NO ONE WINS FOREVER: RETHINKING VOTING, SELF-REFINEMENT, AND DEBATE OF LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) has been established as a paradigm for improving reasoning capabilities through communication among multiple large language models (LLMs). Despite its empirical appeal, the mechanisms underlying multi-agent debate remain underexplored. Existing work often draws largely binary conclusions, characterizing MAD as either effective or ineffective under narrow or even unfair experimental settings. Such analyses fail to capture the rich dynamics of debate, resulting in limited conclusions. In this paper, we develop a simple mathematical model that captures two properties of model behavior: *generation quality* and *answer stability*. Our theory characterizes the complete ordering among three methods: majority voting (MV), self-refinement (SR), and MAD. Instead of drawing a binary conclusion, our analysis suggests that when independent generation favors the correct answer, voting eventually dominates by amplifying this advantage. Otherwise, debate outperforms self-refinement when the correct answer is the most prevalent in the long-run group distribution. We further show that answer stability shapes this collective behavior: stronger interaction increasingly favors answers that the model is more likely to retain under revision. For a fixed number of agents, we also derive stationary accuracy limits and exact finite-budget evaluation formulas. Through extensive experiments across six benchmarks and six LLMs, we find that the proposed theory closely matches empirical results. Guided by these insights, we further propose *gated debate*, which decides whether to use debate or self-refinement at each turn based on a training-free statistic. Results show it outperforms both self-refine and debate baselines at larger model-call budgets. Code and data will be released upon acceptance.

1 INTRODUCTION

Multi-agent debate (MAD) has emerged as a prominent paradigm for extending the capabilities of LLM-based agents across a wide range of tasks, including reasoning (Liang et al., 2024; Du et al., 2023), creative writing (Lin et al., 2025), and cultural debiasing (Tan et al., 2026). As multi-agent systems continue to develop, agent-to-agent debate has become a common coordination primitive (Tang et al., 2024). Despite its growing adoption, however, we still have a limited understanding of when it is beneficial and which capabilities are responsible for its gains (Wu et al., 2025), making the design and deployment of multi-agent systems largely empirical and potentially unstable.

Existing analysis of MAD often draws largely binary conclusions, characterizing MAD as either unconditionally effective or ineffective. On the one hand, some work reports substantial benefits from MAD. Du et al. (2023) demonstrate that multi-agent debate can substantially outperform both single-agent reasoning and self-reflection, while Xu et al. (2026) show that multi-agent debate can also be effective in long-context tasks. In contrast, other studies conclude that MAD itself does not provide meaningful benefits, showing that multi-agent debate can underperform simpler aggregation strategies such as majority voting (Choi et al., 2026), and amplify errors when agents become disproportionately influenced by incorrect peers (Estornell & Liu, 2024). Taken together, the literature presents a puzzling or even conflicting picture.

We argue that these apparently conflicting findings arise in part from two limitations in existing work. First, prior studies are often conducted under relatively narrow experimental settings, covering only a small number of models or tasks with limited generalizability. Second, existing comparisons

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

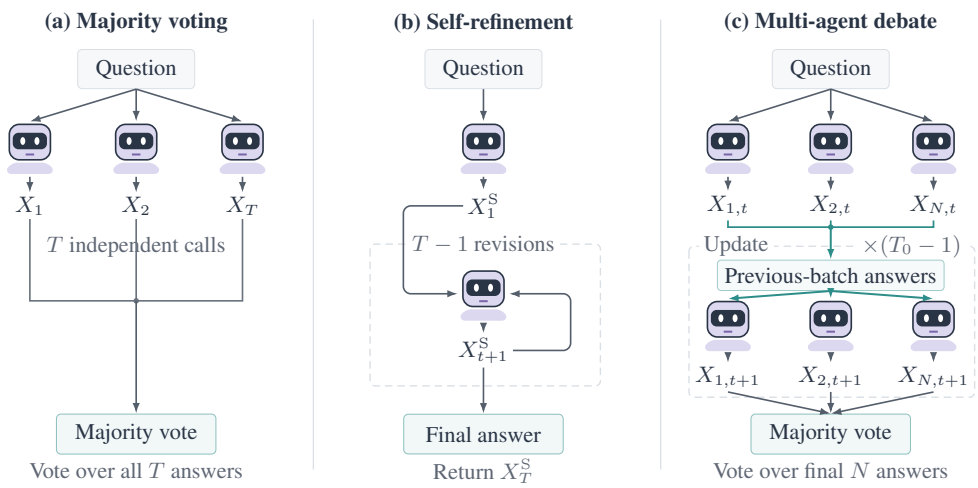


Figure 1: Matched-budget protocols with $T = NT_0$ model calls. MV draws T independent answers. SR makes one initial call and $T - 1$ revisions. MAD initializes N agents and performs $T_0 - 1$ synchronous update rounds, and votes over the final N answers.

often fail to control for inference budget for MAD and baselines, leading to unmatched inference-time computation. This motivates a more fundamental question: rather than asking whether multi-agent debate works, *we ask under the same number of model calls, when should one approach be more accurate than the others?*

To address this question, we develop a simple answer-level model that compares self-refinement (Madaan et al., 2023) (SR), majority voting (Wang et al., 2022) (MV), and multi-agent debate (Du et al., 2023) (MAD) under the same model-call budget (Figure 1). The model separates two properties of agent behavior: *generation quality*, described by the distribution of answers from an independent call, and *answer stability*, described by the answer-dependent probability of retaining an answer rather than redrawing it during revision. We study how these properties, together with group feedback, affect the relative accuracy of the three methods. Our contributions are as follows:

- **Complete eventual ordering.** Under the same model-call budget, we establish a strict ordering of MV, SR, and MAD as both the number of agents and rounds grow, provided the baseline and debate margins are nonzero. If independent generation favors the correct answer, MV wins. Otherwise, MAD wins when the correct answer leads the debate equilibrium, and SR wins when it does not. We compute this equilibrium from generation quality, answer stability, and group interaction through one scalar equation. We further show that interaction favors more stable answers and,identify the threshold at which the correct answer becomes the equilibrium leader.

- **Gated debate design.** Inspired by our theoretical modeling, since interaction can reinforce either correct or incorrect answers, we introduce a gate that decides before each revision whether agents receive their peers' answers. For binary answers, an oracle gate uses the correct label to activate interaction when the group's correct-answer share exceeds its probability under independent generation. We prove that its correct-answer count stochastically dominates that of every policy in the matched interaction family at every finite round. At test time, we use the most likely independent answer as a reference instead of the correct label. We characterize when this gate helps or hurts relative to revision without peer answers.

- **Empirical evaluation.** We conduct extensive experiments with *six* large language models across *six* datasets, leading to three main findings: (1) the relative ranking of the three methods varies substantially across datasets; (2) our theory provides a surprisingly strong fit to the empirical results, as demonstrated by an accuracy-fitting analysis; and (3) the theoretically derived conditions under which multi-agent methods are expected to be effective are well supported by the experimental results.

2 Related work

Independent Voting and Self-Refinement. Independent sampling provides a simple way to spend additional inference compute. Self-consistency demonstrates the effectiveness of this approach by sampling diverse reasoning paths and selecting the most frequent final answer (Wang et al., 2022; 2024). In multi-agent settings, (Choi et al., 2026) further find that voting over initial responses accounts for much of the improvement attributed to debate. In contrast, self-refinement uses additional inference to revise a single evolving answer. Self-Refine implements this process through repeated feedback and revision using the same language model (Madaan et al., 2023). Studies of intrinsic self-correction find that revision without external feedback can fail to improve reasoning and can even degrade initially correct responses (Huang et al., 2024). Broader analyses emphasize the importance of feedback reliability and training assumptions (Kamoi et al., 2024). More recently, Kumar et al. (2025) show that self-correction can be strengthened through multi-turn reinforcement learning. Our model and analysis mainly leverage these two methods as well-established agent baselines, capturing both generation quality and answer stability.

Multi-Agent Debate. Multi-agent debate (MAD) has emerged as a broad family of inference-time methods in which multiple language-model agents exchange information and revise their responses. Early work demonstrated improvements in factuality and reasoning through repeated peer interaction (Du et al., 2023), while adversarial debate was proposed to encourage alternative perspectives when self-reflection becomes trapped in an incorrect solution (Liang et al., 2024). Subsequent research has explored how to preserve useful differences between agents and structure their exchanges, including distinct reasoning strategies (Liu et al., 2025), dynamically sparse communication (Sun et al., 2025), and explicit training for collaboration (Estornell et al., 2025). Recent surveys organize these developments by collaboration mechanisms and debate design choices, including participants, interaction, and agreement protocols (Motger et al., 2026; Tran et al., 2025). Yet interaction can also propagate errors: Wynn et al. (2025) document cases in which agents abandon correct answers in response to incorrect peer reasoning. These mixed findings motivate explaining when group interaction improves final accuracy.

Theory of Multi-Agent Debate. A central theoretical question is what makes interaction favor correct answers. Estornell & Liu (2024) show how similar responses or capabilities can produce static debate dynamics and reinforce shared misconceptions. Under specified Bayesian update assumptions, Choi et al. (2026) characterize agents' beliefs as martingales, with no systematic increase in expected belief in the correct answer. These results suggest that effective debate requires an asymmetry that favors correct answers. Existing interventions seek to create such an asymmetry by preserving correct responses in an oracle setting or using majority opinions to guide updates in practice (Choi et al., 2026), communicating calibrated confidence (Zhu et al., 2026)

We study a different source of asymmetry: *answer stability*. Our model captures *generation quality* through the baseline answer distribution and answer stability through answer-dependent persistence under revision. We show that stability asymmetry makes interaction directional: stronger interaction shifts the equilibrium distribution toward more stable answers. Because stability need not align with correctness, this shift can improve or degrade collective accuracy. We explicitly characterize the equilibrium, derive the interaction threshold above which the correct answer becomes its leader when it is uniquely the most stable, and use these results to compare debate, independent voting, and sequential refinement under the same number of model calls.

3 Problem Setting

We compare majority voting, self-refinement, and multi-agent debate under the same number of model calls. At a high level, *Majority voting (MV)* draws independent answers and selects a plurality. *Self-refinement (SR)* revises one evolving answer and returns its final answer. *Multi-agent debate (MAD)* revises several answers using the current group's answer distribution and votes over the final group. For simplicity, we formulate these protocols at the answer level for a fixed question.

