ConFreeze: Selective Multi-Model Debate through Consensus Freezing

Anonymous ACL submission

Abstract

Multi-model debate enhances large language model reasoning but suffers from prohibitive computational costs and instability risks, where excessive deliberation can overturn correct initial consensus. However, existing research has primarily focused on performance gains, while the efficiency and stability implications of iterative debate remain underexplored. To address these limitations, we formulate Multi-Model Debate as a decision control problem and propose *ConFreeze*, a selective execution mechanism that uses initial vote patterns as a gating signal. When models unanimously agree in the initial round, we freeze the consensus to avoid computational waste and instability risk. When models disagree, we trigger a subsequent round of collaborative refinement where models critique and revise predictions. This allocates debate budget where reasoning conflicts signal improvement potential. To better reflect robustness and comprehensively evaluate debate dynamics, we evaluate not only end-task quality but also stability measures(flip rate, improve/worsen rate) together with token cost. Experiments on ANLI, AdvGLUE, and TruthfulQA demonstrate that *ConFreeze* achieves 29.5%-43.1% token reduction while maintaining accuracy.

Our findings reveal that debate benefits are concentrated almost exclusively in disputed instances, validating initial consensus as a reliable signal for efficient inference control. Code and prompts are available at an anonymized repository anonymous.4open.science/ConFreeze.

1 Introduction

Large language models have demonstrated impressive capabilities, yet single-model reasoning remains brittle, often suffering from hallucinations or incomplete evidence integration (Rawte et al., 2023) (Sahoo et al., 2024). To address this, Multi-Model Debate has been proposed as a mechanism to improve reliability by coordinating multiple model instances to critique and refine predictions (Choi et al., 2025). In this approach, multiple model instances propose answers, critique one another, and optionally revise their predictions (Du et al., 2024; Li et al., 2024; Liang et al., 2023).

While effective (Wei et al., 2022; Wang et al., 2023), current debate frameworks face two critical challenges. First, the computational cost is prohibitive, as token consumption grows multiplicatively with the number of agents and rounds (Smit et al., 2024). Second, and often overlooked, is the risk of instability: excessive deliberation can overturn initially correct answers through persuasive but flawed critiques, a phenomenon we term harmful flips (Smit et al., 2024; Zhang et al., 2025). Wynn et al. (Wynn et al., 2025) find that debate is not uniformly beneficial and can sometimes degrade performance due to over-deliberation and persuasive but incorrect arguments. Amayuelas et al. (Amayuelas et al., 2024) show that persuasive ability can disproportionately influence other agents in debate, including under adversarial settings.

We observe a fundamental asymmetry in consensus dynamics: unanimous agreement typically signals a stable, high-confidence decision, whereas disagreement signals genuine uncertainty. For unanimous cases, additional debate often yields diminishing returns while increasing the risk of false consensus . Conversely, when models disagree, the minority viewpoint may surface overlooked evidence, making collaborative refinement highly beneficial. This suggests that debate resources should be allocated selectively—invested where conflict signals potential improvement, and withheld where early consensus indicates stability .

To operationalize this principle, we formulate multi-model debate as a decision control problem and propose *ConFreeze* as a selective execution mechanism that uses initial voting patterns as a

gating signal. The underlying logic dictates that if models reach unanimous agreement in the initial round, we freeze the consensus to avoid computational waste and instability risks. Conversely, we trigger a subsequent round of collaborative refinement if the models disagree. This design is training-free as it requires no hyperparameter tuning and depends only on the previous round by relying solely on observable vote distributions. Furthermore, the framework is composable by acting as a drop-in wrapper for existing pipelines. To better characterize robustness beyond final-task scores, we evaluate debate with both quality and stability objectives. In addition to standard end-task metrics (accuracy and macro-F1), we report stability metrics: flip rate (whether the final label changes from the previous round to the subsequent round), and improved/worsened rates (whether a change fixes an error or introduces a new one), which more directly quantify harmful vs. beneficial revisions. To quantify deployability, we log per-instance token usage and report token savings under gating. To clarify the trade-off relationship between cost and quality, we record per-instance token usage together with debate histories, and evaluate how much computation can be saved under selective subsequent round execution. Our contributions are as follows:

- We propose *ConFreeze*, a training-free gating mechanism that uses initial vote patterns to selectively execute debate, addressing the dual challenges of prohibitive token costs and prediction instability while enabling drop-in deployment over existing Multi-Model Debate frameworks.
- We introduce fine-grained stability metrics (flip rate, improve rate, and worsen rate) that decompose prediction changes into beneficial and harmful components, enabling explicit measurement of cost-accuracy-stability trade-offs and providing empirical validation that these measures effectively discriminate when debate helps versus harms.
- We conduct a comprehensive evaluation across ANLI, AdvGLUE, and TruthfulQA, demonstrating approximately 29.5%-43.1% token reduction while maintaining accuracy parity against always-debate baselines. Through systematic ablations over team composition, temperature, and prompting, we establish that consensus-based gating generalizes across configurations and that debate benefits concentrate almost entirely on initially disputed instances, validating initial consensus as a reliable gating signal.

