form)**.** *Let* Π *range over text-based deliberation protocols among sampled model agents, and let the pool have a truth-scarce subpool of non-negligible mass. Then, at sample size* $n \geq 150$ *per arm,*

$$D_{\Pi} \;\leq\; D_{\text{ind}},$$

with equality if and only if Π *preserves decision-relevant independence almost surely. Agreement is evidence of truth only insofar as the agents could have agreed without talking; interaction preserves or destroys diagnosticity and does not create it, and the destruction concentrates where independent truth-availability is low.* Falsifier: *a single deliberation protocol whose post-interaction majorities are more diagnostic than matched independent votes on a truth-scarce pool, at* $n \geq 150$ *with statistical separation (disjoint 95% intervals, or a pre-registered one-sided test), refutes the conjecture.*

Remark 9 (status of the evidence). Proposition 2 is the proved restricted case. All measured points go the conjecture's way: weak tier (D: $0.538 \to -0.496$), registered frontier replication E1 ($0.476 \to -0.334$, Holm $p = 6.9 \times 10^{-5}$), and the EXP-1 control, in which a non-interacting cross-model protocol holds diagnosticity at the independent level ($F = 25.9\%$ $[13.2, 44.7]$ against the 23.1% $[16.8, 31.0]$ reference), consistent with the equality case of Proposition 2.

C.5 The Coverage Law

The grounding protocol gives each agent access to a retriever over an evidence store. The law decomposes residual false consensus into a hit term and a miss term.

Definition 6 (coverage, hit, miss)**.** For question q, let $H(q)$ be the event that retrieval surfaces gold (truth-bearing) evidence for q, and let $p = P(H)$ be the *retrieval coverage* measured on the pool. On a miss, $K(q)$ is the number of agents independently anchored on the truth.

Assumption 2 (measured support in parentheses)**.** (A1) *Anchor retention.* An anchored agent retains its answer under contrary social pressure, up to a small defection probability. (Individual adoption of a confident injected falsehood by anchored agents: 7% frontier panel, 12% arithmetic substrate.)

(A2) *Unanchored adoption.* An unanchored agent adopts the most confident type-consistent peer claim. (Groups adopt an injected confident falsehood at 91.0% $[83.8, 95.2]$; a type-inconsistent confident claim was adopted in $0/50$ cases, $z = 11.38$.)

(A3) *Evidence anchors.* When retrieval surfaces gold evidence, agents anchor on it, and residual false consensus conditional on a hit is a small constant $\varepsilon \in [0.02, 0.05]$. (Measured 2–5% conditional on a hit; 5.2% on hits in the live-retrieval arm.)

(A4) *Coverage is independent of anchor count.* $P(H \mid K = k) = p$ for every k. Where this fails, the law holds stratum by stratum with p read as the stratum-specific coverage $p_k = P(H \mid K = k)$; the pool-level rate is then the π_k-weighted average, not the formula with the marginal p. (In the simulated sweeps coverage is assigned by design, so (A4) holds exactly; in live retrieval it fails, and Remark 10(2) shows the consequence.)

Theorem 1 (Coverage Law)**.** *Under* (A1)–(A4)*, the false-consensus rate of grounded deliberation on the stratum of questions with anchor count* $K = k$*, on a pool with coverage* p*, is*

$$\text{FalseRate}(p, k) = p\,\varepsilon + (1 - p)\,\phi(k), \qquad \varepsilon \in [0.02, 0.05],$$

where $\phi(k) = P(\text{false consensus} \mid H^c, K = k)$ *is the ungrounded falsity at anchor count* k*. Under* (A1) *and* (A2)*,* ϕ *is nonincreasing in* k*. Measured values:* $\phi(0) = 1.00$*,* $\phi(1) = 0.69$*,* $\phi(2) = 0.52$ *(registered attenuation experiment, gradient trend* $p = 5.3 \times 10^{-8}$*).*

