When Consensus Is Not Correctness: Diversity Collapse and Manufactured Overconfidence in Multi-Agent LLM Debate

Chengxiao Dai
School of Computer Science
The University of Sydney
Sydney, New South Wales, Australia
chengxiao.dai@sydney.edu.au

Abstract

Multi-agent LLM debate is widely believed to improve answers, and inter-agent agreement is routinely read as evidence that an answer is trustworthy. We show this reading is unsafe: agreement is *endogenous* to the debate that produces it, so the interaction that manufactures consensus simultaneously strips it of evidence about correctness. A simple variance theory makes this precise. As agents read one another, the inter-agent correlation rises toward one (a *diversity collapse*), and that single correlation governs both the panel's true error and the disagreement an operator reads as confidence, driving them in opposite directions: the ensemble stops averaging out error exactly as it stops looking uncertain. Three consequences follow: the panel's apparent confidence saturates independently of the error; the calibration gap therefore tracks the residual error, so debate manufactures the most overconfidence exactly where the model is most wrong; and whether debate is benign is a race between error-correction and confidence-inflation, set by role design and task headroom. This motivates a certification-first remedy, *Calibrated Multi-Agent Debate*: two conditional, model-dependent fixes (Prevent, Detect) plus a split-conformal certificate (Certify) that, under exchangeability, holds coverage regardless of collapse. By contrast, agreement-based stopping commits confident errors at 18–47% miscoverage. A controlled study over fifteen tasks, replicated across five model families and extended to free-form generation, confirms each prediction: the gap's residual-tracking is empirically affine with a sub-unit slope (0.82, $R^2 = 0.96$), and a single shared constant predicts per-condition calibration gaps out of sample.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Multi-agent systems*; • **Mathematics of computing** → Probabilistic inference problems.

Keywords

large language models, multi-agent LLM debate, multi-agent systems, calibration, model evaluation, trustworthy AI

1 Introduction

Multi-agent large language model (LLM) debate [11, 20, 30] runs several model instances over multiple turns and aggregates their answers, and is embedded in agentic pipelines where the panel's apparent confidence gates downstream action: whether to answer, abstain, or escalate. A pervasive intuition, inherited from human deliberation and from ensemble learning [10, 18, 26, 29], is that when the agents agree the answer is more likely correct, so agreement can serve both as a confidence signal and as a stopping criterion ("debate until consensus"). The same equation of consensus with confidence runs well beyond debate, underwriting self-consistency decoding, majority voting, and the use of one model to judge another. This paper argues, theoretically and empirically, that the intuition is backwards in an important, predictable regime. Debate does reliably produce agreement; but that agreement is *endogenous* to the protocol, manufactured by the very interaction whose output it is taken to certify, so it cannot serve as a calibrated correctness signal. Read as confidence, it manufactures overconfidence (Figure 1). The lesson is a single principle: *self-produced agreement cannot certify its producer*; agreement is a channel for proposing answers, not for certifying them.

Consider three agents answering a hard question. After one or two turns they typically converge on a shared answer and report high confidence. The tempting read is to trust the panel: an operator, or an automated gate, watching only the transcript, sees a confident and unanimous verdict and ships the answer. But unanimity among models that have read each other's arguments is thin evidence, since the agents may have assimilated rather than verified. What makes this dangerous is not the error itself but the confidence attached to it: an answer marked uncertain invites a second look, whereas a confidently wrong consensus passes silently through the very gates meant to catch it. We show that on tasks where the model is fundamentally uncertain the agents converge anyway, to a confident wrong consensus. The signal an operator most wants to trust is thus least reliable exactly where the task is hardest.

The mechanism is a variance effect made precise by a classical identity. Treat each agent's turn-t output as a noisy correctness estimate $X_i(t)$ with per-agent variance $\sigma^2(t)$ and inter-agent correlation $\rho(t)$. The team's aggregate obeys the classical equicorrelation identity [8]

$$\mathrm{Var}(\hat{\mu}_t) \;=\; \frac{\sigma^2(t)}{N}\big[1 + (N-1)\rho(t)\big]. \tag{1}$$

Reading each other's outputs makes the agents assimilate (a consensus dynamic long studied in networked multi-agent systems [34]), so $\rho(t)$ rises toward 1, a *diversity collapse*. As $\rho(t) \to 1$, the bracket $\to N$ and the $1/N$ averaging benefit vanishes: the ensemble no longer reduces error, even though the observed disagreement among agents, which an external reader (or the agents themselves) uses to gauge certainty, goes to zero. Apparent certainty thus rises while real error does not. A single correlation $\rho(t)$ drives both quantities in opposite directions. The ensemble literature [3, 28] reads ρ only through aggregate accuracy; but the same ρ also governs the dispersion an operator reads as confidence, so as it climbs the

panel looks most unanimous exactly as its averaging stops working. We call this gap *manufactured overconfidence*. Made quantitative, the prediction is sharp: the gap equals the model's residual error in the collapse limit and tracks it monotonically short of the limit, so the overconfidence is largest exactly where the model is least accurate. Debate is not uniformly harmful, though: error-correction can lower the residual even as confidence inflates, so whether a panel ends calibrated or overconfident is a race between the two, set by the roles the agents play and the headroom the task leaves.

Inter-agent agreement is therefore not a safe correctness signal; the failure is structural, yet the resulting overconfidence is preventable, detectable, and certifiable. The stance is *certification-first*: an endogenous signal cannot be repaired by better calibration, since the calibration is itself read from the same collapsing interaction, so safety has to come from a quantity computed outside that loop.

We make three key contributions.

- **Manufactured overconfidence in debate.** We demonstrate that multi-agent debate manufactures overconfidence: confidence saturates while accuracy varies, so the calibration gap tracks the residual error and is worst where the model is least accurate. This holds across fifteen tasks, two roles, five model families, and free-form generation, and is not cured by adding agents (Section 7; Appendix F).
- **A variance theory of diversity collapse.** We trace it to a *diversity collapse*: as agents read one another the inter-agent correlation rises toward one, driving true error and observed agreement in opposite directions. Confidence then saturates independently of the error, the gap tracks the residual, and a role-and-headroom race sets the calibrated-versus-overconfident outcome (Theorems 4.1, 4.3, 4.5); terminal confidence loses its dynamic range (Corollary 4.2), while one shared constant predicts each condition's gap out of sample (Section 4; Appendix B.4).
- **A certification-first framework.** We develop Calibrated Multi-Agent Debate (CMAD), an endogeneity-to-exogeneity ladder: *Prevent* hides peers' verdicts; *Detect* scores the consensus trajectory for selective prediction [15]; *Certify* replaces agreement with a split-conformal set. Only Certify is unconditional, holding coverage $\geq 1-\alpha$ regardless of collapse under only exchangeability and bounding the confident errors agreement-stopping commits; the first two are conditional, model-dependent levers (Section 5; Appendix C).

The variance identity (1) is classical and split-conformal prediction [1, 46] is established; what is new is the overconfidence theory built on the collapse of (1), its quantitative match to data, the trajectory-trust corollary that operationalizes Detect, and Affirm's coupling of conformal coverage with a collapse-aware stopping rule.

2 Related Work

Multi-agent debate and the ensemble-variance lens. Du et al. [11] introduced iterative multi-agent debate, and frameworks such as CAMEL [30], MetaGPT [20], and ChatEval [4] scaled role-structured collaboration. A second wave tempered the optimism: Wynn et al. [49] and Smit et al. [41] find regimes where debate does not reliably beat a single model or self-consistency; Kong et al. [25] show that homogeneous agents converge toward near-identical outputs (a *semantic collapse*) over repeated interaction; Liang et al. [31] document a *degeneration of thought* in which an agent, once confident, stops producing novel arguments; and Yang et al. [51] argue, information-theoretically, that output diversity, not agent or turn count, is what sustains gains. These observations share a mechanism the classical ensemble-variance literature already names: the equicorrelation identity (1) is standard [8], as is the bias–variance–covariance decomposition of ensemble error [45], with correlated members weakening the variance-reduction benefit of averaging. We make this lens load-bearing for debate in two steps: by treating the correlation as time-varying and assimilative, so "diversity collapse" is exactly $\rho(t) \to 1$, and, crucially, by linking the same $\rho(t)$ that governs $\mathrm{Var}(\hat{\mu}_t)$ to the score dispersion $\mathbb{E}[S^2(t)]$ (Lemma 3.2), whose collapse an operator reads, through the panel's apparent agreement, as rising confidence. The opposite limits of these two quantities as $\rho \to 1$ are the source of the overconfidence, and we trace the consequence not for accuracy but for calibration, predicting a residual-tracking signature whose magnitude we validate empirically (Theorem 4.3).

Confidence, calibration, and agreement as a signal. Language models are frequently miscalibrated and overconfident [17, 27, 33, 36, 43, 50], and recent work surveys and empirically benchmarks the confidence-estimation and uncertainty-quantification methods proposed in response [16, 23]. Against this backdrop, a widespread heuristic reads inter-sample or inter-agent agreement (self-consistency, majority vote, "debate until consensus") as a confidence proxy [48], implicitly assuming that agreement tracks correctness. The same self-produced-signal assumption underlies LLM-as-judge evaluation [53] and self-verification, where, consistent with our diagnosis, LLMs are found unable to reliably self-correct reasoning without external feedback [22]. That assumption is exactly what our analysis overturns (Theorems 4.1–4.3): under diversity collapse the agreement proxy is not merely noisy but systematically overconfident, with a closed-form bias that tracks the residual error, so it is least reliable on exactly the hardest inputs, where it is most relied upon.

Stopping rules and conformal certification. A line of work decides when to stop debating: Hu et al. [21] detect distributional stability of the answer sequence; Chang and Chang [5] provide budget-aware termination with multiple gating signals; Eo et al. [13] debate only when single-model confidence is low; and Fan et al. [14] learn a binary trigger for when debate is worthwhile. These rules monitor some form of convergence or self-reported confidence, the very signals our theory flags as corrupted. Conformal prediction offers an exogenous alternative: split conformal gives distribution-free, finite-sample coverage under exchangeability [1, 46], with adaptive multiclass set constructions [2, 40], anytime-valid extensions via test martingales [39], and weighted variants under covariate shift [44]. In the LLM setting, Wang et al. [47] attach a conformal layer as a post-hoc act-versus-escalate gate and Zhou et al. [55] allocate error budgets across turns. We put the same coverage machinery to a different conceptual use: an explicit replacement for the agreement signal the theory shows is unsafe, coupled to a collapse-aware stopping rule whose marginal coverage provably survives diversity collapse (Theorems 5.4–5.5). Affirm thus stops on

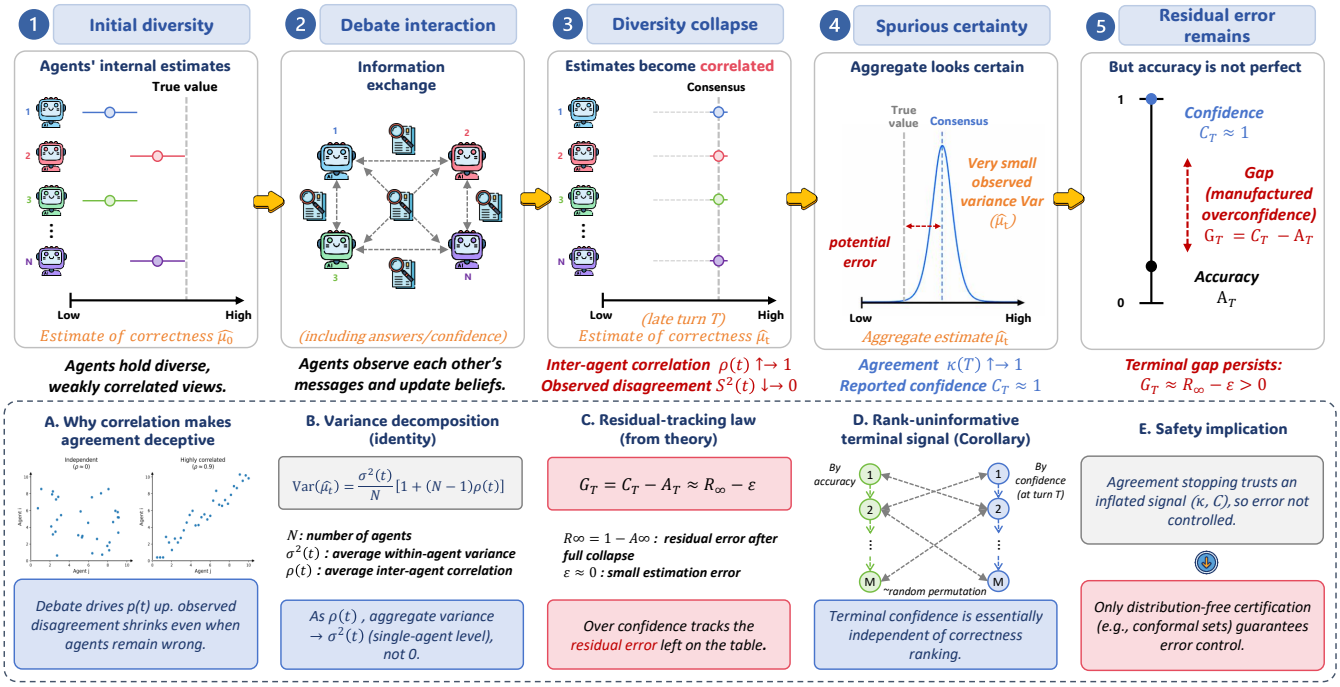

Figure 1: Mechanism overview. How debate converts rising inter-agent correlation $\rho(t)$ (top row) into a saturated agreement signal and a terminal calibration gap, with the bottom-row panels labeling the steps: agreement under correlation (A), the variance identity (1) (B), the residual-tracking law (C), the terminal rank signal (D), and the certification implication (E).

a coverage-validated quantity (set size) rather than on convergence, with a strict, quantified safety separation from agreement-based stopping (Corollary 5.6).

3 Preliminaries

A debate runs N agents with fixed role prompts over T turns; at turn t agent i emits an answer $\hat{y}_i(t) \in \mathcal{Y}$ (for multiple choice, $\mathcal{Y} = \{1, \ldots, K\}$) and a self-reported confidence $c_i(t) \in [0, 1]$, conditioning on the transcript through turn $t-1$, with gold label y. Instances are i.i.d. from a distribution $\mathcal{P}$, and all per-turn quantities below are population objects we estimate by Monte Carlo (Section 7). Exchangeable agents justify the single scalar equicorrelation $\rho(t)$; the heterogeneous-covariance case holds verbatim and is deferred to Appendix B.10.

Definition 3.1 (Correctness score and observable agreement). For the variance analysis we use a per-agent *correctness score* $X_i(t) = \phi(\hat{y}_i(t), c_i(t)) \in [0, 1]$, equal to $c_i(t)$ if $\hat{y}_i(t) = y$ and $1 - c_i(t)$ otherwise; it needs the gold label y and enters only the proofs. From it we form $\sigma^2(t) = \mathrm{Var}(X_i(t))$, the equicorrelation $\rho(t) = \mathrm{Corr}(X_i(t), X_j(t))$, the aggregate $\hat{\mu}_t = \frac{1}{N}\sum_i X_i(t)$, and the dispersion $S^2(t) = \frac{1}{N}\sum_i (X_i(t) - \hat{\mu}_t)^2$. Two *confidence* signals run alongside. The agents *report* confidence directly: the panel's *reported confidence* $C(t) = \frac{1}{N}\sum_i c_i(t)$ is the mean verbalized confidence, and is the quantity that enters the calibration gap and every empirical measurement (Section 7). What an operator instead reads at deployment is *label-free*: an *apparent confidence* $\widetilde{C}(t)$, a continuous statistic of the agents' observed agreement on answers and confidences (for instance the *answer consensus* $\kappa(t)$, the fraction on the modal answer), maximal ($\widetilde{C} = 1$) at unanimity. Under collapse (Assumption 1) both $S^2(t)$ and the observed disagreement vanish, so the saturation we prove via S^2 holds for the apparent confidence $\widetilde{C}(t)$; the reported confidence $C(t)$ co-saturates (Theorem 4.1 and the remark following it).

