From Debate to Decision: Distribution-Free Act-or-Defer Control for Multi-Agent LLM Pipelines

Mengdie Flora Wang [1] **Haochen Xie** [1] **Guanghui Wang** [1] **Aijing Gao** **Guang Yang** [1] **Ziyuan Li** [2]
Qucy Wei Qiu [2] **Fangwei Han** [2] **Hengzhi Qiu** [2] **Yajing Huang** [2] **Bing Zhu** [2] **Jae Oh Woo** [1]

Abstract

Agentic LLM pipelines repeatedly face act-or-defer decisions—whether to answer, route, continue, or escalate—yet deployed heuristics (top-1 confidence, majority vote, consensus stopping) carry no finite-sample risk certificate. We study how to attach a distribution-free validity layer to a black-box agentic pipeline, using multi-agent debate as a case study where naive consensus stopping can confidently commit wrong answers. Conformal Social Choice aggregates verbalized probabilities from N heterogeneous agents via a linear opinion pool and applies split conformal prediction post-hoc on the aggregated belief, producing sets $\mathcal{C}(x)$ with $\Pr[y \in \mathcal{C}(x)] \geq 1 - \alpha$ under exchangeability alone—no per-agent calibration. A hierarchical action policy maps singletons to autonomous action, multi-element sets to escalation, and empty sets to anomaly review. The contribution is a statistical systems framework, not a new conformal theorem: it identifies what to calibrate (the composed output), where to intervene (after aggregation, before action), and how to convert set size into deployment behavior. Across a full domain-by-round stress test on eight MMLU-Pro domains with three heterogeneous agents, the layer maintains near-nominal coverage while set size adapts to task ambiguity and singleton decisions increase as debate progresses; at α=0.05 it intercepts 81.9% of wrong-consensus cases while introducing only 2 new wrong singletons out of 4,158 (240:1 ratio), and preliminary robustness checks on GPQA and LogiQA show similar behavior without retuning.

[1]AWS Generative AI Innovation Center [2]HSBC Holdings Plc., HSBC Technology Center, China. Correspondence to: Mengdie Flora Wang <mengdie.fwang@gmail.com>, Jae Oh Woo <jaeoh.woo@gmail.com>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

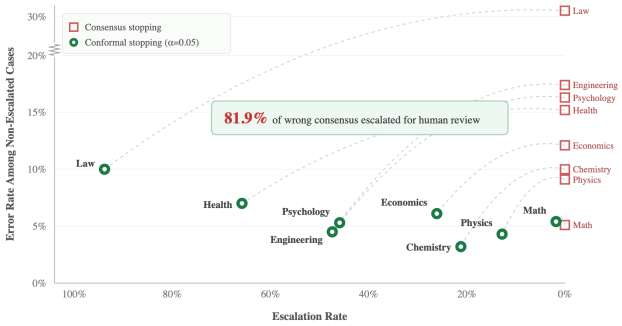

Figure 1. Error-vs-automation Pareto frontier across eight MMLU-Pro domains (α=0.05, round 3). Each domain appears as a matched pair: consensus stopping (red circles) achieves high automation rates (87.7–97.8% singletons) but incurs error rates up to 32.1% on acted cases; conformal stopping (blue squares) shifts the operating point down and left, trading automation for substantially lower error. Arrows connect matched domains; the down-left direction illustrates calibrated abstention intercepting wrong-consensus errors (81.9% of wrong-consensus cases are escalated rather than acted upon).

1. Introduction

Modern agentic pipelines repeatedly face *act-or-defer* decisions: whether to answer now, route to another agent, invoke a tool, continue deliberation, or escalate to a human. Most deployed systems resolve these decisions with heuristic confidence, majority vote, or consensus stopping, none of which provide a finite-sample risk certificate at the point of action. This paper studies attaching a post-hoc *statistical validity layer* to a black-box agentic pipeline so that its act-or-defer decisions carry a distribution-free coverage guarantee.

Debate as a case study. Multi-agent debate (Du et al., 2023; Liang et al., 2024) is a canonical setting where naive stopping fails. Agents deliberate, converge, and a deployment pipeline commits once they agree. But LLM agents conform to perceived majority opinions (Perez et al., 2023; Sharma et al., 2024), and in debate this social reinforcement produces *wrong consensus*—unanimous agreement on an incorrect answer. Empirically (§5.2), 23.9% of initially-disputed questions end in unanimous wrong consensus within four debate rounds. Consensus-based stopping

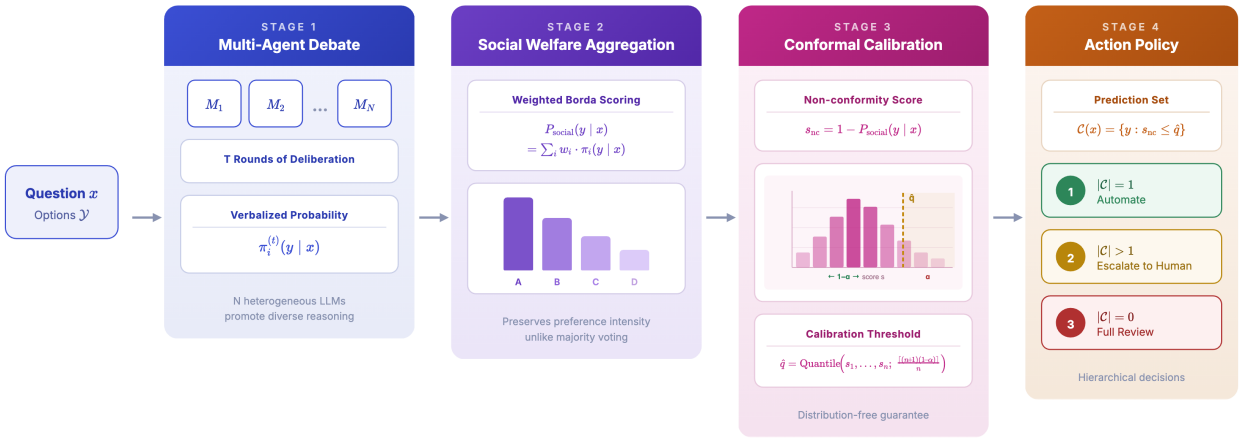

Figure 2. From debate output to deployment decision. Heterogeneous LLM agents debate for T rounds, producing verbalized probability distributions (Stage 1). These are aggregated via a linear opinion pool into social probabilities (Stage 2), calibrated using a held-out set to compute the conformal threshold $\hat{q}$ (Stage 3), and used to construct prediction sets with a hierarchical action policy (Stage 4). The pipeline adds a calibrated refusal layer: $\mathcal{C}(x)$ satisfies a marginal coverage guarantee ($\Pr[y \in \mathcal{C}(x)] \geq 1-\alpha$); singletons trigger autonomous action under the calibrated policy, while larger sets trigger escalation.

(Hu et al., 2025) cannot distinguish correct from incorrect agreement and point-estimate aggregation (Genest & Zidek, 1986) discards the distributional information that could have flagged the error. The failure is not merely one of efficiency: consensus stopping is *actively unsafe* because it commits to wrong consensus with no recourse.

A statistical systems framework. Our premise is that the right object to calibrate in an agentic system is the composed pipeline output, not the individual agents. We introduce **Conformal Social Choice**, a post-hoc validity layer that treats the multi-agent pipeline as a black box and wraps it in a single distribution-free certificate. The layer (i) aggregates verbalized probability distributions from N heterogeneous agents into a collective belief P_{social} via a linear opinion pool, and (ii) applies split conformal prediction (Vovk et al., 2005) on top of P_{social} to produce prediction sets satisfying

$$\Pr\big[y \in \mathcal{C}(x)\big] \geq 1 - \alpha$$

under only exchangeability of calibration and deployment examples. A hierarchical action policy (§3.5) maps singleton sets to *autonomous action*, multi-element sets to *human escalation*, and empty sets to *anomaly review*, converting calibrated uncertainty into an operational deployment rule with a user-adjustable risk budget α.

What is novel is not the conformal theorem. The contribution is a statistical systems framework that answers three questions for an agentic pipeline: (i) *what to calibrate*—not each agent, but the composed pipeline output P_{social}, so heterogeneous or miscalibrated agents do not need individual calibration; (ii) *where to intervene*—after aggregation and before action, so the guarantee covers the horizontal composition of agents → opinion pool → conformal set at each fixed round; (iii) *how to operationalize uncertainty*—set size directly determines act/defer/anomaly, so the guarantee is actionable rather than descriptive. The resulting certificate is a *pipeline-level post-hoc* guarantee (on the composed output), not a per-component or anytime guarantee—a scope we separate and discuss explicitly (§3.4, §6).

Contributions. **(1) A post-hoc validity layer for agentic act-or-defer decisions.** We formulate multi-agent LLM pipelines as black-box systems producing per-option belief distributions and wrap them with split conformal prediction to obtain marginal coverage under exchangeability. **(2) A pipeline-level certificate for heterogeneous agent aggregation.** Rather than requiring each agent to be calibrated, we calibrate the output of the composed pipeline—heterogeneous agents → social aggregation → prediction set—yielding a single distribution-free certificate for the downstream decision layer. **(3) A failure-mode analysis of consensus stopping.** We identify *wrong-consensus convergence*, where agents converge to the same incorrect answer through social reinforcement, and show that consensus-based stopping commits these errors without a validity check. **(4) Empirical evidence that calibrated abstention improves operational safety.** At α=0.05, the layer intercepts 81.9% of wrong-consensus cases while introducing only 2 new wrong singletons (240:1 ratio), and resolved singletons reach 90.0–96.8% accuracy across eight MMLU-Pro domains. Preliminary robustness checks on GPQA and LogiQA suggest similar behavior under the same calibration procedure, without benchmark-specific heuristic tuning.

2. Related Work

Multi-agent debate and stopping. Debate improves factuality and reasoning (Du et al., 2023; Liang et al., 2024); subsequent work refined stopping (Hu et al., 2025), compared debate with voting (Choi et al., 2025), modeled debate via equilibrium (Yi et al., 2025), used cross-examination (Cohen et al., 2023), and observed diminishing returns from larger ensembles (Li et al., 2024; Qian et al., 2025). Iterative belief exchange recalls DeGroot averaging and bounded-confidence models (DeGroot, 1974; Deffuant et al., 2000; Hegselmann & Krause, 2002). None provide distribution-free guarantees at the point of action.

Conformal prediction and distribution-free inference. We build on split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023); closely related are risk-controlling prediction sets (Bates et al., 2021), conformal risk control (Angelopoulos et al., 2024), adaptive prediction sets (Romano et al., 2020; Angelopoulos et al., 2021), covariate-shift methods (Tibshirani et al., 2019; Gibbs & Candès, 2021), and safe anytime-valid inference (Ramdas et al., 2023). LLM applications include QA (Kumar et al., 2023), open-ended generation (Quach et al., 2024), cascaded inference (Fisch et al., 2021), black-box APIs (Su et al., 2024), and selection-to-abstention connections (Fisch et al., 2022; Jin & Candès, 2023). The closest prior method, Debate-as-Optimization (DAO; Wang & Huang, 2024), uses conformal prediction as an intra-debate filter with a decaying threshold, breaking the coverage guarantee and requiring white-box calibration; we instead apply standard split conformal post-hoc on the aggregated collective belief, preserving marginal coverage in a fully black-box setting. Prediction-powered inference (Angelopoulos et al., 2023) is complementary: machine-generated labels preserving validity, analogous to our machine-generated probabilities.

Social choice and verbalized confidence. Yang et al. (2024b) and Choi et al. (2025) treat aggregation as winner selection; we use the linear opinion pool (Genest & Zidek, 1986) as a social-choice-inspired aggregation primitive (not a contribution to social choice theory). Verbalized confidence has been elicited from single agents (Lin et al., 2022; Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024; Yang et al., 2024a); semantic uncertainty (Kuhn et al., 2023) and self-consistency (Wang et al., 2023) are complementary. Our framework treats verbalized probabilities as noisy black-box inputs—biases degrade efficiency, not validity.

