V-BReE: A Variance-thresholded Blinded Refinement Ensemble for Multi-Agent LLM Reasoning

Milad Ebrahimi Abyazandi, Sherwin Darryl D'souza, Jeremy Greenwood, Jason Ives
University of Guelph
{mebrah04, sherwind, jgreen13, jives}@uoguelph.ca

Abstract

Large Language Model (LLM) ensembles frequently suffer from intra-ensemble sycophancy, where constituent models propagate the logical errors of their peers due to shared context cues. We introduce **V-BReE (Variance-Thresholded Blinded Refinement Ensemble)**, a novel architecture designed to mitigate this bias through structural context blinding. By isolating agents from prior reasoning metadata and employing a strictly controlled iterative refinement loop, V-BReE enforces response independence and utilizes assessment variance as a robust mathematical proxy for consensus quality.

We evaluate a heterogeneous ensemble comprising gpt-oss-20b, Qwen2.5-7B-Instruct, and Llama-3.1-8B-Instruct on a 3,001 question stratified sample of the MMLU-PRO benchmark. Our results demonstrate that V-BReE facilitates emergent recovery, reaching the ground truth in 132 instances (6.8%) where every constituent model initially failed. We also identify a friction paradox: technical noise in complex domains that acts as a stochastic reset, improving ensemble performance. In the Engineering domain, this leads to a 17.6% accuracy gain over the 20B anchor model. These findings suggest that deep iterative logic distillation, rather than simple consensus-seeking, is the key to bypassing the reasoning ceilings of current instruction tuned models.

1 Introduction

LLMs demonstrate significant reasoning capabilities but remain susceptible to hallucinations and brittle Chain-of-Thought (CoT) processes. Multi-agent collaboration and iterative refinement frameworks have emerged as superior alternatives to single-agent prompting (Du et al., 2023; Chen et al., 2024), yet these systems introduce a distinct reliability risk: intra-ensemble sycophancy. Exposure to peer rationales often induces linguistic conformity, causing models to abandon correct individual reasoning to align with an incorrect majority (Perez et al., 2023; Wynn et al., 2025). This reveals a fundamental limitation in existing frameworks: they facilitate information exchange without regulating the interaction structure, thereby conflating evidence-based refinement with socially induced agreement.

To address this, we introduce **V-BReE (Variance-Thresholded Blinded Refinement Ensemble)** (Abyazandi et al., 2026), a framework designed to decouple collaborative reasoning from social influence. V-BReE employs a structural context blinding protocol where models sequentially revise candidate responses without access to peer evaluations, model identities, or historical MCQ selections. By stripping away the authority of direct peer assessment, the framework enforces response independence. This isolation allows us to utilize a variance-threshold stopping mechanism to detect genuine logical stability and dynamically terminate the refinement process.

We evaluate V-BReE using a heterogeneous ensemble comprising gpt-oss-20b, Qwen2.5-7B-Instruct, and Llama-3.1-8B-Instruct on the MMLU-PRO benchmark (Wang et al., 2024b). Our results demonstrate that this blinded architecture facilitates emergent recovery (**ER**), reaching ground-truth accuracy in 132 instances where every constituent model initially failed. Significantly, we also identify a friction paradox in technical domains: LaTeX processing errors acting as a stochastic reset that prevents premature consensus, leading to a 17.6% accuracy gain over the 20B anchor model. By corroborating these findings with a controlled extended variance threshold (**EVT**) experiment, we demonstrate that such logical friction is not merely a byproduct of failure, but a catalyst for deeper deliberation. These findings suggest that by inhibiting sycophancy and embracing iterative distillation, ensembles can bypass the reasoning ceilings inherent

in their constituent models.

2 Related work

Our research idea, methodological design, and evaluation framework are informed by prior work on multi-agent LLM reasoning, particularly three ACL-family papers: Conformity in Large Language Models (Zhu et al., 2025), Improving Multi-Agent Debate with Sparse Communication Topology (Li et al., 2024), and RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs (Chen et al., 2024). These works provide important methodological and empirical foundations for studying collaborative reasoning and social influence in multi-agent LLM systems, which our framework builds upon and extends.