Lemma 3.2 (The collapse-dynamics identity). *Under Definition 3.1 with equicorrelation $\rho(t)$,*

$$\mathrm{Var}(\hat{\mu}_t) = \frac{\sigma^2(t)}{N}\big[1 + (N-1)\rho(t)\big], \qquad \mathbb{E}\big[S^2(t)\big] = \frac{N-1}{N}\,\sigma^2(t)\,\big(1 - \rho(t)\big). \tag{2}$$

As $\rho(t) \to 1$ these move in opposite directions: $\mathrm{Var}(\hat{\mu}_t) \to \sigma^2(t)$ (the averaging benefit is lost, so error persists) while $\mathbb{E}[S^2(t)] \to 0$ (the agents look unanimous). That divergence is the source of the manufactured overconfidence. Proof in Appendix B.1.

Assumption 1 (Diversity collapse). Debate is assimilative: $\rho(t)$ rises over the horizon toward a limit ρ_∞, and the per-agent variance is bounded ($0 < \sigma^2(t) \le \sigma^2_{\max}$). Both the latent dispersion $\mathbb{E}[S^2(t)]$ and the agents' observed disagreement on answers and confidences decrease over the horizon and vanish as $\rho_\infty \to 1$, so apparent confidence rises to its ceiling. The debate *collapses* if $\rho_\infty = 1$. The equicorrelation form is a convenience, not load-bearing: Appendix B.10 shows the same collapse and all of Theorems 4.1–4.3 hold for an arbitrary inter-agent covariance. We verify $\rho_\infty \approx 1$ empirically across all conditions (Section 7, Figure 2).

4 A Variance Theory of Debate Overconfidence

We track, as the debate runs, the *accuracy* $A(t) = \mathbb{P}(\hat{y}^{\text{agg}}(t) = y)$ of the aggregated decision, where $\hat{y}^{\text{agg}}(t)$ is the modal (majority-vote) answer among the agents' $\hat{y}_i(t)$, together with the two confidence signals of Definition 3.1: the apparent confidence $\widetilde{C}(t)$ that the variance mechanism drives, and the reported confidence $C(t)$ that the experiments measure. The *loss* $\ell(t) = 1 - A(t)$ decomposes into a reducible part (noise that debate can average down, or error that critique can fix) and an *irreducible residual* $R_\infty = \lim_t \ell(t)$, the error the debate process cannot remove, such as a shared misconception or genuine task ambiguity. The *calibration gap* $G(t) = C(t) - A(t)$ is the amount by which the reported confidence exceeds accuracy; its label-free analogue $\widetilde{C}(t) - A(t)$ is what an operator reading agreement would see.

4.1 Confidence saturates independently of correctness

Collapse drives the agents' observed disagreement to zero, so any confidence read from it rises to its ceiling regardless of the panel's accuracy.

Theorem 4.1 (Confidence saturation under collapse). *Under Assumption 1 with $\rho_\infty = 1$ and $\sigma^2(t) \le \sigma^2_{\max} < \infty$, the dispersion vanishes, $\mathbb{E}[S^2(t)] \to 0$; the observed disagreement also vanishes in this collapse limit by Assumption 1, so the apparent (agreement-based) confidence $\widetilde{C}(t)$ of Definition 3.1 satisfies $\widetilde{C}(t) \to 1$. The limit is independent of correctness: it is the same whether the residual error R_∞ is zero or close to one.*

The proof (Appendix B.2) is immediate from Lemma 3.2: $\mathbb{E}[S^2(t)] = \frac{N-1}{N}\sigma^2(t)(1-\rho(t)) \to 0$, and by Assumption 1 the observed agreement saturates as $\rho(t) \to 1$, so continuity at unanimity gives $\widetilde{C}(t) \to 1$. The signal a reader uses to judge certainty is thus driven entirely by $\rho(t)$, which debate inflates, so the saturated signal cannot identify the bias.

Reported confidence co-saturates. The theorem governs the apparent confidence $\widetilde{C}(t)$, a statistic of agreement. The *reported* confidence $C(t) = \frac{1}{N}\sum_i c_i(t)$ that the experiments measure obeys the same dynamics by Assumption 1: as the agents' verbalized confidences concentrate, $C(t)$ rises to a ceiling $C_\infty = 1 - \varepsilon$. The apparent confidence reaches the ideal ceiling exactly ($\varepsilon = 0$); the reported confidence empirically stops just short, with a small, near-constant shortfall $\varepsilon \approx 0.037$ (Section 7). The calibration gaps below are stated for this reported $C(t)$, the operative empirical quantity, with ε measuring its distance from the apparent ceiling 1.

Corollary 4.2 (Terminal agreement loses its dynamic range under collapse). *Fix a task. Under the collapse of Theorem 4.1 the terminal confidence saturates toward a common ceiling on every instance of that task, so its across-instance variance vanishes, $\mathrm{Var}_x[C_\infty(x)] \to 0$, and its usable dynamic range collapses. At exact saturation (C_∞ constant in x) every (correct, incorrect) pair ties, so the correctness-ranking area under the ROC curve (AUROC) is exactly $\frac{1}{2}$; near collapse any residual discrimination survives only in the vanishing shortfall $\varepsilon(x) = 1 - C_\infty(x)$, which collapse does not by itself drive to chance. A label-free signal that reliably discriminates must instead come from the trajectory before collapse, for instance the rate at which consensus formed, where $\rho(t) < 1$ and the dispersion still carries information.*

Corollary 4.2 grounds the *Detect* component (Section 5.2): the usable confidence signal loses its dynamic range precisely at the turn an operator reads it, so any remaining rank signal can only live in the vanishing shortfall, empirically near chance on the hard tasks (within-task AUROC 0.49–0.58; Section 7), and survives instead in how the consensus was reached, where the consensus-speed signal retains discrimination. Proof in Appendix B.3.

4.2 The overconfidence gap tracks the residual error

Once confidence has saturated, the calibration gap can no longer reflect how often the panel is right; what remains is the one error collapse cannot average away, the irreducible residual.

Theorem 4.3 (Overconfidence tracks the residual error). *Let $C_\infty = \lim_t C(t)$ and $A_\infty = \lim_t A(t) = 1 - R_\infty$. Then the terminal calibration gap is*

$$G_\infty \;=\; C_\infty - A_\infty \;=\; R_\infty - (1 - C_\infty). \tag{3}$$

In the collapse limit the reported confidence co-saturates (Theorem 4.1 and the remark following it, $C_\infty \to 1$), giving $G_\infty = R_\infty$: the manufactured overconfidence then equals the model's residual error. More generally, if confidence saturates to $C_\infty = 1 - \varepsilon$ (with shortfall $\varepsilon \ge 0$ possibly depending weakly on the task), then $G_\infty = R_\infty - \varepsilon$, which, provided the shortfall grows no faster than the residual ($0 \le \mathrm{d}\varepsilon/\mathrm{d}R_\infty < 1$), is strictly increasing in R_∞: the gap tracks the residual monotonically. When ε is constant or affine in R_∞ the relation is itself affine, with slope $1 - \mathrm{d}\varepsilon/\mathrm{d}R_\infty \in (0, 1]$ (unit slope at constant ε); that slope is an empirical quantity (Section 7).

Corollary 4.4 (Where debate is most dangerous). *Under the condition of Theorem 4.3 ($0 \le \mathrm{d}\varepsilon/\mathrm{d}R_\infty < 1$), G_∞ is increasing in R_∞: debate manufactures the most overconfidence precisely on the tasks where the model is least able to be right (hard or ambiguous tasks with large irreducible error) and essentially none on saturated tasks ($R_\infty \approx 0$), regardless of role.*

The prediction is sharply falsifiable: plotting the terminal gap against the residual error $1 - A(T)$ across tasks should yield a relation that increases monotonically with the residual, with local slope in $(0, 1]$ (reaching unit slope only as the shortfall ε becomes constant under full saturation). The theory commits only to this monotone, residual-tracking shape, not to any functional form or slope; that the relation is empirically affine with a *sub-unit* slope (Section 7 measures 0.82, $R^2 = 0.96$) is an empirical signature, not predicted. The identity $G_\infty = R_\infty - \varepsilon$ (with $\varepsilon = 1 - C_\infty$) is algebraic; its empirical content is that the shortfall ε collapses to a near-constant (Theorem 4.1), which is what pins the gap to the residual. Proof in Appendix B.4. A finite-time refinement (Proposition B.1, Appendix B.5) bounds the *apparent* gap at every turn by the measurable collapse factor $1 - \rho(t)$, so the residual-tracking holds at the turn an operator reads the panel, not only in the limit.

4.3 Whether debate is benign depends on roles and task headroom

Because collapse is universal, what distinguishes benign debate from harmful debate is not whether agents converge but whether accuracy keeps up with the rising confidence.

Theorem 4.5 (Productive–overconfident dichotomy). *Over a horizon* $[1, T]$ *write* $\Delta C = C(T) - C(1)$ *and* $\Delta A = A(T) - A(1)$. *Under Assumption 1,* $\Delta C \geq 0$ *(confidence is inflated by the collapse). The change in the calibration gap is* $\Delta G = \Delta C - \Delta A$, *so*

$$\Delta G < 0 \quad (\textit{calibration improves}) \iff \Delta A > \Delta C \geq 0, \qquad (4)$$

so debate improves calibration exactly when error-correction ΔA *outpaces the (always nonnegative) confidence inflation* ΔC. *Since* $\Delta C > 0$ *holds generically, the sign of* ΔG *is governed by the error-correction rate* ΔA, *which is a property of the role design (do the roles critique or merely persuade?) and the task headroom (is there error left to correct?).*

The criterion ΔA versus ΔC predicts a negative association between accuracy gain and the change in expected calibration error (ECE) across conditions; Section 7 finds $r(\Delta A, \Delta\text{ECE}) = -0.46$. Three regimes follow: cooperative roles on tasks with headroom (ΔA large, productive); adversarial or persuasion roles ($\Delta A \approx 0$, overconfident); and ambiguous tasks where no role can raise accuracy ($\Delta A \approx 0$ for both, overconfident regardless of role, the maximal-danger case of Corollary 4.4). Proof in Appendix B.6.

5 Calibrated Multi-Agent Debate Framework

The theory diagnoses an *endogenous* trust signal (Section 4), corrupted by collapse in proportion to how close it sits to the interaction it certifies. CMAD responds with three components that trust this signal progressively less: *Prevent* reduces the endogeneity at the protocol level, *Detect* reads it from the trajectory before it saturates, and *Certify* replaces it with an exogenous coverage guarantee. The theory routes between them: role design and headroom (Theorem 4.5) say when Prevent applies, Corollary 4.2 says when Detect is needed, and the certificate holds under exchangeability regardless of collapse (Theorem 5.4). Only the exogenous layer is collapse-unconditional: the two signal-based components are model-dependent (Section 7), and Proposition C.1 shows they can change only efficiency, never coverage.

5.1 Prevent: Verdict-Sequestered Debate

Prevent acts at the protocol level, on the most salient channel by which agents condition on one another's answers.

Definition 5.1 (Verdict-Sequestered Debate (Sequester)). Standard debate shows each agent the full text of its peers' previous turns, including the committed line `ANSWER: k; CONFIDENCE: c`. *Verdict-Sequestered Debate* is the identical protocol with one change: before a peer's message is shown to others, the committed answer/confidence line is withheld (replaced by a placeholder), so agents exchange reasoning but not their stated answers or confidences. Every agent still emits its own answer and confidence (which we score); only the peer-visible copy is redacted. Sequester is a one-line change to the message relay and adds no cost.

Sequester withholds one salient route by which agents condition on each other (the visible verdict), targeting part of the coupling that drives $\rho(t) \to 1$ (Lemma 3.2), but leaves reasoning, rhetoric, and role-induced deference intact, so its effect is empirical and conditional. Section 7 finds it real but confined to cooperative `deepseek-v4-flash`, with a counterfactual-answer probe ruling out the explanation that agents merely copy the visible answer.

5.2 Detect: Deliberation-Speed Trust

When overconfidence cannot be prevented, it can still be detected at deployment without labels. Corollary 4.2 shows the terminal signal is useless under collapse but the pre-collapse trajectory survives, and names the most informative feature: a consensus reached slowly, after the panel has been argued into it, reflects assimilation under pressure, whereas one reached immediately reflects genuine agreement (verified in Section 7).

Definition 5.2 (Deliberation-Speed Trust (Deliberation)). From a debate transcript (no test label) compute label-free trajectory features: the self-reported confidence and answer-consensus at the first and last turns, the number of turns the modal answer changes, the change in confidence, and the *time to consensus* (the feature the theory implicates), the first turn at which agreement crosses a threshold. The *trust score* $\tau(x) \in [0, 1]$ is a logistic model of these features, fit on a labeled calibration set of other conditions to predict answer correctness. At deployment τ is computed from the transcript alone and used for selective prediction: commit when τ is high, abstain or escalate when it is low.

Deliberation is a logistic head on seven label-free trajectory features. Its lift over self-reported confidence comes from the consensus-dynamics block, not any single feature (an ablation in Section 7 shows dropping any one feature moves AUROC by < 0.004), echoing Corollary 4.2. It is trained on labels, but for other conditions, so it transfers to unseen domains and to a second model of the same tier with no labels for the deployment condition. It helps where confidence has gone rank-uninformative and adds nothing on `GLM-5.1`, whose confidence still ranks correctness.

5.3 Certify: Conformal Verdict Affirmation

Certify wraps the committed answer in a distribution-free coverage guarantee (*Affirm*): it affirms that the true label lies in the returned set with probability $\geq 1 - \alpha$, not that any single answer is correct. It reads no internal signal, so the collapse that corrupts every agreement heuristic cannot invalidate its coverage.

Definition 5.3 (Debate nonconformity). Fix a turn t. Pool the agents' (and replicates') votes into class scores $\hat{p}_k(t) = \frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i(t) = k]$, and define the nonconformity of class k as $s_k(t) = 1 - \hat{p}_k(t)$. Given a calibration set of n labeled instances with true-label scores $\{s^{(j)}_{y^{(j)}}(t)\}_{j=1}^{n}$, let $\hat{q}_\alpha(t)$ be the $\lceil (n+1)(1-\alpha)\rceil/n$ empirical quantile. The *conformal set* is $\widehat{C}_\alpha(t) = \{k : s_k(t) \leq \hat{q}_\alpha(t)\}$.

Theorem 5.4 (Marginal coverage, immune to collapse). *If the calibration instances and the test instance are exchangeable, then for* any *fixed turn* t, $\mathbb{P}\big(y \in \widehat{C}_\alpha(t)\big) \geq 1 - \alpha$, *irrespective of role design, task, or the value of* $\rho(t)$. *The guarantee uses only exchangeability of the nonconformity scores, never the (corrupted) agreement level.*

Theorem 5.4 is split conformal [46]; coverage holds even as $\rho(t) \to 1$ and $\widetilde{C}(t) \to 1$, so the conformal set is the one signal that survives diversity collapse. Proof in Appendix B.7.