3. Method: Conformal Social Choice

We instantiate the pipeline-level validity layer as Conformal Social Choice, a four-stage pipeline that transforms the outputs of a multi-agent debate into prediction sets with marginal coverage guarantees (i.e., population-level, not per-instance). Our goal is not a new conformal theorem. It is to identify the correct object to calibrate in an agentic pipeline (the composed output, not the individual agents), to expose why consensus-based stopping lacks a validity contract, and to show how a post-hoc set-valued layer converts black-box multi-agent outputs into operational act-or-defer decisions.

3.1. Problem Formulation

Consider a multiple-choice question-answering task with input space $\mathcal{X}$ and finite label space $\mathcal{Y} = \{A, B, C, D, \ldots\}$ (we study $|\mathcal{Y}| = 10$ throughout; extensions to open-ended generation require candidate generation followed by conformal filtering and are left to future work). We have access to an ensemble of N LLM agents $\{M_1, \ldots, M_N\}$ that engage in a T-round debate. Given a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$ with ground-truth labels, our goal is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ satisfying the *marginal coverage* guarantee:

$$\Pr\big[Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\big] \geq 1 - \alpha, \tag{1}$$

where $(X_{\text{test}}, Y_{\text{test}})$ is a new exchangeable test example and $\alpha \in (0, 1)$ is the user-specified miscoverage rate. The output $\mathcal{C}(x)$ is not merely a set for evaluation; it is the object used for downstream *action*: a singleton triggers automation, a larger set triggers human escalation, and an empty set flags an anomaly (§3.5). This formulation recasts multi-agent debate from an accuracy-maximization problem into a *decision* problem with calibrated risk control.

3.2. Stage 1: Multi-Agent Debate with Verbalized Probability Elicitation

Given input x, a heterogeneous ensemble of N LLM agents debates for T rounds. At each round $t \in \{1, \ldots, T\}$, every agent M_i receives x and a summary of all agents' responses from round $t{-}1$ (empty for $t{=}1$), and produces a verbalized probability distribution $\pi_i^{(t)}(y \mid x)$ over $y \in \mathcal{Y}$, with $\sum_y \pi_i^{(t)}(y \mid x) = 1$. Rather than relying on token-level log-probabilities (unavailable for many proprietary APIs), each agent is prompted to output explicit numerical probabilities within structured tags (Tian et al., 2023; Xiong et al., 2024); parsed probabilities are clipped to $[0, 1]$ and renormalized, and parsing failures (0.77%; Appendix H) default to uniform. Verbalized scores are noisy proxies for true beliefs (Yang et al., 2024a); crucially, noisy inputs do not break the coverage guarantee of Theorem 3.4—they can only degrade set efficiency.

3.3. Stage 2: Social Probability Aggregation

We aggregate agent beliefs into a *full probability distribution*—not a single vote or argmax—because conformal

calibration (§3.4) requires a continuous score over all labels. We adopt a *linear opinion pool* (Genest & Zidek, 1986), a weighted mixture standard in Bayesian aggregation and social choice theory.

Definition 3.1 (Social Probability)**.** Given agent distributions $\{\pi_i^{(t)}\}_{i=1}^N$ at round t and agent weights $\{w_i\}_{i=1}^N$ with $\sum_i w_i = 1$, the *social probability* for option y is:

$$P_{\text{social}}^{(t)}(y \mid x) = \sum_{i=1}^{N} w_i \cdot \pi_i^{(t)}(y \mid x). \tag{2}$$

Proposition 3.2 (Basic properties of social probability)**.** *The social probability $P_{\text{social}}(\cdot \mid x)$ satisfies normalization, anonymity (under equal weights), neutrality, unanimity, and monotonicity.*

All properties follow from the linearity of weighted sums; proof in Appendix A. Unlike majority voting, which reduces each agent's output to a single vote, the social probability preserves the *intensity* of preferences: an agent assigning 0.6 to option A contributes differently than one assigning 0.99—a distinction that majority voting loses entirely.

Weight strategy. We use uniform weights $w_i = 1/N$ throughout. This is the cleanest assumption-free default: learned or accuracy-based weighting would require extra supervision or validation data, complicating the black-box, post-hoc framing. Empirically, entropy-based weighting (which upweights more confident agents per instance) produces negligible differences (mean $|\Delta\text{coverage}| = 0.4\%$, mean $|\Delta\text{set size}| = 0.06$; Appendix F), confirming that the debate consensus mechanism dominates the aggregation rule.

A margin-robustness property of the social winner under agent-level perturbations—$\|\tilde{P}_{\text{social}} - P_{\text{social}}\|_\infty \le \sum_i w_i \varepsilon_i$ and top-label stability when $\Delta(x) > 2\sum_i w_i \varepsilon_i$—follows from linearity and is given in Appendix A as Theorem A.1.

3.4. Stage 3: Conformal Calibration

We apply split conformal prediction (Vovk et al., 2005) on top of the aggregated social probabilities, transforming heuristic confidence scores into rigorous coverage guarantees.

Non-conformity score. We define a non-conformity score measuring how poorly a candidate label y conforms to the social consensus:

Definition 3.3 (Probability-based non-conformity score)**.**

$$s_{\text{nc}}(x, y) = 1 - P_{\text{social}}(y \mid x). \tag{3}$$

This score is high when the social probability for y is low, indicating that the collective believes y is unlikely. An alternative rank-based cumulative score (analogous to APS; Romano et al., 2020) is defined in Appendix B. We use the probability-based score throughout because it yields a clean threshold interpretation (Proposition A.2: $y \in \mathcal{C}(x)$ iff $P_{\text{social}}(y \mid x) \ge 1 - \hat{q}$), directly linking set membership to social probability and making the automate/escalate decision transparently interpretable.

Calibration procedure. Given the calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$: (1) for each calibration example (x_i, y_i), run the debate pipeline and compute the social probability $P_{\text{social}}(\cdot \mid x_i)$; (2) compute the non-conformity score of the ground truth: $s_i = s_{\text{nc}}(x_i, y_i)$; (3) determine the conformal threshold:

$$\hat{q} = \text{Quantile}\Big(s_1, \ldots, s_n;\ \tfrac{\lceil (n+1)(1-\alpha) \rceil}{n}\Big). \tag{4}$$

The finite-sample correction $\lceil (n+1)(1-\alpha) \rceil / n$ ensures the following guarantee:

Theorem 3.4 (Marginal Coverage (Vovk et al., 2005))**.** *If the calibration and test examples are exchangeable, then the prediction set $\mathcal{C}(x) = \{y \in \mathcal{Y} : s_{\text{nc}}(x, y) \le \hat{q}\}$ satisfies:*

$$\Pr\big[Y_{\textit{test}} \in \mathcal{C}(X_{\textit{test}})\big] \ge 1 - \alpha. \tag{5}$$

This guarantee is *distribution-free*: it requires only exchangeability of calibration and test data, with no assumptions on individual agent calibration. We stress that this is a guarantee on *set coverage*—not on per-instance correctness, singleton accuracy, or conditional coverage within any subgroup.

Pipeline-level post-hoc certificate. Theorem 3.4 applies to the output of the debate-plus-aggregation pipeline at any fixed round t: calibration example x_i is run through t-round debate, aggregated into $P_{\text{social}}^{(t)}(\cdot \mid x_i)$, and scored as $s_i^{(t)} = 1 - P_{\text{social}}^{(t)}(y_i \mid x_i)$. Exchangeability of the scores $\{s_i^{(t)}\}$ follows from exchangeability of the inputs (x_i, y_i), provided the debate procedure up to round t is a fixed function of its input (deterministic, or stochastic with fixed seeds). We use "compositional" in the black-box post-hoc sense: the guarantee applies to the output of the composed pipeline as a whole—heterogeneous agents $\to$ linear opinion pool $\to$ conformal set—not to each internal component separately. The construction respects black-box access: biases or miscalibration in π_i affect set efficiency but not coverage.

Exchangeability in operational terms. The exchangeability requirement is not an assumption about calibrated agents. It is an assumption about calibration and deployment examples being drawn from the same task distribution under a fixed pipeline. Agent miscalibration, biased verbalized probabilities, or heterogeneous reasoning styles affect

efficiency, not validity, as long as the same pipeline is used on calibration and test examples. If prompts, agent composition, aggregation weights, or tool policies are changed after calibration, the certificate no longer applies without recalibration. Under genuine covariate shift, weighted conformal methods (Tibshirani et al., 2019) or adaptive online variants (Gibbs & Candès, 2021) can restore validity.

Scope: per-round vs. anytime-valid stopping. We apply split conformal independently at each round, giving a separate marginal guarantee $\Pr[y \in \mathcal{C}_t(x)] \geq 1 - \alpha$ for each fixed $t \in \{0, \ldots, T\}$. This does not automatically extend to a data-dependent stopping round t^*: when the action policy terminates debate at the first round with $|\mathcal{C}_{t^*}(x)| = 1$, selecting among valid prediction sets based on their properties introduces a form of selection bias that standard split conformal does not control. A Bonferroni correction—calibrating each round at level α/T—gives a conservative simultaneous guarantee $\Pr[\forall t \leq T, y \in \mathcal{C}_t(x)] \geq 1 - \alpha$ at the cost of larger sets, and fully anytime-valid sequential inference via e-processes or martingale tests (Ramdas et al., 2023) would close the gap without this conservatism. Our experiments evaluate the per-round certificate because it isolates the core question of where to place a validity layer in a composed pipeline; this choice is a scope restriction, not an empirical claim about the other two regimes.

Rearranging the inclusion condition gives the equivalent threshold form $\mathcal{C}(x) = \{y : P_{\text{social}}(y \mid x) \geq 1 - \hat{q}\}$ (Proposition A.2 in Appendix A), which we use throughout.

3.5. Stage 4: Prediction Set Construction and Action Policy

For a new test input x, the system executes the debate, computes $P_{\text{social}}(\cdot \mid x)$, and constructs the prediction set $\mathcal{C}(x) = \{y \in \mathcal{Y} : s_{\text{nc}}(x, y) \leq \hat{q}\}$, whose cardinality is bounded by $\lfloor 1/(1-\hat{q}) \rfloor$ (Appendix A). The set size maps directly to an action: $|\mathcal{C}(x)| = 1$ triggers *autonomous action*, $|\mathcal{C}(x)| > 1$ triggers *human-in-the-loop escalation* with the pruned candidate set (shrinking the decision space from $|\mathcal{Y}|$ to $|\mathcal{C}(x)|$), and $|\mathcal{C}(x)| = 0$ triggers *anomaly review*. The conditions on (p_1, p_2) that determine each case, and the implied decision boundaries, are given in Appendix A (Proposition A.4). We stress that the coverage guarantee is marginal: this provides a calibrated risk budget over the population of queries, not a per-instance certificate.

4. Experimental Setup

Dataset. We evaluate primarily on **MMLU-Pro** (Wang et al., 2024), a 10-option multiple-choice benchmark covering eight professional domains: Mathematics (n=1,351), Physics (n=1,299), Chemistry (n=1,132), Law (n=1,101), Engineering (n=969), Economics (n=844), Health (n=818), and Psychology (n=798). The expanded option space (compared to MMLU's 4 options) makes the task substantially harder and better differentiates methods. For out-of-distribution robustness checks, we additionally evaluate on **LogiQA** (Liu et al., 2020) (n=8,678; 4-option logical reasoning) and **GPQA Extended** (Rein et al., 2023) (n=546; 4-option graduate-level science), applying the same pipeline with no per-benchmark tuning. Domain descriptions are provided in Appendix C.