Proof. Condition throughout on $K = k$. The pair $\{H, H^c\}$ partitions the probability space, so total probability gives

$$\mathrm{FalseRate}(p, k) = P(H \mid K = k)\, P(\text{false} \mid H, K = k) + P(H^c \mid K = k)\, P(\text{false} \mid H^c, K = k),$$

and (A4) replaces $P(H \mid K = k)$ by p. By (A3), $P(\text{false} \mid H, K = k) = \varepsilon$, independent of the question. By the definition of ϕ and the assumption that, on a miss, the social configuration enters only through the anchor count ((A1) governs anchored retention, (A2) governs unanchored adoption), $P(\text{false} \mid H^c) = \phi(k)$. For monotonicity: on a miss with k anchored agents, a false consensus requires either the defection of anchored agents (probability bounded by (A1)'s retention rate, weakly decreasing as the number of required defections grows) or a false plurality formed entirely among the $n - k$ unanchored agents, which becomes impossible once k is large enough that unanchored agents cannot reach the majority threshold; both channels are nonincreasing in k. The values of ϕ are measured, not derived: perfect retention (the idealized limit of (A1)) is violated at $k = 2$, since $\phi(2) = 0.52$ exceeds any defection-rate bound, so ϕ is taken from the attenuation experiment rather than from a stronger version of (A1). □

Corollary 1 (coverage threshold)**.** *Fix an anchor count k with $\phi(k) > \varepsilon$ and a target δ with $\varepsilon < \delta < \phi(k)$. Then* $\mathrm{FalseRate}(p, k) < \delta$ *if and only if*

$$p \;>\; p^* = \frac{\phi(k) - \delta}{\phi(k) - \varepsilon} = 1 - \frac{\delta - \varepsilon}{\phi(k) - \varepsilon}.$$

In the idealized limit $\varepsilon \ll \delta$ this reads $p > 1 - \delta/\phi(k)$.

Proof. Solve $p\varepsilon + (1-p)\phi(k) < \delta$ for p; since $\phi(k) > \varepsilon$ by premise, dividing by $\phi(k) - \varepsilon$ preserves the inequality and yields $p > (\phi(k) - \delta)/(\phi(k) - \varepsilon)$. □

Remark 10 (instantiation and consistency with measurement). *Instantiation.* $\delta = 0.10$, $\phi(0) = 1.00$: the idealized rule gives $p > 0.90$. Honoring the measured ε range, the exact threshold is $p > 0.918$ at $\varepsilon = 0.02$ and $p > 0.947$ at $\varepsilon = 0.05$. The engineering rule "coverage above roughly 0.9 for false consensus below 10%" is the optimistic edge; the conservative edge is 0.95.

Consistency. (1) Coverage sweep: observed rates $68, 45, 39, 30, 15\%$ across $p = 0.30 \ldots 0.85$ track $\phi(0)(1 - p)$ at $r = -0.996$ (falsity falls as coverage rises). (2) Live retrieval ($n = 205$ after screening, measured $p = 0.75$): predicted $\mathrm{FalseRate} \approx 25.0\%$; observed 20.5% $[15.5, 26.5]$. The prediction lands inside the interval, and the small overprediction has an identified cause: live misses carry anchor mass, which is a failure of (A4) at pool level (coverage and anchor count are correlated under live retrieval), so the effective miss-conditional falsity is $\phi_{\text{eff}} \approx 0.667$ (observed 66.7% on misses) rather than $\phi(0) = 1.00$; hits behave as (A3) says (5.2% on hits). The four coverage strata track $1 - p$ at $r = 0.81$. (3) Registered attenuation ($n = 240$ TriviaQA, ≥ 1 anchored agent): the falsity-versus-scarcity slope attenuates from 1.13 (s.e. 0.11) to 0.72 (s.e. 0.11), $p = 0.010$, which is $\phi(k) < \phi(0)$ in regression form; miss-conditional false consensus fell from 100% to 63% (Fisher $p = 1.2 \times 10^{-8}$). (4) Grounded deliberation remains less diagnostic than grounded independent voting on the paired comparison ($n = 167$): 12.6% against 4.2% (Fisher $p = 0.0092$); uncovered stratum 17.2% vs 4.6% ($p = 0.013$); covered stratum 7.5% vs 3.7% (not significant). Retrieval shrinks the manufactured mass; it does not repeal Proposition 2.