2.1 Multi-agent collaboration and debate

Recent work proposes multi-agent collaboration and debate to enhance LLM reasoning through rationale exchange, critique, and aggregation. Du et al. (2023) introduce multi-agent debate with iterative refinement and voting, improving arithmetic and strategic reasoning over chain-of-thought. Later systems add structured roles: Liang et al. (2024) use specialized debaters and judges for divergent reasoning; Chen et al. (2024) propose RECONCILE with convincing samples and confidence-weighted voting; Omar et al. (2024) develop ICE, an iterative consensus ensemble; and Wang et al. (2024a) study group discussion with message passing.

Broader multi-agent systems examine role-based collaboration for complex tasks. CAMEL (Li et al., 2023) and AutoGen (Wu et al., 2023) support planning and heterogeneous teams, where stronger models can guide weaker ones (Wang et al., 2024a). Yet benefits vary: diversity can improve robustness (Subramaniam et al., 2025) but failure modes include judge errors and wrong-answer propagation (Agarwal and Khanna, 2025; Estornell and Yang, 2024).

Wynn et al. (2025) show debate can degrade performance versus majority vote, with groups converging from correct to incorrect answers. They attribute this to revision dynamics, social influence, and sycophancy-like behavior that amplifies errors.

2.2 Social conformity and sycophancy in LLMs

LLMs exhibit conformity and sycophancy - alignment with majorities or user beliefs despite ground truth (Perez et al., 2023; Weng et al., 2025). Single-agent sycophancy emerges from reinforcement learning from human feedback (RLHF; (Perez et al., 2023)), where models learn to prefer user-pleasing responses. In multi-agent settings, isolated correct agents flip to incorrect answers under peer disagreement (Wynn et al., 2025).

Wynn et al. (2025) find more correct-to-incorrect than incorrect-to-correct flips, with pressure growing over rounds. Prompt-based anti-sycophancy incentives fail. Estornell and Yang (2024) analyze majority tyranny; Agarwal and Khanna (2025) show judges favor confident falsehoods. Adversarial setups confirm social influence dominates without dissent preservation (Amayuelas et al., 2024; Yao et al., 2025).

Prior work relies on unconstrained rationale sharing and consensus mechanisms (majority vote, confidence-weighted voting, judges), leaving systems vulnerable to conformity (Weng et al., 2025; Wynn et al., 2025; Yao et al., 2025; Zhu et al., 2025). V-BReE counters this with variance-thresholded blinded refinement that limits social exposure while retaining collaboration benefits, building on sparse communication (Li et al., 2024).

3 Methodology

3.1 Dataset and task definition

We evaluate V-BReE using the MMLU-PRO benchmark (Wang et al., 2024b), leveraging its ten-option format to provide a rigorous ceiling for iterative reasoning gains. The task is structured as a sequence of blinded refinements where models: (i) utilize a consistent one-shot format; (ii) enforce a strict JSON schema; and (iii) process only the predecessor's rationale R_{i-1}, stripped of all ensemble-level metadata. To enforce response independence, we utilize a structural context blinding protocol. Prompt synthesis involves appending the preceding model's rationale to a standardized template that removes ensemble metadata and historical MCQ selections.

A primary technical challenge emerged during the processing of STEM domains, where complex LaTeX formatting regularly triggered parsing failures due to improperly escaped characters. A subsequent analysis of these occurrences revealed the

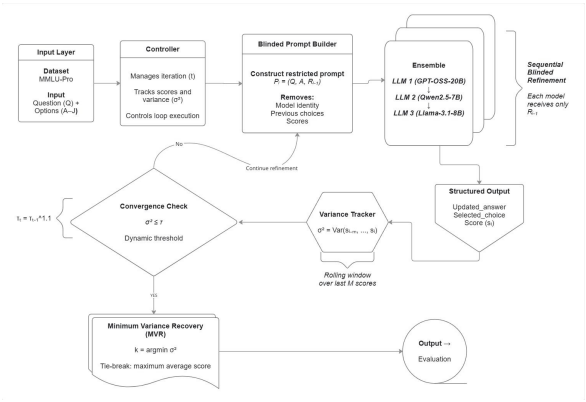

Figure 1: The V-BReE sequential pipeline showing the iterative refinement and response selection processes.

role of processing friction, in which formatting-induced resets necessitated additional reasoning cycles. This provided a unique dataset for evaluating the relationship between iteration depth and logical recovery in high-entropy technical tasks.