Baselines. We use *consensus stopping* as the primary comparator: commit to an answer once all agents agree—the dominant deployment rule in debate systems. Point-estimate baselines (greedy decoding, self-reflection, majority voting, debate + majority voting) are reported in Appendix Table 4; debate + majority voting achieves the highest accuracy (66.7–94.2%). We do not claim conformal abstention uniformly dominates tuned heuristic rules; the claim is that it supplies a finite-sample, distribution-free validity contract where heuristics provide only tuned operating points. A detailed comparison against Fixed τ=0.95, calibration-tuned top-1 probability, calibration-tuned entropy, and vote-disagreement baselines (Appendix J) shows all methods achieve comparable wrong-consensus interception (79–82%), but conformal leads overall (81.9%) while providing the coverage certificate that no baseline offers.

Metrics. We report *marginal coverage* (fraction of test instances where $y \in \mathcal{C}(x)$; target $\geq 1-\alpha$), *average set size* $\frac{1}{m}\sum_j |\mathcal{C}(x_j)|$ (smaller is better, conditioned on coverage), *singleton rate* (fraction with $|\mathcal{C}(x)|$=1), and *singleton accuracy* (accuracy among instances where $|\mathcal{C}(x)|$=1). Small finite-sample deviations below the nominal coverage target on individual domains are expected and do not violate the marginal guarantee, which holds in expectation over the random calibration split.

Implementation. Our ensemble consists of three heterogeneous LLMs accessed via AWS Bedrock—Claude Haiku 4.5 (Anthropic), DeepSeek-R1 (DeepSeek), and Qwen-3 32B (Alibaba)—selected to maximize architectural and training diversity. Debate runs for T=4 rounds (indexed 0–3, where round 0 denotes the independent initial responses before inter-agent exchange) with temperature 0.7. For conformal calibration, each domain is split 50/50 into calibration and test sets, and we evaluate at $\alpha \in \{0.05, 0.10\}$. The threshold $\hat{q}$ is computed independently per domain and round. Full implementation details (prompting, parsing, hyperparameters) are in Appendix C.

5. Results

We organize results around three empirical claims directly tied to the workshop's themes: (1) **Marginal cov-**

erage holds under domain heterogeneity (§5.1)—the distribution-free guarantee of Theorem 3.4 is empirically validated across eight MMLU-Pro domains without domain-specific heuristic tuning. (2) **Consensus stopping fails compositionally** (§5.2)—debate itself introduces a failure mode (wrong-consensus convergence) that no point-estimate stopping rule can detect. (3) **Calibrated abstention intercepts failures** (§5.3)—the validity layer rejects 81.9% of wrong-consensus cases at α=0.05, with a singleton-accuracy selection effect of up to 22.1pp over consensus stopping.

5.1. Empirical Validation of Marginal Coverage

We intentionally report the full domain-by-round-by-α matrix—including two OOD benchmarks in the rightmost columns—rather than only averages, because the central question is whether a pipeline-level validity layer remains stable under heterogeneous task difficulty and evolving debate dynamics. Table 1 supports three takeaways.

(i) Coverage stability. At α=0.05 (target 95%) empirical coverage ranges from 93.0% to 97.6% across all 40 domain-round cells (8 MMLU-Pro + 2 OOD); at α=0.10 it clusters in 87.6–93.2%. No domain-specific heuristic tuning is applied, yet per-domain coverage stays within ±3pp of target in every cell—including the OOD benchmarks LogiQA and GPQA, which use a different task format ($|\mathcal{Y}|$=4) and were not seen during any design decision.

(ii) Difficulty-adaptive set size. At α=0.05 and round 3, Math produces average set size 1.01 while Law produces 6.91 (among 10 options)—calibrated conservatism rather than uniform abstention. Domain adaptivity emerges entirely from calibration without any domain-specific heuristic tuning.

(iii) Debate sharpens without breaking validity. As rounds progress, singleton rates grow (e.g., Physics 47.8%→87.2% at α=0.05) while coverage stays on target: debate aggregates evidence and the conformal layer converts that evidence into calibrated act-or-defer decisions rather than inflated confidence. The favorable regime reverses at α=0.01 through a calibration-threshold saturation artifact (Appendix F.1); $\alpha \in [0.05, 0.10]$ is the practical operating range.

Out-of-distribution robustness. The rightmost two columns of Table 1 show LogiQA and GPQA Extended under the same pipeline with no per-benchmark tuning. Coverage stays within ±2pp of target at both α levels; singleton accuracy on resolved cases is 89.2% / 91.7% at α=0.05 (Table 2); and wrong-consensus rejection is 90.8% / 87.3% (Table 3). We read this as consistency evidence across benchmark rather than a strong generality claim.

5.2. Why Point-Estimate Stopping Fails: Wrong-Consensus Convergence

A key practical question is when to stop the debate. The most natural heuristic—*consensus-based* stopping (terminate when all agents agree)—is fast but fundamentally unsafe, because debate itself introduces a systematic failure mode: agents may unanimously agree on the *wrong* answer through social reinforcement, a pattern we call *wrong-consensus convergence*.

Consensus stopping is fast but error-prone. Consensus-based early stopping terminates extremely early (average round 1.36–1.81 across domains), stopping 87.7–97.8% of samples before the final round (Table 2). However, the accuracy of consensus-stopped samples varies widely (67.9–94.9%), and consensus provides *no coverage guarantee*. In Engineering, consensus stops samples by round 1.80 on average with 82.6% accuracy—meaning 17.4% of "confidently" stopped predictions are wrong, with no mechanism to flag them.

Wrong-consensus convergence explains why consensus fails. Among the 1,963 inference samples where agents initially *disagree* at round 0, 469 (23.9%) converge to unanimous wrong consensus by round 3—nearly one in four initially-disputed questions (Table 3). The risk varies by domain: Law (34.8%) and Psychology (33.3%) are most susceptible, while Math (12.1%) is most resilient. Including cases where agents already agreed on a wrong answer at round 0, total wrong consensus at round 3 is 586 out of 4,158 (14.1%). Consensus stopping treats all unanimous agreement equally, and therefore *commits* to these wrong-consensus errors with no recourse.

5.3. Calibrated Abstention Intercepts Failures

Table 2 shows the overall result: conformal stopping yields singleton accuracy 90.0–96.8% across all eight MMLU-Pro domains—up to 22.1pp higher than consensus stopping (Law: 90.0% vs. 67.9%)—at the cost of resolving fewer cases automatically. The same pattern holds on two OOD benchmarks (bottom of Table 2): the largest gains appear on the hardest tasks (GPQA +16.3pp, LogiQA +13.7pp). Table 3 explains the mechanism: conformal prediction rejects 81.9% of wrong-consensus errors on MMLU-Pro while over-rejecting only 31.9% of correct-consensus cases, and introduces just 2 new wrong singletons out of 4,158 samples (a 240:1 error-prevention ratio; Appendix G). The same pattern holds on OOD benchmarks (bottom of Table 3): interception is even higher on LogiQA (90.8%) and GPQA (87.3%), where wrong consensus is more prevalent (34.5% vs. 23.9%). Figure 3 illustrates two concrete cases where unanimous wrong consensus is caught by the conformal layer. The singleton-accuracy gain is therefore a selection ef-

		MMLU-Pro (8 domains, $\lvert\mathcal{Y}\rvert$=10)																LogiQA		GPQA	
		Eng.		Law		Chem.		Phys.		Math		Econ.		Hlth.		Psy.		($\lvert\mathcal{Y}\rvert$=4)		($\lvert\mathcal{Y}\rvert$=4)	
Metric	Rd	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10
$\hat{q}$	0	.975	.952	.980	.945	.962	.784	.950	.769	.750	.605	.937	.764	.970	.904	.954	.890	.917	.857	.923	.897
	1	.987	.960	.990	.967	.982	.702	.923	.651	.718	.359	.974	.744	.984	.924	.967	.933	.957	.917	.974	.950
	2	.990	.963	.994	.977	.987	.669	.957	.492	.704	.229	.980	.823	.990	.930	.978	.944	.970	.937	.987	.957
	3	.995	.980	.997	.984	.990	.733	.973	.537	.740	.126	.980	.852	.992	.940	.983	.950	.977	.947	.990	.963
Cov. (%)	0	93.8	87.6	97.6	89.5	95.0	89.0	95.1	89.4	94.5	88.6	96.0	90.5	96.3	91.4	94.7	88.7	95.2	88.6	95.6	93.0
	1	95.5	90.5	96.2	88.9	95.8	88.0	93.5	88.5	94.8	91.6	95.5	88.2	96.1	91.4	93.5	88.5	96.0	89.3	97.1	93.0
	2	94.8	88.5	96.5	90.0	95.9	88.0	94.0	87.8	94.5	92.9	94.1	88.2	95.8	89.7	93.0	88.5	96.2	88.9	97.1	91.2
	3	95.0	90.5	96.4	90.7	95.6	87.8	94.2	88.3	94.5	93.2	93.6	88.2	95.6	89.0	93.5	88.7	96.2	88.9	94.9	90.5
$\lvert\mathcal{C}\rvert$	0	4.40	3.08	6.99	3.76	2.57	1.37	2.09	1.33	1.26	0.92	1.87	1.23	3.73	1.71	2.59	1.50	2.81	1.82	2.86	2.56
	1	3.92	2.40	6.49	3.65	1.91	1.06	1.26	1.02	1.06	0.96	1.82	1.08	3.53	1.56	2.23	1.49	2.92	1.89	2.81	2.45
	2	3.63	2.00	6.79	3.65	1.75	1.02	1.27	1.00	1.02	0.97	1.73	1.08	3.93	1.42	2.27	1.44	2.98	1.87	2.82	2.22
	3	3.84	2.13	6.91	3.94	1.81	1.01	1.27	1.00	1.01	0.98	1.65	1.06	4.13	1.38	2.45	1.44	3.00	1.88	2.75	2.14
Sing. Rate	0	26.8	32.0	1.5	6.4	43.3	65.7	47.8	69.2	74.9	92.0	52.1	77.3	20.3	50.4	39.6	62.7	14.6	38.8	17.6	23.4
	1	45.2	56.5	4.0	14.3	70.8	92.8	81.8	97.5	94.1	95.6	65.4	91.5	28.4	60.6	49.6	70.9	22.7	46.0	29.3	38.1
	2	50.5	64.3	5.5	18.5	77.0	97.5	86.0	99.5	97.6	97.5	70.1	93.1	31.6	68.9	52.4	74.7	24.8	48.4	31.9	48.7
	3	52.6	67.2	6.2	20.9	78.8	98.8	87.2	99.8	98.2	97.8	73.9	94.3	34.2	72.4	54.1	76.2	25.7	49.6	35.2	51.6
Sing. Acc.	0	96.9	96.8	100	91.4	97.5	93.8	98.4	92.2	96.4	96.1	97.3	91.7	91.6	92.7	96.8	92.4	91.0	87.6	95.8	93.8
	1	96.6	92.4	100	81.8	95.5	78.4	92.8	81.0	89.2	87.5	87.5	73.3	100	88.1	87.5	84.8	89.1	84.6	90.6	82.5
	2	84.6	79.0	62.5	69.6	97.1	77.8	88.9	53.9	83.3	69.2	75.0	57.1	84.6	79.4	90.9	93.3	80.7	75.0	100	89.7
	3	70.0	92.9	75.0	92.3	90.0	42.9	100	100	100	100	75.0	80.0	72.7	71.4	100	83.3	81.1	74.5	66.7	75.0

Table 1. Full stress-test matrix: 8 MMLU-Pro domains ($\lvert\mathcal{Y}\rvert$=10) plus 2 OOD benchmarks (LogiQA and GPQA Extended, both $\lvert\mathcal{Y}\rvert$=4), all under the same pipeline (Haiku + DeepSeek-R1 + Qwen-3 32B) with no per-benchmark tuning. Four debate rounds, two risk levels (α=0.05 left, α=0.10 right). $\hat{q}$ = calibration threshold; **Cov.** = empirical coverage (%); $\lvert\mathcal{C}\rvert$ = average prediction set size; **Sing. Rate** = cumulative singleton fraction (%); **Sing. Acc.** = accuracy among samples whose conformal set reached singleton at that round (%). Coverage stays within ±3pp of target across all 40 domain-round cells. Entropy weighting in Appendix Table 6.