Fix a question and an answer set $[K] = \{1, \dots, K\}$, $K \geq 2$, and let $g \in [K]$ be the correct answer. An independent model call returns $b \in [K]$ with probability p_b, where $p = (p_1, \dots, p_K)$ is the *baseline distribution*. This distribution describes *generation quality* for the fixed question.

Majority Voting. For a budget of T calls, majority voting draws T independent answers from p. If $M_T(b)$ is the number of draws that return b, then $M_T \sim \mathrm{Multinomial}(T, p)$. MV returns a plurality of these T answers.

Self-Refinement. Self-refinement maintains one evolving answer. Let X_t^{S} be its answer after the t-th call, with $X_1^{\mathrm{S}} \sim p$. If the current answer is a, SR retains it with probability $1-\alpha_a^{\mathrm{S}}$ and otherwise resamples from p. Thus its answer-level transition rule is

$$P^{\mathrm{S}}(b \mid a) := \Pr(X_{t+1}^{\mathrm{S}} = b \mid X_t^{\mathrm{S}} = a) = (1-\alpha_a^{\mathrm{S}})\mathbf{1}\{a=b\} + \alpha_a^{\mathrm{S}} p_b. \tag{1}$$

The retention probability $1-\alpha_a^{\mathrm{S}}$ captures the model's stability of answer a under self-refinement. Note that a redraw can produce a again, so α_a^{S} is not the probability of observing a changed answer. In addition, we write $\alpha^{\mathrm{S}} = (\alpha_1^{\mathrm{S}}, \ldots, \alpha_K^{\mathrm{S}})$.

Multi-Agent Debate. Multi-agent debate maintains N evolving answers. Let $X_{i,t}$ be agent i's answer after round t. The count and empirical fraction of answer b are $n_t(b) := \sum_{i=1}^{N} \mathbf{1}\{X_{i,t} = b\}$, and $\mu_{t,b} := \frac{n_t(b)}{N}$, respectively. Conditional on the previous round, agents update independently and simultaneously. An agent with answer a retains it with probability $1-\alpha_a^{\mathrm{D}}$. Otherwise, it redraws from p with probability $1-\rho$, or repeats a uniformly chosen member of the previous group with probability ρ:

$$P_\rho^{\mathrm{D}}(b \mid a, \mu) = (1-\alpha_a^{\mathrm{D}})\mathbf{1}\{a=b\} + \alpha_a^{\mathrm{D}}\big[(1-\rho)p_b + \rho\mu_b\big]. \tag{2}$$

The vector $1-\alpha^{\mathrm{D}}$ captures *answer stability* during debate, while ρ measures how strongly a redraw follows the group. In general, α^{S} and α^{D} may differ.

Matched Model-Call Budget. For fair comparison, we use the same budget for all three methods (Figure 1). Specifically, let debate use $N \geq 2$ agents for $T_0 \geq 1$ rounds, and set $T = NT_0$. Initial answers count toward this common budget. Under this budget, SR makes $T-1$ revisions after its initial draw and returns X_T^{S}. MAD draws N independent initial answers, updates each answer for $T_0 - 1$ rounds, and votes over the final N answers. MV uses all T calls as independent final votes.

For a count vector n, let $W(n) := \arg\max_b n(b)$ be the set of plurality winners. Ties are broken uniformly, so the conditional probability of selecting g is $v_g(n) := \mathbf{1}\{g \in W(n)\}/|W(n)|$. The accuracies of three methods are

$$A_{\mathrm{MV}}(T) := \mathbb{E}[v_g(\mathbf{M}_T)], \quad A_{\mathrm{SR}}(T) := \Pr(X_T^{\mathrm{S}} = g), \quad A_{\mathrm{MAD}}(N, T_0; \rho) := \mathbb{E}[v_g(\mathbf{n}_{T_0})]. \tag{3}$$

At $\rho = 0$, MAD becomes N independent refinement chains of length T_0 followed by a vote; SR instead spends all NT_0 calls on one chain. The matched budget counts the number of model calls.

Throughout the theory, we fix $K, p, \alpha^{\mathrm{S}}, \alpha^{\mathrm{D}}, \rho$ for this question and assume $p_b > 0$, $0 < \alpha_b^{\mathrm{S}}, \alpha_b^{\mathrm{D}} \leq 1$, and $0 \leq \rho < 1$ for all $b \in [K]$.

4 Theoretical Analysis

In this section, we study the question: *which method wins under a matched budget?* We first introduce two long-run answer distributions. The SR chain has a stationary distribution, while the large-group debate dynamics have an equilibrium distribution. Both can be computed from generation quality p and the corresponding answer stability parameters. The debate equilibrium also depends on interaction strength ρ. We then use these distributions to compare the methods as $N, T_0 \to \infty$, with all parameters fixed.

4.1 Long-Run Distributions and Method Comparison

SR Stationary Distribution. SR has a stationary distribution

$$s_b^{\mathrm{S}} := \frac{p_b/\alpha_b^{\mathrm{S}}}{\sum_{j=1}^{K} p_j/\alpha_j^{\mathrm{S}}}, \qquad b \in [K]. \tag{4}$$

Here s_b^{S} is the long-run probability that SR outputs answer b, whereas p_b is the probability of generating b in an independent model call. The factor $1/\alpha_b^{\mathrm{S}}$ measures its mean wait until the next redraw. Consequently, $A_{\mathrm{SR}}(T) \to s_g^{\mathrm{S}} \in (0,1)$ as $T \to \infty$ (see Appendix A.1).

MAD Equilibrium Distribution. Let $\alpha^{\mathrm{D}}_{\min} := \min_b \alpha^{\mathrm{D}}_b$. The feedback level z_ρ is determined by the scalar equation

$$\Phi(z) := \sum_{b=1}^{K} \frac{\alpha^{\mathrm{D}}_b p_b}{\alpha^{\mathrm{D}}_b - z}, \quad 0 \le z < \alpha^{\mathrm{D}}_{\min}, \qquad \Phi(z_\rho) = \frac{1}{1-\rho}. \tag{5}$$

The solution z_ρ is unique. In the large-group limit, answer b has equilibrium share

$$u^\star_{\rho,b} := \frac{p_b/(\alpha^{\mathrm{D}}_b - z_\rho)}{\sum_{j=1}^{K} p_j/(\alpha^{\mathrm{D}}_j - z_\rho)}, \qquad b \in [K]. \tag{6}$$

In particular, $z_\rho = \rho \sum_b \alpha^{\mathrm{D}}_b u^\star_{\rho,b}$ is the equilibrium fraction of agents that redraw from the previous round's group distribution in each update. When $\rho = 0$, we have $z_0 = 0$ and $u^\star_{0,b} \propto p_b/\alpha^{\mathrm{D}}_b$. Thus, revision alone can make an initially less probable answer the unique equilibrium leader. Increasing ρ shifts the equilibrium distribution toward more stable answers, in the sense that $u^\star_{\rho,a}/u^\star_{\rho,b}$ increases whenever $\alpha^{\mathrm{D}}_a < \alpha^{\mathrm{D}}_b$.

Correct-Answer Margins. We define the correct answer's margins in the baseline and debate equilibrium:

$$d_p := p_g - \max_{b \ne g} p_b, \qquad d^\star_\rho := u^\star_{\rho,g} - \max_{b \ne g} u^\star_{\rho,b}. \tag{7}$$

A positive margin means that g is the unique leading answer. The first margin describes generation quality relative to rival answers. The second also depends on debate stability α^{D} and interaction ρ.

The following theorem gives our main result.

Theorem 4.1 (Complete eventual ordering)**.** *Under the standing assumptions in Section 3, suppose $d_p \ne 0$ and $d^\star_\rho \ne 0$. Along every sequence with $N, T_0 \to \infty$, we have*

$$A_{\mathrm{MV}}(NT_0) \longrightarrow \mathbf{1}\{d_p > 0\}, \quad A_{\mathrm{SR}}(NT_0) \longrightarrow s^{\mathrm{S}}_g, \quad A_{\mathrm{MAD}}(N, T_0; \rho) \longrightarrow \mathbf{1}\{d^\star_\rho > 0\}. \tag{8}$$

Moreover, there are finite N_0 and $T_{0,\min}$ such that the strict ordering in Table 1 holds whenever $N \ge N_0$ and $T_0 \ge T_{0,\min}$.

Table 1: Eventual ordering under $T = NT_0$ model calls. Parameters are fixed, and both N and T_0 are sufficiently large.

Baseline margin	Debate margin	Strict ordering	Winner
$d_p > 0$	$d^\star_\rho > 0$	MV > MAD > SR	MV
$d_p > 0$	$d^\star_\rho < 0$	MV > SR > MAD	MV
$d_p < 0$	$d^\star_\rho > 0$	MAD > SR > MV	MAD
$d_p < 0$	$d^\star_\rho < 0$	SR > MAD > MV	SR

Interpretation. The Theorem 4.1 gives a direct choice under a matched large budget. If a fresh call favors g, voting eventually wins by amplifying its baseline advantage. Otherwise, MAD wins when debate makes g the leading answer, and SR wins when debate still favors an incorrect answer. The SR redraw rates determine the interior limit s^{S}_g and can affect how large a budget is needed, but they do not change the four eventual orderings. In particular, MAD can recover a correct answer that is initially less frequent than a rival, provided that it leads at equilibrium. If neither the baseline nor the debate equilibrium favors g, SR retains a positive chance of returning it while the other two accuracies approach zero.

The conditions can be checked from the model parameters. By (6),

$$d^\star_\rho > 0 \iff p_g \alpha^{\mathrm{D}}_b - p_b \alpha^{\mathrm{D}}_g + (p_b - p_g) z_\rho > 0 \quad \text{for every } b \ne g. \tag{9}$$

Thus one scalar root and $K - 1$ comparisons determine whether debate favors the correct answer. The final accuracies themselves need not be computed.

4.2 PROOF OVERVIEW

MV concentrates around p, and the SR chain converges to (4). For MAD, define $\bar{\alpha}(u) := \sum_b \alpha_b^{\mathrm{D}} u_b$ as the average redraw probability at group distribution u. Averaging one synchronous update gives $\mathbb{E}[\mu_{t+1} \mid \mu_t] = F_\rho(\mu_t)$, where

$$F_\rho(u)_b = (1 - \alpha_b^{\mathrm{D}})u_b + \bar{\alpha}(u)\big[(1-\rho)p_b + \rho u_b\big]. \tag{10}$$

The first term accounts for agents that retain answer b, and the second for agents that redraw and obtain b. The population map has the unique attracting fixed point $u_\rho^\star$.