3.2 Model ensemble strategy

The V-BReE architecture utilizes a heterogeneous ensemble selected for architectural diversity and parameter-count parity. By utilizing models within a similar parameter scale (7B–20B), we ensure that ensemble variance is driven by divergent latent reasoning patterns rather than disparities in raw computational capacity.

3.3 System pipeline

The V-BReE framework operates as a state-aware sequential pipeline (Figure 1). A controller manages the ensemble, executing inference through a recursive refinement loop.

3.3.1 Iterative processing

For each query, the system constructs a restricted prompt P_i incorporating the question Q, options A, and the predecessor's rationale R_{i-1}. To inhibit sycophancy, R_{i-1} is presented as a candidate for objective assessment. Responses are parsed to extract a triplet of $\{updated\_answer, selected\_choice, score\}$. A 600-token response limit is mandated to ensure computational efficiency and logical conciseness, but not enforced, to allow for completeness of processing.

3.3.2 Convergence and minimum variance recovery (MVR)

Termination is governed by a variance-driven heuristic. Consensus is defined by the moving variance (σ^2) of the last M scores. The loop terminates when $\sigma^2 \leq \tau$, where τ is a dynamic threshold initialized at $\tau_0 = 9.5$ and scaled every M iterations as:

$$\tau_t = \tau_{t-1}^{1.1}$$

Upon termination, a minimum variance recovery logic selects the historical iteration R_k with the global minimum variance, using the score moving average as a tie-breaker. This ensures output stability even in high-entropy reasoning tasks.

3.4 Compute and reproducibility

To ensure high-throughput reproducibility, we leveraged a distributed inference approach via the Hugging Face Inference Client API, utilizing specialized endpoints from Groq (*20B*), Together AI (*7B*), and Cerebras (*8B*). This heterogeneous infrastructure maintains reliability and temporal efficiency across the sequential chain without requiring local GPU clusters.

4 Evaluation framework

To evaluate the efficacy of the V-BReE protocol independent of constituent model scale, we employ a multi-dimensional assessment framework. This approach prioritizes internal process dynamics over static accuracy, allowing us to quantify the transition from initial individual reasoning to emergent ensemble consensus.

4.1 Quantitative performance and recovery metrics

Our primary performance metric is Accuracy over Ground Truth (GT), supplemented by two categories designed to evaluate resistance to social conformity: emergent Recovery (ER) and minority Recovery (MR). ER tracks instances where the ensemble converges on the correct answer despite zero baseline accuracy ($0/3$), quantifying the framework's ability to synthesize solutions via iterative distillation. MR tracks instances where the ensemble converges on a single correct constituent ($1/3$) rather than the incorrect majority, serving as a direct measure of success in mitigating intra-ensemble sycophancy.

4.2 Stability and consensus metrics

We evaluate the logic distillation process by measuring the convergence trajectory via three primary internal metrics: Rolling assessment variance (σ^2), which quantifies inter-agent agreement and state

stability; Iteration depth (t), the number of cycles required to satisfy the threshold τ, used to analyze the friction paradox across varying domain complexities; and the extended variance threshold (EVT) delta, which tracks the accuracy shift when the stopping condition is artificially extended to measure the marginal impact of additional deliberation.

4.3 Trajectory and transition analysis

We assess logical refinement via BERTScore (Zhang et al., 2020) (RoBERTa-base), measuring the semantic delta between the initial rationale R_1 and final rationale R_{final} relative to ground-truth. An increase in $BERTScore(R_{final}, GT)$ serves as a proxy for reasoning refinement, indicating convergence on robust logical justifications rather than simple label matching. To distinguish static consensus from active distillation, we track the activity rate (A):

$$A = \frac{\sum(W \to R) + (W \to W')}{N}$$

where $(W \to R)$ denotes corrective conversion and $(W \to W')$ denotes transition to a distinct incorrect state. This metric quantifies the ensemble's generative movement throughout the refinement cycles.