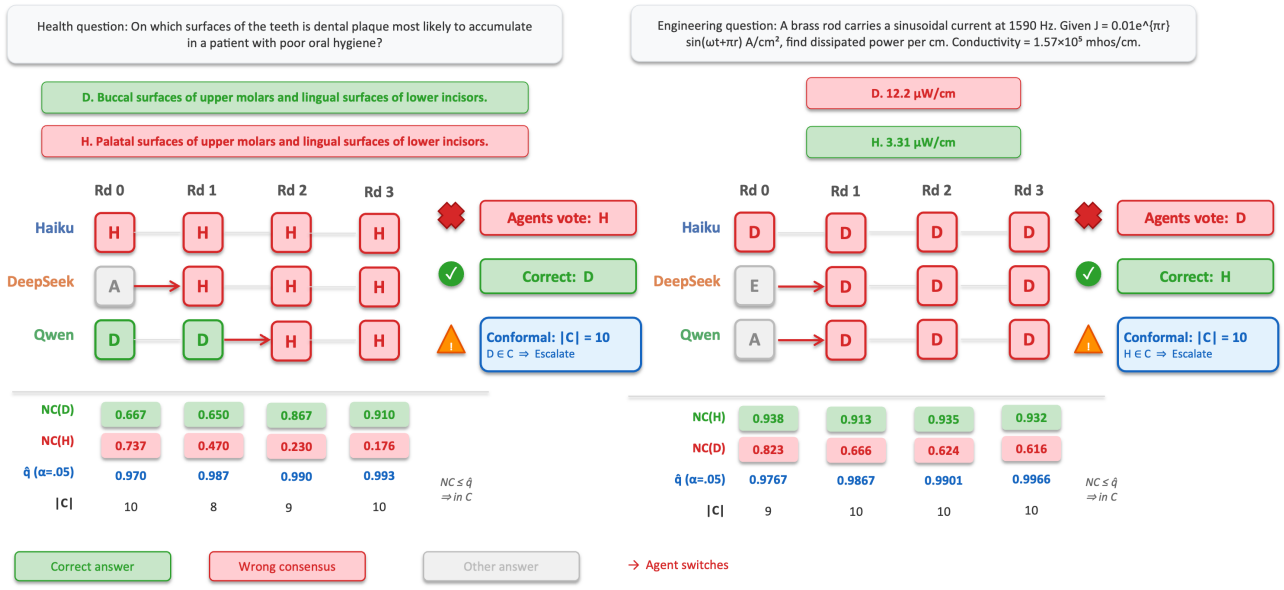

Figure 3. Calibrated refusal in action (α=0.05). In both examples, agents unanimously converge to the wrong answer through social reinforcement. Consensus alone would commit this error to an automated action; conformal prediction keeps the correct answer inside the set and forces escalation to human review.

		Avg Round		Singleton %		Singleton Acc.		
Domain	n	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**ΔAcc**
Engineering	485	1.80	1.67	90.1	52.6	82.6	95.5	+12.9
Law	551	1.81	2.24	87.7	6.2	67.9	90.0	+22.1
Chemistry	566	1.63	1.57	95.2	78.8	90.0	96.8	+6.8
Physics	650	1.55	1.53	95.1	87.2	90.9	95.7	+4.8
Math	676	1.43	1.29	97.8	98.2	94.9	94.6	−0.3
Economics	422	1.36	1.46	96.0	73.9	87.9	93.9	+6.0
Health	409	1.50	1.66	93.4	34.2	84.8	93.0	+8.2
Psychology	399	1.39	1.38	96.7	54.1	83.7	94.7	+11.0
Out-of-distribution benchmarks (same pipeline, no retuning)								
LogiQA	4339	1.51	2.38	97.1	25.7	75.6	89.2	+13.7
GPQA (ext.)	273	1.84	2.21	93.4	35.2	75.3	91.7	+16.3

Table 2. Safer but more conservative stopping (α=0.05, round 3). Compared with consensus stopping, conformal stopping resolves fewer cases automatically but yields substantially higher accuracy on the cases it does resolve. The accuracy gain is a selection effect explained by Table 3. The bottom section shows two OOD benchmarks under the same pipeline with no retuning: the largest gains appear on the hardest benchmarks (GPQA +16.3pp, LogiQA +13.7pp).

		Wrong-Consensus Convergence			Wrong Consensus		Correct Consensus	
Domain	n	**Disagree**	**→WC**	**WC%**	**Count**	**Rejected%**	**Count**	**Rejected%**
Engineering	485	312	75	24.0	83	88.0	375	41.1
Law	551	388	135	34.8	162	97.5	344	95.9
Chemistry	566	288	55	19.1	63	82.5	494	18.8
Physics	650	313	50	16.0	58	70.7	570	8.4
Math	676	256	31	12.1	35	11.4	631	0.6
Economics	422	135	35	25.9	50	62.0	362	26.5
Health	409	145	46	31.7	67	88.1	335	69.3
Psychology	399	126	42	33.3	68	91.2	323	43.3
All	**4,158**	**1,963**	**469**	**23.9**	**586**	**81.9**	**3,434**	**31.9**
Out-of-distribution benchmarks (same pipeline, no retuning)								
LogiQA	4339	1810	635	35.1	1042	90.8	3139	74.2
GPQA (ext.)	273	190	55	28.9	63	87.3	190	61.6

Table 3. Wrong-consensus convergence and conformal interception at round 3 (α=0.05). **→WC** = converge to unanimous wrong consensus. **Rejected%** = fraction where $|\mathcal{C}| > 1$ prevents automated action. Conformal prediction intercepts 81.9% of wrong-consensus errors on MMLU-Pro while over-rejecting 31.9% of correct ones. The bottom section shows two OOD benchmarks: interception is even higher on LogiQA (90.8%) and GPQA (87.3%), where wrong consensus is more prevalent.

fect of calibrated abstention, not a reasoning improvement—the agents themselves are unchanged by the layer.

Automation cost. On Math conformal stopping offers no advantage (−0.3pp); on Law it resolves only 6.2% of samples, escalating the remaining 93.8% to human review—not a failure but calibrated refusal in a high-ambiguity domain, with the operating point adjustable via α (Appendix I).

6. Discussion: Toward Statistical Control in Agentic Systems

What is guaranteed today. At any fixed debate round t, the composed pipeline (heterogeneous agents → linear opinion pool → conformal set) satisfies marginal coverage $\Pr[y \in \mathcal{C}_t(x)] \geq 1 - \alpha$ under exchangeability of calibration and deployment examples. No individual agent needs to be calibrated; miscalibrated or biased verbalized probabilities affect set efficiency, not validity. The certificate is actionable: the hierarchical action policy (§3.5) converts set size directly into automation, escalation, or anomaly review, with the operating point controlled by α.

What is not guaranteed. Coverage is marginal, not conditional, and the singleton-accuracy gains in §5.3 are a selection effect of calibrated abstention, not improved reasoning—the agents themselves are unchanged. The certificate is per-round, not anytime: a data-dependent stopping rule that terminates at the first singleton round introduces selection bias standard split conformal does not control, and we explicitly do *not* claim anytime validity unless paired with Bonferroni or e-process/martingale variants (§3.4). This is not a failure but a boundary of the statistical contract: the certificate is valid for a fixed calibrated pipeline and a fixed intervention point.

Deployment contract. The certificate binds a frozen pipeline to a frozen task distribution: any change in prompts, agents, weights, or tool policies after calibration requires recalibration—an operational invariant for agentic systems, where pipelines evolve rapidly. Although our empirical instantiation is multi-agent multiple-choice debate, the statistical interface only requires a frozen pipeline outputting a score over candidate actions; debate acts as a controlled testbed for the broader act-or-defer problem.

Extensions. Natural next steps include anytime-valid sequential stopping via e-processes or martingale tests (Ramdas et al., 2023); online conformal methods (Gibbs & Candès, 2021) for non-stationary deployment; weighted conformal methods (Tibshirani et al., 2019) for covariate shift; conditional coverage via group-balanced calibration or conformal risk control (Bates et al., 2021; Angelopoulos et al., 2024); and candidate-generation-plus-filtering for open-ended outputs, complementing prediction-powered inference (Angelopoulos et al., 2023) by extending distribution-free validity to decision-time act-or-defer control.

7. Conclusion

We presented Conformal Social Choice, a post-hoc validity layer that attaches a distribution-free act-or-defer guarantee to a black-box agentic LLM pipeline by calibrating the composed output rather than the individual agents. Two takeaways carry beyond the debate instantiation: consensus stopping is unsafe (the layer intercepts 81.9% of wrong-consensus cases while introducing only 2 new wrong singletons out of 4,158), and set size becomes an operational signal mapping to autonomous action, human escalation, or anomaly review. The framework is a starting point for anytime-valid, compositional, and open-ended extensions.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *International Conference on Learning Representations (ICLR)*, 2024.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Choi, H. K., Zhu, X., and Li, S. Debate or vote: Which yields better decisions in multi-agent large language models?, 2025. URL https://arxiv.org/abs/2508.17536.

Cohen, R., Hamri, M., Geva, M., and Globerson, A. LM vs LM: Detecting factual errors via cross examination, 2023. URL https://arxiv.org/abs/2305.13281.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

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

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations (ICLR)*, 2021.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Conformal prediction sets with limited false positives, 2022. URL https://arxiv.org/abs/2202.07650.

Genest, C. and Zidek, J. V. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Hegselmann, R. and Krause, U. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5 (3), 2002.

Hu, T., Tan, Z., Wang, S., Qu, H., and Chen, T. Multi-agent debate for llm judges with adaptive stability detection, 2025. URL https://arxiv.org/abs/2510.12697.

Jin, Y. and Candès, E. J. Selection by prediction with conformal p-values. In *Journal of Machine Learning Research*, 2023.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation of natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering, 2023. URL https://arxiv.org/abs/2305.18404.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need, 2024. URL https://arxiv.org/abs/2402.05120.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate, 2024. URL https://arxiv.org/abs/2305.19118.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Qian, C., Xie, Z., Wang, Y., Liu, W., Zhu, K., Xia, H., Dang, Y., Du, Z., Chen, W., Yang, C., et al. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 41488–41505, 2025.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling, 2024. URL https://arxiv.org/abs/2306.10193.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Romano, Y., Sesia, M., and Candès, E. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2024. URL https://arxiv.org/abs/2310.13548.

Su, J., Luo, J., Wang, H., and Cheng, L. API is enough: Conformal prediction for large language models without logit-access, 2024. URL https://arxiv.org/abs/2403.01216.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

Tibshirani, R. J., Barber, R. F., Candès, E., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

Vovk, V., Gammerman, A., and Shafer, G. Algorithmic learning in a random world. New York, 2005. Springer.

Wang, S. and Huang, L. Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction, 2024. URL https://arxiv.org/abs/2406.12197.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs, 2024. URL https://arxiv.org/abs/2306.13063.

Yang, D., Tsai, Y.-H. H., and Yamada, M. On verbalized confidence scores for LLMs, 2024a. URL https://arxiv.org/abs/2412.14737.

Yang, J. C., Dailisan, D., Korecki, M., Hausladen, C. I., and Helbing, D. LLM voting: Human choices and AI collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1696–1708, October 2024b. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31758. URL http://dx.doi.org/10.1609/aies.v7i1.31758.

Yi, X., Zhou, Z., Cao, C., Niu, Q., Liu, T., and Han, B. From debate to equilibrium: Belief-driven multi-agent llm reasoning via Bayesian Nash equilibrium, 2025. URL https://arxiv.org/abs/2506.08292.

A. Proofs

Proof of Proposition 3.2. All five properties follow from the linearity of the weighted sum $P_{\text{social}}(y \mid x) = \sum_{i=1}^{N} w_i \, \pi_i(y \mid x)$.