2 Methodology

2.1 Overview

Figure 1 shows the overview of *ConFreeze*, consisting of three stages: initial round prediction and consensus detection, consensus-freeze gating, and selective refinement and aggregation. We begin by having each model independently examine the input and produce a prediction with a brief rationale ①. We then aggregate these predictions via majority voting to obtain the initial consensus ②, and compute a unanimity flag indicating whether all models agree ③. For consensus-freeze gating, we make a gating decision based on the unanimity flag ④: if unanimous, we freeze the prediction and skip subsequent rounds; otherwise, we trigger collaborative refinement. Instances are partitioned into two sets ⑤: the frozen set (unanimous) directly outputs initial predictions as final answers, while the eligible set (disagreement) proceeds to further deliberation. For selective refinement, we trigger a subsequent round for the eligible set ⑥, where each model reconsiders its prediction given peer rationales from the initial round. The frozen set skips this round entirely. Finally, we aggregate predictions ⑦: frozen instances use initial majority vote, while eligible instances use majority voting over subsequent round predictions. We now describe each stage in more detail.

2.2 Stage 1: Initial Round Prediction and Consensus Detection

The first stage of *ConFreeze* collects independent predictions from each model and detects consensus patterns to inform the subsequent gating decision. This stage consists of three steps: 1) each model independently generates a prediction with rationale; 2) predictions are aggregated via majority voting; 3) a unanimity flag is computed to identify instances with full agreement. These signals enable *ConFreeze* to distinguish stable unanimous instances from uncertain disagreement cases in the next stage.

Independent Prediction Generation. *ConFreeze* prompts each model A_k (where $k \in \{1, \ldots, K\}$) to independently examine

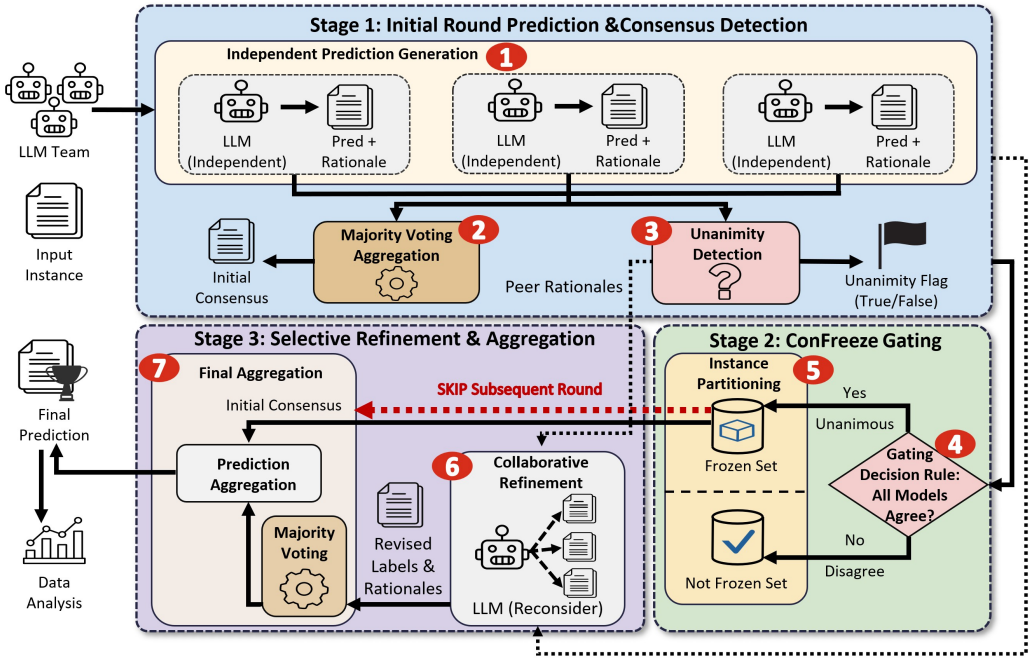

Figure 1: Overview of *ConFreeze*.

instance x_i and produce both a categorical prediction $\hat{y}_{i,k}^{(1)} \in \mathcal{Y}$ and a brief rationale explaining its reasoning. The input prompt contains only the task instruction and instance content (e.g., premise-hypothesis pair for NLI, or question-choices for multiple-choice QA), with no visibility into other models' reasoning. This ensures that initial round outputs reflect genuine model beliefs before any coordination takes place. The independent reasoning format provides both the necessary label predictions for voting and the rationales for subsequent critique.