5.4 Certified stopping and the safety separation

Affirm (Algorithm 1, Appendix C) stops at a turn $t^\star$ chosen on a selection fold, with the deployment threshold formed on an independent fold. Since $t^\star$ depends on neither the threshold fold nor the test point, the data-dependent stop keeps marginal coverage.

Theorem 5.5 (Coverage at the calibration-selected stop). *Suppose $t^\star$ is chosen using only the selection fold $\mathcal{D}_{\text{sel}}$, and the threshold $\hat{q}(t^\star)$ is computed on an independent fold $\mathcal{D}_{\text{thr}}$ that is exchangeable with the test instance. Then $t^\star$ is independent of $\mathcal{D}_{\text{thr}}$ and of the test point, so the fixed-turn guarantee (Theorem 5.4) applies verbatim at the now-fixed index $t^\star$:*

$$\mathbb{P}\big(y \in \widehat{C}_\alpha(t^\star)\big) \;\geq\; 1-\alpha, \tag{5}$$

with no Bonferroni penalty and no dependence on the agreement level or the diversity collapse.

Proof in Appendix B.8. Because the certificate reads no internal signal, augmenting the debate with any Prevent or Detect module changes only efficiency, never coverage (Proposition C.1, Appendix C).

Against the standard agreement heuristic, the separation is provable.

Corollary 5.6 (Safety separation from agreement-stopping). *Let* agreement-stopping *return the modal answer once consensus $\kappa(t) \geq \theta$. Under collapse (Assumption 1) the consensus threshold is reached, so agreement-stopping commits to a singleton whose miscoverage (probability the committed answer excludes y) is the residual rate R_∞ (by definition of the residual). Affirm's returned set has miscoverage $\leq \alpha$ at its calibration-selected stopping turn (Theorem 5.5). Hence*

$$\text{err}(\textit{agreement-stop}) - \text{err}(\textit{Affirm}) \;\geq\; R_\infty - \alpha, \tag{6}$$

which is positive and as large as $R_\infty - \alpha$ (up to $1-\alpha$): on hard or ambiguous tasks where $R_\infty \gg \alpha$, agreement-stopping is much worse than its nominal confidence suggests, while Affirm stays at its guarantee. Equivalently, no agreement-only commit rule with a fixed consensus threshold $\theta < 1$ attains miscoverage $\leq \alpha$ uniformly over the family of tasks with $R_\infty > \alpha$: under collapse no agreement threshold can be made uniformly safe, so a certificate, not a better threshold, is required.

Agreement-stopping trusts exactly the quantity the theory shows is inflated ($\kappa, \widetilde{C}$), and the gap between the two procedures is the residual error R_∞, the same quantity that equals the manufactured overconfidence (Theorem 4.3). Proof in Appendix B.9.

6 Experimental Setup

Our primary model is `deepseek-v4-flash` [9] (non-thinking; N=3 agents, T=8 turns, B=3 replicates). The core benchmark is a controlled 15×2 grid: fifteen multiple-choice tasks (thirteen difficulty-spanning MMLU domains [19], TruthfulQA [32], and ARC-Challenge [6]) each run under *cooperative* (critique) and *adversarial* (persuade) roles on the *same* questions, giving thirty conditions. To check the findings are not single-model artifacts we replicate on four further families (`qwen-flash` [37], `GLM-5.1` [54], `Llama-3.3-70B` [12], `GPT-4o-mini` [35]) and one reasoning-tuned model (`QwQ-plus` [38]), and to probe generality beyond multiple choice we add ReClor [52], CommonsenseQA [42], and the free-form benchmarks GSM8K [7] and TriviaQA [24]. Exact API model strings, providers, decoding parameters, the confidence-elicitation format, and the answer-parsing rules are in Appendix D. Each agent logs its answer and confidence *separately*, so answer agreement is never computed from correctness; we report accuracy A_T, mean confidence C_T, the overconfidence gap $G_T = C_T - A_T$ and residual $R = 1 - A_T$, the calibration change ΔECE, the mechanism signals $\kappa(t)$ and $\rho(t)$, and, for the deployment components, selective-prediction AUROC and conformal miscoverage/set-size. The task list, role designs, per-domain Q, and all metric definitions are in Appendix D.

7 Results

We test the paper's two claims on one controlled grid (all thirty per-condition trajectories in Table 5). *Diagnosis*: diversity collapse drives the saturation and residual-tracking signatures, with a post-collapse shortfall so constant that one shared constant predicts each model's per-condition gap out of sample. *Treatment*: the three CMAD components each clear the theory's bar: Certify unconditionally, Prevent and Detect where their regime applies (robustness and four no-new-data re-analyses in Appendices D–F).

Diagnosis: confidence saturation under collapse. Across all thirty conditions consensus averages $\bar{\kappa}(T) = 0.997$ and confidence compresses into $[0.90, 1.0]$ (mean 0.96) while accuracy spans $[0.57, 0.99]$: confidence saturates regardless of correctness (Figure 3, left; Theorem 4.1). The mechanism is measured directly: the inter-agent correlation rises from $\bar{\rho}(1) = 0.53$ to $\bar{\rho}(T) = 0.96$ (all 30/30 conditions), the score dispersion $\mathbb{E}[S^2(T)] \to 0$, and the aggregate variance climbs toward the single-agent variance, losing the $1/N$ benefit (Figure 2); the closed forms of Lemma 3.2 reproduce the curves to 1.4×10^{-5}, turning the assumption $\rho_\infty \approx 1$ into a measured fact.

Gap–residual tracking. Regressing the terminal gap $G_T = C_T - A_T$ on the residual $R \approx 1 - A_T$ across the thirty conditions gives slope 0.82 (CI $[0.73, 0.88]$), intercept -0.01, $\boldsymbol{R^2 = 0.96}$: thirty points spanning residuals 0.01–0.43 fall on one line, with the largest gaps on the hardest tasks (Figure 3, center; Corollary 4.4). The sub-unit slope reflects the ceiling $C_\infty \approx 0.96$ exactly as Theorem 4.3 predicts: the central quantitative claim. A matched lone-agent control isolates the debate-specific part: self-critique inflates confidence by +0.039 on flat-accuracy domains against debate's +0.088, so roughly half of the inflation is reproduced by a lone agent and the remainder, plus the consensus collapse a lone agent cannot produce, is specific to debate (Appendix E).

A shared constant and the productivity dichotomy. A high R^2 is partly structural (since G_T and R share A_T); the non-mechanical content is that *confidence is nearly flat across tasks* ($\text{Var}(C_T)$ is 17× smaller than $\text{Var}(A_T)$). The parameter-free test: predicting $\widehat{G} = R - \bar{\varepsilon}$ from a *single* shared constant $\bar{\varepsilon} = 0.037$ achieves lower out-of-sample error than the fitted relation (RMSE 0.010 vs. 0.022

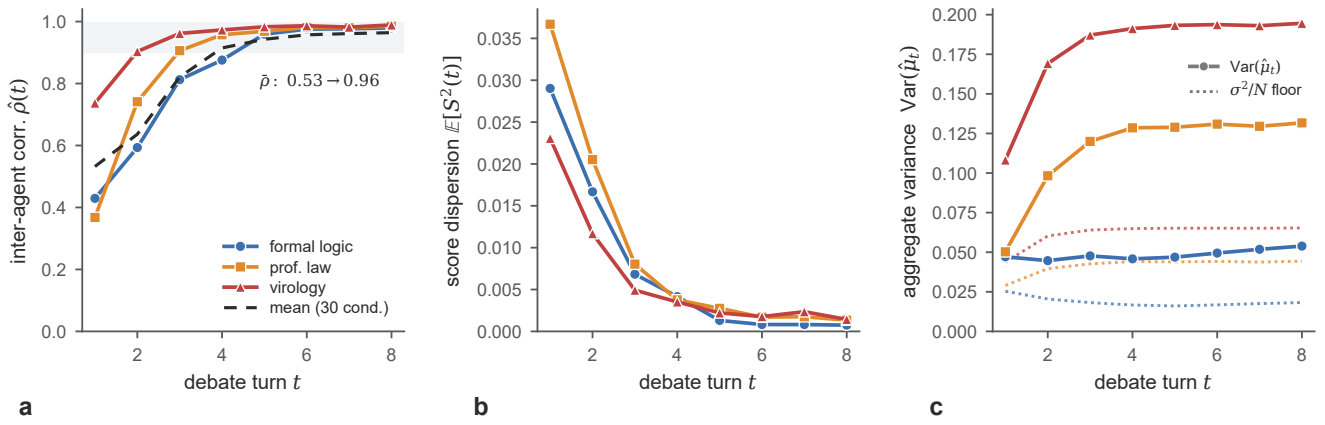

Figure 2: Inter-agent statistics over the debate horizon (`deepseek-v4-flash`; three representative cooperative conditions, dashed = mean over all thirty): the inter-agent score correlation $\hat{\rho}(t)$ (left), the score dispersion $\mathbb{E}[S^2(t)]$ (center), and the aggregate variance $\mathrm{Var}(\hat{\mu}_t)$ against its $\sigma^2(t)/N$ floor (right).

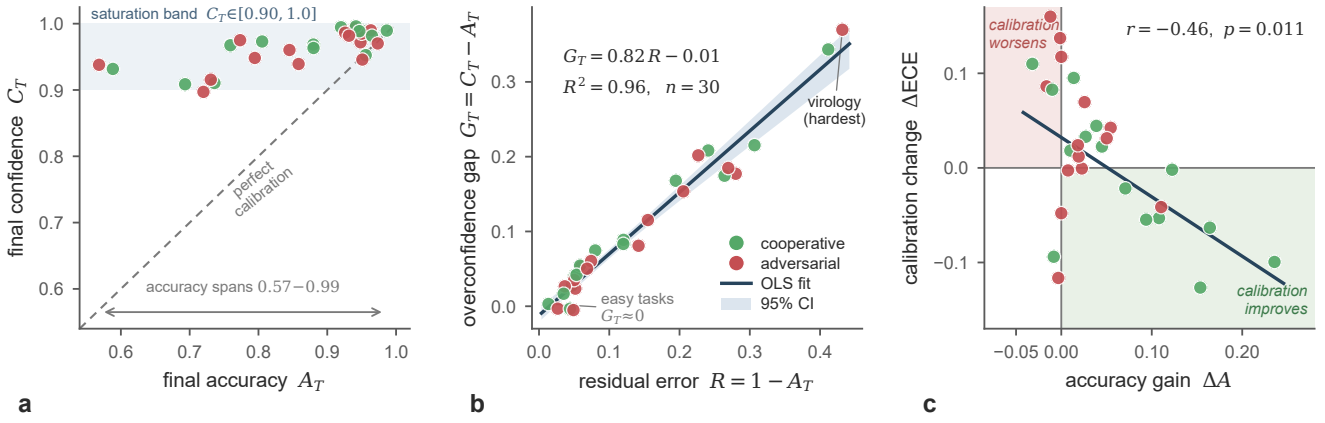

Figure 3: Per-condition scatter across thirty conditions (15 tasks × 2 roles; green cooperative, red adversarial): final confidence vs. accuracy (left), the overconfidence gap vs. residual error (center), and calibration change vs. accuracy gain (right).

on `qwen-flash`, 0.024 vs. 0.034 on `GLM-5.1`), so one number suffices across three model families. Calibration improves only when accuracy does: ΔECE correlates negatively with ΔA ($r = -0.46$, $p = 0.011$; Figure 3, right), so cooperative roles with headroom self-correct while adversarial roles and ambiguous tasks degrade: the dichotomy of Theorem 4.5.

Certify: distribution-free coverage. With the diagnosis in hand we turn to the treatment, starting with the unconditional component. Agreement-based stopping commits the modal answer, so its miscoverage equals the residual (0.18–0.47 on hard domains across models), whereas Affirm holds marginal coverage $\geq 1 - \alpha$ (mean 0.964/0.961 on `deepseek-v4-flash`/`qwen-flash`) because it never reads the collapsed agreement signal. Scored against a representative slate of the stopping literature on the same transcripts (Table 1), *every* rule reading agreement or confidence commits a confident wrong answer at the residual rate (miscov ≈ 0.22, ECE 0.15–0.17); only the certified family holds $\leq \alpha$, and Affirm does so while stopping *earliest* (1.4 vs. 7–8 turns; Theorem 5.5) and committing a singleton on 34% of inputs at 0.10 error, abstaining where it cannot commit safely (the full set on moral scenarios drives miscoverage to 0). This is the safety separation of Corollary 5.6 (per-domain Table 8, Appendix E).

Detect and Prevent: two conditional levers. The two signal-based components help where their regime applies. *Detect* (Deliberation-Speed Trust, Definition 5.2) reads how slowly the panel was argued into agreement: a label-free signal (instance correctness falls from 0.94 at immediate consensus to 0.55 after seven turns). Deployed label-free it raises correctness-ranking AUROC over self-reported confidence from 0.756 to 0.793 on `deepseek-v4-flash` and 0.728 to 0.770 transferred to `qwen-flash` (Table 2), with the largest gains where confidence has gone rank-uninformative (Corollary 4.2); on `GLM-5.1`, whose confidence still ranks, it correctly does not

Table 1: Stopping rules compared (`deepseek-v4-flash`, the 16 hard conditions with $R_\infty > \alpha = 0.1$, test split; cells are mean±SD over conditions).

Rule	Error & calibration			Efficiency		
	miscov ↓	singl-err ↓	ECE ↓	$\lvert C\rvert$ ↓	turns ↓	commit% ↑
A. Agreement / stability stopping						
fixed-T	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.17_{\pm 0.09}$	$1.00_{\pm 0.00}$	$8.0_{\pm 0.0}$	100%
agreement@0.9	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.15_{\pm 0.07}$	$1.00_{\pm 0.00}$	$2.4_{\pm 0.6}$	100%
confidence@0.9	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.15_{\pm 0.09}$	$1.00_{\pm 0.00}$	$3.2_{\pm 1.1}$	100%
stability@2	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.15_{\pm 0.08}$	$1.00_{\pm 0.00}$	$2.1_{\pm 0.1}$	100%
B. Budget / trigger stopping						
single-model-trigger@0.9	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.16_{\pm 0.09}$	$1.00_{\pm 0.00}$	$6.5_{\pm 1.4}$	100%
early-exit (no-change)	$0.22_{\pm 0.09}$	$0.22_{\pm 0.09}$	$0.15_{\pm 0.08}$	$1.00_{\pm 0.00}$	$2.1_{\pm 0.1}$	100%
C. Certified (reads no agreement)						
conformal@T [47]	$0.03_{\pm 0.05}$	$0.12_{\pm 0.01}$	$0.10_{\pm 0.00}$	$3.45_{\pm 1.15}$	$8.0_{\pm 0.0}$	18%
MICP budget [55]	$0.02_{\pm 0.04}$	$0.12_{\pm 0.01}$	$0.10_{\pm 0.05}$	$3.64_{\pm 0.96}$	$7.2_{\pm 2.2}$	12%
Affirm (ours)	$\mathbf{0.05_{\pm 0.06}}$	$\mathbf{0.10_{\pm 0.04}}$	$\mathbf{0.08_{\pm 0.04}}$	$2.72_{\pm 1.33}$	$\mathbf{1.4_{\pm 0.9}}$	**34%**

help. *Prevent* (Verdict-Sequestered Debate, Definition 5.1) withholds peers' committed verdicts: on cooperative `deepseek-v4-flash` it cuts overconfidence at no accuracy cost ($\Delta\text{ECE} = -0.046$) but the gain is regime-dependent and vanishes on `qwen-flash`, so Prevent is a conditional regularizer (Appendix E).