(1) Normalization. Because each π_i is a valid distribution,

$$\begin{aligned}\textstyle\sum_y P_{\text{social}}(y \mid x) &= \textstyle\sum_y \sum_i w_i \pi_i(y \mid x) \\ &= \textstyle\sum_i w_i \underbrace{\sum_y \pi_i(y \mid x)}_{=\,1} = 1.\end{aligned}$$

Non-negativity is immediate since $w_i \geq 0$ and $\pi_i(y \mid x) \geq 0$.

(2) Anonymity. Under equal weights $w_i = 1/N$, the social probability becomes $P_{\text{social}}(y \mid x) = \frac{1}{N} \sum_i \pi_i(y \mid x)$, which is invariant to any permutation of the agent indices.

(3) Neutrality. Let $\rho : \mathcal{Y} \to \mathcal{Y}$ be any relabeling of answer options. Under ρ, each agent's distribution transforms as $\pi_i^{\rho}(\rho(y) \mid x) = \pi_i(y \mid x)$. Therefore,

$$\begin{aligned}P^{\rho}_{\text{social}}(\rho(y) \mid x) &= \textstyle\sum_i w_i \pi_i^{\rho}(\rho(y) \mid x) \\ &= \textstyle\sum_i w_i \pi_i(y \mid x) \\ &= P_{\text{social}}(y \mid x).\end{aligned}$$

(4) Unanimity. If $\pi_i(y^\star \mid x) \geq \pi_i(y \mid x)$ for all agents i and all $y \neq y^\star$, then

$$\begin{aligned}&P_{\text{social}}(y^\star \mid x) - P_{\text{social}}(y \mid x) \\ &\quad = \textstyle\sum_i w_i \big[\pi_i(y^\star \mid x) - \pi_i(y \mid x)\big] \geq 0.\end{aligned}$$

(5) Monotonicity. If agent j increases its mass on $y^\star$ by $\delta > 0$ while all other quantities remain fixed, then $P_{\text{social}}(y^\star \mid x)$ increases by $w_j \delta > 0$. □

Theorem A.1 (Margin robustness)**.** *Let $\Delta(x) = P_{\text{social}}(y_1 \mid x) - P_{\text{social}}(y_2 \mid x)$ be the margin between the top two labels. If perturbed distributions satisfy $\|\widetilde{\pi}_i(\cdot \mid x) - \pi_i(\cdot \mid x)\|_\infty \leq \varepsilon_i$, then $\|\widetilde{P}_{\text{social}}(\cdot \mid x) - P_{\text{social}}(\cdot \mid x)\|_\infty \leq \sum_i w_i \varepsilon_i$. If $\Delta(x) > 2 \sum_i w_i \varepsilon_i$, the top label is unchanged.*

Proof of Theorem A.1. *Part 1: ℓ_∞ bound.* For each label $y \in \mathcal{Y}$, the triangle inequality gives

$$\begin{aligned}&\big|\widetilde{P}_{\text{social}}(y \mid x) - P_{\text{social}}(y \mid x)\big| \\ &= \big|\textstyle\sum_i w_i [\widetilde{\pi}_i(y \mid x) - \pi_i(y \mid x)]\big| \\ &\leq \textstyle\sum_i w_i \, |\widetilde{\pi}_i(y \mid x) - \pi_i(y \mid x)| \\ &\leq \textstyle\sum_i w_i \, \varepsilon_i.\end{aligned}$$

Taking the maximum over y yields $\|\widetilde{P}_{\text{social}}(\cdot \mid x) - P_{\text{social}}(\cdot \mid x)\|_\infty \leq \sum_i w_i \varepsilon_i$.

Part 2: Winner stability. Let y_1, y_2 be the top two labels under P_{social}. The perturbed gap satisfies

$$\begin{aligned}&\widetilde{P}_{\text{social}}(y_1 \mid x) - \widetilde{P}_{\text{social}}(y_2 \mid x) \\ &\quad \geq \big[P_{\text{social}}(y_1 \mid x) - \textstyle\sum_i w_i \varepsilon_i\big] \\ &\qquad - \big[P_{\text{social}}(y_2 \mid x) + \textstyle\sum_i w_i \varepsilon_i\big] \\ &\quad = \Delta(x) - 2 \textstyle\sum_i w_i \varepsilon_i.\end{aligned}$$

When $\Delta(x) > 2 \sum_i w_i \varepsilon_i$, this quantity is strictly positive, so y_1 remains the top label under the perturbed social probability. □

Proposition A.2 (Threshold form of the prediction set)**.** *For $s_{\text{nc}}^{\text{prob}}(x, y) = 1 - P_{\text{social}}(y \mid x)$, the conformal prediction set at threshold $\hat{q}$ satisfies:*

$$\begin{aligned}\mathcal{C}(x) &= \{y : s_{\text{nc}}^{\text{prob}}(x, y) \leq \hat{q}\} \\ &= \{y : P_{\text{social}}(y \mid x) \geq 1 - \hat{q}\}. \qquad (6)\end{aligned}$$

Proof of Proposition A.2. By definition, $y \in \mathcal{C}(x)$ if and only if the non-conformity score satisfies

$$\begin{aligned}&s_{\text{nc}}^{\text{prob}}(x, y) \leq \hat{q} \\ &\iff 1 - P_{\text{social}}(y \mid x) \leq \hat{q} \\ &\iff P_{\text{social}}(y \mid x) \geq 1 - \hat{q}.\end{aligned}$$

Therefore $\mathcal{C}(x) = \{y \in \mathcal{Y} : P_{\text{social}}(y \mid x) \geq 1 - \hat{q}\}$. □

Corollary A.3 (Cardinality bound on the prediction set)**.** *Under Proposition A.2, $|\mathcal{C}(x)| \leq \lfloor 1/(1 - \hat{q}) \rfloor$.*

Proof of Corollary A.3. Let $m = |\mathcal{C}(x)|$. By Proposition A.2, every label in $\mathcal{C}(x)$ satisfies $P_{\text{social}}(y \mid x) \geq 1 - \hat{q}$. Since P_{social} is a valid probability distribution (Proposition 3.2),

$$1 \;\geq\; \sum_{y \in \mathcal{C}(x)} P_{\text{social}}(y \mid x) \;\geq\; m \cdot (1 - \hat{q}).$$

Rearranging gives $m \leq 1/(1 - \hat{q})$, and since m is an integer, $m \leq \lfloor 1/(1 - \hat{q}) \rfloor$. □

Proposition A.4 (Conditions for singleton automation)**.** *Let $p_k = P_{\text{social}}(y_k \mid x)$ with $p_1 \geq p_2 \geq \cdots$, $\Delta(x) = p_1 - p_2$, and $\tau = 1 - \hat{q}$. Then: (1) $|\mathcal{C}(x)| = 1$ (automation) iff $p_1 \geq \tau$ and $p_2 < \tau$; equivalently, $p_1 \geq \tau$ and $\Delta(x) > p_1 - \tau$. A simpler sufficient condition is $p_1 \geq \tau$ and $\Delta(x) > \hat{q}$. (2) $|\mathcal{C}(x)| \geq 2$ (escalation) iff $p_2 \geq \tau$. (3) $|\mathcal{C}(x)| = 0$ (full review) iff $p_1 < \tau$.*

Proof of Proposition A.4. By Proposition A.2, $\mathcal{C}(x) = \{y : P_{\text{social}}(y \mid x) \geq \tau\}$ where $\tau = 1 - \hat{q}$. Write $p_k = P_{\text{social}}(y_k \mid x)$ with $p_1 \geq p_2 \geq \cdots$.

(1) Singleton condition. $|\mathcal{C}(x)| = 1$ iff exactly one label meets the threshold τ. Since p_1 is the largest probability, this is equivalent to $p_1 \geq \tau$ and $p_2 < \tau$. Using $\Delta(x) = p_1 - p_2$, the condition $p_2 < \tau$ rewrites as $p_1 - \Delta(x) < \tau$, i.e., $\Delta(x) > p_1 - \tau$.

For the simpler sufficient condition, suppose $\Delta(x) > \hat{q} = 1 - \tau$. Then, since $p_1 \leq 1$:

$$p_2 = p_1 - \Delta(x) \leq 1 - \Delta(x) < 1 - \hat{q} = \tau.$$

Therefore, if additionally $p_1 \geq \tau$, we have $|\mathcal{C}(x)| = 1$.

(2) Multiple candidates. $|\mathcal{C}(x)| \geq 2$ iff at least two labels satisfy the inclusion criterion. Since p_2 is the second-largest probability, this is equivalent to $p_2 \geq \tau$.

(3) Empty set. $|\mathcal{C}(x)| = 0$ iff no label meets the threshold. Since $p_1 \geq p_k$ for all k, this is equivalent to $p_1 < \tau$. □

B. Rank-Based Cumulative Score

As an alternative to the probability-based score (Definition 3.3), one can define a rank-based cumulative non-conformity score analogous to the Adaptive Prediction Sets (APS) score of Romano et al. (2020).

Definition B.1 (Rank-based cumulative score)**.** Let σ be a permutation of $\mathcal{Y}$ such that $P_{\text{social}}(\sigma(1) \mid x) \geq P_{\text{social}}(\sigma(2) \mid x) \geq \cdots$. For option y at rank r:

$$s_{\text{nc}}^{\text{rank}}(x, y) = \sum_{j=1}^{r} P_{\text{social}}(\sigma(j) \mid x). \tag{7}$$

The rank-based score captures ordinal information: even if two options have similar probabilities, the one ranked lower requires more cumulative mass to reach and thus receives a higher non-conformity score. In this work, we use the probability-based score throughout; comparison of the two scores is left to future work.

C. Experimental Setup Details

Domain descriptions. The eight MMLU-Pro domains are chosen to span a range of reasoning types. *Mathematics* (n=1,351) tests formal reasoning and calculation. *Physics* (n=1,299) requires scientific reasoning with quantitative analysis. *Chemistry* (n=1,132) involves domain-specific knowledge with procedural reasoning. *Law* (n=1,101) tests interpretive reasoning over complex legal rules. *Engineering* (n=969) combines applied problem-solving across multiple disciplines. *Economics* (n=844) requires analytical reasoning over market and policy concepts. *Health* (n=818) covers biomedical knowledge with clinical reasoning. *Psychology* (n=798) involves behavioral science with theoretical reasoning.

Baseline descriptions. *Single Agent (Greedy)* uses standard top-1 prediction from each individual LLM, providing a lower bound on ensemble performance. *Single Agent Self-Reflection* has a single agent iteratively review and revise its answer over k rounds, testing whether multi-round reasoning without diversity suffices. *Majority Voting* has each agent cast a vote for its top-1 answer; the plurality winner is selected (Wang et al., 2023). *Debate + Majority Voting* runs multi-agent debate for T rounds, with final answers aggregated by majority vote (Du et al., 2023).

Agent configuration. The three models were selected to maximize architectural and training diversity. Claude Haiku 4.5 is a compact model from the Claude family. DeepSeek-R1 is a reasoning-specialized model trained with reinforcement learning. Qwen-3 32B is an open-weight model from a distinct training pipeline. All models are accessed via AWS Bedrock.

Debate configuration. We run T=4 rounds for all multi-agent methods. At each round, agents receive the full summary of the previous round's responses. The generation temperature is set to 0.7 with top-p sampling at 1.0 and a maximum token budget of 4,096 per response. A structured system prompt enforces the output format, requiring step-by-step reasoning within `<reasoning>` tags and a probability distribution within `<answer>` tags.

Probability parsing. Verbalized probabilities are extracted from agent responses using regex-based parsing that prioritizes content within `<answer>` tags. The parser handles various formatting artifacts (e.g., markdown bold markers, escape characters). Extracted values are clipped to $[0, 1]$ and renormalized to sum to one. When parsing fails entirely for an agent, a uniform distribution is substituted (see Section 3.2). Across all 8,312 samples (calibration + test) $\times$ 3 agents $\times$ 4 rounds = 99,744 agent-round responses, the overall parse failure rate (fallback to uniform) is 0.77%, confirming that structured prompting yields reliable probability elicitation (Appendix H).