C.6 The anchor threshold: a step, not a slope

Empirical Regularity 1 (first-anchor threshold). On the three-agent core pool, the probability that a consensus is true, by anchor count, is

$$P(\text{true consensus} \mid K = k) = 0.06,\ 0.71,\ 0.88 \quad \text{at } k = 0, 1, 2,$$

with logits $-2.75, 0.90, 1.99$. The step odds ratios are

$$\mathrm{OR}_{0\to1} = \frac{0.71/0.29}{0.06/0.94} \approx 38.4 \approx 39, \qquad \mathrm{OR}_{1\to2} = \frac{0.88/0.12}{0.71/0.29} \approx 3.0 \approx 2.9.$$

The natural model is a threshold (first-anchor) model, $\mathrm{logit}\, P(k) = \beta_0 + \beta_1 \mathbf{1}\{k \geq 1\} + \beta_2 \mathbf{1}\{k \geq 2\}$ with $\beta_1 \approx 3.65$, $\beta_2 \approx 1.10$, saturated at three points and fitting them exactly. Its reading is the Anchor Principle in miniature: the first independently held truth breaks the symmetry among fabricated claims; the second adds little, because the consensus was already decided by the first.

Remark 11 (why a smooth logistic misleads here). Fitting $\operatorname{logit} P(k) = \alpha + \beta k$ to the three points (unweighted, on the logit scale) gives $\beta = 2.37$, a per-anchor odds ratio of $e^{\beta} \approx 10.7$. That number estimates neither step; it is the geometric mean $\sqrt{38.4 \times 3.0} \approx 10.7$, forced by the linear-in-k form (with three equally spaced points, the fitted slope equals half the total logit range regardless of where the jump sits). The compression understates the first-anchor effect by a factor of ≈ 3.6 in the odds multiplier and overstates the second by the same factor, and its fitted probabilities are $0.089, 0.511, 0.918$ against the observed $0.06, 0.71, 0.88$: the largest error sits at $k = 1$, the operationally decisive case, where the smooth fit says a coin flip and the data say 71%. An analyst using the smooth coefficient to price the value of injecting one correct, confident anchor (measured at $100\% \to 1\%$ false consensus, Fisher $p = 2.2 \times 10^{-57}$) would underprice it by an order of magnitude in odds. Status: empirical regularity plus a model-comparison remark. Per-k counts were not retained at this granularity, so no intervals are reported for the three probabilities. The five-agent replication shows the same shape at the level of unanimity (97.8% $[88.4, 99.6]$ false with no independently correct agent, 30.0% with one, 16.7% with two; trend $p = 9 \times 10^{-10}$), and the registered attenuation experiment gives the miss-conditional gradient $100/69/52\%$ at $k = 0/1/2$ (trend $p = 5.3 \times 10^{-8}$).

Remark 12 (operative versus nominal anchors: the two-panel evidence). The threshold model treats $K(q)$ as the count of agents holding the truth. EXP-6's two arms measured the refinement of what "holding" means. The registered prediction P6.2 (a strictly smaller diagnosticity gap on GSM8K than on MMLU-hard) relies on the substrate-level reading. On the local weak 3B panel it failed, reversed, and the reversal is robust (identical verdicts under the original and the repaired judging pipeline): $+58.8$pp on GSM8K against $+26.8$pp on MMLU (DiD bootstrap $[-0.482, -0.152]$); on the operative reading, a 3B panel that cannot reliably verify its own arithmetic at run time gains no operative anchor from the nominally re-derivable substrate, and the collapse there is the program's largest (80.0% deliberated falsity). On the resumed mid-capability panel the same prediction *confirms*: $+1.0$pp on GSM8K against $+15.8$pp on MMLU-hard (DiD $[0.045, 0.252]$), because that panel can execute the GSM8K verification (95.5% of its independent majorities are true) and cannot execute the MMLU-hard one. Replace