Lemma 4.2 (Equilibrium and finite-group approximation). *For constants $C_1, C_2 < \infty$ and $0 \le \chi < 1$ depending only on the fixed parameters, the debate fractions satisfy, for every N and $t \ge 1$,*

$$\mathbb{E}\|\mu_t - u_\rho^\star\|_1 \le C_1 \chi^{t-1} + C_2\sqrt{K/N}. \tag{11}$$

The map F_ρ has $u_\rho^\star$ as its unique fixed point, and every deterministic trajectory converges to it.

The first term bounds the effect of initialization and decays with the number of rounds. The second bounds finite-group sampling error uniformly over rounds. The proof in Appendix B.1 uses a contraction after reweighting the answer shares to keep these errors from accumulating. With $d_\rho^\star \neq 0$, a final vote can give a different correctness decision from the equilibrium only when the empirical fractions deviate by at least $|d_\rho^\star|$. This proves the MAD limit in (8) for any joint growth of N and T_0. The two quantities need not grow at comparable rates.

When MV and MAD both approach zero or both approach one, their limits alone do not give a strict ordering. Each debate agent has conditional probability at least $q_b := (1-\rho)\alpha_{\min}^{\mathrm{D}} p_b > 0$ of producing b in the final round. A unanimous round therefore has probability at least q_b^N, whether b is correct or incorrect. By contrast, the probability that MV selects against the baseline leader is at most a constant times $e^{-d_p^2 N T_0/2}$. Increasing T_0 separates these errors uniformly in N, even when the two limiting accuracies agree. Appendix B.2 gives the bounds and completes the proof of Table 1.

4.3 HOW INTERACTION CHANGES THE LEADING ANSWER

To understand when interaction changes the leading answer, consider the equilibrium odds between the correct answer g and any rival $b \neq g$:

$$\frac{u_{\rho,g}^\star}{u_{\rho,b}^\star} = \frac{p_g}{p_b}\frac{\alpha_b^{\mathrm{D}} - z_\rho}{\alpha_g^{\mathrm{D}} - z_\rho}. \tag{12}$$

The first factor captures the baseline preference under independent generation, while the second captures the effects of answer stability and interaction. Increasing interaction raises the equilibrium ratio in favor of the answer with the smaller redraw probability, as formalized below.

Proposition 4.3 (Effect of increasing interaction). *The root z_ρ increases strictly with ρ. Hence the odds in (12) increase when $\alpha_g^{\mathrm{D}} < \alpha_b^{\mathrm{D}}$, decrease when the inequality is reversed, and remain p_g/p_b when the two redraw probabilities are equal. If g has the uniquely smallest debate redraw probability, it is the unique equilibrium leader for all $\rho < 1$ sufficiently close to one.*

Thus interaction favors answers that are more stable under revision, regardless of whether they are correct. If $\alpha_g^{\mathrm{D}} < \alpha_b^{\mathrm{D}}$ for all $b \neq g$, there is an explicit interaction threshold. Put

$$z_c := \max\left(\{0\} \cup \left\{\frac{p_b\alpha_g^{\mathrm{D}} - p_g\alpha_b^{\mathrm{D}}}{p_b - p_g} : b \neq g,\ p_b > p_g\right\}\right), \qquad \rho_c := 1 - \frac{1}{\Phi(z_c)}. \tag{13}$$

For $0 < \rho < 1$, g leads the debate equilibrium exactly when $\rho > \rho_c$. Appendix B.3 proves the proposition and threshold. At $\rho = 0$, $u_{0,b}^\star \propto p_b/\alpha_b^{\mathrm{D}}$. When the two revision vectors are also equal, $u_{0,b}^\star = s_b^{\mathrm{S}}$, but this equality concerns individual answers, whereas MAD votes over N final answers. The threshold changes the eventual winning protocol when $d_p < 0$. When $d_p > 0$, MV wins even if interaction makes g the debate leader.

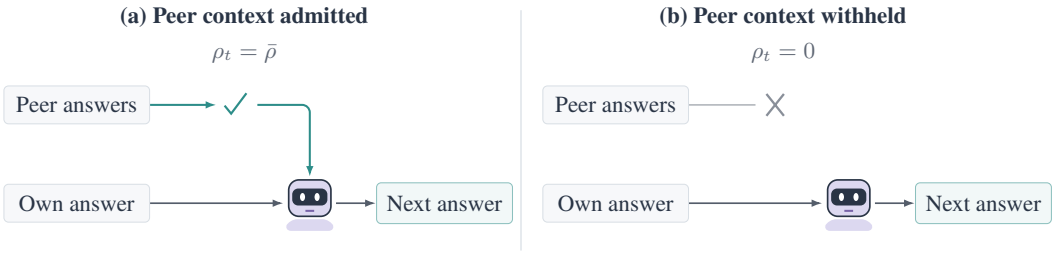


Figure 2: Overview of proposed gated debate. The gate controls peer information for each agent's next update, while the own-answer path remains active in both cases, and every agent still makes one revision call.

Fixed number of agents. When N stays fixed and $T_0 \to \infty$, the final debate vote has a stationary accuracy $a_{N,\rho} \in (0,1)$, obtained by averaging v_g over the stationary law of the finite count chain. SR approaches s_g^{S}. If $d_p > 0$, MV eventually exceeds both, whereas if $d_p < 0$, it eventually falls below both. MAD and SR are ordered by the sign of $a_{N,\rho} - s_g^{\mathrm{S}}$ when it is nonzero. Agents share a history, so this count law is generally different from N independent draws from $u_\rho^\star$. Appendix B.4 provides the count transition kernel, the proof, and exact finite-budget accuracy formulas.

5 Gated Debate

The analysis in Section 4 shows that interaction can favor either correct or incorrect answers, depending on their stability. This motivates a gate that determines whether agents receive the previous round's answers at each update. Our analysis assumes that withholding these answers sets $\rho = 0$ in (2), while leaving p and α^{D} unchanged.

Definition 5.1 (Gated debate). *Fix $\bar{\rho} \in (0,1]$ as the interaction strength when peer answers are provided. At each round t, the gate either supplies those answers, giving effective weight $\rho_t = \bar{\rho}$, or withholds them, giving $\rho_t = 0$. It decides before the next update using the history through round t. At $\rho_t = 0$, all agents still revise using the same α^{D} and fresh-draw distribution p. For the theoretical comparison, we also allow policies to select $\rho_t \in [0, \bar{\rho}]$ from the history, with optional randomness independent of the next update. This interval describes a comparison family, not a claim that intermediate weights can be set by a prompt. All policies update agents independently and synchronously by (2), share the initial law, and use the final plurality rule and budget $T = NT_0$ from Section 3. The constant weights 0 and $\bar{\rho}$ give matched no-group and always debate.*

Definition 5.2 (Binary oracle gate). *The oracle uses the correct label to choose*

$$\rho_t^\star = \bar{\rho}\,\mathbf{1}\{n_t(g)/N > p_g\}. \tag{14}$$

Its final answer still uses the common plurality rule.

The cutoff follows directly from (2):

$$\mathbb{E}[n_{t+1}(g) \mid n_t, \rho_t] - \mathbb{E}[n_{t+1}(g) \mid n_t, \rho_t = 0] = \rho_t(\mu_{t,g} - p_g) \sum_{b \in [K]} n_t(b)\alpha_b^{\mathrm{D}}. \tag{15}$$

Compared with $\rho_t = 0$, interaction with $\rho_t > 0$ increases the expected correct count in the next round when $\mu_{t,g} > p_g$ and decreases it when $\mu_{t,g} < p_g$. The magnitude of this change scales with the expected number of redrawing agents.

Theorem 5.3 (Multi-round optimality of the binary oracle gate). *Assume $K = 2$. Consider the oracle gate and any policy π in the matched policy family of Definition 5.1, with the same initial count distribution. Let $n_t^\star(g)$ and $n_t^\pi(g)$ denote their respective correct-answer counts at round t. For every finite t and $k \in \{0, \dots, N\}$,*

$$\Pr(n_t^\star(g) \geq k) \geq \Pr(n_t^\pi(g) \geq k). \tag{16}$$

Consequently, the oracle's final binary plurality accuracy is at least that of always debate, matched no-group, and every other policy in this family.

Table 2: Overall accuracy (%) with fixed call budgets. Each number is averaged over all *six* LLMs. Bold is the overall best, while the underline indicates the best in the group. * indicates that Gated debate outperforms Always debate under the same call budget. Full results are in Appendix F.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	Super-GPQA
Majority voting	10	2nd 68.24	1st 61.48	1st 84.14	1st 63.86	2nd 92.90	2nd 47.13
Self-refine	10	3rd 67.58	3rd 59.22	3rd 82.42	3rd 56.16	3rd 92.12	3rd 46.55
Always debate	10	1st 68.66	2nd 60.56	2nd 83.37	2nd 62.04	1st 92.97	1st 47.23
Matched no-group	10	68.28	60.68	83.38	61.17	92.47	47.47
Gated debate	10	68.66	*61.12	83.22	61.65	92.72	*47.24
Majority voting	20	2nd 68.39	1st 61.60	1st 84.23	2nd 62.46	1st **93.17**	1st 47.33
Self-refine	20	3rd 67.36	3rd 58.80	2nd 82.78	3rd 58.05	3rd 91.97	3rd 46.46
Always debate	20	1st 68.83	2nd 61.00	3rd 82.67	1st 62.93	2nd 92.86	2nd 47.33
Matched no-group	20	68.58	60.69	83.85	60.50	92.71	47.09
Gated debate	20	*68.93	*61.14	*83.42	*63.32	92.75	*47.49
Majority voting	40	2nd 68.35	1st **62.56**	1st **84.26**	2nd 61.70	1st 93.12	2nd 47.19
Self-refine	40	3rd 67.21	3rd 58.30	3rd 82.49	3rd 60.63	3rd 92.14	3rd 46.52
Always debate	40	1st **69.19**	2nd 60.74	2nd 83.11	1st 63.50	2nd 92.60	1st 47.48
Matched no-group	40	68.79	60.17	83.62	60.54	92.84	**47.80**
Gated debate	40	68.90	*61.14	*83.36	***64.25**	*92.92	*47.60

Interpretation. The sign of the one-step gain in (15) explains the oracle's choice. To extend this comparison across rounds, we construct a coupling under which the oracle's correct count remains at least as large as that of any policy. Appendix C provides the proof. Since the correct answer g is unavailable at test time, we use the most likely answer under the baseline distribution $\mathbf{p}$ as a proxy.