Majority Voting Aggregation. After obtaining independent predictions from all K models, *ConFreeze* aggregates these votes into an initial round consensus using *equal-weight majority voting*. For each label $\ell \in \mathcal{Y}$, we compute the vote share: $p_{i,\ell} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[\hat{y}_{i,k}^{(1)} = \ell]$, $\sum_\ell p_{i,\ell} = 1$ The initial round majority label is then $\hat{y}_{\text{maj},i}^{(1)} = \arg\max_\ell p_{i,\ell}$. With an odd number of models (we use $K = 3$ throughout), ties are avoided by construction. Majority voting provides a more robust baseline than selecting a single model's output while remaining computationally lightweight.

We also compute vote margin and entropy (Shannon, 1948) (Settles, 2009) from the initial-round vote distribution to characterize agreement strength for later offline analysis and threshold sweeps(metrics defined in Section 3.3).

Unanimity Detection. Leveraging the initial round majority label and vote distribution, we detect consensus patterns for gating: unanimous agreement indicates stable, high-confidence instances with minimal debate benefit and potential instability risk. *ConFreeze* thus computes a lightweight, ground-truth-free unanimity flag: $u_i = \mathbf{1}[\max_\ell p_{i,\ell} = 1]$, which identifies instances where all K models produce identical predictions ($p_{i,\ell} = 1$ for some ℓ). Instances with $u_i = 1$ are frozen, while $u_i = 0$ proceed to collaborative refinement.

2.3 Stage 2: Consensus-Freeze Gating

This stage determines which instances undergo collaborative refinement, using a lightweight, training-free gating mechanism that allocates computation only to initially uncertain cases.

Gating Decision Rule. We formalize gating as a binary decision $g_i \in \{0, 1\}$ computed after the initial round: $g_i = 1$ triggers the subsequent round, while $g_i = 0$ freezes the initial round decision. Our consensus-freeze policy uses a simple unanimity-based rule: $g_i = 0$ if $u_i = 1$ (all models agree), and $g_i = 1$ if $u_i = 0$ (models disagree). Equivalently, $g_i = \mathbf{1}[u_i = 0]$. The core principle is: unanimous agreement signals stability; disagreement signals genuine uncertainty.

When $g_i = 0$ (freeze), we lock the initial round majority label $\hat{y}_{\text{maj},i}^{(1)}$ as the final prediction and skip the subsequent round entirely. When $g_i = 1$ (pro-

Table 1: Evaluation datasets.

Dataset	Split	Size	Label Type	Task Description
ANLI-R3	dev	1200	3-way NLI	Adversarially constructed natural language inference
TruthfulQA-MC1	val	684	Multi-choice	Factuality testing under misleading prompts
AdvGLUE-QQP	test	142	Binary	Adversarial paraphrase detection
AdvGLUE-SST-2	test	262	Binary	Adversarial sentiment analysis
AdvGLUE-QNLI	test	266	Binary	Adversarial QA-based entailment
AdvGLUE-MNLI	test	242	3-way NLI	Adversarial natural language inference

ceed), we trigger collaborative refinement where each model reconsiders its prediction in light of peer rationales. This rule is training-free, requires no ground truth, and can be applied as a drop-in wrapper to any debate pipeline.

Instance Partitioning and Execution Paths. Based on the gating decisions, we partition the dataset $\mathcal{N}$ into two disjoint subsets: $\mathcal{U} = \{x_i \in \mathcal{N} : u_i = 1\}$ (unanimous) and $\mathcal{D} = \{x_i \in \mathcal{N} : u_i = 0\}$ (disagreement). Instances in $\mathcal{U}$ directly use the initial round majority label $\hat{y}^{(1)}_{\text{maj},i}$ as the final prediction, while instances in $\mathcal{D}$ proceed to collaborative refinement in the subsequent round.

This partitioning determines the coverage (fraction of instances proceeding to refinement), which directly impacts token savings. This instance-level routing allocates debate budget precisely where initial disagreement signals genuine uncertainty, while protecting stable unanimous decisions from unnecessary perturbation.

In deployment, frozen instances physically skip the subsequent round to save cost. For analysis purposes, we may emulate subsequent rounds on frozen instances to study counterfactuals while maintaining the same final predictions.

2.4 Stage 3: Selective Refinement and Aggregation.

After partitioning instances into frozen and eligible sets, *ConFreeze* enters the final stage where eligible instances undergo collaborative refinement while frozen instances remain locked. This stage consists of two steps: 1) prompt models to reconsider their predictions in light of peer feedback; 2) aggregate revised predictions to produce final outputs.