$$K(q) = \textstyle\sum_i \mathbb{1}\{i \text{ correct in isolation on } q\} \quad \text{with} \quad K_{\mathrm{op}}(q) = \textstyle\sum_i \mathbb{1}\{i \text{ can verify on } (i,q)\},$$

where "can verify" is to be measured by verification accuracy under test, not asserted by substrate choice. This program did not measure K_{op} independently: verification capability is inferred from each panel's independent accuracy on the substrate and from the direction of the P6.2 verdicts. The two arms are therefore two views of the same statement that strongly support the operative-anchor account without measuring it directly; a direct test requires per-seat verification accuracy measured outside deliberation (the Anchor Principle's falsifier, Section 2). Status: the substrate-level prediction was confirmed on one registered panel and reversed on the other, in the directions the operative form accounts for (the form was adopted after the reversal); both verdicts are ledger entries.

C.7 The diagnosticity–coverage tradeoff

EXP-1 showed that a non-interacting cross-model protocol restores independent-level diagnosticity at the price of deciding on very few questions. This section makes the tradeoff precise as a frontier, with a proof sketch.

Definition 7 (non-manufacturing protocol)**.** Fix a pool Q and an independent-vote ensemble: each agent's votes are generated without access to any other agent's outputs (e.g., m samples per agent, plurality across all $n \times m$ votes, as in SC-5x3). For question q, let $\hat{y}(q)$ be the ensemble plurality ($\perp$ if none) and $Z(q) = \mathbf{1}\{\hat{y}(q) \neq \perp,\ \hat{y}(q) \neq y^*(q)\}$ its falsity indicator. A *non-manufacturing protocol* is a decision rule $s(q) \in \{\mathrm{decide}, \mathrm{abstain}\}$ measurable with respect to non-interacting information only, outputting $\hat{y}(q)$ when it decides. Its coverage, conditional falsity, and diagnosticity are

$$\gamma(s) = P(s = \mathrm{decide}), \qquad F(s) = \frac{E[\mathbf{1}\{s = \mathrm{decide}\}\, Z]}{\gamma(s)}, \qquad D(s) = 1 - 2F(s).$$

Proposition 3 (tradeoff frontier; proof sketch)**.** *Among non-manufacturing protocols on a fixed ensemble, every achievable pair (γ, D) satisfies $D \leq D^*(\gamma)$, where*

$$D^*(\gamma) = 1 - 2F^*(\gamma), \qquad F^*(\gamma) = \frac{1}{\gamma} \min_{s:\, \gamma(s) = \gamma} E\big[\mathbf{1}\{s = \mathrm{decide}\}\, Z\big],$$

and $D^*(\gamma)$ is nonincreasing in γ. The frontier is attained by threshold rules that decide on the questions with the lowest posterior falsity $P(Z = 1 \mid \textit{features})$.

Proof sketch. The coverage-constrained problem is to choose a (possibly randomized) set of mass γ minimizing the false-decision mass $E[\mathbf{1}\{\text{decide}\}\, Z]$. If a rule abstains on an event e_1 and decides on an event e_2 with $P(Z \mid e_2) > P(Z \mid e_1)$, swapping the two decisions weakly reduces the false mass at unchanged coverage (an exchange argument; for continuous posteriors this is the Neyman–Pearson likelihood-ratio form, with posterior falsity as the statistic). Hence an optimal rule is a threshold on posterior falsity. For monotonicity, take $\gamma' > \gamma$: any rule covering mass γ', restricted to its γ-mass of lowest-falsity decisions, is feasible at γ, so $F^*(\gamma) \le F^*(\gamma')$ and $D^*(\gamma) \ge D^*(\gamma')$. □