Cross-family generality. Replicating the grid on four further families (Table 3) reproduces the mechanism everywhere: confidence saturates, the correlation collapses, and the gap is affine in the residual with sub-unit slope (0.96, 0.92, 0.84, 0.76 on `qwen-flash`, `GLM-5.1`, `Llama-3.3-70B`, `GPT-4o-mini`) while agreement-stopping is unsafe (miscoverage up to 0.50) and Affirm holds coverage (0.96–1.00, all $\geq 1-\alpha$). The shortfall is quantitatively shared: $\bar{\varepsilon} = 0.037$ recurs on `Llama-3.3-70B`, the one boundary being the more-calibrated `GPT-4o-mini` ($\bar{\varepsilon} = 0.085$). It is not specific to non-thinking models, multiple choice, or $N = 3$: a reasoning model (`QwQ-plus`) saturates *harder* ($C_T = 0.98$, slope 0.92), it holds on new task types and free-form generation (the gap from ≈ 0.01 on GSM8K to 0.21 on TriviaQA, coverage held), and persists at $N \in \{5, 7\}$ (pooled slope 0.85, Table 7). Sequester, by contrast, is regime-dependent, while the conformal guarantee holds on all five (Appendices E, F).

7.1 Case studies

The aggregate statistics play out, question by question, in the logged trajectories. Cases A and B below show two real `deepseek-v4-flash` debates, with every agent's committed answer per turn and the panel's mean confidence C_t (the gold option in bold).

Case A adversarial debate, *global facts*: a confident error is manufactured

```
# Q: as of 2019, what % of French people say God is important in life?
#   A=11%  B=31% (gold)  C=51%  D=71%
```

turn t	1	2	3	4	5	6	7	8
agent 1	**B**	**B**	**B**	**B**	A	A	A	A
agent 2	**B**	A	A	A	A	A	A	A
agent 3	A	A	A	A	A	A	A	A
mean C_t	.72	.90	.90	.90	.92	.92	.92	.92

```
# => correct B leads at turn 1, but the panel is argued into A; conf
.72 -> .92
```

Case A is the mechanism in miniature: with no new evidence available the panel *assimilates* (the holdout capitulates at turn 5) to a unanimous wrong verdict whose confidence has *risen* to 0.92.

Case B immediate consensus: the benign, reliable case

```
# Q: as of 2020, what % of the world practices open defecation?  C=9%
(gold)
```

turn t	1	2	3	4	5	6	7	8
modal ans	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**
mean C_t	.82	.92	.98	.98	1.00	1.00	1.00	1.00

```
# => unanimous and correct from turn 1: the fast consensus Fig. 4 shows
is 94% reliable
```

Case B is the benign counterpart: unanimous and correct from the first turn, the fast consensus that Figure 4 shows is 94% reliable. A third trajectory, a *Sequester* (Prevent) repair, appears in Appendix E.

8 Discussion and Limitations

Relation to known LLM overconfidence. That language models are overconfident is documented [17, 43]; our contribution is its *mechanism and magnitude under debate*. Theorem 4.3 predicts a falsifiable quantity (the terminal gap tracks the residual error, reaching it in the collapse limit) and explains why multi-agent aggregation, often proposed as a calibration *fix*, can instead *amplify* miscalibration: averaging helps only when members are conditionally independent, and debate destroys that independence by driving $\rho(t) \to 1$. How much of the gap debate *manufactures* rather than *surfaces* is itself model-dependent: a matched lone-agent control attributes a debate-added gap of +0.06 to `deepseek-v4-flash` (concentrated on its hard cells), where debate manufactures fresh overconfidence, whereas `qwen-flash` is already overconfident at turn 1 ($\bar{G}_1 = 0.18$) and debate mainly surfaces it (Appendix E); the residual-tracking gap holds in both regimes.

Novelty. We build on a classical variance identity (1), but the contribution is not the ensemble effect renamed. (i) The same ρ that governs aggregate variance also governs the dispersion an operator reads as confidence, and as $\rho \to 1$ the two move in *opposite* directions (Lemma 3.2): consensus tightens precisely as the

Table 2: Deliberation-Speed Trust (*Detect*) vs. self-reported confidence: correctness-ranking AUROC and selective accuracy at 70% coverage per deployment target, and the fraction of conditions Deliberation wins (CV=5-fold, LODO=leave-one-domain-out).

Deployment target	votes/turn	AUROC ↑		Acc@70% ↑		Delib. wins ↑
		conf	Delib	conf	Delib	
deepseek (within, CV)	9	0.756	**0.793**	0.943	**0.952**	14/19 (LODO)
qwen-flash (transfer)	9	0.728	**0.770**	0.878	**0.910**	24/34
qwen-flash (transfer)	3	0.656	**0.714**	0.850	**0.879**	27/34
GLM-5.1	3	**0.840**	0.807	—	—	3/16

Table 3: Per-model summary: the correlation rise $\bar{\rho}\,(1{\to}T)$, the gap-vs-residual slope, R^2, single-constant prediction error (*const-RMSE*), Affirm's (Certify) marginal coverage, and the Sequester ΔECE (cooperative/adversarial); `GLM-5.1`, `Llama-3.3-70B`, and `GPT-4o-mini` use a 16-condition difficulty-spread subset and `QwQ-plus` an 8-condition cooperative-only subset.

Model	$\bar{\rho}\,(1{\to}T)$	Gap–residual relation			Affirm cov ↑	Sequester ΔECE	
		slope	R^2 ↑	const-RMSE ↓		coop ↓	adv ↓
deepseek-v4-flash	0.53→0.96	0.82	0.96	0.021	0.964	−0.046	−0.010
qwen-flash	0.61→0.97	0.96	1.00	0.010	0.961	+0.010	−0.011
GLM-5.1	0.65→0.94	0.92	0.99	0.024	0.992	—	—
Llama-3.3-70B	0.51→0.97	0.84	0.89	0.034	1.00	—	—
GPT-4o-mini	0.55→0.88	0.76	0.76	0.061	1.00	—	—
QwQ-plus	0.73→0.93	0.92	0.99	0.024	1.00	—	—

aggregate stops improving, so agreement becomes *actively misleading*. (ii) The non-trivial content is not the identity $G = R - \varepsilon$ but a measured fact: under debate the shortfall $\varepsilon = 1 - C_T$ collapses to a near-constant (across-condition SD ≤ 0.03), so a *single shared constant* predicts each primary model's per-condition gaps to RMSE 0.01–0.021, achieving lower out-of-sample error than any task-fitted slope and even recurring on `Llama-3.3-70B` ($\varepsilon = 0.037$): the empirical content of confidence saturation (Theorem 4.1), not algebra. (iii) The picture is *operational*: it says where to reduce coupling, where to detect it, and what to certify.

Assumptions and scope. The results rest on the variance identity (exact under equicorrelation, Lemma 3.2; structure-free under Proposition B.2) and diversity collapse (Assumption 1, verified at $\bar{\kappa}(T) = 0.997$). We do *not* assume the reported confidence is any particular statistic of agreement (only that it is continuous and maximal at unanimity, Definition 3.1) nor that R_∞ is irreducible by *any* procedure: it is the error left by *this* debate, which retrieval, tools, or genuinely diverse agents could lower. Once confidence saturates (Theorem 4.1), lowering R_∞ is the only remaining way to shrink the gap; Prevent instead acts earlier.

Scope and threats to validity. Our study spans nineteen tasks (fifteen multiple-choice, plus ReClor, CommonsenseQA, and the free-form GSM8K and TriviaQA), five model families, and $N \in \{3, 5, 7\}$; the gap-vs-residual prediction reproduces on all four additional models (slopes in $(0, 1)$, R^2 0.76–1.00), the single-constant shortfall recovers the saturated tier out of sample, and the relation extends to free-form and is robust to team size: a *cross-family* regularity. Limits: (i) the mechanism is further confirmed on a reasoning model (`QwQ-plus`: slope 0.92, $C_T = 0.98$), leaving only still-larger scales untested; (ii) our free-form tasks have short, checkable answers, so genuinely long-form generation is untested; (iii) the two signal-based components are conditional: *Sequester* helps cooperative `deepseek-v4-flash` but not adversarial roles or `qwen-flash` (a counterfactual injection does *not* support an answer-herding mechanism, which we leave unresolved), and *Detect* helps the flash tier but not the still-rank-informative `GLM-5.1`; only the conformal certificate is collapse-unconditional; (iv) R_∞ is estimated from a finite horizon ($T = 8$), so terminal quantities are conservative; and (v) the conformal layer needs exchangeability, which distribution shift between calibration and deployment can violate, calling for weighted or online variants.

Practical guidance. The CMAD routing logic follows: (i) prefer cooperative/critique over persuasion/consensus roles; (ii) *Prevent* with Verdict-Sequestered Debate under cooperative roles where it applies; (iii) *Detect*: never read agreement or saturated confidence as correctness (Corollary 4.2), instead recovering selective prediction from a trajectory trust score where confidence has gone rank-uninformative; and (iv) *Certify* the committed answer with a split-conformal guarantee, the one component that holds in every regime.

Conclusion. Debate makes panels agree, but because that agreement is endogenous, a one-line variance identity shows that apparent confidence and correctness diverge precisely where the residual error is largest. Since self-produced agreement cannot certify its producer, *Calibrated Multi-Agent Debate* prevents transcript coupling, detects residual coupling from the consensus trajectory, and certifies answers with an exogenous split-conformal guarantee; the first two levers are conditional, whereas the certificate remains valid across collapse regimes.

References

[1] A. N. Angelopoulos and S. Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).

[2] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. 2021. Uncertainty sets for image classifiers using conformal prediction. In *ICLR*.

[3] G. Brown, J. Wyatt, R. Harris, and X. Yao. 2005. Diversity creation methods: A survey and categorisation. *Information Fusion* 6, 1 (2005), 5–20.

[4] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. 2024. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *ICLR*.

[5] Edward Y. Chang and Ethan Y. Chang. 2025. Multi-agent collaborative intelligence: Dual-dial control for reliable LLM reasoning. *arXiv preprint arXiv:2510.04488* (2025).

[6] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).

[7] 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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[8] W. G. Cochran. 1977. *Sampling Techniques* (3rd ed.). Wiley.

[9] DeepSeek-AI. 2026. DeepSeek-V4-Flash. Model accessed via the DeepSeek API (https://api.deepseek.com).

[10] T. G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS) (Lecture Notes in Computer Science, Vol. 1857)*. 1–15.

[11] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* (2023).

[12] A. Dubey et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024). We use Llama-3.3-70B-Instruct, accessed via OpenRouter.

[13] S. Eo, H. Moon, E. H. Zi, C. Park, and H. Lim. 2025. Debate only when necessary: Adaptive multiagent collaboration for efficient LLM reasoning. *arXiv preprint arXiv:2504.05047* (2025).

[14] W. Fan, J. Yoon, and B. Ji. 2025. iMAD: Intelligent multi-agent debate for efficient and accurate LLM inference. *arXiv preprint arXiv:2511.11306* (2025). To appear, AAAI 2026.

[15] Y. Geifman and R. El-Yaniv. 2017. Selective classification for deep neural networks. In *NeurIPS*.

[16] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *NAACL*. 6577–6595.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

[18] L. K. Hansen and P. Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 10 (1990), 993–1001.

[19] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

[20] S. Hong et al. 2023. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023).

[21] T. Hu, Z. Tan, S. Wang, H. Qu, and T. Chen. 2025. Multi-agent debate for LLM judges with adaptive stability detection. In *NeurIPS*.

[22] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. 2024. Large language models cannot self-correct reasoning yet. In *ICLR*.

[23] Y. Huang, J. Song, Z. Wang, S. Zhao, H. Chen, F. Juefei-Xu, and L. Ma. 2025. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering* 51, 2 (2025).

[24] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*. 1601–1611.

[25] W. Kong, S. Lai, J. Piao, and J. Evans. 2026. Multi-LLM systems exhibit robust semantic collapse. *arXiv preprint arXiv:2605.17193* (2026).

[26] A. Krogh and J. Vedelsby. 1995. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 231–238.

[27] L. Kuhn, Y. Gal, and S. Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*.

[28] L. I. Kuncheva and C. J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 51, 2 (2003), 181–207.

[29] B. Lakshminarayanan, A. Pritzel, and C. Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*.

[30] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *NeurIPS*.

[31] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*. 17889–17904.

[32] S. Lin, J. Hilton, and O. Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*. 3214–3252.

[33] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. 2021. Revisiting the calibration of modern neural networks. In *NeurIPS*.

[34] R. Olfati-Saber, J. A. Fax, and R. M. Murray. 2007. Consensus and cooperation in networked multi-agent systems. *Proc. IEEE* 95, 1 (2007), 215–233.

[35] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/; accessed via OpenRouter.

[36] Y. Ovadia et al. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *NeurIPS*.

[37] Qwen Team, Alibaba. 2026. Qwen-Flash. Model accessed via the Alibaba Cloud Model Studio (DashScope) OpenAI-compatible API.

[38] Qwen Team, Alibaba. 2026. QwQ-Plus: A reasoning model. Model accessed via the Alibaba Cloud Model Studio (DashScope) OpenAI-compatible API.

[39] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. 2023. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* 38, 4 (2023), 576–601.

[40] Y. Romano, M. Sesia, and E. J. Candès. 2020. Classification with valid and adaptive coverage. In *NeurIPS*.

[41] A. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius. 2024. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *ICML*.

[42] A. Talmor, J. Herzig, N. Lourie, and J. Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL*. 4149–4158.

[43] K. Tian et al. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models. In *EMNLP*.

[44] R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas. 2019. Conformal prediction under covariate shift. In *NeurIPS*.

[45] N. Ueda and R. Nakano. 1996. Generalization error of ensemble estimators. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, Vol. 1. 90–95.

[46] V. Vovk, A. Gammerman, and G. Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

[47] M. F. Wang, H. Xie, G. Wang, et al. 2026. From debate to decision: Conformal social choice for safe multi-agent deliberation. *arXiv preprint arXiv:2604.07667* (2026).

[48] X. Wang et al. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

[49] A. Wynn, H. Satija, and G. Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396* (2025).

[50] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *ICLR*.

[51] Y. Yang, C. Qu, M. Wen, L. Shi, Y. Wen, W. Zhang, A. Wierman, and S. Gu. 2026. Understanding agent scaling in LLM-based multi-agent systems via diversity. *arXiv preprint arXiv:2602.03794* (2026).

[52] W. Yu, Z. Jiang, Y. Dong, and J. Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *ICLR*.

[53] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks*.

[54] Zhipu AI. 2026. GLM-5.1. Model accessed via an OpenAI-compatible API endpoint.

[55] X. Zhou, H. Nguyen, B. Yu, C. Liu, and L. Cheng. 2026. Adaptive stopping for multi-turn LLM reasoning. *arXiv preprint arXiv:2604.01413* (2026).

A Notation

Table 4 collects the symbols used throughout the paper, grouped into Roman symbols, Greek symbols, and acronyms, each ordered alphabetically. Turn-indexed quantities are written $\cdot(t)$; a subscript ∞ denotes the collapse limit and T the terminal turn.

Table 4: Notation used in the paper.