Collaborative Refinement. For each instance $i \in \mathcal{D}$ (the disagreement, eligible set), *ConFreeze* triggers a subsequent round in which each model A_k reconsiders its initial prediction after observing the peer rationales from the initial round. Each model now receives additional context: the original input x_i together with the rationales produced by all K models in the initial round. The model is then prompted to reconsider, producing a revised prediction $\hat{y}^{(2)}_{i,k}$. This design encourages explicit cross-checking: models can challenge unsupported claims, reconcile conflicting interpretations, and correct earlier mistakes using peer evidence and counterarguments.

Final Prediction Aggregation After collecting subsequent round predictions from all models on eligible instances, *ConFreeze* applies equal-weight majority voting to produce the final prediction: $y^*_i = \hat{y}^{(1)}_{\text{maj},i}$ if $i \in \mathcal{U}$, and $y^*_i = \arg\max_\ell \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}[\hat{y}^{(2)}_{i,k} = \ell]$ if $i \in \mathcal{D}$. Frozen instances ($\mathcal{U}$) directly use the initial round majority vote, while eligible instances ($\mathcal{D}$) use majority voting over subsequent round predictions.

3 Experimental Setup

3.1 Evaluation Datasets

We select six benchmarks with categorical outputs and challenging instances where multi-model debate can provide value. ANLI-R3 (Nie et al., 2020) provides three-way NLI with challenging annotation rounds, AdvGLUE (Wang and et al., 2021) contributes adversarially perturbed classification tasks (QQP, SST-2, QNLI, MNLI), and TruthfulQA-MC1 (Lin et al., 2022) tests robustness on multiple-choice questions. We follow each benchmark's official splits and evaluation protocol. Table 1 summarizes dataset characteristics.

3.2 Model Configuration

We use three-model teams to enable majority voting. Our model pool includes gpt-4o-mini and gpt-3.5-turbo-0125 from OpenAI, claude-3-haiku-20240307 from Anthropic, deepseek-chat from DeepSeek, and llama-3.1-8b-instant from Meta. This selection spans different capability levels and cost points while covering both proprietary APIs and open-weight models to reflect realistic deployment scenarios.

To probe how model diversity affects consensus quality, we evaluate three team configurations: (i) Homogeneous (gpt-3.5-turbo $\times$3); (ii)

Table 2: Main Results on Efficiency and Performance.

Dataset	Method	Inference Efficiency			Reasoning Performance	
		Tokens/item	Saving	Coverage	Accuracy (%)	Macro-F1 (%)
ANLI-R3	GPT-3.5	263.7	–	–	37.4	38.5
	GPT-4o-mini	282.7	–	–	37.3	38.8
	DeepSeek-V3	299.7	–	–	39.6	39.7
	Initial Round	846.2	–	0	39.1	40.0
	ConFreeze	**1,757.6**	**29.5%**	**53.7%**	42.6	43.0
	Full Debate	2,493.6	base	100%	42.4	43.2
AdvGLUE	GPT-3.5	97.0	–	–	60.6	59.1
	GPT-4o-mini	99.2	–	–	64.9	65.0
	DeepSeek-V3	145.5	–	–	70.2	70.7
	Initial Round	341.7	–	0	65.0	65.7
	ConFreeze	**744.2**	**43.1%**	**65.1%**	73.0	72.4
	Full Debate	1,308.7	base	100%	74.0	73.6
TruthfulQA	GPT-3.5	166.2	–	–	64.4	–
	GPT-4o-mini	191.0	–	–	78.3	–
	DeepSeek-V3	221.1	–	–	90.2	–
	Initial Round	578.3	–	0	80.3	–
	ConFreeze	**1,224.3**	**40.3%**	**58.6%**	89.7	–
	Full Debate	2,051.3	base	100%	88.7	–

Moderately-Diverse (gpt-4o-mini, gpt-3.5-turbo, deepseek-chat); (iii) Highly-Diverse (gpt-3.5-turbo, claude-3-haiku, llama-3.1-8b-instant).

3.3 Evaluation Metrics

We evaluate performance across four dimensions: vote dispersion signals, end-task quality, stability, and cost efficiency. For each instance, y is the ground truth, $\hat{y}^{(1)}_{\text{maj}}$ is the initial round majority label, and y^* is the final output. Under *ConFreeze*, we partition instances into frozen unanimous subset $\mathcal{U}$ and eligible disagreement subset $\mathcal{D}$, where $|\mathcal{U}| + |\mathcal{D}| = N$.

vote-dispersion signals. To identify which instances benefit from debate, we compute lightweight vote-dispersion signals from the initial round. Using the vote shares $p_{i,\ell}$ defined in Section 2.2, we defined vote margin and the vote entropy: $m_i = p_{(1)} - p_{(2)}$, $H_i = -\sum_{\ell \in \mathcal{Y}} p_{i,\ell} \log p_{i,\ell}$ These signals characterize agreement strength without requiring ground truth: low margin or high entropy indicates initial disagreement, suggesting potential benefit from collaborative refinement. We use these signals for gating decisions and offline threshold sweeps.