Definition 5.4 (Baseline-mode gate)**.** *Choose* $\widehat{g} \in \arg\max_{b\in[K]} p_b$*, breaking ties deterministically. For a threshold* $\tau \geq 0$ *fixed before evaluation, choose*

$$\rho_t^{\text{gate}} = \bar{\rho}\,\mathbf{1}\{\mu_{t,\widehat{g}} - p_{\widehat{g}} > \tau\}. \tag{17}$$

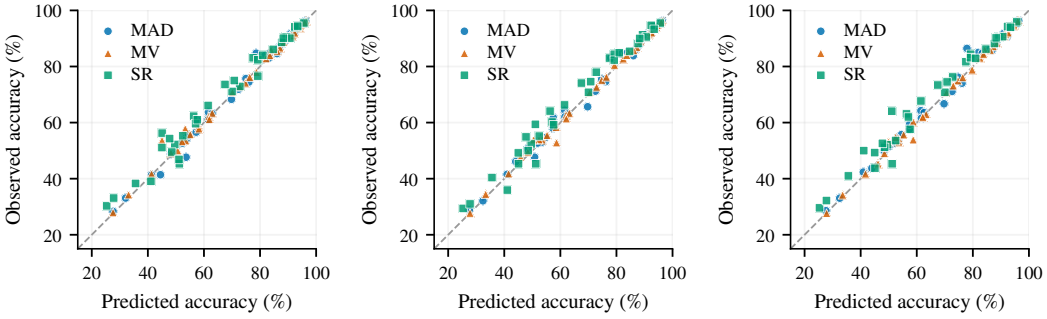
For each question, the reference $\widehat{g}$ *remains fixed, while the gate is recomputed each round.*

Implications for algorithm design. The binary result suggests a simple rule: supply peer answers when the group's share of the correct answer exceeds its probability under independent generation. Since the correct answer is unavailable at test time, Gated Debate uses the baseline-mode gate in Definition 5.4. It takes the most likely answer under p as a fixed reference and supplies peer answers only when its current group share exceeds its baseline probability by the threshold τ. Otherwise, agents continue revising without peer answers, using the same number of model calls. The theoretical guarantee applies to the binary oracle, while this practical rule is defined for any number of answer options. In the next section, we compare its final accuracy with always debate and matched no-group revision across models, benchmarks, and matched call budgets.

6 Experiments and Analysis

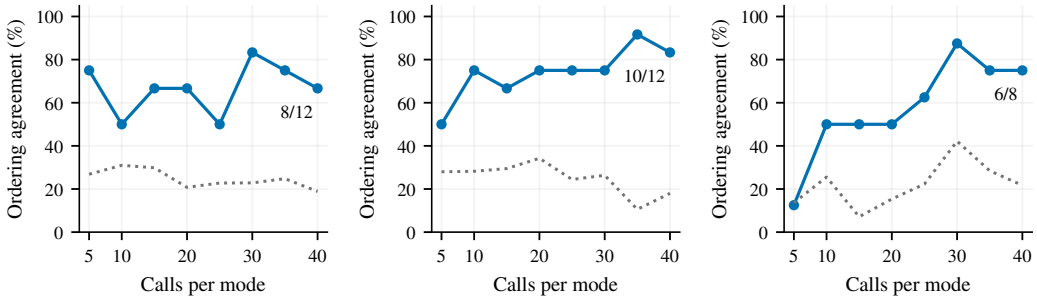
We run large-scale experiments on six models and six datasets, and fit the accuracy and order of three methods using the theory. Full details are in Appendix G.

Overall Results. Table 2 shows the results of all experiments (detail settings in Appendix D). As shown, the ordering among MV, SR, and MAD varies across settings. Concretely, debate achieves the best performance on three datasets and ranks second on the others when the budget is 40, while SR ranks third in most settings. This result motivates our analysis: the advantage of multi-agent debate is not a binary conclusion. Furthermore, Gated debate outperforms vanilla debate in most cases when the call budget is 40, demonstrating the effectiveness of the proposed method. It achieves the best performance less frequently when the call budget is 10 due to noise, which is also aligned



(a) 10 calls (b) 20 calls (c) 40 calls

Figure 3: Predicted versus observed final-answer accuracy for MAD, MV, and SR at matched call budgets. Each panel contains 33 model–benchmark pairs.



(a) **Qwen3.5-4B** (b) **Kimi-K2.6** (c) **GPT-5.4-mini**

Figure 4: Ordering agreement across call budgets for three representative models.

with our theory that Gated debate works best with sufficient rounds. Full accuracy tables at 10, 20, and 40 calls are in Appendix F (Tables 3–8).

Accuracy predictions. To understand the effectiveness of the proposed mathematical model, we run a prediction test in which we first fit the parameters, such as α, ρ, to predict the final accuracy (Appendix E) We conduct extensive evaluation on 33 model–benchmark runs (27,189 model–question pairs). As shown in Figure 3, the predicted accuracy aligns surprisingly well with the empirical accuracy, showing the correctness and potential of the proposed theory.

Ordering agreement. To understand whether empirical results align with Table 1, we compute $d_p > 0$ and d_ρ and compare the ranking with the actual ranking. As shown in Figure 4, the fitted results are much higher than random ranking (dotted lines), showing the effectiveness of the d values. Interestingly, the accuracy improves with a higher number of calls, also aligning with our theoretical analysis. Full results can be found in Appendix G.2.

7 Discussion and Conclusion

In this paper, we develop an answer-level analysis that characterizes when multi-agent debate eventually outperforms majority voting and self-refinement under the same model-call budget. We show that their relative accuracy depends on *generation quality* and *answer stability*: stronger interaction favors more stable answers, whether correct or incorrect. Experiments across six LLMs and six benchmarks show varying method rankings and close agreement between fitted-model and observed accuracies, providing empirical support for this account of debate.

AI USE STATEMENT

In this work, we used generative AI tools to assist in proving mathematical claims, including checking and refining intermediate proof steps, and in implementing algorithms. We did not use generative AI tools for other research tasks requiring disclosure where applicable such as algorithm design, ideation or results interpretation. Additionally, we used generative AI tools to aid and polish the writing of the manuscript, including improving clarity, grammar, and presentation.

All AI-assisted mathematical reasoning was independently reviewed and verified by the authors, and all AI-assisted writing was reviewed and edited to ensure that it accurately reflects the authors' intended arguments and findings. We take responsibility for the final content of this work, including all text, mathematical claims, proofs, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release all code, data, and traces upon acceptance. We have already listed the prompt, parameters, data examples, and settings in the appendix for the purpose of reproduction. Furthermore, we construct the code in a flexible way so that we can rapidly adapt to newer data or models.

REFERENCES

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38: 101732–101764, 2026.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. Acc-collab: An actor-critic approach to multi-agent llm collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 90558–90581, 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. Creativity in llm-based multi-agent systems: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27572–27595, 2025.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations*, volume 2025, pp. 24666–24725, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Quim Motger, Marc Oriol, Jordi Marco, and Xavier Franch. Multi-agent debate strategies: Survey, taxonomy, and challenges. *arXiv preprint arXiv:2607.26212*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9503–9523, 2025.

Qian Tan, Lei Jiang, Yuting Zeng, Shuoyang Ding, and Xiaohua Xu. Mitigating cultural bias in llms via multi-agent cultural debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8600–8612, 2026.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19916–19939, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Haolun Wu, Zhenkun Li, and Lingyao Li. Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*, 2025.

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

Zach Xu, Shang Zhu, Jue Wang, Junlin Wang, Ben Athiwaratkun, Chi Wang, James Y Zou, and Ce Zhang. When does divide and conquer work for long context llm? a noise decomposition framework. In *International Conference on Learning Representations*, volume 2026, pp. 87799–87821, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, 2026.

A Auxiliary results

We collect the bounds for independent voting and self-refinement, and the properties of the debate update used to prove Theorem 4.1. Throughout the appendix, the assumptions of Section 3 hold unless a result explicitly uses the matched policy family in Definition 5.1. When a bound has zero update steps, we use $0^0 = 1$.

A.1 Independent voting and self-refinement

Lemma A.1 (Bounds for MV and SR)**.** *For every $T \geq 1$, if $d_p > 0$, then*

$$1 - A_{\mathrm{MV}}(T) \leq (K-1)e^{-Td_p^2/2};$$

if $d_p < 0$, then $A_{\mathrm{MV}}(T) \leq e^{-Td_p^2/2}$. The distribution s^{S} in (4) is the unique stationary distribution of (1), and

$$\left|A_{\mathrm{SR}}(T) - s_g^{\mathrm{S}}\right| \leq \left(1 - \min_b \alpha_b^{\mathrm{S}}\right)^{T-1}, \qquad 0 < s_g^{\mathrm{S}} < 1.$$

Proof. For $d_p > 0$, fix $b \neq g$ and write X_i for the i-th MV draw. Each vote contributes a random variable $\mathbf{1}\{X_i = g\} - \mathbf{1}\{X_i = b\} \in [-1, 1]$ with mean $p_g - p_b \geq d_p$. Hoeffding's inequality gives

$$\Pr(M_T(b) \geq M_T(g)) \leq e^{-Td_p^2/2}.$$

An incorrect plurality decision implies this event for some $b \neq g$, so a union bound proves the first claim. If $d_p < 0$, choose a rival with $p_b = \max_{j \neq g} p_j$. Selecting g requires $M_T(g) \geq M_T(b)$; the same inequality gives the second claim.