Remark 13 (what the proposition does and does not say). The frontier is defined relative to a fixed independent-vote ensemble. Two escape routes exist, and both are observed. (i) *Improve the ensemble*: retrieval grounding or more competent agents change Z itself and shift the whole frontier; this is the Coverage Law's channel. (ii) *Leave the constraint set*: interacting protocols manufacture agreement, buying coverage above the independent agreement mass; Proposition 2 prices this escape, and the price is paid in diagnosticity. What Proposition 3 rules out is a third option: within non-manufacturing protocols, no cleverness in the abstention rule beats the posterior-threshold frontier.

EXP-1 instantiation (Tier-2 canonical panel, $n = 300$ per arm, 15 completions per question in each arm). SC-5x3 (five samples per agent, cross-model plurality, no interaction) attains conditional falsity 25.9% $[13.2, 44.7]$ against the independent-vote reference 23.1% $[16.8, 31.0]$: diagnosticity restored to the independent level (pre-registered P1.2, confirmed). The cost is coverage: only $27/300 = 9.0\%$ $[6.3, 12.8]$ of questions yield a non-abstention plurality, and pool accuracy falls to 6.7% $[4.4, 10.1]$. Single-model SC-15 sits at the other end: far higher coverage, conditional falsity 46.2% $[39.0, 53.7]$, accuracy 31.0% $[26.0, 36.4]$. One qualification applies: deliberation itself reaches 34.7% $[29.5, 40.2]$ accuracy on the canonical pool, numerically above matched-compute SC-15 with overlapping intervals (pre-registered P1.3 was not confirmed), while its agreements are the least diagnostic of all (74.8% false). Accuracy and diagnosticity decouple under interaction; the frontier statement concerns diagnosticity, and the paper's claims are about diagnosticity.

C.8 The anchor-mass account

Proposition 4 (Anchor-mass decomposition)**.** *Let π_k be the fraction of questions on which exactly k of the three seats answer correctly in isolation (the* anchor mass *at k), and let r_k be the probability that deliberation converts a k-anchor question to a true consensus. Because the divergent-start screen excludes questions all three seats answer correctly, $\pi_3 = 0$ by design; independent two-of-three voting is then correct exactly on the $k = 2$ mass, so its accuracy is π_2, and deliberation's accuracy is $\sum_k \pi_k r_k$. The net effect of interaction decomposes as*

$$\Delta \;=\; \underbrace{\pi_1 r_1}_{\textit{lone-anchor rescue}} \;+\; \underbrace{\pi_0 r_0}_{\textit{manufactured truth}} \;-\; \underbrace{\pi_2\,(1 - r_2)}_{\textit{corruption of anchored items}} \quad .$$

All quantities are measured directly from the deposit (`scaled_study.jsonl`, $n = 300$ canonical): the anchor masses are $\pi_0 = 0.577$, $\pi_1 = 0.287$, $\pi_2 = 0.137$, and the per-k true-consensus rates are $r_0 = 0.052$ (9/173), $r_1 = 0.698$ (60/86), $r_2 = 0.878$ (36/41). The identity reproduces both headline accuracies: independent $\pi_2 = 0.137$ (direct: $41/300 = 0.137$) and deliberated $\sum_k \pi_k r_k = 0.350$ (direct majority-of-finals convention: $104/300 = 0.347$; the 0.3-point difference between the cluster-label and majority-of-finals conventions is disclosed, not smoothed). The decomposition makes the paper's central tension quantitative: the rescue term $\pi_1 r_1 = 0.200$ is over six times the manufactured-truth term $\pi_0 r_0 = 0.030$, so on this pool deliberation is a net answerer, while Section 3.1 shows that the agreement it manufactures on the $k = 0$ mass is where false consensus concentrates. The earlier reconstructed masses ($0.540/0.323/0.137$), computed before the deposit was queryable, differ by a $k = 0/k = 1$ boundary classification on ~ 3.4 points of items; the deposit values above supersede them (ledger entry 65).