End-task quality. We report initial accuracy $\text{Acc}_{\text{R1}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}^{(1)}_{\text{maj},i} = y_i]$, final accuracy $\text{Acc}_{\text{final}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y^*_i = y_i]$, and Macro-F1 computed on final outputs.

Stability metrics. Standard accuracy cannot reveal whether debate improves through error correction or introduces instability. To decompose prediction changes, we define three per-instance indicators:

$$\text{flip}_i = \mathbf{1}[\hat{y}^{(1)}_{\text{maj},i} \neq y^*_i] \quad (1)$$

$$\text{imp}_i = \mathbf{1}[\hat{y}^{(1)}_{\text{maj},i} \neq y_i \wedge y^*_i = y_i] \quad (2)$$

$$\text{wor}_i = \mathbf{1}[\hat{y}^{(1)}_{\text{maj},i} = y_i \wedge y^*_i \neq y_i] \quad (3)$$

Aggregating over subset $\mathcal{S}' \subseteq \{1, \ldots, N\}$ yields flip rate $\text{Flip}(\mathcal{S}') = \frac{1}{|\mathcal{S}'|}\sum_{i \in \mathcal{S}'} \text{flip}_i$, improve rate $\text{Improve}(\mathcal{S}')$, and worsen rate $\text{Worsen}(\mathcal{S}')$ (defined analogously). Unless stated, rates are global (denominator N). Reporting rates separately for $\mathcal{U}$ versus $\mathcal{D}$ isolates the effect of gating from debate effectiveness.

Cost and coverage. We define *coverage* as the fraction of instances that proceed to the subsequent round: $\text{Cov} = |\mathcal{D}|/N$, where N is the total dataset size, measuring the fraction of instances proceeding to collaborative refinement. Lower coverage indicates greater token savings. We log per-instance token usage and report average tokens per item.

4 Result Analysis

We structure our analysis around three research questions to systematically evaluate the efficiency-accuracy trade-off, the underlying stability dynamics, and the robustness of *ConFreeze* to design variations.

4.1 Efficiency-Performance Trade-off Analysis (RQ1)

Table 2 evaluates *ConFreeze* against single-model and full debate baselines.

Table 3: Stability Analysis and Attribution of Decision Revisions.

Dataset	Policy	Global Dynamics (%)			Unanimous (U) (%)		Disagreement (D) (%)	
		Flip	Imp	Wor	Imp_U	Wor_U	Imp_D	Wor_D
ANLI-R3	Full Debate	14.4	7.7	4.3	0.4	0.2	14.3	8.1
	ConFreeze	14.4	7.8	3.9	0.0	0.0	14.5	7.3
TruthfulQA	Full Debate	13.1	9.2	0.8	0.5	0.0	22.8	2.1
	ConFreeze	13.1	10.0	0.7	0.0	0.0	24.1	1.6
AdvGLUE	Full Debate	14.1	11.2	2.3	3.0	0.4	27.3	6.1
	ConFreeze	13.3	10.0	3.2	0.0	0.0	28.5	8.5

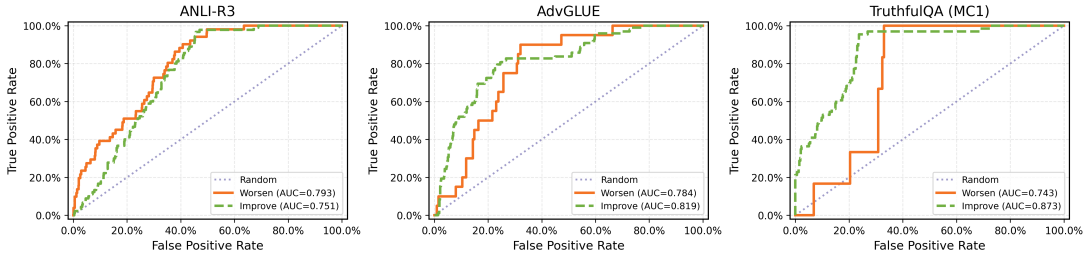

Figure 2: ROC curves for vote entropy (H) predicting revision outcomes (Improve/Worsen). Vote margin (m) yields nearly identical performance and is provided in the accompanying artifact for clarity.

ConFreeze substantially reduces token consumption while maintaining comparable accuracy. Compared to full debate, *ConFreeze* achieves 29.5% token reduction on ANLI, 43.1% on AdvGLUE, and 40.3% on TruthfulQA. These savings are achieved by limiting the second-stage coverage to 53.7% on ANLI, 65.1% on AdvGLUE, and 58.6% on TruthfulQA.