For SR, substitution of (4) into (1) shows that, for every b,

$$s_b^{\mathrm{S}}(1 - \alpha_b^{\mathrm{S}}) + p_b \sum_a s_a^{\mathrm{S}} \alpha_a^{\mathrm{S}} = s_b^{\mathrm{S}}.$$

Every transition contains the same baseline component with weight at least $\min_b \alpha_b^{\mathrm{S}}$. Coupling two chains through that component at each update bounds their probability of disagreement after $T-1$ updates by $(1 - \min_b \alpha_b^{\mathrm{S}})^{T-1}$. This proves uniqueness and the accuracy bound. Positivity of p gives $0 < s_g^{\mathrm{S}} < 1$. □

A.2 The debate update and its equilibrium

Lemma A.2 (One-step sampling error)**.** *For every $t \geq 1$,*

$$\mathbb{E}[\mu_{t+1} \mid \mu_t] = F_\rho(\mu_t), \qquad \mathbb{E}\big[\|\mu_{t+1} - F_\rho(\mu_t)\|_1 \mid \mu_t\big] \leq \sqrt{K/N}.$$

Proof. Grouping the agents by their current answers and applying (2) gives the conditional mean (10). Conditional on the old state, the N next answers are independent. The sum of their coordinate variances, after division by N^2, is at most $1/N$. The bound follows from $\|x\|_1 \leq \sqrt{K}\|x\|_2$ and conditional Cauchy–Schwarz. The conditional quantities depend on the old state only through its counts, so the same bound holds conditional on μ_t. □

Lemma A.3 (Equilibrium from the scalar root)**.** *Equation (5) has a unique solution $z_\rho \in [0, \alpha^{\mathrm{D}}_{\min})$. The vector $u^\star_\rho$ in (6) is the unique fixed point of F_ρ, and $z_\rho = \rho\bar{\alpha}(u^\star_\rho)$.*

Proof. The function Φ is continuous and strictly increasing on $[0, \alpha^{\mathrm{D}}_{\min})$, with $\Phi(0) = 1$ and $\Phi(z) \to \infty$ as $z \uparrow \alpha^{\mathrm{D}}_{\min}$. Thus the root exists and is unique, including $z_0 = 0$. From (5),

$$\Phi(z) = 1 + z\sum_b \frac{p_b}{\alpha^{\mathrm{D}}_b - z}, \qquad \bar{\alpha}(u^\star_\rho) = \frac{\Phi(z_\rho)}{\sum_b p_b/(\alpha^{\mathrm{D}}_b - z_\rho)}.$$

These identities imply $z_\rho = \rho\bar{\alpha}(u^\star_\rho)$. Consequently,

$$(\alpha^{\mathrm{D}}_b - \rho\bar{\alpha}(u^\star_\rho))u^\star_{\rho,b} = (1-\rho)\bar{\alpha}(u^\star_\rho)p_b,$$

which is equivalent to $F_\rho(u^\star_\rho) = u^\star_\rho$.

Conversely, any fixed point u has $\bar{\alpha}(u) > 0$ and, by the same equation, $(\alpha^{\mathrm{D}}_b - \rho\bar{\alpha}(u))u_b = (1-\rho)\bar{\alpha}(u)p_b > 0$ for every b. Hence $\rho\bar{\alpha}(u) < \alpha^{\mathrm{D}}_{\min}$. Multiplying the resulting expression for u_b by α^{D}_b and summing gives $\Phi(\rho\bar{\alpha}(u)) = 1/(1-\rho)$. Uniqueness of the scalar root and normalization now give $u = u^\star_\rho$. □

The next lemma controls repeated updates. Its coordinate change is used only in the proof of the finite-group approximation.

Lemma A.4 (Contraction after reweighting)**.** *There are constants $C < \infty$ and $0 \le \chi < 1$, depending only on the fixed parameters, and a bijection Ψ of the probability simplex such that Ψ and Ψ^{-1} are C-Lipschitz in ℓ_1 and*

$$\|\Psi(F_\rho(u)) - \Psi(u^\star_\rho)\|_1 \le \chi\|\Psi(u) - \Psi(u^\star_\rho)\|_1 \quad \textit{for every probability vector } u. \tag{18}$$

Proof. Write

$$B = I - \mathrm{diag}(\alpha^{\mathrm{D}}) + (1-\rho)p(\alpha^{\mathrm{D}})^\top, \quad \ell_b = \frac{\alpha^{\mathrm{D}}_b}{\alpha^{\mathrm{D}}_b - z_\rho}, \quad \Lambda = 1 - z_\rho.$$

The population map satisfies $F_\rho(u) = Bu + \rho\bar{\alpha}(u)u$. The fixed-point equation and (5) imply $Bu^\star_\rho = \Lambda u^\star_\rho$ and $\ell^\top B = \Lambda\ell^\top$. All entries of B are positive because $(1-\rho)p_b\alpha^{\mathrm{D}}_a > 0$. Define

$$\Psi(u) = \frac{\mathrm{diag}(\ell)u}{\ell^\top u}, \qquad R = \Lambda^{-1}\,\mathrm{diag}(\ell)B\,\mathrm{diag}(\ell)^{-1}.$$

The matrix R has positive entries and column sums one, and $R\Psi(u^\star_\rho) = \Psi(u^\star_\rho)$. Direct substitution yields

$$\Psi(F_\rho(u)) = \frac{\Lambda}{\Lambda + \rho\bar{\alpha}(u)}R\Psi(u) + \frac{\rho\bar{\alpha}(u)}{\Lambda + \rho\bar{\alpha}(u)}\Psi(u).$$

Since R is a positive column-stochastic matrix, for probability vectors x, y,

$$\|Rx - Ry\|_1 \le \big(1 - K\min_{a,b} R_{ba}\big)\|x - y\|_1.$$

The inequality follows by subtracting $\min_{a,b} R_{ba}$ from every entry: the removed matrix annihilates $x - y$, and the remaining columns sum to $1 - K\min_{a,b} R_{ba}$. Thus (18) holds with

$$\chi = 1 - K\min_{a,b} R_{ba}\frac{\Lambda}{\Lambda + \rho\max_b \alpha^{\mathrm{D}}_b} \in [0, 1).$$

Finally, reweighting by the positive entries of ℓ and then normalizing is a bijection. Both this map and its inverse are C-Lipschitz with $C = 2\max_b \ell_b/\min_b \ell_b$. □

B Proof of Theorem 4.1

We first prove the approximation used in Section 4.2, then use its equilibrium margin to order the three methods. The last two subsections prove the interaction and fixed-group claims in Section 4.

B.1 Finite-group approximation

Lemma 4.2 (restated). *The map F_ρ has $u^\star_\rho$ as its unique fixed point, and every deterministic trajectory converges to it. For constants $C_1, C_2 < \infty$ and $0 \le \chi < 1$ depending only on the fixed parameters, every N and $t \ge 1$ satisfy*

$$\mathbb{E}\|\mu_t - u^\star_\rho\|_1 \le C_1 \chi^{t-1} + C_2\sqrt{K/N}.$$

Proof. Lemma A.3 gives the unique fixed point. Lemma A.4 gives convergence of every deterministic trajectory. For the constants and map in that lemma, Lemma A.2 yields

$$\begin{aligned}&\mathbb{E}\big[\|\Psi(\mu_{t+1}) - \Psi(u^\star_\rho)\|_1 \mid \mu_t\big]\\ &\quad\le \chi\|\Psi(\mu_t) - \Psi(u^\star_\rho)\|_1 + C\sqrt{K/N}.\end{aligned}$$

The initial transformed distance is at most two. Iterating this inequality and applying the inverse Lipschitz bound gives

$$\mathbb{E}\|\mu_t - u^\star_\rho\|_1 \le 2C\chi^{t-1} + \frac{C^2}{1-\chi}\sqrt{K/N}.$$

This proves the claim for any initial count distribution. □

B.2 Complete eventual ordering

If $d^\star_\rho \neq 0$, a final group distribution within $|d^\star_\rho|$ in ℓ_1 of $u^\star_\rho$ makes the same correctness decision: every rival remains below g when the margin is positive, and a rival remains above g when it is negative. Markov's inequality and Lemma 4.2 therefore give

$$\big|A_{\mathrm{MAD}}(N, T_0; \rho) - \mathbf{1}\{d^\star_\rho > 0\}\big| \le \frac{C_1\chi^{T_0-1} + C_2\sqrt{K/N}}{|d^\star_\rho|}. \tag{19}$$

We also need a strict comparison when MV and MAD share a limit of zero or one. Recall from Section 4.2 that $q_b = (1-\rho)\alpha^{\mathrm{D}}_{\min} p_b > 0$. Every agent has conditional probability at least q_b of returning b in the last batch. If $T_0 = 1$, the initial draw has probability $p_b \ge q_b$. Conditional independence gives

$$A_{\mathrm{MAD}}(N, T_0; \rho) \ge q_g^N, \qquad 1 - A_{\mathrm{MAD}}(N, T_0; \rho) \ge q_b^N \quad (b \neq g). \tag{20}$$

Indeed, unanimity on g ensures success and unanimity on any other answer ensures failure. In contrast, Lemma A.1 gives an MV error of at most $(K-1)e^{-d_p^2 NT_0/2}$ when $d_p > 0$ and an MV success probability of at most $e^{-d_p^2 NT_0/2}$ when $d_p < 0$. Consequently, for any fixed wrong answer b, sufficient conditions for the respective MV–MAD comparisons are

$$\begin{aligned}d_p > 0: &\quad T_0 > \frac{2}{d_p^2}\big(\log(1/q_b) + \log(K-1)\big),\\ d_p < 0: &\quad T_0 > \frac{2}{d_p^2}\log(1/q_g).\end{aligned} \tag{21}$$

For every $N \ge 2$, the respective conditions give

$$\begin{aligned}(K-1)e^{-d_p^2 NT_0/2} &< (K-1)^{1-N} q_b^N \le q_b^N,\\ e^{-d_p^2 NT_0/2} &< q_g^N.\end{aligned}$$

By (20), MV is strictly above MAD in the first case and MAD is strictly above MV in the second case.

Theorem 4.1 (restated). *Under the assumptions of Section 3, suppose $d_p \neq 0$ and $d^\star_\rho \neq 0$. Along every sequence with $N, T_0 \to \infty$,*

$$\begin{aligned}A_{\mathrm{MV}}(NT_0) &\to \mathbf{1}\{d_p > 0\},\\ A_{\mathrm{SR}}(NT_0) &\to s^{\mathrm{S}}_g,\\ A_{\mathrm{MAD}}(N, T_0; \rho) &\to \mathbf{1}\{d^\star_\rho > 0\}.\end{aligned}$$

There are finite $N_0, T_{0,\min}$ such that the strict ordering in Table 1 holds whenever $N \ge N_0$ and $T_0 \ge T_{0,\min}$.

Proof. Lemma A.1 proves the MV and SR limits; (19) proves the MAD limit along every joint growth sequence. Since $0 < s_g^{\mathrm{S}} < 1$, these bounds also give finite $N_0, T_{0,\min}$ for which any method with limit one is above SR and any method with limit zero is below SR, uniformly for all larger N, T_0.