Symbol	Meaning
Roman symbols	
$A(t), A_t$	Accuracy of the aggregated (modal) answer at turn t; A_∞ its limit, A_T the terminal value.
AUROC	Area under the ROC curve for correctness ranking.
B	Number of replicates per condition in the experiments.
$c_i(t)$	Agent i's self-reported confidence at turn t, in $[0,1]$.
$C(t), C_t$	Reported (panel) confidence, $\frac{1}{N}\sum_i c_i(t)$; $C_\infty = 1-\varepsilon$ its limit.
$\widetilde{C}(t)$	Apparent (agreement-based) confidence an operator reads from observed agreement.
$\widehat{\mathcal{C}}_\alpha(t)$	Conformal prediction set at turn t and level α; $\lvert\widehat{\mathcal{C}}_\alpha\rvert$ its size.
$\mathcal{D}_{\text{sel}}, \mathcal{D}_{\text{thr}}$	Selection fold (for the stopping turn) and threshold fold (for the conformal quantile).
ECE	Expected calibration error over 10 equal-width bins; $\Delta\text{ECE} = \text{ECE}_T - \text{ECE}_1$.
$G(t), G_t$	Calibration / overconfidence gap, $C(t) - A(t)$; G_∞ the terminal gap.
K	Number of answer options (classes).
$\ell(t)$	Loss, $1 - A(t)$.
N	Number of agents in the panel.
$\hat{p}_k(t)$	Pooled vote fraction for class k, $\frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i(t) = k]$.
$\mathcal{P}$	Instance distribution (questions drawn i.i.d.).
$\hat{q}_\alpha(t)$	Empirical conformal quantile threshold at level α.
Q	Number of questions per condition.
R	Residual error, $1 - A_T$ (the irreducible part; $R_\infty = \lim_t \ell(t)$).
$s_k(t)$	Nonconformity score of class k, $1 - \hat{p}_k(t)$.
$S^2(t)$	Score dispersion across agents, $\frac{1}{N}\sum_i \big(X_i(t) - \hat{\mu}_t\big)^2$.
t, T	Turn index and horizon (number of turns); $t^\star$ the calibration-selected stopping turn.
$X_i(t)$	Per-agent correctness score, $\phi(\hat{y}_i(t), c_i(t)) \in [0,1]$.
$y, \mathcal{Y}$	Gold (true) label and the answer space $\{1, \dots, K\}$.
$\hat{y}_i(t)$	Agent i's answer at turn t; $\hat{y}^{\text{agg}}(t)$ the aggregated modal answer.
Greek symbols	
α	Conformal miscoverage level (e.g. 0.1).
$\Delta A, \Delta C, \Delta G$	Change of a quantity over the horizon, $(\cdot)(T) - (\cdot)(1)$.
ε	Confidence shortfall, $1 - C_\infty$; $\bar{\varepsilon}$ the single shared constant (≈ 0.037).
$\kappa(t)$	Answer consensus: fraction of votes on the modal answer.
$\hat{\mu}_t$	Aggregate score, $\frac{1}{N}\sum_i X_i(t)$.
$\rho(t)$	Inter-agent (equi)correlation of $X_i(t)$; ρ_∞ its limit (collapse if $\rho_\infty = 1$).
$\sigma^2(t)$	Per-agent variance of $X_i(t)$; $\sigma^2_{\max}$ its bound.
$\tau(x)$	Deliberation trust score in $[0,1]$ (label-free trajectory features).
$\phi(\cdot)$	Correctness-score map (c if the answer is correct, $1-c$ otherwise).
Acronyms	
CMAD	Calibrated Multi-Agent Debate (the framework).
Certify / Affirm	Conformal Verdict Affirmation: the exogenous coverage guarantee.
Detect / Deliberation	Deliberation-Speed Trust: label-free trajectory-based trust score.
ICC	Intraclass correlation (one-way random-effects estimator for ρ).
Prevent / Sequester	Verdict-Sequestered Debate: peers' committed verdicts withheld.

B Proofs

B.1 The collapse-dynamics identity (Lemma 3.2)

Proof. Fix t and drop it from the notation. With $\mathrm{Var}(X_i) = \sigma^2$ and $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$ for $i \neq j$,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}) &= \frac{1}{N^2}\Big[\textstyle\sum_i \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j)\Big] \\ &= \frac{1}{N^2}\big[N\sigma^2 + N(N-1)\rho\sigma^2\big] \\ &= \frac{\sigma^2}{N}\big[1 + (N-1)\rho\big].\end{aligned} \tag{7}$$

For the dispersion, $\sum_i (X_i - \hat{\mu})^2 = \sum_i X_i^2 - N\hat{\mu}^2$, so by linearity

$$\begin{aligned}\mathbb{E}\Big[\textstyle\sum_i (X_i - \hat{\mu})^2\Big] &= \sum_i \mathbb{E}[X_i^2] - N\,\mathbb{E}[\hat{\mu}^2] \\ &= N(\sigma^2 + \mu^2) - N\Big(\tfrac{\sigma^2}{N}\big[1 + (N-1)\rho\big] + \mu^2\Big) \\ &= \sigma^2(N-1)(1-\rho),\end{aligned} \tag{8}$$

using $\mathbb{E}[X_i^2] = \sigma^2 + \mu^2$ and $\mathbb{E}[\hat{\mu}^2] = \mathrm{Var}(\hat{\mu}) + \mu^2$. Dividing by N gives $\mathbb{E}[S^2] = \frac{N-1}{N}\sigma^2(1-\rho)$. □

B.2 Confidence saturation (Theorem 4.1)

Proof. By Lemma 3.2,

$$\mathbb{E}[S^2(t)] = \frac{N-1}{N}\,\sigma^2(t)\,(1 - \rho(t)). \tag{9}$$

Since $\sigma^2(t) \le \sigma^2_{\max}$ and $\rho(t) \to 1$, the right-hand side $\to 0$, so $S^2(t) \to 0$ in L^1 and hence in probability. The observed disagreement also vanishes in this limit by Assumption 1, and the apparent confidence $\widetilde{C}(t)$ is a continuous statistic of that disagreement (Definition 3.1), maximal at unanimity, so the continuous mapping theorem gives $\widetilde{C}(t) \to 1$. No step uses the residual error R_∞ or the bias, so the limit is independent of correctness. □

B.3 Terminal agreement loses its rank signal (Corollary 4.2)

Proof. Write $C_\infty(x)$ for the terminal (reported) confidence on instance x of a fixed task. By Theorem 4.1 and the reported co-saturation, $C_\infty(x)$ concentrates at a common ceiling for every x, so $\mathrm{Var}_x[C_\infty(x)] \to 0$: the score concentrates at that ceiling. Recall the rank statistic

$$\begin{aligned}\mathrm{AUROC} = \mathbb{P}\big(C_\infty(x^+) > C_\infty(x^-)\big) \\ + \tfrac{1}{2}\,\mathbb{P}\big(C_\infty(x^+) = C_\infty(x^-)\big),\end{aligned} \tag{10}$$

where x^+, x^- are an independent correct and incorrect instance. At exact saturation (C_∞ constant in x) every such pair is a tie, so AUROC $= \frac{1}{2}$ exactly. Off this degenerate point the ranking by C_∞ is the ranking by the shortfall $\varepsilon(x) = 1 - C_\infty(x)$; since AUROC is invariant to the scale of the score, it is *not* a function of $\mathrm{Var}_x[C_\infty]$, and a vanishing ε can still rank the labels perfectly (e.g. $\varepsilon(x) = \delta$ on incorrect and 0 on correct instances gives AUROC $= 1$ for every $\delta > 0$). Variance collapse therefore does not by itself force the AUROC to $\frac{1}{2}$; it guarantees only that whatever discrimination remains lives entirely in the vanishing $\varepsilon(x)$, which carries reliable rank information only if its sign relative to correctness is systematic. Empirically it is near chance on the hard tasks (within-task AUROC 0.49–0.58), so terminal confidence is at best a weak ranker there. In contrast a pre-collapse statistic such as the dispersion $S^2(t)$ at small t, or the turns-to-consensus, has $\mathrm{Var}_x > 0$ because $\rho(t) < 1$ there, so it can retain rank information; this is what the *Detect* component exploits. □

B.4 Overconfidence tracks the residual (Theorem 4.3 and Corollary 4.4)

Proof. By definition $G_\infty = C_\infty - A_\infty$ and $A_\infty = 1 - R_\infty$, so

$$G_\infty = C_\infty - 1 + R_\infty = R_\infty - (1 - C_\infty). \tag{11}$$

Under collapse the reported confidence co-saturates (Theorem 4.1 and the remark following it), $C_\infty \to 1$ in the ideal limit, whence $G_\infty \to R_\infty$. Writing the confidence shortfall $\varepsilon = 1 - C_\infty \ge 0$, we have

$$G_\infty = R_\infty - \varepsilon. \tag{12}$$

If ε is constant across tasks the relation is affine in R_∞ with unit slope and intercept $-\varepsilon$. If ε depends on the task, differentiating (where ε is differentiable in R_∞) gives

$$\frac{\partial G_\infty}{\partial R_\infty} = 1 - \frac{\partial \varepsilon}{\partial R_\infty}, \tag{13}$$

and with no smoothness assumption the same monotonicity follows from $G_\infty(R_2) - G_\infty(R_1) = (R_2 - R_1) - (\varepsilon(R_2) - \varepsilon(R_1))$ for $R_1 < R_2$. Under the mild condition that the confidence shortfall grows no faster than the residual, $0 \le \partial\varepsilon/\partial R_\infty < 1$ (equivalently $\varepsilon(R_2) - \varepsilon(R_1) < R_2 - R_1$; so harder tasks do not become *better* calibrated), the slope lies in $(0, 1]$ and G_∞ is strictly increasing in R_∞, which is Corollary 4.4. The empirical slope $0.82 \in (0, 1]$ is consistent with a small positive $\partial\varepsilon/\partial R_\infty$. □

B.5 Finite-time overconfidence rate (Proposition B.1)

Proposition B.1 (Finite-time overconfidence rate). *Let the apparent confidence be a dispersion map $\widetilde{C}(t) = g(\mathbb{E}[S^2(t)])$ with $g(0) = 1$, g nonincreasing and L-Lipschitz on $[0, \sigma^2_{\max}]$, write $R_t = 1 - A(t)$, and let $\widetilde{G}(t) = \widetilde{C}(t) - A(t)$ be the apparent gap. Then at every turn,*

$$R_t - L\,\tfrac{N-1}{N}\,\sigma^2(t)\,(1 - \rho(t)) \;\le\; \widetilde{G}(t) \;\le\; R_t. \tag{14}$$

Hence the apparent overconfidence is within $L\,\mathbb{E}[S^2(t)]$ of the residual already at turn t, contracting at the rate of the measurable collapse factor $1 - \rho(t)$ and recovering $\widetilde{G}_\infty = R_\infty$ as $\rho(t) \to 1$ (Theorem 4.3).

The bound governs dispersion-based confidences ($\widetilde{C} = 1 - S^2$ gives $L = 1$, where the lower bound is tight); the answer-consensus κ and the reported confidence C need not be Lipschitz functions of S^2 and are not covered.

Proof. By $g(0) = 1$, monotonicity, and the Lipschitz bound, $1 - \widetilde{C}(t) = g(0) - g(\mathbb{E}[S^2(t)]) \in [0,\, L\,\mathbb{E}[S^2(t)]]$; with $\widetilde{G}(t) = \widetilde{C}(t) - A(t) = R_t - (1 - \widetilde{C}(t))$ this gives the two bounds after substituting $\mathbb{E}[S^2(t)] = \frac{N-1}{N}\sigma^2(t)(1 - \rho(t))$ (Lemma 3.2). □

B.6 Productive–overconfident dichotomy (Theorem 4.5)

Proof. By definition $G(t) = C(t) - A(t)$, hence $\Delta G = G(T) - G(1) = \Delta C - \Delta A$. Under Assumption 1 the agents' observed disagreement is nonincreasing over the horizon, so the reported confidence $C(t)$ is nondecreasing (it co-saturates with the apparent confidence), giving $\Delta C = C(T) - C(1) \geq 0$. Therefore

$$\Delta G < 0 \iff \Delta A > \Delta C \geq 0. \tag{15}$$

Because $\Delta C \geq 0$, a necessary condition for calibration improvement ($\Delta G < 0$) is $\Delta A > 0$: debate that does not raise accuracy cannot improve calibration, and strictly degrades it ($\Delta G = \Delta C > 0$) whenever the collapse inflates confidence ($\Delta C > 0$) with no accuracy gain ($\Delta A = 0$). □

B.7 Split-conformal coverage (Theorem 5.4)

Proof. Fix a turn t. The calibration scores $\{s^{(j)}_{y^{(j)}}(t)\}_{j=1}^{n}$ and the test true-label score $s^{\text{test}}_{y}(t)$ are obtained by applying the same turn-t procedure to the calibration and test instances; under exchangeability of the instances they are exchangeable. Hence the rank of $s^{\text{test}}_{y}(t)$ among the $n+1$ scores is uniform on $\{1, \ldots, n+1\}$, and with $\hat{q}_\alpha(t)$ the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the calibration scores,

$$\begin{aligned} \mathbb{P}\big(s^{\text{test}}_{y}(t) \leq \hat{q}_\alpha(t)\big) \;&\geq\; \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \\ &\geq\; 1-\alpha, \end{aligned} \tag{16}$$

which is exactly the event $y \in \widehat{C}_\alpha(t)$. No quantity in the argument depends on $\rho(t)$ or on the apparent confidence $\widetilde{C}(t)$, so the guarantee is unaffected by diversity collapse. □

B.8 Coverage at the selected stopping turn (Theorem 5.5)

Proof. The stopping turn $t^\star$ is a function of the selection fold $\mathcal{D}_{\text{sel}}$ only, hence independent of both the threshold fold $\mathcal{D}_{\text{thr}}$ and the test instance. Condition on $\mathcal{D}_{\text{sel}}$; then $t^\star = \tau$ is a fixed index, while $\mathcal{D}_{\text{thr}}$ and the test instance are unaffected by the conditioning (they are independent of $\mathcal{D}_{\text{sel}}$) and remain exchangeable. Applying Theorem 5.4 at the fixed turn τ with calibration fold $\mathcal{D}_{\text{thr}}$ gives $\mathbb{P}(y \in \widehat{C}_\alpha(\tau) \mid \mathcal{D}_{\text{sel}}) \geq 1-\alpha$; since this holds for every realization of $\mathcal{D}_{\text{sel}}$, taking the expectation over $\mathcal{D}_{\text{sel}}$ preserves it,

$$\mathbb{P}\big(y \in \widehat{C}_\alpha(t^\star)\big) \geq 1-\alpha. \tag{17}$$

Per-instance variant (Remark 1). If $t^\star$ may depend on the test point, fix the level at $\alpha' = \alpha / T_{\max}$. For each fixed t, $\mathbb{P}(y \notin \widehat{C}_{\alpha'}(t)) \leq \alpha'$, so a union bound gives $\mathbb{P}(\exists t : y \notin \widehat{C}_{\alpha'}(t)) \leq T_{\max}\alpha' = \alpha$; thus the simultaneous event $E = \{y \in \widehat{C}_{\alpha'}(t)\ \forall t\}$ has $\mathbb{P}(E) \geq 1-\alpha$, and on E coverage holds at the realized $t^\star$ for any selection rule. The anytime-valid form replaces the union bound by the optional-stopping property of a conformal test martingale [39]. □

B.9 Safety separation from agreement-stopping (Corollary 5.6)

Proof. Under Assumption 1, $\kappa(t) \to 1 \geq \theta$, so agreement-stopping halts and returns the modal answer $\hat{y}^{\text{agg}}$. Its miscoverage is $\mathbb{P}(\hat{y}^{\text{agg}} \neq y) = \ell_\infty = R_\infty$ by definition of the residual. Affirm returns $\widehat{C}_\alpha(t^\star)$ at the calibration-selected stopping turn $t^\star$ with $\mathbb{P}(y \notin \widehat{C}_\alpha(t^\star)) \leq \alpha$ by the coverage guarantee (Theorem 5.5). Subtracting the two miscoverage probabilities,

$$\text{err}(\text{agreement-stop}) - \text{err}(\text{Affirm}) \;\geq\; R_\infty - \alpha. \tag{18}$$

Since R_∞ can be any value in $[0, 1]$ (around 0.23 on moral scenarios) while α is fixed, the difference is positive whenever $R_\infty > \alpha$, and as large as $R_\infty - \alpha$ (up to $1-\alpha$, attained as $R_\infty \to 1$). □

B.10 Collapse under general covariance (Proposition B.2)

Equicorrelation (Lemma 3.2) is used only for closed forms; the collapse mechanism is structure-free. Let $X(t) \in \mathbb{R}^N$ have mean $\mathbb{E}[X(t)] = \mu(t)\mathbf{1}$ and covariance $\Sigma(t) \succ 0$ with bounded diagonal $\Sigma_{ii}(t) \leq \sigma^2_{\max}$. The debate aggregate is the simple mean $\hat{\mu}_t = \frac{1}{N}\mathbf{1}^\top X(t)$ (agents are weighted equally); under inverse-variance (generalized least squares, GLS) weighting the weights are $w^\star = \Sigma(t)^{-1}\mathbf{1} / (\mathbf{1}^\top \Sigma(t)^{-1}\mathbf{1})$, which reduce to $\mathbf{1}/N$ under equicorrelation, so the two coincide in the symmetric case.