Conformal calibration. For each domain, a 50/50 random split of the data produces calibration and test sets (exchangeability is maintained by random shuffling). We evaluate at miscoverage levels $\alpha \in \{0.05, 0.10\}$, corresponding to target coverage rates of 95% and 90%. The conformal threshold $\hat{q}$ is computed independently for each domain and each debate round, enabling analysis of how coverage and set size evolve across rounds.

D. Point-Estimate Baselines

Table 4 contextualizes our framework against standard point-estimate methods. Debate + majority voting shows steady accuracy gains across rounds, from 67.6–92.2% at round 0 (static majority vote) to 80.0–94.2% by round 3, outperforming greedy decoding and self-reflection. Most improvement occurs in round 1, with diminishing returns thereafter. However, point-estimate methods provide no mechanism to flag uncertain or incorrect predictions. The conformal approach of §5.1 addresses this gap: singleton predictions achieve 90.0–96.8% accuracy while non-singletons are escalated, yielding a risk-aware decision pipeline that point estimates cannot provide.

E. Consensus-Based Early Stopping Details

Table 5 provides per-round statistics for consensus-based early stopping, showing how samples are distributed across stopping rounds.

F. Ablation: Uniform vs. Entropy-Based Weighting

Our social probability aggregation (Eq. 2) admits different weighting strategies for combining agent probability distributions. We compare two strategies:

- **Uniform**: $w_i = 1/n$ for all agents.
- **Entropy**: $w_i(x) = e^{-\lambda H(P_i)} / \sum_j e^{-\lambda H(P_j)}$, where $H(P_i) = -\sum_y P_i(y \mid x) \log P_i(y \mid x)$ is the Shannon entropy of agent i's distribution, and $\lambda = 1$.

The entropy weighting upweights agents that are more confident (lower entropy) on each individual instance, in the spirit of epistemic social choice theory.

Table 6 presents the full entropy-weighted results; the corresponding uniform-weighted figures are in Table 1. Table 7 summarizes the round 3 (final round) differences (entropy − uniform) across all eight domains and both miscoverage levels. The differences are small: the mean absolute change in coverage is 0.6%, in set size 0.06, and in accuracy 0.4%. The largest set-size change is 0.37, and most domain–α combinations show coverage and accuracy shifts well under 1 percentage point.

This result suggests that after multi-round debate, agents converge to similar distributions regardless of initial confidence differences, making the weighting strategy largely irrelevant. The debate process itself—rather than the aggregation rule—is the primary driver of uncertainty reduction. This is consistent with findings in the multi-agent debate literature, where deliberation tends to produce consensus that is robust to the choice of aggregation rule (Du et al., 2023).

F.1. Stringent Coverage: $\alpha = 0.01$ Results

Table 8 presents detailed conformal prediction results at $\alpha = 0.01$ (target coverage: 99%) across all eight MMLU-Pro domains. Unlike the $\alpha = 0.05$ and $\alpha = 0.10$ settings reported in the main body (§5.1), this stringent confidence level reveals a qualitatively different behavior: *prediction set sizes increase across debate rounds*, often approaching the full option set of 10.

Why do prediction sets grow at stringent α? The key mechanism is *calibration threshold saturation*. The conformal threshold $\hat{q}$ is set as the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the calibration nonconformity scores $s_i = 1 - P_{\text{social}}(y_i \mid x_i)$. At $\alpha = 0.01$, this corresponds to approximately the 99th percentile—effectively the worst-case calibration example.

As debate progresses, agents become increasingly confident: the social choice distribution P_{social} concentrates more mass on the majority answer. For the majority of examples where the agents are *correct*, this sharpening reduces nonconformity scores toward zero. However, for the small fraction of examples where agents are *confidently wrong*, the nonconformity score $s = 1 - P_{\text{social}}(y^* \mid x)$ approaches 1.0, since the true label y^* receives near-zero probability.

At $\alpha = 0.05$ or $\alpha = 0.10$, the calibration threshold is determined by a less extreme quantile (95th or 90th), which is not dominated by these confidently-wrong tail cases. But at $\alpha = 0.01$, the 99th-percentile quantile is precisely these worst-case examples, pushing $\hat{q} \to 1.0$. When $\hat{q} = 1.0$, the prediction set $\mathcal{C}(x) = \{y : s(x, y) \leq \hat{q}\} = \{y : 1 - P_{\text{social}}(y \mid x) \leq 1\}$ includes *all* options, as every candidate trivially satisfies the inclusion criterion.

This explains the striking pattern in Table 8: Math, which achieves the highest accuracy (>91%), maintains small prediction sets ($|\mathcal{C}| \approx 3$) through rounds 0–2 because even the 99th-percentile calibration example retains some probability on the correct answer. But by round 3, $\hat{q}$ reaches 1.0 and the set size jumps to 9.76. In contrast, domains like Law (60–66% accuracy) already have $\hat{q} = 1.0$ by round 1, as the larger fraction of errors produces extreme nonconformity scores earlier in the debate.

This phenomenon highlights an inherent tension between debate-driven confidence sharpening and stringent coverage guarantees: the same mechanism that improves accuracy (consensus formation) also amplifies the nonconformity scores of remaining errors, inflating the calibration threshold. At moderate α levels, this tradeoff is favorable—set sizes decrease while maintaining coverage. At very stringent α, the tail behavior dominates, rendering the prediction sets uninformative. This suggests that $\alpha \in [0.05, 0.10]$ represents a practical sweet spot for conformal social choice in

Method	Model(s)	Engineering	Law	Chemistry	Physics	Math	Economics	Health	Psychology
Greedy	Claude Haiku	70.1	60.5	83.2	84.8	90.2	83.8	75.9	80.8
	DeepSeek-R1	58.4	63.9	77.2	78.4	88.4	82.8	76.8	79.7
	Qwen-3 32B	50.5	39.4	62.4	62.0	66.8	76.4	68.8	70.3
Self-Reflection (Rd 3)	Claude Haiku	72.7	59.5	86.6	86.2	92.1	85.2	76.0	80.6
	DeepSeek-R1	78.3	66.3	**88.7**	87.4	93.0	85.6	80.1	82.1
	Qwen-3 32B	56.8	39.4	69.5	72.2	78.5	78.1	68.6	72.6
Debate + Maj. Vote	H+D+Q (Rd 0)	67.6	60.8	84.0	84.0	92.2	85.7	79.3	81.5
	H+D+Q (Rd 1)	77.9	64.8	88.2	89.3	93.9	87.0	80.1	81.5
	H+D+Q (Rd 2)	79.6	65.8	88.7	89.8	94.0	87.2	80.3	**82.3**
	H+D+Q (Rd 3)	**80.0**	**66.7**	88.6	**89.9**	**94.2**	**87.4**	**80.7**	**82.3**

Table 4. Point-estimate baselines (%) on MMLU-Pro. **H**=Claude Haiku, **D**=DeepSeek-R1, **Q**=Qwen-3 32B. Self-reflection and debate report final-round (round 3) accuracy. Rd 0 corresponds to static majority voting (no debate). Debate + majority voting is the strongest point-estimate method, but point estimates provide no mechanism to flag incorrect predictions; the set-valued conformal approach (§5.1) addresses this gap.

Domain	R0	R1	R2	R3
Engineering	36.3%	40.9%	14.3%	8.5%
Law	34.2%	38.3%	14.7%	12.8%
Chemistry	53.2%	35.9%	7.2%	3.7%
Physics	52.3%	35.2%	8.2%	4.4%
Math	60.8%	31.8%	4.8%	2.6%
Economics	69.2%	21.7%	5.5%	3.7%
Health	63.1%	23.7%	8.4%	4.8%
Psychology	69.2%	20.2%	6.9%	3.8%

Table 5. Distribution of samples across stopping rounds for consensus-based early stopping. The majority of samples stop at rounds 0–1, with domain difficulty reflected in the fraction reaching the final round.

multi-agent debate settings.

G. Sycophantic Convergence and Conformal Safety Analysis

Multi-agent debate improves accuracy through deliberation, but it also carries a systemic risk: agents may converge to unanimous agreement on the *wrong* answer through social reinforcement, a pattern consistent with known sycophantic tendencies in LLMs. We provide a detailed analysis of this phenomenon and quantify how conformal prediction mitigates—or, in rare cases, compounds—the risk.

G.1. Sycophantic Convergence Across Rounds

Table 9 reports, for each domain, the number of inference samples where agents initially disagree (no unanimous consensus at round 0) but later converge to unanimous wrong consensus. We use the inference split (second half of each domain's dataset) to ensure alignment with the conformal prediction evaluation.

By round 3, nearly one in four initially-disputed questions (23.9%) ends in unanimous wrong consensus. The risk varies substantially by domain difficulty: Law (34.8%) and Psychology (33.3%) are most susceptible, while Math (12.1%) is most resilient. This monotonically increasing trend across rounds confirms that extended debate amplifies sycophantic convergence—agents do not merely fail to correct errors but actively converge toward them.

G.2. Wrong-Consensus Rejection by Conformal Prediction

We next ask: when agents reach unanimous wrong consensus, does conformal prediction catch the error? For each round, we identify all inference samples with unanimous wrong consensus and check whether the conformal prediction set has $|\mathcal{C}| > 1$ (correctly flagging the error for human review) or $|\mathcal{C}| = 1$ (letting the wrong answer through as a singleton).

Table 3 shows that at α=0.05, conformal prediction correctly flags **81.9%** of wrong-consensus cases across all domains. The rejection rate is highest in domains where prediction sets tend to be large: Law achieves 97.5% rejection, as the inherent difficulty keeps sets far from singleton even when agents superficially agree. In contrast, Math (20.0%) and Physics at α=0.10 (3.4%) show low rejection rates—in these high-accuracy domains, agents' probability distributions are so concentrated that even wrong answers produce near-singleton sets.

G.3. Conformal-Introduced Wrong Singletons

An important counter-risk is that conformal prediction can *introduce* wrong singletons that would not exist under consensus-based stopping. This occurs when agents do *not* unanimously agree on the wrong answer (so consensus-based stopping would escalate), but the social probability distribution concentrates enough on a wrong answer that the conformal set shrinks to $|\mathcal{C}| = 1$ containing that wrong