Increase $T_{0,\min}$, if necessary, so that the relevant condition in (21) holds. If $d_p > 0$, MV also strictly exceeds MAD, giving the first two rows of Table 1. If $d_p < 0$, MAD strictly exceeds MV, giving the last two rows. The thresholds are finite because all parameters and the nonzero margins are fixed. □

B.3 Interaction and the equilibrium leader

Proposition 4.3 (restated). *The root z_ρ increases strictly with ρ. The equilibrium odds $u^\star_{\rho,g}/u^\star_{\rho,b}$ increase when $\alpha_g^{\mathrm{D}} < \alpha_b^{\mathrm{D}}$, decrease when the inequality is reversed, and remain p_g/p_b when the two redraw probabilities are equal. If g has the uniquely smallest debate redraw probability, it is the unique equilibrium leader for $\rho < 1$ sufficiently close to one.*

Proof. The right side of (5) is strictly increasing in ρ, while Φ is strictly increasing. Hence z_ρ increases strictly, and $z_\rho \uparrow \alpha^{\mathrm{D}}_{\min}$ as $\rho \uparrow 1$. By (12),

$$\frac{\mathrm{d}}{\mathrm{d}z} \log\left(\frac{p_g}{p_b}\frac{\alpha_b^{\mathrm{D}} - z}{\alpha_g^{\mathrm{D}} - z}\right) = \frac{\alpha_b^{\mathrm{D}} - \alpha_g^{\mathrm{D}}}{(\alpha_g^{\mathrm{D}} - z)(\alpha_b^{\mathrm{D}} - z)}.$$

This proves the three odds comparisons. If g uniquely minimizes α_b^{D}, then the weight $p_g/(\alpha_g^{\mathrm{D}} - z_\rho)$ diverges as $\rho \uparrow 1$, while all rival weights stay bounded. Thus $u^\star_{\rho,g} \to 1$. □

For completeness, the leadership condition in (9) follows by cross-multiplying the positive denominators in (6). Suppose $\alpha_g^{\mathrm{D}} < \alpha_b^{\mathrm{D}}$ for all $b \neq g$. A rival with $p_b \le p_g$ is always below g. For $p_b > p_g$, $u^\star_{\rho,g} > u^\star_{\rho,b}$ is equivalent to

$$z_\rho > \frac{p_b\alpha_g^{\mathrm{D}} - p_g\alpha_b^{\mathrm{D}}}{p_b - p_g}.$$

Every threshold on the right is less than α_g^{D}. Maximizing these thresholds with zero gives z_c in (13); strict leadership for $0 < \rho < 1$ is equivalent to $z_\rho > z_c$, or $\rho > \rho_c$. At $\rho = 0$, (6) gives $u^\star_{0,b} \propto p_b/\alpha_b^{\mathrm{D}}$; when $\alpha^{\mathrm{S}} = \alpha^{\mathrm{D}}$, this equals s_b^{S} from (4).

B.4 Fixed groups and finite budgets

Let $\Omega_{N,K} = \{n \in \{0,\dots,N\}^K : \sum_b n(b) = N\}$. For $n, m \in \Omega_{N,K}$, write $Q_\rho(m \mid n) = \Pr(n_{t+1} = m \mid n_t = n)$. An array $L = (L_{ab})$ records the number of agents changing from answer a to answer b. Conditional independence in (2) gives the exact count transition:

$$Q_\rho(m \mid n) = \sum_{\substack{L_{ab}\ge 0:\ \sum_b L_{ab} = n(a) \\ \sum_a L_{ab} = m(b)}} \left(\prod_a \frac{n(a)!}{\prod_b L_{ab}!}\right) \prod_{a,b} \left[P_\rho^{\mathrm{D}}(b \mid a, n/N)\right]^{L_{ab}}. \tag{22}$$

Here the L_{ab} are nonnegative integers. The initial count law is

$$\nu_1(n) = \frac{N!}{\prod_b n(b)!} \prod_b p_b^{n(b)}.$$

Every conditional agent update contains a common baseline draw with weight at least $(1-\rho)\alpha^{\mathrm{D}}_{\min}$. Consequently,

$$Q_\rho(m \mid n) \ge \left[(1-\rho)\alpha^{\mathrm{D}}_{\min}\right]^N \nu_1(m) > 0.$$

Subtracting this common component from each transition row contracts total variation by a factor at most $1 - [(1-\rho)\alpha^{\mathrm{D}}_{\min}]^N$. The finite count chain therefore has a unique stationary law $\Pi_{N,\rho}$ with

positive mass on every count state, and its law converges to $\Pi_{N,\rho}$ as $T_0 \to \infty$. In particular, the fixed-group accuracy in Section 4.3 is

$$a_{N,\rho} = \sum_{n \in \Omega_{N,K}} \Pi_{N,\rho}(n) v_g(n) \in (0,1). \tag{23}$$

Positivity follows by considering all-correct and all-incorrect count states. Lemma A.1 gives $A_{\mathrm{SR}}(NT_0) \to s_g^{\mathrm{S}}$ and the corresponding zero-or-one limit for MV. Because both limits in (23) and for SR are interior, MV eventually exceeds both when $d_p > 0$ and falls below both when $d_p < 0$. When $a_{N,\rho} \neq s_g^{\mathrm{S}}$, convergence orders MAD and SR by the sign of their difference.

The same count transition gives finite-budget accuracies without an asymptotic approximation. Let P^{S} also denote the matrix whose row a and column b contain $P^{\mathrm{S}}(b \mid a)$. With the current count in the row index of Q_ρ and the next count in its column index,

$$\begin{aligned} A_{\mathrm{MAD}}(N, T_0; \rho) &= \sum_{n,m \in \Omega_{N,K}} \nu_1(n) \, [Q_\rho^{T_0 - 1}](m \mid n) \, v_g(m), \\ A_{\mathrm{SR}}(T) &= \sum_{a \in [K]} p_a \, [(P^{\mathrm{S}})^{T-1}](g \mid a). \end{aligned} \tag{24}$$

Here P_0^{D} likewise denotes the row-stochastic matrix from (2) at $\rho = 0$. For $\rho = 0$, the N debate agents are independent refinement chains: their final counts have a multinomial distribution with N draws from $p^\top (P_0^{\mathrm{D}})^{T_0 - 1}$. The MV formula is the multinomial average in (3).

C Proofs for Gated Debate

All results in this section use $K = 2$ and the matched comparison in Definition 5.1. Let b denote the answer other than g. The weights may be selected using the past, but are fixed before the next updates.

C.1 Proof of Theorem 5.3

Theorem 5.3 (restated). *Assume $K = 2$. Consider the oracle gate and any policy π in the matched policy family of Definition 5.1, with the same initial count distribution. Let $n_t^\star(g)$ and $n_t^\pi(g)$ denote their respective correct-answer counts at round t. For every finite t and $k \in \{0, \dots, N\}$,*

$$\Pr(n_t^\star(g) \geq k) \geq \Pr(n_t^\pi(g) \geq k). \tag{25}$$

Consequently, the oracle's final binary plurality accuracy is at least that of always debate, matched no-group, and every other policy in this family.

Proof. Suppose the group has $n_t(g) = m$ and uses weight r. By (2), the probabilities that an agent returns g are

$$\begin{aligned} P_r^{\mathrm{D}}(g \mid g, n_t/N) &= 1 - \alpha_g^{\mathrm{D}} + \alpha_g^{\mathrm{D}}[p_g + r(m/N - p_g)], \\ P_r^{\mathrm{D}}(g \mid b, n_t/N) &= \alpha_b^{\mathrm{D}}[p_g + r(m/N - p_g)]. \end{aligned} \tag{26}$$

Both increase with the probability that a redraw returns g. For any redraw probability $z \in [0,1]$, an agent already holding g is at least as likely to return it as an agent holding b:

$$(1 - \alpha_g^{\mathrm{D}} + \alpha_g^{\mathrm{D}} z) - \alpha_b^{\mathrm{D}} z = (1 - \alpha_g^{\mathrm{D}})(1 - z) + (1 - \alpha_b^{\mathrm{D}}) z \geq 0. \tag{27}$$

Thus, if one group has at least as many correct answers as another and its redraw is at least as likely to return g, pair agents so that every correct agent in the second group is paired with a correct agent in the first. Using the same independent uniform number for both agents in each pair makes the first group's next correct count at least the second's, almost surely.

At count m, the oracle's probability of redrawing g is

$$p_g + \bar{\rho} \max\{m/N - p_g, 0\}.$$

It is nondecreasing in m and at least $p_g + r(m/N - p_g)$ for every $r \in [0, \bar{\rho}]$ at the same count. Couple the initial counts to be equal. If the oracle has at least as many correct answers after batch

t, the preceding coupling orders the counts after batch $t+1$, even when the other policy selected r using the whole history and independent auxiliary randomness. Induction gives the stated upper-tail inequalities. The binary plurality reward $v_g(n)$ is nondecreasing in $n(g)$, so the same order holds for final accuracy. □

D EXPERIMENT SETUP

We evaluate majority voting (MV), self-refinement (SR), always-debate (MAD), matched no-group debate, and gated debate with fixed model calls; among them, the matched no-group debate is a control experiment that omits peer answers; gated debate is the proposed revision of always-debate. **Models and benchmarks.** Our experiments cover Qwen3.5-4B, Qwen3.5-9B, Qwen3.6-35B-A3B (Team, 2026), Kimi-K2.6 (Team et al., 2026), GLM-5.3 [1], and GPT-5.4-mini [2], across MMLU-Pro (Yue et al., 2025), MMMU (Yue et al., 2024), MedQA [3], GPQA-Diamond (Rein et al., 2023), HellaSwag (Zellers et al., 2019), and SuperGPQA (Du et al., 2026). MMMU includes image inputs.

E FITTING OF PARAMETERS

We use the experimental results to fit the following parameters. The parameters are then used to predict the final performance of the model.

p_b	Probability of answer b in independent calls,
α_a^S	Probability of redraw in SR when the current answer is a,
ρ	Weight of the peer-answer distribution when redrawing in MAD,
α_a^D	Redraw probability in MAD when the current answer is a,
τ	Threshold determining whether the gate is open or closed.

We estimate p from repeated independent calls, α^S from no-peer revisions where the model sees its previous answer and is asked to reconsider, and (α^D, ρ) from revisions that additionally include peer answers. The redraw parameters are fitted by maximum likelihood from the full next-answer transitions, since a redraw may return the same answer. The threshold τ is selected on the development set, and all parameters are fixed before testing.