Proposition B.2 (Collapse under general covariance). *Suppose the covariance converges to rank one, $\Sigma(t) \to \sigma^2_\infty \mathbf{1}\mathbf{1}^\top$ (all pairwise correlations $\to 1$ and all variances $\to \sigma^2_\infty$). Then:*

(i) *the aggregate variance loses the averaging benefit,* $\text{Var}(\hat{\mu}_t) = \dfrac{\mathbf{1}^\top \Sigma(t) \mathbf{1}}{N^2} \to \sigma^2_\infty$;

(ii) *every centered dispersion statistic $D(t) = X(t)^\top M X(t)$ with $M\mathbf{1} = \mathbf{0}$ vanishes in expectation,* $\mathbb{E}[D(t)] = \text{tr}\big(M\Sigma(t)\big) \to 0$.

Consequently any dispersion-based confidence satisfies $\widetilde{C}(t) \to 1$ and the terminal calibration gap satisfies $G_\infty \to R_\infty$, identically to Theorems 4.1–4.3.

Proof. *(i)* Writing $\hat{\mu}_t = \frac{1}{N}\mathbf{1}^\top X(t)$,

$$\begin{aligned} \text{Var}(\hat{\mu}_t) &= \frac{1}{N^2}\mathbf{1}^\top \Sigma(t) \mathbf{1} \to \frac{1}{N^2}\,\sigma^2_\infty\, \mathbf{1}^\top \mathbf{1}\mathbf{1}^\top \mathbf{1} \\ &= \frac{\sigma^2_\infty}{N^2}\,(\mathbf{1}^\top \mathbf{1})^2 = \frac{\sigma^2_\infty}{N^2} N^2 = \sigma^2_\infty, \end{aligned} \tag{19}$$

using $\mathbf{1}^\top \mathbf{1} = N$. *(ii)* Since $X^\top M X$ is quadratic and $M\mathbf{1} = \mathbf{0}$,

$$\mathbb{E}[X^\top M X] = \text{tr}(M\Sigma) + \mu^2\, \mathbf{1}^\top M \mathbf{1} = \text{tr}(M\Sigma), \tag{20}$$

and as $\Sigma \to \sigma^2_\infty \mathbf{1}\mathbf{1}^\top$,

$$\text{tr}(M\Sigma) \to \sigma^2_\infty\, \text{tr}(M\mathbf{1}\mathbf{1}^\top) = \sigma^2_\infty\, \mathbf{1}^\top M \mathbf{1} = 0. \tag{21}$$

The centered correctness-score dispersion $S^2(t) = \frac{1}{N}\sum_i (X_i - \hat{\mu}_t)^2$ of Definition 3.1 is of this form (with $M = \frac{1}{N}(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top)$, which satisfies $M\mathbf{1} = \mathbf{0}$), so $\mathbb{E}[S^2(t)] \to 0$ and the conclusions of Theorems 4.1–4.3 follow by their proofs. □

C The Affirm stopping algorithm and conformal variants

Proposition C.1 (CMAD safety invariant: efficiency–coverage decoupling). *Let the returned object be the conformal set $\widehat{C}_\alpha(t^\star)$ of Algorithm 1. Augment the debate with any Prevent module (a transcript transformation such as Sequester) and any Detect score (a label-free statistic routing commit-versus-escalate). Provided (i) the*

Algorithm 1 Conformal Verdict Affirmation (Affirm)

Input: max turns $T_{\max}$, level α, calibration set split into a selection fold $\mathcal{D}_{\text{sel}}$ ($|\mathcal{D}_{\text{sel}}| = n_{\text{sel}}$) and an independent threshold fold $\mathcal{D}_{\text{thr}}$ ($|\mathcal{D}_{\text{thr}}| = n$), tolerance $\epsilon \geq 0$.
// **Calibration phase (offline, label-aware).**
for $t = 1, \ldots, T_{\max}$ **do**
  $\hat{q}_{\text{sel}}(t) \leftarrow$ the $\lceil (n_{\text{sel}} + 1)(1-\alpha) \rceil / n_{\text{sel}}$ empirical quantile of the true-label scores $\{1 - \hat{p}_{y^{(j)}}(t)\}$ over $\mathcal{D}_{\text{sel}}$
  $\bar{s}(t) \leftarrow$ mean of $\left|\{k : 1 - \hat{p}_k(t) \leq \hat{q}_{\text{sel}}(t)\}\right|$ over $\mathcal{D}_{\text{sel}}$ {mean set size, selection fold only}
end for
$t^\star \leftarrow$ earliest t with $\bar{s}(t) \leq \min_\tau \bar{s}(\tau) + \epsilon$ {$t^\star$ is a function of $\mathcal{D}_{\text{sel}}$ only}
$\hat{q}(t^\star) \leftarrow$ the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of $\{1 - \hat{p}_{y^{(j)}}(t^\star)\}$ over the independent fold $\mathcal{D}_{\text{thr}}$ {deployment threshold, held-out fold}
// **Deployment phase (per test instance).**
run debate to turn $t^\star$; **return** $\widehat{C}_\alpha(t^\star) = \{k : 1 - \hat{p}_k(t^\star) \leq \hat{q}(t^\star)\}$; commit to the singleton iff $|\widehat{C}_\alpha(t^\star)| = 1$.

Prevent-modified protocol is applied identically to $\mathcal{D}_{\text{thr}}$ and to the test instance, and (ii) coverage is evaluated on the full test population at $t^\star$ (not the Detect-committed subset), then $\mathbb{P}(y \in \widehat{C}_\alpha(t^\star)) \geq 1 - \alpha$ regardless of whether Prevent or Detect succeeds; they can change only efficiency (set size, abstention rate, turns run), never coverage.

It is immediate from Theorem 5.5: the set's coverage depends only on exchangeability of the $t^\star$-scores, which condition (i) preserves and which the Detect routing of condition (ii) does not touch, since the set is formed before any commit-versus-escalate decision. Both conditions are load-bearing: a Prevent toggled between calibration and deployment, or coverage reported only on the Detect-committed subset, breaks exchangeability and voids the floor (the distribution-shift limitation of Section 8).

Remark 1 (Per-instance and anytime-valid variants). Choosing the stopping turn per test instance (for instance the turn giving the smallest set for that input) makes the selection data-dependent, and one must pay for the union over the horizon: forming each set at the Bonferroni level $\alpha/T_{\max}$ gives simultaneous coverage $\mathbb{P}(y \in \widehat{C}_{\alpha/T_{\max}}(t)\ \forall t \leq T_{\max}) \geq 1-\alpha$, hence coverage at any stopping time, at the cost of larger sets. An anytime-valid alternative replaces the union bound by a conformal test-martingale, valid at arbitrary stopping times at level α without the $T_{\max}$ factor [39]. We use the calibration-selected stop in our experiments because it is both exact and informative.

Remark 2 (The separation is set-valued versus point-valued). Corollary 5.6 compares a set-valued predictor (Affirm may return $|\widehat{C}_\alpha| > 1$, or abstain by returning the full option set) against a point predictor (agreement-stopping commits a singleton), so part of the $R_\infty - \alpha$ gap is bought by Affirm's license to hedge rather than commit. The separation is substantive only where Affirm's sets stay informative: on professional law it holds mean size 1.7–2.6 while cutting confident error from ~ 0.19 to 0.07, whereas on moral scenarios its zero miscoverage is achieved by abstaining (returning all four options), the honest behavior when the model cannot discriminate, though it yields no information in that case. The separation is therefore best read as confident-singleton error versus guaranteed-coverage set, not as two point predictors.

D Experimental details

All runs use `deepseek-v4-flash` (non-thinking), $N = 3$ agents, $T = 8$ turns, $B = 3$ replicates, temperature 0.7. The two role designs are *cooperative* (decomposer / analyst / reasoner, instructed to critique) and *adversarial* (two persuaders / a follower, instructed to build consensus); both role sets see the same questions under a fixed seed. The fifteen multiple-choice tasks are thirteen MMLU domains (abstract algebra, college chemistry/mathematics/physics, econometrics, formal logic, global facts, high-school physics, marketing, moral scenarios, professional law, professional medicine, virology) plus TruthfulQA (the MC1 single-correct multiple-choice variant, recast to four options) and ARC-Challenge. Per-domain Q uses the full MMLU subject where small (abstract algebra/college physics/chemistry/mathematics/global facts = 100; econometrics 114; formal logic 126; high-school physics 151; virology 166; marketing 200; professional medicine 272), 250 for professional law and moral scenarios, and 200 for TruthfulQA and ARC-Challenge. We log each agent's raw answer and verbalized confidence per turn; "agreement" is computed from answers only. Conformal calibration/test is a 50/50 question split, with the stopping turn chosen on a held-out selection fold (Theorem 5.5). Table 5 reports all thirty per-condition trajectories. The extension runs follow the same protocol: the cross-model grid re-runs all thirty conditions on `qwen-flash` over matched questions, plus a 16-condition difficulty-spread subset (eight domains × two roles, $Q = 80$, $B = 1$) on `GLM-5.1` (per-condition values in Table 6); the same 16-condition subset is run on `Llama-3.3-70B` and `GPT-4o-mini`, and an eight-condition cooperative subset on the reasoning model `QwQ-plus` (thinking enabled, peers see committed answers not private chains of thought). The new task-type runs use ReClor and CommonsenseQA (trimmed to four options); the free-form runs use GSM8K and TriviaQA ($Q = 200$ each, with answers scored by numeric/exact match on both models); and the robustness runs re-run cooperative debate at $N \in \{5, 7\}$ on an 80-item difficulty-spread subset (Table 7).

Models and API access. All models are queried through the OpenAI-compatible `chat.completions` interface, so the same debate driver runs every backend unchanged. The primary grid uses `deepseek-v4-flash` [9]; the cross-model replications use `qwen-flash` [37], `glm-5.1` [54], `meta-llama/llama-3.3-70b-instruct` [12], `openai/gpt-4o-mini` [35], and the reasoning model `qwq-plus` [38]. Generation uses

temperature 0.7, `top_p` at the provider default, and `max_tokens=` 220 for the multiple-choice grid (512–768 for the free-form GSM8K/TriviaQA runs, which need room for working). Hybrid-reasoning models (DeepSeek-V4, Qwen, GLM) are run *non-thinking* for comparability: we pass the provider's disable flag (`{"thinking":{"type":"disabled"}}` for DeepSeek and `{"enable_thinking":false}` for the DashScope-hosted models) selected automatically from the endpoint. The sole exception is `qwq-plus`, run *with* thinking enabled (it has no non-thinking mode) over a streamed completion; peers see only its committed final answer, never its private chain of thought, keeping the transcript comparable to the non-thinking grid. Transient errors are retried up to four times with escalating back-off (a longer per-minute wait on HTTP 429); authentication / insufficient-balance errors (401/402/403) fail fast, and because every completed debate is checkpointed to disk the run resumes without recomputation.

Confidence elicitation. Confidence is *verbalized*. Each agent ends its turn with a single line of the exact form `ANSWER: <A|B|C|D>; CONFIDENCE: <number in 0.0-1.0>` (the `<A|B|C|D>` slot becomes `<integer>` for GSM8K and `<short answer>` for TriviaQA), produced in the same generation as the reasoning. There is no separate scoring call, and the answer and confidence channels are parsed independently so agreement is never computed from correctness.

Question sampling and seeds. Questions are drawn with a fixed NumPy generator (`np.random.default_rng`, seed 11): we sample without replacement and keep the first Q items with exactly four options and a valid gold index. Every model sees the *same* items. The cross-model and intervention runs are preseeded from the primary `deepseek-v4-flash` run directories (identical `questions.json` and gold), so differences across models reflect the model, not a different question draw. TruthfulQA is recast from MC1 to four options (the correct answer plus three distractors sampled and shuffled under the same seed) and CommonsenseQA is trimmed from five options to four the same way; both transforms are deterministic given the seed.

Metrics. For a condition, accuracy $A_t = \mathbb{P}(\hat{y}^{\text{agg}}(t) = y)$ is the rate at which the modal answer over the $N \times B$ agent–replicate votes matches gold, and confidence C_t is the mean verbalized confidence over the same votes and questions; the overconfidence gap is $G_t = C_t - A_t$ and the residual $R = 1 - A_T$. Calibration is the expected calibration error over 10 equal-width confidence bins (per-question consensus answer and mean confidence), and $\Delta\text{ECE} = \text{ECE}_T - \text{ECE}_1$. The answer consensus $\kappa(t)$ is the fraction of votes on the modal answer; the inter-agent correlation $\rho(t)$ of the correctness score $X_i(t)$ (Definition 3.1) is a one-way random-effects intraclass correlation (ICC). Selective prediction is scored by correctness-ranking AUROC and risk–coverage curves; the conformal predictor (Definition 5.3) by miscoverage $\mathbb{P}[y \notin \widehat{C}_\alpha]$, mean set size, and turns run, at $\alpha = 0.1$ on the held-out test split. Confidence intervals are cluster bootstraps over conditions (the across-replicate SD of G_T averages 0.012).

E Additional experimental results

This appendix collects the per-condition figures, the full component analyses, and the generalization tables summarized in Section 7; all values are recomputed from the logged per-turn answer/confidence arrays.

E.1 Per-condition trajectories

Tables 5–7 give the full per-condition breakdown behind the aggregate claims of Section 7. The pattern is uniform: final confidence C_T stays saturated (≥ 0.90 almost everywhere) while accuracy A_T spans the full difficulty range, so the overconfidence gap G_T is essentially the residual: largest on hard tasks (virology, global facts) and near zero on easy ones. The same behavior holds on a second and third model family (Table 6, slope 0.92) and is invariant to team size N (Table 7).