Beyond achieving a substantial reduction in token cost, *ConFreeze* also preserves near-identical accuracy performance compared to the full debate baseline. To statistically validate this equivalence, we conducted paired McNemar exact tests on the accuracy results of *ConFreeze* versus full debate across all datasets, revealing no significant divergence between the two approaches. The observed p-values across ANLI (0.940), AdvGLUE (0.392), and TruthfulQA (0.648) all exceed conventional significance thresholds ($alpha = 0.05$), demonstrating that *ConFreeze* maintains strict accuracy parity with the full debate baseline and imposes no measurable performance penalty. Quantitatively, the final accuracy of *ConFreeze* deviates by only 0.3–1.0 percentage points from full debate across all datasets: ANLI achieves 42.6% versus 42.4%, AdvGLUE achieves 73.0% versus 74.0%, and TruthfulQA achieves 89.7% versus 88.7%. Macro-F1 scores exhibit a similarly stable trend.

Furthermore, *ConFreeze* consistently outperforms both single-model and 1-round debate baselines while approaching the performance ceiling of full debate. On ANLI, for instance, *ConFreeze* boosts the accuracy from 39.6% (best single model) and 39.1% (with initial debate round) to 42.3%, which attests to its effective multi-model coordination mechanism. Identical performance patterns are observed across AdvGLUE and TruthfulQA, confirming that *ConFreeze* retains the core advantages of multi-model debate while strategically allocating deliberation resources exclusively to instances where initial inter-model disagreement signals potential for performance improvement.

4.2 Stability and Attribution Analysis (RQ2)

In RQ2, we investigate the mechanical basis of *ConFreeze* by asking whether initial voting patterns reliably signal the cost-benefit trade-off of debate.

Discriminative Power and Asymmetry. We first evaluate whether vote dispersion signals like entropy H and margin m can predict revision outcomes. Treating revision as a binary classification task, Figure 2 demonstrates that entropy achieves strong discriminative power across datasets. For instance, on ANLI-R3, entropy attains an AUC of 79.3% for predicting worsening flips and 75.1% for improvements as these values significantly exceed random baselines. Margin signals yield nearly identical results.

This predictive validity stems from the stark asymmetry revealed in the attribution analysis shown in Table 3. In the unanimous subset $\mathcal{U}$, revisions are predominantly unnecessary because full debate results in minimal bidirectional shifts, ex-

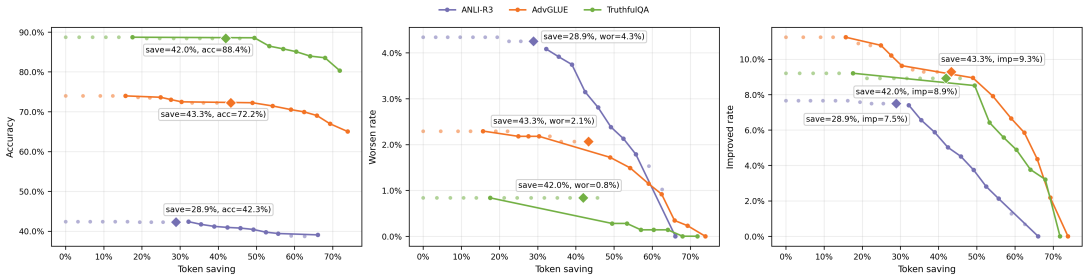

Figure 3: Budget-accuracy Pareto frontiers using vote entropy (H). Vote margin-based sweeps show similar trends (see artifact).

hibiting insignificant magnitudes of both improvement and worsening. By freezing $\mathcal{U}$, *ConFreeze* eliminates this instability. Conversely, corrective utility is heavily concentrated in the disagreement subset $\mathcal{D}$. For example, on TruthfulQA under *Full Debate*, the improvement rate is 22.8% in $\mathcal{D}$ versus 0.5% in $\mathcal{U}$ ($\approx 45.6\times$). *ConFreeze* then freezes $\mathcal{U}$ (0.0% by design) and focuses refinement on $\mathcal{D}$, where improvement reaches 24.1%. This confirms that dispersion signals effectively isolate the high-yield instances that warrant collaborative refinement.

Budget-Accuracy Frontier. To generalize beyond the binary gate, we simulate a continuous budget allocation policy by selectively debating the top-q instances ranked by uncertainty. Figure 3 plots the resulting trade-offs. We observe a clear diminishing returns pattern where accuracy remains stable over a broad mid-saving region of 20% to 50% token reduction before dropping sharply. This indicates that low-uncertainty instances contribute minimal marginal value to final quality. Notably, the *ConFreeze* operating points represented by diamond markers consistently align with the Pareto-efficient frontier of these curves (Censor, 1977). This suggests that our training-free unanimity rule is not merely a heuristic but a near-optimal strategy that maximizes efficiency without compromising the reasoning gains of the eligible subset.