C.9 WHAT IS PROVED AND WHAT IS MEASURED

Statement	Content	Status
Lemma 1	Condorcet majority: $P(\text{Maj correct}) > p$, monotone in n, exponential convergence	Theorem (standard; proof given)
Lemma 2	Unanimity with dispersed errors concentrates posterior on truth, rate $\log(p/\bar{q})$; heterogeneous form in the following remark	Theorem (List–Goodin direction; i.i.d. premise stated)
Proposition 1	$F_{\text{del}} > F_{\text{ind}}$ iff $(1-w)(f_{\text{new}} - F_{\text{ind}}) > w(F_{\text{ind}} - f_{\text{inh}})$; sufficient condition verified by measurement	Theorem; all three terms measured ($w = 0.443$, $f_{\text{inh}} = 0.706$, $f_{\text{new}} = 0.781$; identity reproduces F_{del})
Proposition 2	Restricted DAL: (T1)–(T4) $\Rightarrow$ $D_{\text{del}} \leq D_{\text{ind}}$, equality iff decision-relevant independence preserved; (T4) holds with margin (measured $\mu = 0.70$ against 0.25)	Theorem (from model assumptions)
Conjecture 1	DAL, general form; falsifier: $D_{\Pi} > D_{\text{ind}}$ at $n \geq 150$ on a truth-scarce pool	Conjecture; supported by weak tier, registered E1, EXP-1 equality case
Theorem 1	Coverage Law: $\text{FalseRate}(p, k) = p\varepsilon + (1-p)\phi(k)$; ϕ nonincreasing	Theorem given (A1)–(A4); $\varepsilon \in [0.02, 0.05]$, $\phi = 1.00/0.69/0.52$ measured; (A4) is exact in the sweeps and fails at pool level under live retrieval
Corollary 1	$\text{FalseRate} < \delta$ iff $p > 1 - (\delta - \varepsilon)/(\phi(k) - \varepsilon)$, for $\varepsilon < \delta < \phi(k)$; $\delta = 0.10$, $\phi(0) = 1$: $p > 0.90$ idealized, 0.918–0.947 with measured ε	Corollary; validated by sweep ($r = -0.996$) and live retrieval (20.5% vs 25.0% predicted)
Empirical Regularity 1	First-anchor threshold: step ORs ≈ 39 then ≈ 2.9; smooth logistic compresses to geometric-mean OR ≈ 11 and misfits $k = 1$	Empirical regularity; no theorem
Proposition 3	Diagnosticity–coverage frontier $D^*(\gamma)$ nonincreasing; threshold rules optimal; interaction escapes the constraint set at the price Proposition 2 prices	Proposition with proof sketch (exchange argument)
Proposition 4	Anchor-mass decomposition with $\pi_3 = 0$ by the screen; measured masses $\pi = (0.577, 0.287, 0.137)$ reproduce both headline accuracies exactly	Deposit-verified identity (canonical $n = 300$)

Integrity notes. The registered E1 identifiers (OSF, anonymized view-only link: `https://osf.io/a5t3v/overview?view_only=be06a28afaf74a3f8cbc02d33cd337da`) are used throughout; the earlier generation-mixed frontier run is a separate experiment and none of its numbers appear in this appendix. EXP-6 (GSM8K and MMLU-hard) has since executed on two panels, and where this appendix uses its outcomes (Remark 12), the resumed arm's basis is the repaired-judge registered analysis and the local arm's verdicts are invariant under the repair. The anchor masses in Proposition 4 are computed directly from the deposit (`tier2/anchor_mass.py`), superseding the earlier reconstructed values (ledger entry 65). The core study holds 301 records on 300 questions (one resume-time duplicate); the canonical basis deduplicates by question, keeping the first occurrence, and every number in this appendix is reported on the executed canonical basis ($n = 300$; headline accuracy $104/300 = 34.7\%$ majority-of-finals, 0.350 cluster-label).