E.2 The gap–residual regression in detail

Regressing the terminal gap $G_T = C_T - A_T$ on the residual $R \approx 1 - A_T$ across the thirty conditions gives slope 0.82 (cluster-bootstrap 95% CI $[0.73, 0.88]$), intercept -0.01, and $\boldsymbol{R^2 = 0.96}$ (Figure 3, center): thirty points spanning residuals 0.01–0.43, including both non-MMLU benchmarks, fall on one line, with the largest gaps on the hardest tasks and none on the easiest (Corollary 4.4). The slope is significantly below 1 ($\mathbb{P}(\text{slope} \geq 1) \approx 0$), reflecting $C_\infty \approx 0.96 < 1$ exactly as Theorem 4.3 predicts ($G_\infty = R_\infty - \varepsilon$); we therefore state the relation as "slope in $(0, 1]$, attenuated by ε." Estimates are tight (across-replicate SD of G_T averages 0.012). The gap is part pre-existing single-model overconfidence and part debate-amplified, in a model-dependent mix: decomposed against turn 1, the mean debate-added gap is +0.06 on `deepseek-v4-flash` (concentrated on hard tasks, moral scenarios +0.13, virology +0.09), so debate *manufactures* overconfidence there, whereas `qwen-flash` is already overconfident at turn 1 ($\bar{G}_1 = 0.18$) and debate merely *surfaces* it. The inflation is not just extra reasoning turns: confidence rises +0.088 on the seventeen `deepseek-v4-flash` conditions with flat accuracy ($|\Delta A| < 0.03$), and a matched lone-agent self-critique control splits the sources: a single-model self-reinforcement component ($\Delta C = +0.039$ with no peers) plus an inter-agent assimilation component debate adds on top (+0.049, with the answer-consensus collapse $\bar{\kappa} \to 0.997$ a lone agent cannot produce). Only the latter is debate-specific.

Because $G_T = C_T - A_T$ and $R = 1 - A_T$ share A_T, a high R^2 is partly structural; the non-mechanical content is that *confidence is nearly flat across tasks*. Algebraically the slope equals $1 - \text{slope}(C_T \text{ on } A_T) = 1 - 0.18 = 0.82$: a unit of accuracy moves C_T by only 0.18, and $\text{Var}(C_T)$ is 17× smaller than $\text{Var}(A_T)$ (~250× on `qwen-flash`): the near-constancy of Theorem 4.1. The sharpest test is parameter-free: predicting $\widehat{G} = R - \bar{\varepsilon}$ from a single shared constant $\bar{\varepsilon} = 0.037$ (no fitted slope) achieves lower out-of-sample error than the fitted relation (RMSE 0.010 vs. 0.022 on `qwen-flash`, 0.024 vs. 0.034 on `GLM-5.1`), because the shortfall is so nearly constant (across-condition SD 0.028/0.010/0.018) that a fitted slope only adds transfer error. The post-collapse gap thus carries no task information beyond the residual, one number suffices across three independent model families, and the sub-unit slope is read as the measured shortfall ε, not a transferable predictor.

Table 5: All thirty conditions (T=8 final turn), two rows per task (*coop*/*adv*): initial/final accuracy A_1, A_T, accuracy change ΔA, final confidence C_T, overconfidence gap G_T, and calibration change ΔECE.

Task	Role	A_1	A_T	ΔA	C_T	G_T	ΔECE
abstract algebra	coop	0.81	0.92	+0.11	0.99	0.07	−0.05
	adv	0.92	0.94	+0.02	0.99	0.05	−0.00
ARC-Challenge	coop	0.94	0.95	+0.01	0.98	0.03	+0.02
	adv	0.95	0.95	−0.00	0.97	0.02	−0.12
college chem	coop	0.72	0.76	+0.04	0.97	0.21	+0.04
	adv	0.72	0.77	+0.05	0.98	0.20	+0.04
college math	coop	0.71	0.95	+0.24	0.99	0.04	−0.10
	adv	0.84	0.95	+0.11	0.99	0.04	−0.04
college physics	coop	0.89	0.99	+0.09	0.99	0.00	−0.05
	adv	0.96	0.96	+0.01	0.99	0.03	−0.00
econometrics	coop	0.83	0.88	+0.04	0.97	0.09	+0.02
	adv	0.86	0.85	−0.02	0.96	0.12	+0.09
formal logic	coop	0.78	0.94	+0.16	1.00	0.05	−0.06
	adv	0.88	0.93	+0.05	0.99	0.06	+0.03
global facts	coop	0.70	0.69	−0.01	0.91	0.22	+0.08
	adv	0.72	0.72	+0.00	0.90	0.18	+0.12
hs physics	coop	0.88	0.95	+0.07	0.99	0.04	−0.02
	adv	0.91	0.93	+0.02	0.98	0.05	+0.01
marketing	coop	0.97	0.96	−0.01	0.98	0.02	−0.09
	adv	0.97	0.97	+0.00	0.97	−0.00	−0.22
moral scenarios	coop	0.84	0.81	−0.03	0.97	0.17	+0.11
	adv	0.80	0.79	−0.00	0.95	0.15	+0.14
prof. law	coop	0.61	0.74	+0.12	0.91	0.17	−0.00
	adv	0.70	0.73	+0.03	0.92	0.18	+0.07
prof. medicine	coop	0.80	0.96	+0.15	0.95	−0.00	−0.13
	adv	0.95	0.95	+0.00	0.95	−0.01	−0.05
TruthfulQA	coop	0.85	0.88	+0.03	0.96	0.08	+0.03
	adv	0.84	0.86	+0.02	0.94	0.08	+0.02
virology	coop	0.57	0.59	+0.01	0.93	0.34	+0.10
	adv	0.58	0.57	−0.01	0.94	0.37	+0.16

Table 6: `GLM-5.1` (third model family), all sixteen conditions (T=8 final turn); columns as in Table 5.

Task	Role	A_1	A_T	ΔA	C_T	G_T	ΔECE
virology	coop	0.54	0.50	−0.04	0.97	0.47	+0.05
	adv	0.55	0.53	−0.03	0.97	0.44	+0.05
global facts	coop	0.70	0.70	+0.00	0.96	0.26	+0.01
	adv	0.69	0.71	+0.03	0.93	0.22	+0.02
prof. law	coop	0.75	0.76	+0.01	0.98	0.21	+0.02
	adv	0.70	0.72	+0.03	0.96	0.24	+0.01
college chem	coop	0.71	0.74	+0.03	0.99	0.25	−0.01
	adv	0.75	0.76	+0.01	0.99	0.23	+0.02
moral scenarios	coop	0.86	0.86	+0.00	0.99	0.12	+0.00
	adv	0.86	0.84	−0.03	0.98	0.15	+0.03
econometrics	coop	0.86	0.89	+0.02	0.99	0.10	−0.01
	adv	0.88	0.88	+0.00	0.99	0.12	+0.02
formal logic	coop	0.94	0.96	+0.03	1.00	0.04	−0.02
	adv	0.91	0.96	+0.05	1.00	0.04	−0.04
hs physics	coop	0.91	0.97	+0.06	0.99	0.02	−0.03
	adv	0.96	0.97	+0.01	0.99	0.02	−0.00

E.3 Per-domain conformal comparison (Certify)

Table 8 compares Affirm against fixed-T and agreement/confidence stopping (test split, $\alpha = 0.1$). On the *hard* domains ($R_\infty \gg \alpha$), agreement stopping commits a confident wrong answer 18–23% of the time, whereas Affirm holds miscoverage $\leq \alpha$: on professional law it cuts the confident-error rate from 0.18–0.19 to 0.07 with still-informative sets (mean size 1.7–2.6), and on moral scenarios it drives miscoverage to 0 by *abstaining* (returning the full option

Table 7: Team-size robustness (cooperative debate, T=8): final accuracy A_T, confidence C_T, overconfidence gap G_T, and residual $R = 1 - A_T$ at $N \in \{3, 5, 7\}$ (N=3 uses the full-subject question set, $N \in \{5, 7\}$ a matched 80-item subset).

Domain	N	Q	A_T	C_T	G_T	R
virology	3	166	0.59	0.93	0.34	0.41
	5	80	0.55	0.94	0.39	0.45
	7	80	0.53	0.94	0.42	0.47
prof. law	3	250	0.74	0.91	0.17	0.26
	5	80	0.80	0.92	0.12	0.20
	7	80	0.80	0.91	0.11	0.20
college math	3	100	0.95	0.99	0.04	0.05
	5	80	0.95	0.99	0.04	0.05
	7	80	0.97	0.99	0.02	0.03
hs physics	3	151	0.95	0.99	0.04	0.05
	5	80	0.98	0.99	0.01	0.02
	7	80	0.98	0.99	0.01	0.02

set), the correct response when the model cannot discriminate, while also stopping far earlier than fixed-T (1–4 vs. 8 turns). On the *easy* domains it matches the baselines, mildly over-covering (empirical coverage 0.96–0.99 vs. nominal 0.90, from finite-sample discreteness); the safety advantage is confined to $R_\infty > \alpha$ as predicted, with an empirical safety gap of +0.11–+0.23 matching the $R_\infty - \alpha$ separation of Corollary 5.6. The guarantee is model-agnostic (marginal coverage averages 0.964 (`deepseek-v4-flash`) and 0.961 (`qwen-flash`), both $\geq 1 - \alpha$) whereas agreement stopping incurs miscoverage equal to the residual, up to 0.43–0.47 on the hardest domains. Because Affirm never reads the agreement signal, the diversity collapse that corrupts the consensus heuristic leaves its coverage untouched.

E.4 Consensus speed and Deliberation-Speed Trust (Detect)

The theory has a sharp label-free corollary: a consensus reached *immediately* reflects genuine agreement, whereas one the panel is slowly argued into reflects assimilation under pressure, so correctness should fall as the number of turns needed to reach consensus grows. It does, steeply and near-monotonically: pooling all n=6458 instances, the probability the panel's committed answer is correct drops from 0.94 when consensus is immediate to 0.55 (near chance) when it takes seven turns (Figure 4). "How long the panel argued before agreeing" is a correctness signal hidden inside the trajectory, and it is precisely the slow, forced consensus that the Detect intervention targets.

The finding is the basis of a deployable method. Fitting the Deliberation head (Definition 5.2) and pooling over all `deepseek-v4-flash` conditions (19 tasks × 2 roles; n=6458 questions), it raises the correctness-ranking AUROC from confidence's 0.756 to 0.793 and lowers selective error at every coverage (Figure 5, left; accuracy at 70% coverage 0.943 → 0.952). The gain is concentrated where Corollary 4.2 bites: within the hard domains confidence ranks correctness near chance (AUROC 0.49 virology, 0.58 abstract algebra), while Deliberation, reading chiefly the consensus *speed*, recovers real discrimination. It also transfers: trained on the other domains and deployed *label-free* on a held-out domain it beats confidence in 14/19 domains, with the largest gains where confidence collapses (GSM8K 0.49 → 0.80, TriviaQA 0.67 → 0.83).

The cross-model picture exposes the method's predicted boundary (Table 2, Figure 5, right). Trained on `deepseek` and deployed label-free on the matched `qwen-flash` grid, Deliberation raises AUROC from 0.728 to 0.770 (24/34 conditions) and 70%-coverage accuracy from 0.878 to 0.910; restricting both models to a single replicate (three votes/turn) *widens* the edge to +0.058 AUROC (27/34), ruling out a vote-count explanation. On `GLM-5.1` Deliberation does *not* help (AUROC 0.807 vs. 0.840, 3/16): the predicted limit, not a failure; GLM's confidence still *ranks* correctness (0.84), so there is no broken signal to repair. Deliberation is thus a theory-grounded *Detect* lever for the saturated-confidence regime, with the Certify guarantee holding regardless of tier.

E.5 Verdict-Sequestered Debate (Prevent) and the counterfactual injection

We compare Sequester (Definition 5.1) against standard debate on six domains × both roles, matched questions (Figure 6). Three findings. *(i) Accuracy is preserved everywhere* (mean change +0.007; the scatter sits on or above the diagonal). *(ii) The calibration repair is a clean double dissociation by role*: under *cooperative* roles Sequester sharply reduces overconfidence (mean ΔECE = −0.046; moral scenarios 0.169 → 0.050, abstract algebra 0.075 → 0.019, TruthfulQA 0.088 → 0.046), but under *adversarial* roles it barely moves (mean −0.010). *(iii) The mechanism is not simple answer-copying*: answer consensus $\bar{\kappa}$ stays saturated under ablation in every regime, and the counterfactual injection below shows visible verdicts steer a panel most under adversarial roles, the regime where Sequester helps *least*. We therefore read Sequester not as localizing overconfidence to an answer channel but as a *conditional protocol regularizer* that removes one high-risk visible-verdict exposure at no accuracy cost; role design (Theorem 4.5) is the complementary lever for the adversarial regime, and the conformal certificate the universal one.

The effect is *regime-dependent* rather than channel-localizing: on `qwen-flash` the cooperative gain vanishes (ΔECE = +0.010 vs. −0.046 on `deepseek`), matching `deepseek`'s adversarial roles. Answer consensus $\bar{\kappa}$ stays saturated under ablation in every regime (including cooperative `deepseek`, where Sequester *does* help) and the per-cell correlation between $\Delta\bar{\kappa}$ and the calibration gain is insignificant ($r = -0.39$, $p = 0.21$), so the regimes are separated by the calibration *outcome*, not a measurable break in consensus.

Table 8: Method comparison (test split, $\alpha = 0.1$): miscoverage $\mathbb{P}$[truth excluded], Affirm mean set size, and turns run, per domain and role.

Role	R_∞ (miscov)	fixed-T	agreement	Affirm	
		miscov	miscov	miscov	$\lvert set \rvert$/turns
moral scenarios					
coop	0.21	0.21	0.21	**0.00**	4.0 / 1
adv	0.23	0.23	0.23	**0.00**	4.0 / 1
prof. law					
coop	0.18	0.18	0.18	**0.07**	2.6 / 1
adv	0.19	0.19	0.19	**0.07**	1.7 / 1
formal logic					
coop	0.05	0.05	0.06	0.06	1.0 / 3
adv	0.08	0.08	0.08	0.05	1.1 / 2
hs physics					
coop	0.05	0.05	0.07	0.08	1.0 / 2
adv	0.07	0.07	0.07	0.08	1.0 / 2
prof. medicine					
coop	0.03	0.03	0.03	0.07	1.0 / 4
adv	0.02	0.02	0.02	0.02	1.0 / 1

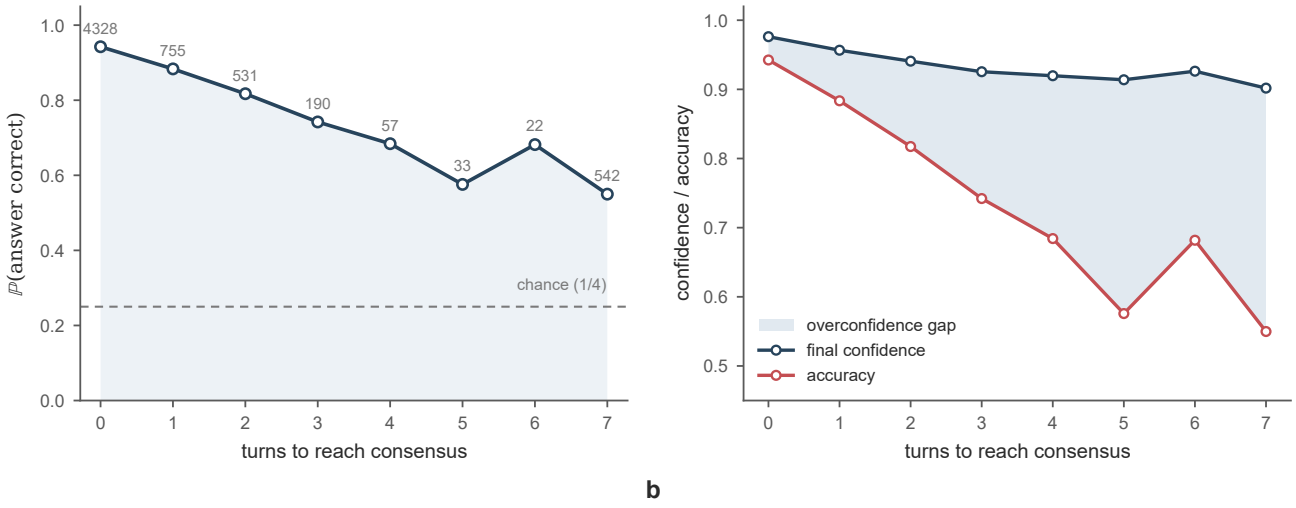

Figure 4: Consensus speed and correctness (n=6458, all conditions): (a) $\mathbb{P}$(answer correct) versus the number of turns to reach consensus (bar labels are instance counts), and (b) final confidence and the overconfidence gap (shaded) versus turns to consensus.