4.3 Ablation studies (RQ3)

This RQ investigates the design sensitivity of the framework. We systematically ablate decoding temperature and team composition as well as subsequent-round prompting strategies to evaluate their impact on task performance and stability attribution alongside computational cost.

Temperature Sensitivity As shown in Table 4, we vary the temperature with the mechanism held constant. Across temperatures ranging from 0.3 to 0.9 the final accuracy stays within a narrow range which indicates that *ConFreeze* is not brittle to moderate sampling changes. However the composition of revisions shifts significantly. At lower temperatures the mechanism becomes more conservative which reduces both corrections and harmful flips but also yields lower final accuracy. At temperatures of 0.7 and 0.9 *ConFreeze* attains the best trade-off by achieving higher improvement rates at comparable worsening rates leading to the strongest final accuracy. Token cost is relatively stable across temperatures suggesting that temperature mainly affects how revisions happen rather than the overall routing coverage.

Team Composition Varying the team composition reveals a direct relationship between diversity and performance. Using a Homogeneous team substantially reduces subsequent-round coverage and tokens per item but also shrinks correction capacity and degrades final performance. In contrast a Diverse team increases coverage and yields the strongest final performance driven by a larger fraction of beneficial revisions. However this diversity also makes revisions more aggressive as both change rates and worsening rates rise while token cost per item increases. Overall team diversity primarily modulates the size of the disagreement subset where a larger eligible set exposes more instances to both revision and risk.

Prompting Strategy Replacing the plain prompt with a richer critique-style prompt substantially degrades ANLI performance. Final accuracy drops and the overall delta becomes negative, mirrored in revision composition where worsening rates exceed improvement rates and harmful flips dominate. The richer prompt is also more expensive, indicating that increasing deliberation complexity without guardrails can be both costlier and less reliable.

We qualitatively attribute this failure to three recurring mechanisms. First is rationale conta-

Table 4: Design sensitivity ablations on ANLI under *ConFreeze*. Team settings: Homogeneous (gpt-3.5-turbo ×3); (ii) Moderately-Diverse (gpt-4o-mini, gpt-3.5-turbo, deepseek-chat); (iii) Highly-Diverse (gpt-3.5-turbo, claude-3-haiku, llama-3.1-8b-instant).

Ablation	Setting	Cov (%)	Acc (%)	Macro-F1 (%)	Change (%)	Imp (%)	Wor (%)	Tok/item
Temp	*T=0.3*	51.4	42.5	42.7	14.1	6.8	4.2	1701.7
Temp	*T=0.5*	51.4	41.6	42.1	12.4	5.9	3.6	1707.7
Temp	***T=0.7***	**53.7**	**42.6**	**43.1**	**14.4**	**7.8**	**3.9**	**1757.6**
Temp	*T=0.9*	50.8	42.7	43.1	14.8	7.9	4.1	1692.6
Team	*Moderately-Diverse*	53.7	42.6	43.1	14.4	7.8	3.9	1757.6
Team	***Highly-Diverse***	**60.2**	**46.1**	**45.4**	**23.2**	**12.2**	**6.3**	**1957.0**
Team	*Homogeneous*	20.5	39.3	39.7	7.6	4.3	2.3	1077.0
Prompt	*R2-rich*	51.7	37.5	40.4	10.8	3.4	4.7	2318.5
Prompt	***R2-plain***	**53.7**	**42.6**	**43.1**	**14.4**	**7.8**	**3.9**	**1757.6**

gion where models defer to confident but wrong rationales even if logically invalid. Second is a forced revision bias where instructions act as latent commands that pressure models to update correct answers to appear responsive. Third is shallow critique where models overfit to surface cues and superficial heuristics rather than core reasoning. These failures explain why prompting sensitivity matters as the prompt defines the update rule. Accordingly our default setting uses a minimal peer-aware prompt that permits revision but does not pressure it. The full text of these prompts is available in the accompanying artifact.

5 Related Work

Debate was proposed by Irving et al. (Irving et al., 2018) as an alignment protocol in which two agents argue and a judge selects the better argument, motivated by the idea that evaluating arguments can be easier than producing them. Adapting this paradigm to LLMs, Du et al. (Du and et al., 2023) reported factuality/reasoning gains from MAD with a mediated judge on diverse tasks, and Liang et al. (Liang et al., 2023) encouraged divergent thinking via debate to surface alternative solutions under controlled setups. Alongside these protocols, Zheng et al. (Zheng and et al., 2023) examined LLM-as-a-judge and documented position/verbosity biases, advising caution when using automated judges to arbitrate multi-agent outputs. Beyond explicitly adversarial debate, Wang et al. (Zheng and et al., 2023) introduced Self-Consistency to marginalize over multiple chains of thought, Yao et al. (Yao and et al., 2023) proposed Tree-of-Thoughts for deliberate search over intermediate states, and Shinn et al. (Shinn et al., 2023) presented Reflexion to add episodic self-feedback where each improves solution quality by aggregating diverse reasoning traces.