E.1 SUPPLEMENTARY WITHIN-QUESTION FITTING

We also fit question-specific redraw probabilities using the early part of each saved trajectory. This is a separate, retrospective analysis: it does not replace the frozen parameters used for the accuracy predictions and ordering figures in the main text. The independent-answer distribution $\widehat{p}_j$ for question j, including its previously selected additive smoothing, is kept fixed. No additional model calls are made, and the gate threshold is not refitted.

Fitting and evaluation split. We process each model–benchmark run separately and fit MAD and SR as separate families. For each target question, updates 1–3 are used for fitting; initialization is not an update. With five MAD agents, this gives 15 fitting transitions per question, followed by 20 reserved transitions from updates 4–7. For single-chain SR, the first three revisions are fitted and revisions 4–39 are reserved. All agents in a MAD round remain on the same side of the split. The separate controlled-calibration records are pooled across their original micro-train and micro-dev labels, then divided within each question and previous-answer label into approximately 80% fitting and 20% reserved records, using seed 42. A label with only one trial remains in fitting. We use the retained valid-response calibration export; previously discarded invalid calls are not reconstructed.

Shared fit. For a transition r from answer a to answer b, the MAD likelihood is

$$K^{\mathrm{D}}_{j,r}(b \mid a) = (1-\alpha^{\mathrm{D}}_{j,a})\mathbf{1}\{b=a\} + \alpha^{\mathrm{D}}_{j,a}\big[(1-\rho)\widehat{p}_j(b) + \rho\mu_{j,r}(b)\big],$$

[1] https://z.ai/blog/glm-5.3
[2] https://openai.com/index/introducing-gpt-5-4-mini-and-nano/
[3] https://ukgovernmentbeis.github.io/inspect_evals/evals/medqa/index.html

where $\mu_{j,r}$ is the recorded distribution of all five answers immediately before the update. SR uses the same form with $\rho = 0$ and its own redraw probabilities. For each family $F \in \{\mathrm{D}, \mathrm{S}\}$, we first fit a shared center

$$\overline{\alpha}^F_{j,a} = \mathrm{clip}\big(\sigma\big(\beta^F_0 + \beta^F_1[-\log \widehat{p}_j(a)]\big), 0.001, 1 - 10^{-6}\big), \qquad \sigma(x) = (1 + e^{-x})^{-1}.$$

We minimize the pooled next-answer negative log likelihood over the new fitting records only. MAD uses the controlled-calibration fitting records and target prefixes; SR uses its target prefixes. The coefficients are shared within a run and family, with $\beta_0 \in [-15, 15]$, $\beta_1 \in [-10, 10]$, and MAD $\rho \in [0, 0.999]$. We use bounded L-BFGS-B from four initializations and retain the converged solution with the lowest objective. Refitting the center avoids reusing parameters that may have seen responses now reserved for evaluation.

Local fit. We then hold the shared center and pooled ρ fixed. For each question, family, and previous answer a, let $\mathcal{T}^F_{j,a}$ contain its fitting transitions and write $\alpha^F_{j,a} = \sigma(z)$. We minimize

$$-\sum_{r \in \mathcal{T}^F_{j,a}} \log K^F_{j,r}(b_r \mid a) + \lambda\left[z - \mathrm{logit}(\overline{\alpha}^F_{j,a})\right]^2, \qquad z \in [\mathrm{logit}(0.001), \mathrm{logit}(1 - 10^{-6})].$$

The one-dimensional search evaluates a 33-point grid, refines candidate minima by bounded scalar optimization, and also checks both boundaries and the shared center. We report all three penalties $\lambda \in \{0.1, 1, 10\}$; none is chosen using gap signs or reserved-response scores. If $\mathcal{T}^F_{j,a}$ is empty, the answer retains its shared-center estimate and is marked as having no local support. Thus a complete parameter vector need not mean that every answer-specific parameter was identified from local observations.

Gap calculation and reserved-response evaluation. After fitting, we compute $\widehat{d}_p = \widehat{p}_j(g) - \max_{b \neq g} \widehat{p}_j(b)$ and $\widehat{d}_\star = \widehat{u}^\star_j(g) - \max_{b \neq g} \widehat{u}^\star_j(b)$, where $\widehat{u}^\star_j$ is the MAD mean-field equilibrium under the fitted local redraw probabilities and pooled ρ. The correct label g is used only for these subsequent correctness-based calculations, not to fit parameters or assign the new splits. Reserved responses are scored by next-answer negative log likelihood and multiclass Brier score, conditional on their recorded previous states. These are one-step continuation diagnostics, not newly computed final-vote accuracy predictions or an unseen-question evaluation.

F Full benchmark results at 10, 20, and 40 calls

Table 3: Qwen3.5-4B: final-answer accuracy (%) at matched call budgets.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (6)
Test questions (n)		1000	1000	1000	64	1000	1000	
Independent voting								
MV	10	49.23	**53.30**	**70.82**	**53.65**	**88.65**	27.93	**57.26**
MV	20	49.48	**53.96**	**70.65**	47.66	**89.10**	27.73	**56.43**
MV	40	49.03	**54.80**	**70.70**	45.31	88.35	27.68	55.98
Single-chain refinement								
SR	10	**51.10**	49.50	66.10	39.06	88.60	**30.30**	54.11
SR	20	49.20	50.00	66.30	35.94	88.20	**29.40**	53.17
SR	40	49.30	51.20	67.70	**50.00**	88.20	29.60	**56.00**
Group protocols (5 agents)								
Always debate	10	48.79	52.65	68.30	41.41	88.35	28.34	54.64
Always debate	20	**50.70**	52.70	65.65	46.09	88.05	28.85	55.34
Always debate	40	50.00	52.60	66.70	43.75	88.00	28.60	54.94
Matched no-group	10	49.72	52.80	68.95	46.09	88.20	28.32	55.68
Matched no-group	20	50.59	52.20	69.85	46.09	88.30	27.88	55.82
Matched no-group	40	49.81	51.70	69.50	44.53	**89.15**	28.67	55.56
Gated debate	10	49.74	52.85	67.10	43.75	87.80	28.02	54.88
Gated debate	20	50.07	52.10	67.55	**49.22**	88.05	28.47	55.91
Gated debate	40	**50.33**	52.60	66.20	42.19	88.40	**30.46**	55.03

Table 4: Qwen3.5-9B: final-answer accuracy (%) at matched call budgets.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (6)
Test questions (n)		1000	1000	1000	64	1000	1000	
Independent voting								
MV	10	**55.78**	**58.63**	**76.20**	50.00	92.00	**34.30**	**61.15**
MV	20	**55.50**	**58.36**	**76.28**	**53.91**	**92.05**	**34.50**	**61.76**
MV	40	55.68	**60.30**	**76.10**	53.12	**92.05**	**34.22**	**61.91**
Single-chain refinement								
SR	10	54.30	55.30	73.60	**56.25**	90.10	33.10	60.44
SR	20	54.90	55.20	74.10	45.31	90.00	31.00	58.42
SR	40	52.60	53.60	73.40	43.75	90.20	32.20	57.63
Group protocols (5 agents)								
Always debate	10	55.50	56.60	74.45	47.66	**92.25**	33.09	59.92
Always debate	20	54.60	58.05	74.60	47.66	91.65	32.10	59.78
Always debate	40	55.75	57.70	74.05	51.56	90.70	33.05	60.47
Matched no-group	10	54.51	56.80	73.95	46.88	90.50	34.09	59.45
Matched no-group	20	55.08	57.30	75.60	52.34	90.30	31.26	60.31
Matched no-group	40	**55.91**	55.95	74.65	46.09	90.65	32.80	59.34
Gated debate	10	54.43	57.45	73.60	49.22	91.30	32.40	59.73
Gated debate	20	54.46	57.90	74.45	48.44	90.95	33.28	59.91
Gated debate	40	55.67	57.95	75.60	**54.69**	91.10	31.81	61.14

Table 5: Qwen3.6-35B-A3B: final-answer accuracy (%) at matched call budgets.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (6)
Test questions (n)		1000	1000	1000	64	1000	1000	
Independent voting								
MV	10	**63.42**	61.17	**85.50**	57.81	**95.75**	**41.73**	67.56
MV	20	63.40	61.40	85.75	52.86	**95.80**	**41.80**	66.84
MV	40	62.93	62.00	85.85	53.91	**96.20**	41.73	67.10
Single-chain refinement								
SR	10	59.60	61.00	84.00	46.88	94.40	38.30	64.03
SR	20	60.20	59.20	84.90	59.38	93.30	40.40	66.23
SR	40	62.00	57.60	82.90	**64.06**	94.00	41.00	66.93
Group protocols (5 agents)								
Always debate	10	63.24	61.10	84.60	**60.94**	95.45	41.30	**67.77**
Always debate	20	63.05	**62.05**	83.85	61.72	95.60	41.55	67.97
Always debate	40	63.55	61.55	**86.05**	59.38	95.25	**42.35**	68.02
Matched no-group	10	62.76	61.45	85.25	53.12	94.80	40.54	66.32
Matched no-group	20	63.12	61.50	85.10	55.47	95.25	40.70	66.86
Matched no-group	40	63.17	60.95	85.60	53.12	95.20	42.34	66.73
Gated debate	10	63.30	**61.80**	85.40	57.03	95.30	40.80	67.27
Gated debate	20	**63.89**	61.95	**85.85**	**62.50**	95.45	41.32	**68.49**
Gated debate	40	**63.86**	**62.10**	85.75	60.94	95.65	40.95	**68.21**

Table 6: Kimi-K2.6: final-answer accuracy (%) at matched call budgets.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (6)
Test questions (n)		1000	1000	1000	64	1000	1000	
Independent voting								
MV	10	74.15	**72.83**	86.65	**57.81**	96.30	53.57	**73.55**
MV	20	75.15	**72.70**	86.90	**54.69**	96.35	**53.87**	**73.28**
MV	40	74.95	**73.15**	**87.15**	53.12	96.20	**54.35**	**73.15**
Single-chain refinement								
SR	10	75.00	71.10	86.10	45.31	95.50	52.20	70.87
SR	20	74.60	70.80	85.40	45.31	95.50	52.40	70.67
SR	40	74.50	70.80	86.20	45.31	95.90	52.10	70.80
Group protocols (5 agents)								
Always debate	10	**75.70**	71.90	86.80	54.69	96.20	53.80	73.18
Always debate	20	75.50	71.20	86.10	**54.69**	96.20	53.05	72.79
Always debate	40	**76.20**	71.10	85.90	**54.69**	96.30	53.20	72.90
Matched no-group	10	75.25	71.65	86.30	52.34	96.05	53.29	72.48
Matched no-group	20	74.41	71.75	86.80	50.00	96.30	53.79	72.17
Matched no-group	40	74.90	72.10	86.75	53.91	96.35	53.69	72.95
Gated debate	10	75.50	72.40	**87.55**	53.12	**96.45**	**54.55**	73.26
Gated debate	20	**75.90**	72.60	**87.10**	50.78	**96.65**	53.71	72.79
Gated debate	40	75.30	71.90	86.55	53.91	**96.60**	53.65	72.98

Table 7: GLM-5.3 (low reasoning): final-answer accuracy (%) at matched call budgets.

Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (5)
Test questions (n)		1000	—	1000	59	1000	1000	
Independent voting								
MV	10	82.85	—	91.10	77.97	93.75	62.30	81.59
MV	20	82.75	—	91.10	80.51	94.10	62.75	82.24
MV	40	83.05	—	90.90	78.81	94.45	61.90	81.82
Single-chain refinement								
SR	10	83.10	—	90.60	72.88	94.00	62.30	80.58
SR	20	83.00	—	91.30	77.97	94.30	64.10	82.13
SR	40	81.70	—	90.50	76.27	93.90	63.10	81.09
Group protocols (5 agents)								
Always debate	10	84.00	—	**91.65**	84.75	**94.30**	63.75	83.69
Always debate	20	**84.55**	—	**91.70**	**83.05**	94.40	64.40	**83.62**
Always debate	40	**85.00**	—	**91.80**	**86.44**	94.40	64.25	**84.38**
Matched no-group	10	82.50	—	90.90	**88.14**	94.20	64.67	**84.08**
Matched no-group	20	83.42	—	91.00	76.27	**94.50**	64.45	81.93
Matched no-group	40	83.52	—	91.05	79.66	93.90	**65.67**	82.76
Gated debate	10	**84.25**	—	90.95	84.75	**94.30**	**64.82**	83.81
Gated debate	20	84.30	—	91.35	**83.05**	94.35	**64.87**	83.58
Gated debate	40	84.55	—	91.55	84.75	**94.60**	65.42	84.17

Table 8: GPT-5.4-mini: final-answer accuracy (%) at matched call budgets.

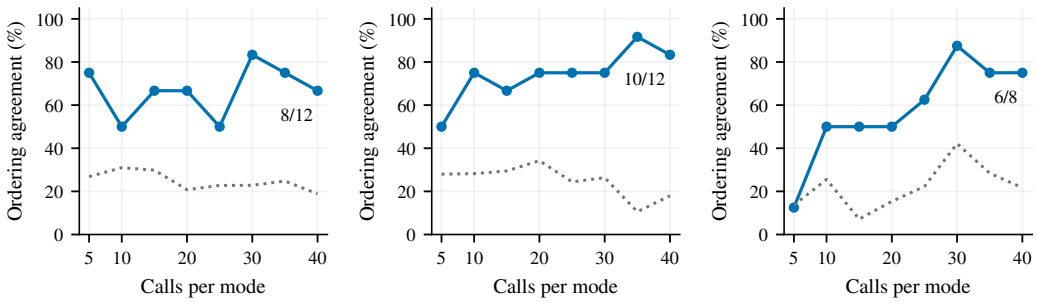
Method	Calls	MMLU-Pro	MMMU	MedQA	GPQA-Diamond	HellaSwag	SuperGPQA	Average (5)
Test questions (n)		931	—	939	64	940	918	
Independent voting								
MV	10	84.00	—	94.57	**85.94**	90.96	62.96	**83.68**
MV	20	84.03	—	**94.73**	85.16	**91.60**	63.34	83.77
MV	40	84.43	—	**94.83**	85.94	91.49	63.24	83.98
Single-chain refinement								
SR	10	82.38	—	94.14	76.56	90.11	63.07	81.25
SR	20	82.28	—	94.68	84.38	90.53	61.44	82.66
SR	40	83.14	—	94.25	84.38	90.64	61.11	82.70
Group protocols (5 agents)								
Always debate	10	84.72	—	94.41	82.81	**91.28**	63.13	83.27
Always debate	20	84.59	—	94.14	84.38	91.28	64.05	83.69
Always debate	40	84.64	—	94.14	85.16	90.96	63.45	83.67
Matched no-group	10	**84.96**	—	**94.94**	80.47	91.06	**63.89**	83.07
Matched no-group	20	84.85	—	**94.73**	82.81	**91.60**	**64.49**	83.70
Matched no-group	40	**85.45**	—	94.20	85.94	**91.81**	**63.64**	84.21
Gated debate	10	84.75	—	94.73	82.03	91.17	62.85	83.11
Gated debate	20	**84.96**	—	94.20	**85.94**	91.06	63.29	**83.89**
Gated debate	40	83.67	—	94.52	**89.06**	91.17	63.31	**84.35**

G PROMPTS AND ALGORITHM USED IN THE EXPERIMENTS

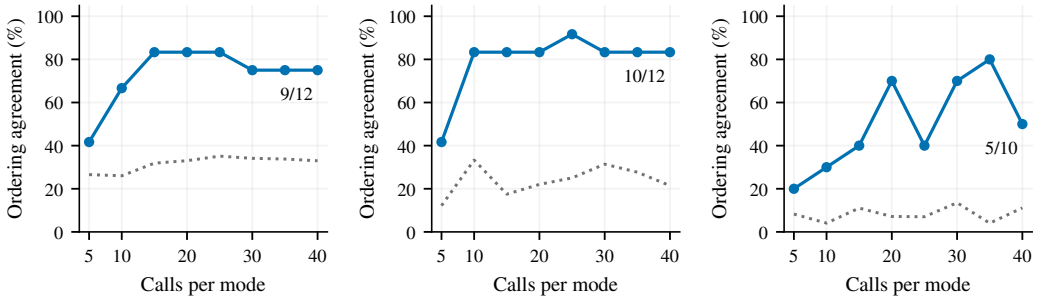
The experiments use three user-message templates, with a shared system message and output-format instruction. Braced fields below are placeholders, not literal text sent to the model. `{question}` is the benchmark question; `{choices}` lists the available options, one per line as `A. option text`, in label order. `{legal replies}` lists exactly the permitted responses for that question, one per line as `FINAL: A`, `FINAL: B`, and so on. Thus the number of permitted labels follows the question rather than a fixed four-option format. `{answer format}` is replaced verbatim by the common output-format block below.

Table 9: Prompt selection for calibration and test calls. Gate decisions are made by the experiment controller, not requested from the LLM.

Call type	User-message template
Independent-distribution estimation; all MV calls	Independent answer
Initial answers for group methods and SR	Independent answer
Micro-train and micro-dev revisions	Peer-informed revision
Always-debate updates	Peer-informed revision
Matched no-group updates; SR revisions	Independent revision
Gated-debate updates	Peer-informed revision if open; independent revision if closed
Random-gate updates ($30\%, 50\%, 70\%$)	Same open/closed templates as gated debate



(a) **Qwen3.5-4B** (b) **Kimi-K2.6** (c) **GPT-5.4-mini**



(d) **Qwen3.5-9B** (e) **Qwen3.6-35B** (f) **GLM-5.3 (low)**

Figure 5: Ordering agreement across call budgets for all models.

G.1 Message construction

Each call is a fresh system/user exchange, without a preceding conversation transcript. Revision prompts supply only the agent's previous answer letter and, when applicable, four peer answer letters; they do not supply peer explanations or the correct answer. Test-time group updates are synchronous: all five agents receive labels from the preceding round, with a seeded permutation of the four peers' display order. In micro-train and micro-dev, the same peer-informed template is applied to controlled four-peer configurations rather than a live debate trajectory. Lower-budget results reuse saved calls and require no additional prompt variants.

For MMMU, references such as `<image 1>` in the question or choices are replaced at their positions by the corresponding image content in the user message. Text-only questions use the same templates without image parts. Model-specific chat renderers or API message formats encode these messages. For non-thinking Tinker calls, generation is additionally prefilled with `FINAL:`; the prefix is retained in the returned answer. The instruction to omit explanations constrains the requested output and does not by itself disable a provider's internal reasoning mode.

G.2 Ordering Results

We list the figures of the ordering results of six models in Figure 5.

G.3 Gated debate

We implement the practical baseline-mode gate from Definition 5.4 as follows.

Algorithm: baseline-mode gated revision.

Input: $p, \alpha^{\mathrm{D}}, \bar{\rho}, \tau, N, T_0$ and initial answers $X_{1,1}, \ldots, X_{N,1}$, sampled independently from p unless an initial state is supplied. Updates use Equation (2).

1. Set $\widehat{g}$ to a deterministically tie-broken mode of p.
2. For $t = 1, \ldots, T_0 - 1$:
   (a) Form
   $$n_t(b) = \sum_{i=1}^{N} \mathbf{1}\{X_{i,t} = b\}, \qquad \mu_{t,b} = \frac{n_t(b)}{N}, \qquad b \in [K].$$
   (b) Set
   $$s_t = \mu_{t,\widehat{g}} - p_{\widehat{g}}, \qquad G_t = \mathbf{1}\{s_t > \tau\}, \qquad \rho_t^{\mathrm{gate}} = \bar{\rho} G_t.$$
   (c) Independently for each agent, conditional on the current state, draw
   $$X_{i,t+1} \sim P^{\mathrm{D}}_{\rho_t^{\mathrm{gate}}}(\cdot \mid X_{i,t}, \mu_t), \qquad i = 1, \ldots, N.$$
3. Form n_{T_0} from $X_{1,T_0}, \ldots, X_{N,T_0}$ and select uniformly among the plurality winners of n_{T_0}.

Output: the selected label. Every update uses N calls, including when $G_t = 0$; the total budget $T = NT_0$ includes the N initialization calls.

The procedure is defined for any K. A calibrated implementation replaces p in the reference and score by a frozen estimate $\widehat{p}$, fixing $\bar{\rho}$ and τ before evaluation. In a prompt-based implementation, G_t controls whether peer information is supplied. The dominance results compare policies that share the baseline and redraw probabilities and differ only in the interaction weight; prompt-based implementations are evaluated empirically.