Metric	Rd	Engineering .05	Engineering .10	Law .05	Law .10	Chemistry .05	Chemistry .10	Physics .05	Physics .10	Math .05	Math .10	Economics .05	Economics .10	Health .05	Health .10	Psychology .05	Psychology .10
$\hat{q}$	0	.992	.979	.985	.960	.986	.850	.973	.815	.785	.507	.957	.796	.980	.917	.962	.911
	1	.989	.970	.992	.970	.986	.696	.943	.641	.769	.343	.980	.793	.987	.928	.971	.938
	2	.992	.967	.995	.978	.990	.663	.962	.488	.703	.178	.983	.824	.990	.935	.980	.944
	3	.995	.980	.997	.984	.990	.734	.968	.535	.801	.111	.984	.852	.992	.940	.983	.950
Coverage (%)	0	95.3	86.6	96.2	90.4	95.2	89.2	94.8	88.8	93.9	89.2	96.0	90.8	97.3	90.5	94.7	89.0
	1	94.6	90.3	96.5	88.4	94.9	87.5	93.8	88.0	94.8	90.5	95.7	88.4	96.3	91.0	93.5	88.5
	2	94.8	89.3	95.8	90.0	95.6	86.8	94.2	87.4	94.2	91.0	94.5	88.2	96.3	89.7	93.0	88.5
	3	94.6	90.7	95.8	89.8	95.0	86.8	93.5	87.7	94.4	91.1	94.1	88.2	96.1	88.8	93.0	88.7
$\lvert\mathcal{C}\rvert$	0	5.02	3.04	6.48	4.03	2.82	1.40	2.06	1.30	1.22	0.96	1.97	1.22	3.92	1.71	2.58	1.59
	1	3.77	2.41	6.40	3.58	1.87	1.04	1.28	1.02	1.06	0.98	1.93	1.09	3.68	1.54	2.26	1.55
	2	3.81	1.98	6.59	3.62	1.74	1.01	1.28	1.00	1.02	0.98	1.79	1.07	4.16	1.44	2.25	1.46
	3	3.81	2.13	6.54	3.83	1.78	1.01	1.25	1.00	1.02	0.99	1.75	1.07	4.23	1.37	2.33	1.49
Singleton Rate (%)	0	20.6	31.8	1.1	7.1	36.2	65.7	47.4	71.4	78.4	95.9	49.8	79.4	17.6	50.9	40.6	59.4
	1	44.9	56.5	4.0	15.4	69.1	95.4	81.1	97.4	94.4	97.6	61.1	91.2	24.4	61.4	48.6	67.7
	2	49.5	65.2	6.5	19.2	76.0	98.6	85.2	99.7	97.9	98.2	67.3	93.4	28.6	67.5	53.1	72.2
	3	52.0	67.6	7.4	21.2	78.1	99.3	87.4	99.8	98.4	98.5	70.1	94.1	31.1	72.9	54.6	73.7
Singleton Acc. (%)	0	100.0	97.4	100.0	89.7	97.6	93.5	98.4	92.0	95.8	93.1	98.1	92.2	94.4	92.3	95.7	92.4
	1	94.1	92.5	100.0	82.6	95.2	75.6	92.2	76.9	90.7	75.0	85.4	74.0	96.4	83.7	93.8	90.9
	2	86.4	73.8	64.3	66.7	97.4	66.7	88.9	73.3	75.0	75.0	80.8	55.6	88.2	84.0	88.9	94.4
	3	75.0	91.7	80.0	90.9	91.7	50.0	78.6	100.0	100.0	50.0	75.0	66.7	70.0	77.3	83.3	66.7

Table 6. Conformal prediction results with **entropy-based weighting** (λ=1) and early stopping on MMLU-Pro. Columns: $\hat{q}$ = calibration threshold; **Coverage** = empirical coverage (%); $|\mathcal{C}|$ = average prediction set size; **Singleton Rate** = cumulative fraction of samples resolved as singletons by that round (%); **Singleton Acc.** = accuracy among samples whose conformal set reached singleton at that round (%). Rows 0–3 correspond to debate rounds. Differences against uniform weighting are negligible (mean $|\Delta\text{Coverage}| = 0.6\%$, mean $|\Delta\text{Set Size}| = 0.06$, mean $|\Delta\text{Singleton Acc.}| = 0.4\%$; see Table 7 for per-domain breakdown).

Domain	α	ΔCoverage	ΔSet Size	ΔSing. Acc.
Engineering	.05	−0.4	−0.03	−0.2
	.10	+0.2	+0.00	−0.2
Law	.05	−0.5	−0.37	−0.7
	.10	−0.9	−0.11	−0.7
Chemistry	.05	−0.5	−0.03	+0.3
	.10	−1.1	+0.00	−0.7
Physics	.05	−0.6	−0.02	+0.0
	.10	−0.6	+0.00	−0.6
Math	.05	−0.2	+0.01	−0.1
	.10	−2.1	+0.01	−2.5
Economics	.05	+0.5	+0.10	+0.0
	.10	+0.0	+0.01	+0.0
Health	.05	+0.5	+0.10	+0.0
	.10	−0.2	−0.01	−0.2
Psychology	.05	−0.5	−0.12	+0.0
	.10	+0.0	+0.05	+0.0

Table 7. Final-round differences between entropy and uniform weighting (entropy − uniform) in coverage (%), average set size, and singleton accuracy (%). Differences are small: mean $|\Delta\text{Coverage}| = 0.6\%$, mean $|\Delta\text{Set Size}| = 0.06$, mean $|\Delta\text{Singleton Acc.}| = 0.4\%$, max $|\Delta\text{Coverage}| = 2.1\%$. The debate consensus mechanism dominates the aggregation weighting strategy.

answer.

Table 10 reveals that this risk is minimal. At α=0.05 and round 3, only **2 out of 4,158** inference samples (0.05%) have conformal-introduced wrong singletons. At α=0.10, the rate is slightly higher at 8 samples (0.2%), concentrated in Chemistry and Physics where α=0.10 produces very small sets. Two patterns emerge:

1. The rate *decreases* across rounds: as debate refines social probabilities, the residual uncertainty on wrong answers increases, pushing them out of singleton sets.
2. The rate is higher at α=0.10 than α=0.05: tighter sets are more likely to collapse to a wrong singleton.

G.4. Net Safety Balance

The net effect of conformal prediction on safety is overwhelmingly positive. At α=0.05 and round 3, conformal prediction:

- **Catches** 480 out of 586 wrong-consensus errors (81.9% rejection rate);
- **Introduces** only 2 wrong singletons that consensus-based stopping would have caught.

The ratio of errors prevented to errors introduced is approximately **240:1**. At α=0.10, conformal catches 325 wrong-consensus errors while introducing 8 new ones—a ratio of approximately **41:1**. These ratios confirm that conformal calibration provides a substantial net safety improvement over consensus-based stopping, with negligible downside risk.

G.5. Out-of-Distribution Replication

We reproduce the wrong-consensus and conformal-rejection analysis on the two OOD benchmarks introduced in §5.1 (rightmost columns of Table 1). The same calibration proce-

		Engineering	Law	Chemistry	Physics	Math	Economics	Health	Psychology
$\hat{q}$	0	.997	.994	.997	.997	.990	.993	.997	.989
	1	.997	1.00	.999	.000	.997	.997	1.00	.997
	2	1.00	1.00	1.00	1.00	.999	1.00	1.00	1.00
	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Coverage (%)	0	99.0	99.3	99.3	99.4	99.0	99.3	99.3	99.8
	1	98.6	100.0	99.3	99.4	99.6	99.0	100.0	99.5
	2	99.6	100.0	100.0	99.8	99.1	100.0	100.0	100.0
	3	99.6	100.0	100.0	99.8	99.3	100.0	100.0	100.0
$\lvert\mathcal{C}\rvert$	0	7.89	8.59	7.08	7.50	3.25	5.43	8.02	5.97
	1	6.23	9.20	6.01	6.25	3.00	5.37	8.84	7.25
	2	7.97	9.20	8.26	8.28	2.58	7.31	8.84	8.66
	3	7.97	9.20	8.26	8.28	4.03	7.31	8.84	8.66
Singleton Rate (%)	0	3.9	0.0	4.4	4.0	46.2	16.1	2.4	7.8
	1	19.6	0.0	17.7	18.3	55.8	21.1	2.4	10.0
	2	19.6	0.0	17.7	18.3	65.4	21.1	2.4	10.0
	3	19.6	0.0	17.7	18.3	65.4	21.1	2.4	10.0
Singleton Acc. (%)	0	100.0	—	100.0	100.0	99.7	100.0	100.0	100.0
	1	97.4	—	100.0	98.9	100.0	100.0	—	100.0
	2	—	—	—	—	93.8	—	—	—
	3	—	—	—	—	—	—	—	—
Accuracy (%)	0	65.77	60.62	82.86	83.54	91.72	85.78	80.44	81.45
	1	77.32	64.07	86.57	89.08	94.08	86.97	81.91	81.20
	2	79.79	64.97	87.46	89.08	93.93	86.49	82.15	81.70
	3	79.79	66.06	87.28	89.54	94.23	86.73	82.89	81.70

Table 8. Conformal prediction results at $\alpha = 0.01$ (99% target coverage) across debate rounds 0–3 on MMLU-Pro. Unlike $\alpha \in \{0.05, 0.10\}$, prediction set sizes *increase* with debate rounds due to *calibration threshold saturation*: as debate sharpens confidence, the few remaining errors produce extreme non-conformity scores that push $\hat{q} \to 1.0$ at the 99th percentile, collapsing all labels into the prediction set.

		→ Wrong Consensus (%)		
Domain	**Init. Disagr.**	**Rd 1**	**Rd 2**	**Rd 3**
Engineering	312	37 (11.9)	65 (20.8)	75 (24.0)
Law	388	90 (23.2)	125 (32.2)	135 (34.8)
Chemistry	288	29 (10.1)	45 (15.6)	55 (19.1)
Physics	313	31 (9.9)	44 (14.1)	50 (16.0)
Math	256	19 (7.4)	30 (11.7)	31 (12.1)
Economics	135	24 (17.8)	32 (23.7)	35 (25.9)
Health	145	24 (16.6)	36 (24.8)	46 (31.7)
Psychology	126	24 (19.0)	37 (29.4)	42 (33.3)
All	1,963	278 (14.2)	414 (21.1)	469 (23.9)

Table 9. Wrong-consensus convergence among initially disagreeing cases. **Init. Disagr.** = inference samples without unanimous consensus at round 0. Columns show the count (and rate) of samples converging to unanimous *wrong* consensus by each round. The monotonically increasing trend confirms that extended debate amplifies sycophantic convergence: by round 3, nearly one in four initially-disputed questions (23.9%) ends in wrong consensus.

	$\alpha = 0.05$				$\alpha = 0.10$			
Domain	**Rd 0**	**Rd 1**	**Rd 2**	**Rd 3**	**Rd 0**	**Rd 1**	**Rd 2**	**Rd 3**
Engineering	0	0	0	0	0	0	0	0
Law	0	0	0	0	1	0	0	0
Chemistry	1	0	0	0	14	19	8	2
Physics	0	0	0	0	24	19	12	6
Math	17	9	2	2	31	3	0	0
Economics	0	0	0	0	11	3	1	0
Health	0	0	0	0	2	0	0	0
Psychology	0	0	0	0	2	0	0	0
All	18	9	2	2	85	44	21	8
% of n	0.4	0.2	0.0	0.0	2.0	1.1	0.5	0.2

Table 10. New wrong singletons introduced by conformal prediction without prior unanimous wrong consensus: cases where agents do *not* all agree on the wrong answer, but the conformal set collapses to $|\mathcal{C}|$=1 containing the wrong answer. Counts per round; **% of** n relative to total inference samples (n=4,158). At α=0.05 round 3, conformal catches 480 wrong-consensus errors while introducing only 2 new wrong singletons (ratio 240:1). At α=0.10 the ratio is 41:1.

dure is applied per benchmark with a 50/50 split; numbers in Tables 11 and 12 refer to the inference half. The qualitative pattern from MMLU-Pro replicates: wrong-consensus convergence is substantial (34.5% aggregate at round 3, vs. 23.9% on MMLU-Pro), and conformal prediction intercepts most of it (90.6% at α=0.05; 80.8% at α=0.10).

G.6. Early Stopping on OOD Benchmarks

We reproduce the consensus-vs-conformal early-stopping comparison (Table 2) on the two OOD benchmarks at

		Wrong-Consensus Convergence			Wrong Consensus		Correct Consensus	
Benchmark	n	**Disagr.**	**→WC**	**WC%**	**Count**	**Rej.%**	**Count**	**Rej.%**
GPQA (extended)	273	190	55	28.9	63	87.3	190	61.6
LogiQA	4339	1810	635	35.1	1042	90.8	3139	74.2
All	—	**2000**	**690**	**34.5**	**1105**	**90.6**	**3329**	**73.5**

Table 11. Wrong-consensus convergence and conformal interception at round 3 on two OOD benchmarks (α=0.05). Counts are over the inference half of each benchmark after a 50/50 calibration split. **Disagr.** = inference samples with no consensus at round 0. **→WC** = samples converging to unanimous wrong consensus by round 3. **Rej.%** = fraction of wrong- or correct-consensus cases where $|\mathcal{C}| > 1$ prevents automated action.