A *counterfactual-answer injection* (rewriting peers' committed answer to a fixed wrong option at high confidence, reasoning kept; three MMLU domains, N=3, T=8) confirms visible verdicts are causally potent: adoption of the planted answer rises by +0.25 (coop) and +0.66 (adv) on `deepseek` and +0.29 (coop) on `qwen`, with the panel then 0.8–0.96 confident *in the wrong answer*. But steering is *strongest* under adversarial roles and *weakest* under cooperative `deepseek`, the regime where Sequester helps most (these roles are instructed to defer to confident peers, so the ordering is partly built in). We therefore retire the "cooperative herds by copying the answer" reading: the committed-answer channel is causally real but its pull tracks role design, not the calibration-gain pattern, and why Sequester restores cooperative calibration without lowering $\bar{\kappa}$ remains open. The conformal guarantee, which never reads agreement, is unaffected: coverage holds on all five model families (mean 0.961 `qwen-flash`, 0.992 `GLM-5.1`, $\approx$ 1.00 `Llama-3.3-70B` and `GPT-4o-mini`).

E.6 Generalization: new task types and free-form generation

Beyond the knowledge/science multiple-choice tasks, we add two further task *types*, logical reasoning (ReClor) and commonsense (CommonsenseQA, recast to four options), under the identical protocol, to check that the mechanism holds outside factual recall

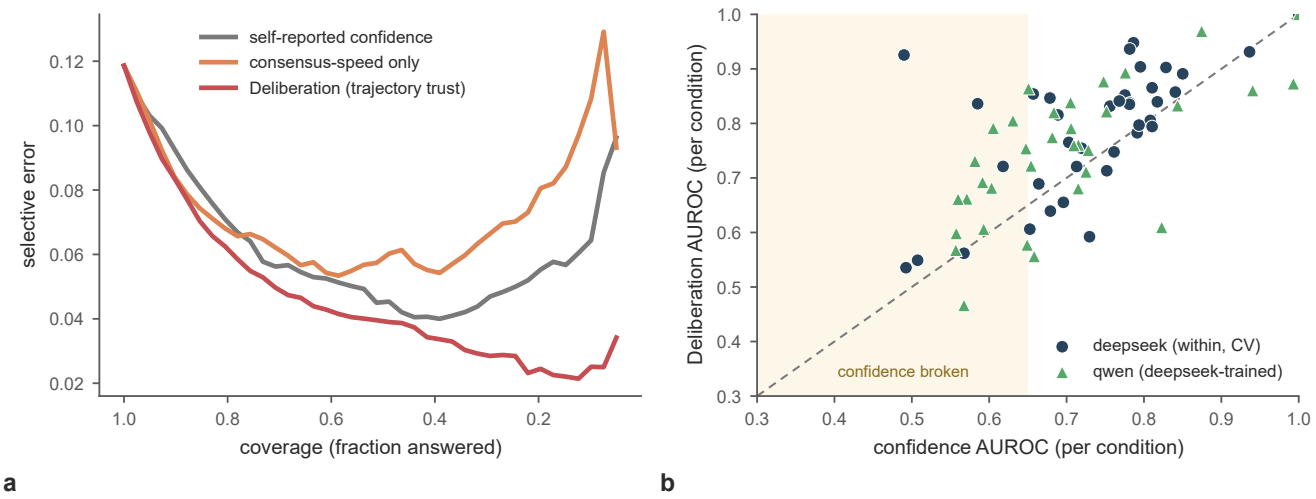

Figure 5: Deliberation-Speed Trust (*Detect*). Left: risk–coverage on `deepseek-v4-flash` for the Deliberation score, self-reported confidence, and a consensus-speed-only signal. Right: per-condition correctness-ranking AUROC, confidence (x) vs. Deliberation (y), for `deepseek` (within-model CV, circles) and `qwen-flash` (deepseek-trained, label-free transfer, triangles).

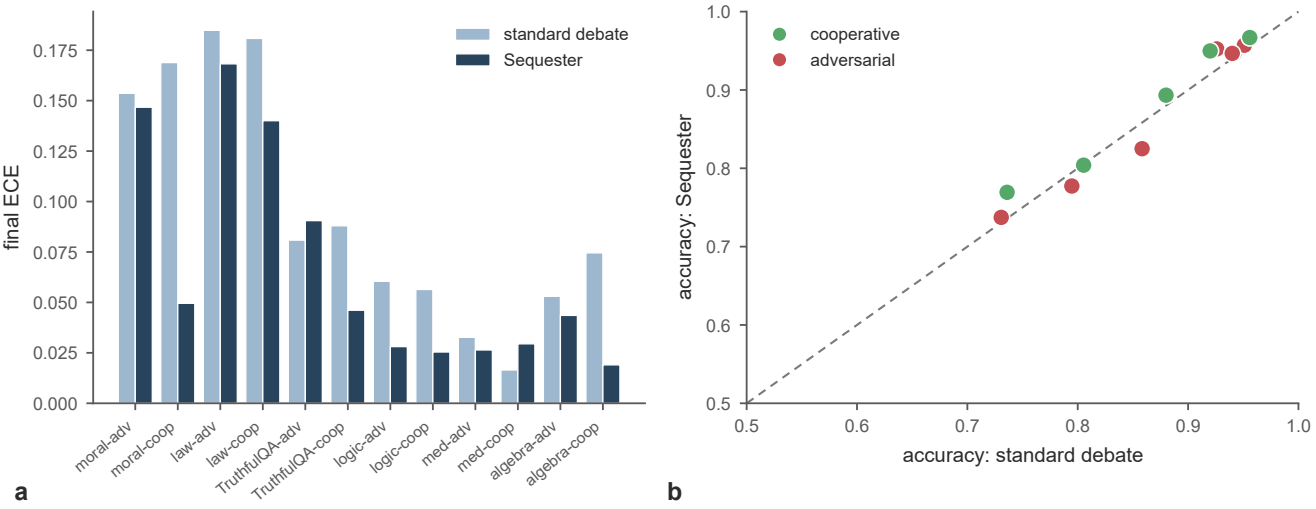

Figure 6: Verdict-Sequestered Debate (Sequester) vs. standard debate, six domains × two roles, matched questions: Left: final ECE; Right: accuracy (green cooperative, red adversarial).

(Table 9). ReClor is high-accuracy with a small residual, so confidence saturates ($C_T \approx 0.95$) and the overconfidence gap is near zero ($G_T \leq 0.02$), with Sequester neutral; CommonsenseQA carries a larger residual and a correspondingly larger gap ($G_T = 0.08$–0.10), and Sequester removes part of it under cooperative roles ($\Delta\text{ECE} = -0.06$). All four conditions fall on the same gap-vs-residual line, so the mechanism is not specific to factual-recall multiple choice.

To rule out a fixed-option artifact, we run two *free-form* benchmarks under the identical protocol on both models (GSM8K (numeric) and TriviaQA (open-domain, short-text), the free-form tasks of Du et al. [11] and the self-consistency line [48]) with agents emitting a free answer string (numeric/exact match) instead of a letter. The predictions hold (Table 10): confidence still saturates ($C_T \in [0.96, 1.00]$ over all eight conditions) while accuracy spans $[0.75, 0.99]$, so the answer-consensus collapse ($\bar{\kappa}(T) \geq 0.996$) reaches numeric *and* textual answers alike, and the gap again tracks the residual, tiny on GSM8K ($G_T \approx 0.01$–0.03) and large on TriviaQA

Table 9: New task types, ReClor (logical reasoning) and CommonsenseQA (commonsense), on `deepseek-v4-flash`: same protocol and columns as Table 5, plus the Sequester ΔECE.

Role	A_1	A_T	ΔA	C_T	G_T	ΔECE	Sequester ΔECE
ReClor (logic)							
coop	0.88	0.94	+0.05	0.96	+0.02	−0.05	+0.00
adv	0.93	0.94	+0.01	0.95	+0.01	−0.10	+0.02
CommonsenseQA							
coop	0.85	0.86	+0.01	0.94	+0.08	+0.06	−0.06
adv	0.85	0.84	−0.02	0.94	+0.10	+0.02	−0.04

Table 10: Free-form tasks (GSM8K numeric math, TriviaQA open-domain QA; both models, T=8, Q=200 each): accuracy A_T, confidence C_T, gap G_T, residual R, the Sequester ΔECE, and split-conformal coverage.

Task	Role	A_T	C_T	G_T	R	Sequester ΔECE	cov
`deepseek-v4-flash`							
GSM8K	coop	0.98	1.00	+0.01	0.02	+0.00	0.96
	adv	0.99	1.00	+0.01	0.01	+0.01	0.99
TriviaQA	coop	0.88	0.99	+0.10	0.12	−0.01	0.91
	adv	0.89	0.98	+0.09	0.11	−0.01	0.90
`qwen-flash`							
GSM8K	coop	0.97	1.00	+0.03	0.03	+0.01	0.94
	adv	0.97	1.00	+0.03	0.03	−0.00	0.94
TriviaQA	coop	0.76	0.97	+0.21	0.24	+0.02	1.00
	adv	0.75	0.96	+0.21	0.25	+0.04	1.00

($G_T \approx 0.21$ at residual 0.24–0.25 for `qwen-flash`), on the same line as the multiple-choice grid. The Sequester regime-dependence persists: on the informative TriviaQA cells withholding the verdicts helps on `deepseek` (ΔECE = −0.01) but not `qwen` (+0.02–+0.04), while GSM8K is too easy ($R \le 0.03$) to matter. Split-conformal coverage holds throughout (0.90–1.00).

E.7 A further debate trajectory

Beyond the body's Cases A and B (Section 7.1), one further trajectory shows a *Sequester* (Prevent) repair on a real `deepseek-v4-flash` debate, with every agent's committed answer per turn and the panel's mean confidence C_t (the gold option in bold).

Case C *Sequester* (Prevent) repairs the error: cooperative, *global facts*

```
# Q: about what % of the global population was literate in 1950?  B=56% (gold)
turn t          1     2     3     4     5     6     7        8
standard ans    A     A     A     A     A     A     A        A
standard C_t   .87   .93   .95   .95   .95   .95   .95      .95
Sequester ans   B     B     B     B     B     B     B        B
Sequester C_t  .85   .82   .83   .82   .80   .87   .85      .85
# => withholding the committed-answer line flips the panel from
confident-wrong A to correct B
```

Together with the body's Cases A and B, Case C traces the CMAD response to one failure mode. *Deliberation* (Detect) distrusts the slow consensus of Case A and trusts the fast one of Case B; *Sequester* (Prevent) removes the visible-verdict exposure that drives Case A and repairs Case C; and *Affirm* (Certify), never reading the agreement signal, would return a calibrated answer set rather than commit to the wrong A. Such confident flips are not anecdotal: they recur across the hard adversarial conditions (e.g. nine in moral scenarios, five in professional law).

F Robustness re-analyses

All checks below are recomputed from the logged per-turn answer/-confidence arrays; none requires new model calls.

Calibration-metric robustness (binning and proper scoring). The overconfidence gaps and the Sequester effect are not artifacts of the 10-bin ECE. Recomputing the gap on the hard cells at {5, 10, 15, 20} equal-width bins and at equal-mass (adaptive) bins changes the value by ≤ 0.004 (e.g. moral scenarios 0.169 at every bin count; virology 0.370), and the binning-free Brier score tells the same story (moral 0.173, virology 0.377): confidence is so compressed into $[0.9, 1.0]$ that the gap is bin-insensitive. The cooperative Sequester effect keeps its sign across all bin counts ($\Delta\text{ECE}_{\text{coop}}$ from −0.055 at 5 bins to −0.028 at 20 bins).

Consensus speed survives a within-task difficulty control. The pooled consensus-speed trend (correctness falls with turns-to-consensus) could in principle be driven by difficulty confounding (hard items both take longer to converge and are wrong more often). Controlling within task, fitting the slope of $\mathbb{P}(\text{correct})$ on turns-to-consensus *separately per domain*, leaves the slope negative in 15/15 tasks on both models (mean −0.046). Within a single hard domain the effect persists (virology: 0.63 → 0.55 from immediate to ≥ 5-turn consensus), so the signal is not merely an across-task difficulty proxy.

Post-hoc recalibration. A natural baseline is to temperature-/Platt-/isotonic-recalibrate the verbalized confidences. On a 50/50 within-condition split these do reduce held-out ECE on the hard cells (mean 0.23 → 0.05). The point is precisely that they require a *labeled in-domain calibration set*, which deployment lacks, and which the whole problem (telling a confident-wrong consensus from a confident-right one) denies us; Sequester (a label-free protocol change) and Affirm (a coverage certificate) do not. We therefore report recalibration as context, not as a baseline that Sequester or Affirm must beat on raw ECE.

Affirm coverage under cross-domain shift. Affirm's guarantee assumes calibration/test exchangeability, which holds within a task (our 50/50 split: mean coverage 0.98) but can break under distribution shift. Calibrating Affirm's threshold on one domain and testing on a *different* one drops mean coverage to 0.89 (minimum 0.50 on the most mismatched pair). This quantifies limitation (v): under domain shift the split-conformal layer needs its weighted or online variants. The *trivial* fixes do not genuinely recover it: pooling or leave-one-domain-out calibration restores coverage to 1.00, but only by enlarging the conformal set toward the full option set (mean size → 4.0 of 4, vs. 3.25 for the within-domain split): over-coverage by abstention, not informative recovery. A *non-trivial* partial fix follows the theory: because the usable difficulty signal survives collapse only in the trajectory (Corollary 4.2), we run a Mondrian

conformal stratified by a label-free trajectory feature (turns-to-consensus, binned fast/medium/slow) taking a per-stratum threshold on the calibration domain and assigning each test instance to its stratum by the same feature. Over all ordered cross-domain pairs of the eight difficulty-spread domains this lifts mean coverage from the naive 0.93 to 0.95 and the worst-pair minimum from 0.52 to 0.65 *while keeping the sets informative* (mean size 3.2, no larger than the within-domain 3.25 and far from the trivial 4.0): Pareto-better than both naive and pooled calibration. It does not restore the full $1-\alpha$ floor on the most mismatched pair, so genuine cross-domain coverage still needs reweighting by calibration-domain similarity (weighted/online conformal), which we leave to future work. Affirm's informativeness (at its in-distribution operating point) is not trivial abstention: across the hard cells it returns singletons or small sets (mean size 1.7–2.6 on professional law) and reserves full-set abstention for the genuinely non-discriminable case (moral scenarios).