In parallel, Li et al. propose CAMEL (Li and et al., 2023), a role-playing framework where LLM agents hold complementary, persistent roles, stabilizing interaction and broadening exploration. Chen et al. (Chen and et al., 2023) present AgentVerse, a programmable setting that systematizes role assignment, tool use, and conversation rules (e.g., planner/solver/critic) with reusable backbones and evaluation hooks. Wu et al. introduce AutoGen (Wu and et al., 2023), a general-purpose library of programmable agents supporting multi-round group chats, function calls/REPL, and orchestration utilities. While these systems primarily target end-task accuracy or exploration, few quantify the robustness costs of multi-round coordination. Our study fills this gap by measuring when collaboration helps versus harms and by introducing lightweight guardrails that curb instability while preserving genuine debate gains.

6 Conclusion

In this work, we reframe multi-model debate not merely as a technique for performance boosting, but as a controllable, budget-aware decision process suited for real-world deployment. Our investigation exposes a fundamental asymmetry in the marginal utility of deliberation: improvement potential is concentrated almost exclusively in instances of initial disagreement, whereas debating unanimous instances yields diminishing returns and incurs risks of instability via "harmful flips." *ConFreeze* leverages this insight through a training-free, consensus-based gating mechanism. Across ANLI, AdvGLUE, and TruthfulQA, *ConFreeze* reduces token usage by 29.5%–43.1% while maintaining statistically indistinguishable accuracy. Overall, it provides a practical protocol to improve the efficiency of multi-model collaboration without sacrificing reliability.

7 Limitations

Stochasticity and Temporal Consistency of Black-box APIs. Our experimental framework relies on both open-weight models and proprietary APIs. While we enforce strict control over inference hyperparameters (e.g., temperature, decoding strategy) to maximize reproducibility, the opaque nature of commercial APIs introduces unavoidable risks of model drift and temporal inconsistency. Updates to the backend architecture or safety alignment strategies by providers may alter model behavior over time, meaning our empirical findings represent a snapshot of model capabilities that may not perfectly transfer to future API versions.

Fidelity of Efficiency Metrics. We utilize token consumption and second-round coverage as vendor-agnostic proxies for computational budget. However, these metrics serve as abstract heuristics rather than holistic measures of deployment cost. They fail to account for tokenizer heterogeneity across providers, hardware-specific latency, or the pricing dynamics of batched inference. Consequently, our cost analysis should be interpreted as illustrating compute-matched comparative trends rather than providing precise dollar-equivalent economic estimates.

Generalization across Task Distributions and Agent Topologies. Although we validate our approach across three diverse datasets with ablation studies on prompting and temperature, our scope remains bounded by a fixed team size and specific interaction protocols. Future work is required to verify whether the observed trade-offs hold under broader task distributions and more complex aggregation mechanisms.

Data and content considerations. Our experiments use publicly available benchmark datasets, which may contain sensitive or offensive text. We do not collect new personal data, and we do not attempt to identify individuals. For the released anonymized artifact, we minimize the risk of exposing identifying information by (i) storing only dataset identifiers/indices and model outputs when possible, (ii) removing any run-time metadata that could be user-specific, and (iii) applying lightweight pattern-based screening (e.g., emails, phone numbers, and other common PII markers) to the logged prompts/outputs, with redaction or exclusion when such strings are detected. In the paper, we also avoid qualitative examples that contain explicit slurs or uniquely identifying details.

References

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. 2024. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6929–6948, Miami, Florida, USA. Association for Computational Linguistics.

Yair Censor. 1977. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization*, 4:41–59.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Xiaoheng Chen and et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv preprint arXiv:2308.10848*.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *Preprint*, arXiv:2508.17536. NeurIPS 2025 Spotlight.

Yilun Du and et al. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *Preprint*, arXiv:1805.00899.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *Preprint*, arXiv:2402.06782.

Guohao Li and et al. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*.

Ray Li, Tanishka Bagade, Kevin Martinez, Flora Yasmin, Grant Ayala, Michael Lam, and Kevin Zhu. 2024. A debate-driven experiment on llm hallucinations and accuracy. *arXiv preprint arXiv:2410.19485*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of ACL*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, page 476–483. IEEE.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Noah Shinn, Federico Cassano, Arnav Labash, and Joelle Pineau. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. *Preprint*, arXiv:2311.17371.

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

Yizhong Wang and et al. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qi Wu and et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

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

Shunyu Yao and et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, and 1 others. 2025. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *Preprint*, arXiv:2502.08788.

Lianmin Zheng and et al. 2023. Llm-as-a-judge: Evaluating llm-as-a-judge with pairwise preference data. *arXiv preprint arXiv:2306.05685*.