		Wrong-Consensus Convergence			Wrong Consensus		Correct Consensus	
Benchmark	n	**Disagr.**	**→WC**	**WC%**	**Count**	**Rej.%**	**Count**	**Rej.%**
GPQA (extended)	273	190	55	28.9	63	76.2	190	39.5
LogiQA	4339	1810	635	35.1	1042	81.1	3139	55.8
All	—	**2000**	**690**	**34.5**	**1105**	**80.8**	**3329**	**54.9**

Table 12. Same analysis as Table 11 at α=0.10. Aggregate rejection of wrong-consensus cases drops to 80.8% (vs. 90.6% at α=0.05), consistent with a user-adjustable risk/automation trade-off.

		Avg Round		Singleton %		Singleton Acc.		
Benchmark	n	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**ΔAcc**
GPQA (extended)	273	1.84	2.21	93.4	35.2	75.3	91.7	+16.3
LogiQA	4339	1.51	2.38	97.1	25.7	75.6	89.2	+13.7

Table 13. Consensus vs. conformal early stopping on the two OOD benchmarks (α=0.05, round 3). Same column definitions as Table 2: **Avg Round** = average 0-indexed stopping round; **Singleton %** = fraction of samples resolved as singletons; **Singleton Acc.** = accuracy among resolved singletons; **ΔAcc** = conformal minus consensus singleton accuracy.

both miscoverage levels. Tables 13 and 14 report the same columns as the main-body table. Conformal stopping trades automation for substantially higher singleton accuracy (GPQA: +16.3pp at α=0.05; LogiQA: +13.7pp), matching the direction of the MMLU-Pro result.

H. Response Parsing Failure Analysis

A small fraction of model responses failed to parse into a valid answer option. Table 15 reports the complete failure rate for each model across all eight domains, aggregated over all debate rounds. Overall, 764 out of 99,744 total responses (0.77%) could not be parsed. Claude-Haiku exhibited the lowest failure rate (0.02%), while DeepSeek-R1 showed the highest (2.05%). As described in Section 3.2, these failures are handled by substituting a uniform distribution over the label set for the affected agent-round, so all samples remain in the calibration and evaluation pipeline. Because the failure rate is negligible (0.77%), the uniform substitutions do not materially affect the reported coverage guarantees.

		Avg Round		Singleton %		Singleton Acc.		
Benchmark	n	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**Consensus**	**Conformal**	**ΔAcc**
GPQA (extended)	273	1.84	1.90	93.4	51.6	75.3	88.7	+13.3
LogiQA	4339	1.51	1.67	97.1	49.6	75.6	86.2	+10.7

Table 14. Consensus vs. conformal early stopping on the two OOD benchmarks (α=0.10, round 3). Columns as in Table 13. Relaxing α from 0.05 to 0.10 increases the conformal singleton rate and lowers the average stopping round, at the cost of smaller singleton-accuracy gains.

I. Domain-Level Analysis

Why do some domains resist singleton convergence? The prediction set sizes reveal a clear pattern tied to domain reasoning type. *Formal-reasoning* domains (Math, Physics, Chemistry) converge rapidly to singletons: Math reaches 98.2% singleton rate at α=0.05 by round 3, reflecting that these domains have unambiguous correct answers that debate can reliably identify. *Interpretive-reasoning* domains (Law, Engineering, Health) maintain large sets: Law's average set size of 6.99 reflects genuine ambiguity among 10 options, where multiple answers may appear plausible under different legal interpretations. This domain-adaptive behavior emerges entirely from the calibration procedure, without any domain-specific heuristic tuning.

Trade-off: reliability vs. automation. The gains in singleton accuracy come at the cost of automation coverage. On Math, where consensus is already reliable (94.9%), conformal stopping offers no accuracy advantage (−0.3pp) while resolving a comparable fraction of cases (98.2% vs. 98.8%). Conversely, on Law, conformal stopping improves accuracy by 22.1pp but resolves only 6.2% of samples—the remaining 93.8% are escalated to human review. This is not a weakness but an intended feature: on genuinely ambiguous domains, the framework correctly identifies that automated action is unsafe. The operating point is user-adjustable: increasing α from 0.05 to 0.10 raises Law's singleton rate from 6.2% to 20.9% while maintaining 90.7% rejection of wrong-consensus cases.

J. Abstention Baseline Comparison

A natural question is whether simpler abstention rules catch roughly the same wrong-consensus cases as conformal prediction. We compare against three heuristic baselines at α=0.05, all using the aggregated social probability P_{social}: (1) **Fixed** τ=0.95: act if $\max_y P_{\text{social}}(y \mid x) \geq 0.95$ (no calibration); (2) **Top-1 Prob (cal-tuned)**: act if $\max_y P_{\text{social}}(y \mid x) \geq \tau$, with τ tuned per round on the calibration half for ≥95% accuracy; (3) **Entropy (cal-tuned)**: act if $H(P_{\text{social}}(\cdot \mid x)) \leq \tau$, similarly calibration-tuned. Vote k=2 (act if ≥2 of 3 agents agree) is omitted because unanimous consensus trivially satisfies the 2-of-3 condition, yielding 0% interception on all consensus cases.

The cal-tuned baselines share conformal's calibrate-then-apply structure but differ in two ways: (i) they use *label-independent* scores producing binary act/abstain decisions, whereas conformal uses a *label-dependent* nonconformity score producing *set-valued* predictions; (ii) conformal computes $\hat{q}$ as a quantile with finite-sample correction providing a marginal coverage guarantee, while the baselines find τ by empirical sweep with no transfer guarantee.

Table 15. Response parsing failure rates across models and domains (all rounds combined). Each cell shows failures / total responses (%).

	Law		Health		Economics		Psychology		Chemistry		Engineering		Math		Physics		Overall
Model	Fail	Rate	Fail	Rate	Fail	Rate	Fail	Rate	Fail	Rate	Fail	Rate	Fail	Rate	Fail	Rate	Rate
Claude-Haiku	1/4,404	0.02%	1/3,272	0.03%	0/3,376	0.00%	0/3,192	0.00%	2/4,528	0.04%	0/3,876	0.00%	2/5,404	0.04%	1/5,196	0.02%	0.02%
DeepSeek-R1	128/4,404	2.91%	41/3,272	1.25%	54/3,376	1.60%	37/3,192	1.16%	118/4,528	2.61%	129/3,876	3.33%	67/5,404	1.24%	108/5,196	2.08%	2.05%
Qwen-32B	2/4,404	0.05%	4/3,272	0.12%	5/3,376	0.15%	0/3,192	0.00%	18/4,528	0.40%	21/3,876	0.54%	16/5,404	0.30%	9/5,196	0.17%	0.23%
All Models: 764 / 99,744																	0.77%

Wrong-consensus interception. Table 16 reports the round-3 interception rates. All four methods achieve comparable overall WC interception (79–82%). Conformal leads at **81.9%** with moderate correct-consensus over-rejection (31.9%). Fixed τ has the lowest over-rejection (**26.8%**) but catches fewer errors (79.4%). The cal-tuned baselines match conformal's WCR but over-reject substantially more correct cases (33–35%).

Domain case studies. On *Health* (67 wrong-consensus, 335 correct-consensus cases), Entropy achieves 100% WC interception but rejects 97% of correct consensus—abstaining on nearly everything. Conformal intercepts 88.1% while over-rejecting 69.3%, providing a more balanced operating point. On *Chemistry* (63 wrong-consensus, 494 correct-consensus), conformal leads all methods on WC interception (82.5%) while keeping CC over-rejection moderate (18.8%); the cal-tuned baselines drop to 62–67% WCR with only marginally lower CCR (5–6%).

The value is the guarantee, not the curve. All baselines trace similar interception-vs-over-rejection tradeoffs. Conformal's contribution is the *coverage certificate*: $\hat{q}$ ensures $\Pr[y \in \mathcal{C}(x)] \geq 1-\alpha$ under exchangeability, while baseline thresholds provide no such assurance. Empirically, the cal-tuned thresholds do not transfer reliably—e.g., Physics test accuracy drops to 91.8–91.9% from a 95% calibration target.

K. Future Work

Conformal Social Choice provides a marginal coverage guarantee for closed-set classification with a fixed agent ensemble. Relaxing each of these constraints opens concrete research directions.

Conditional coverage. Our marginal guarantee does not ensure coverage within specific subgroups (e.g., hard questions or particular domains). Bayesian posterior bounds on singleton error rates—for instance, modeling $\Pr[\text{error} \mid |\mathcal{C}|=1]$ via a Beta posterior updated from calibration data—could complement the marginal guarantee with an instance-conditional risk certificate. More broadly, group-balanced calibration and conformal risk control offer principled routes toward conditional coverage.

Learned weighting. Our entropy-based weighting ablation (Appendix F) shows negligible differences from uniform weights after debate convergence; however, accuracy-based or learned weighting schemes optimized on a separate validation split may improve set efficiency on domains where agent quality varies substantially.

Score comparison. We define a rank-based cumulative score in Appendix B but use only the probability-based score in experiments. A systematic comparison of score functions, including RAPS-style regularization (Angelopoulos et al., 2021), may yield tighter prediction sets.

Broader tasks and ensembles. Evaluating on graduate-level benchmarks (e.g., GPQA), open-ended generation tasks, and larger or more diverse ensembles (varying N, model families, and agent roles) would further test the generality of the framework.

Online and adaptive calibration. In deployment, the exchangeability assumption may be violated by distribution shift. Online conformal methods that update $\hat{q}$ as new labeled data arrives could maintain coverage guarantees in non-stationary settings.

L. Limitations

The guarantee is marginal, not conditional or per-instance, and applies at a fixed calibrated intervention point rather than to arbitrary data-dependent stopping unless combined with Bonferroni or anytime-valid variants. The empirical study uses multi-agent multiple-choice debate as a controlled testbed; extending the same interface to open-ended agentic workflows requires candidate generation followed by conformal filtering. Finally, the certificate binds a frozen pipeline to a frozen task distribution, so prompt, agent, weight, or tool-policy changes require recalibration.

		WC Convergence			Conformal		Fixed τ=.95		Top-1 (cal)		Entropy (cal)	
Domain	n	**Disagr.**	**→WC**	**WC%**	**WC R**	**CC R**	**WC R**	**CC R**	**WC R**	**CC R**	**WC R**	**CC R**
Engineering	485	312	75	24.0	88.0	41.1	83.1	25.6	89.2	43.7	91.6	46.7
Law	551	388	135	34.8	97.5	95.9	88.3	75.3	98.8	97.4	98.8	99.1
Chemistry	566	288	55	19.1	82.5	18.8	77.8	10.5	66.7	5.9	61.9	5.3
Physics	650	313	50	16.0	70.7	8.4	69.0	10.0	41.4	3.5	43.1	3.9
Math	676	256	31	12.1	11.4	0.6	60.0	6.3	20.0	0.3	17.1	0.2
Economics	422	135	35	25.9	62.0	26.5	52.0	25.7	66.0	37.6	66.0	37.8
Health	409	145	46	31.7	88.1	69.3	82.1	53.4	95.5	86.6	100.0	97.0
Psychology	399	126	42	33.3	91.2	43.3	91.2	44.9	92.6	53.6	92.6	55.7
All	**4,158**	**1,963**	**469**	**23.9**	**81.9**	<u>31.9</u>	79.4	**26.8**	79.7	33.5	<u>80.0</u>	35.1

Table 16. Wrong-consensus interception at round 3 (α=0.05). **WC R** = wrong-consensus rejection rate (%); **CC R** = correct-consensus over-rejection rate (%). In the aggregate row, **bold** marks the best value and <u>underline</u> the second best (highest WC R, lowest CC R). Conformal achieves the highest overall WC interception (**81.9%**) while Fixed τ achieves the lowest CC over-rejection (**26.8%**). Vote k=2 is omitted: unanimous consensus trivially satisfies the 2-of-3 agreement condition.