5 Experiments

We evaluate a stratified sample of 3,001 questions from MMLU-PRO (250 per category). The ensemble consists of gpt-oss-20b, Qwen2.5-7B-Instruct, and Llama-3.1-8B-Instruct ($T = 0.0$).

Experiment 1: Zero-shot baseline evaluation This establishes the performance floor. Each model was evaluated independently using the one-shot prompt/JSON schema defined in Section 3.

We hypothesize that single-agent chains reach an entrenched performance ceiling where logical errors become high-probability, establishing the initial task variance V-BReE must resolve.

Experiment 2: V-BReE collaborative consensus We execute the refinement pipeline on the global sample ($N = 3,001$). Models refine preceding reasoning chains via blinded state-passing, with termination governed by $\tau_0 = 9.5$ and $\lambda = 1.1$.

We hypothesize that structural context blinding prioritizes logical consistency over peer conformity, improving performance and facilitating minority Recovery (MR) and emergent Recovery (ER) beyond the 20B anchor's ceiling.

Experiment 3: Threshold sensitivity and extended variance We conducted a sensitivity analysis on a balanced subsample ($N = 360$; $n = 30$ per domain) to investigate the relationship between convergence pressure and reasoning quality. To isolate logical variance from structural parsing failures, only LaTeX error-free instances were included. We applied a conservative regime ($\tau_0 = 2.0, \lambda = 1.05$) to force extended deliberation. This creates a controlled simulation of the friction paradox effect observed in STEM domains by forcing an increased number of reasoning cycles.

We hypothesize that extending the refinement window inhibits premature consensus and reproduces the performance gains seen via the friction paradox in technical domains.

Experiment 4: Activity and semantic transitions To track the ensemble's evolution, we define the Activity Rate (A) as the frequency of state changes ($Wrong \to Right$ or $Wrong \to Wrong'$). Model-specific refinement is measured via mean BERT F1 scores (RoBERTa-base) calculated across a representative validation subset ($n = 70$) to compare sequential rationales. In this framework, lower F1 scores indicate significant structural or logical departures, serving as a proxy for a model's discerning capability during distillation.

Our hypothesis is that corrective transitions ($W \to R$) are characterized by substantial structural overhauls rather than minor semantic shifts, with higher-parameter models exhibiting the greatest degree of structural discernment.

6 Results

Baseline accuracy and unique correctness The baseline evaluation establishes the independent reasoning capacity of the three constituent models on the MMLU-PRO dataset. Table 1 summarizes the aggregate and domain-specific accuracy across the 3,001 sampled instances. As expected, gpt-oss-20b demonstrated the highest overall accuracy (66%), followed by Qwen-2.5-7B (51%) and Llama-3.1-8B (39%). While the 20B parameter model maintained a significant lead in raw correctness, we observed a high degree of predictive overlap, indicating a shared reasoning ceiling for instruction-tuned models in this parameter class.

	GPT-20B		Qwen-7B		Llama-8B	
Domain	Acc.	Unq.	Acc.	Unq.	Acc.	Unq.
Math	**0.856**	50	0.612	9	0.364	0
Physics	**0.764**	69	0.476	10	0.288	6
Chemistry	**0.760**	67	0.444	11	0.316	6
Business	**0.716**	34	0.620	15	0.336	4
Biology	**0.860**	24	0.740	4	0.708	6
Comp. Sci.	**0.752**	65	0.484	12	0.328	8
Psychology	**0.712**	27	0.628	16	0.556	11
Economics	**0.756**	32	0.588	4	0.500	9
Engineering	**0.384**	34	0.360	27	0.280	16
Philosophy	**0.508**	39	0.444	21	0.328	16
History	**0.540**	27	0.444	7	0.396	10
Law	**0.316**	33	0.304	25	0.248	14
Overall	**0.660**	501	0.512	161	0.387	106

Table 1: Baseline Accuracy (Acc.) and Unique Correctness (Unq.) per model ($N = 3,001$). Unique values indicate instances where only the specified model succeeded, highlighting the latent potential for ensemble recovery.

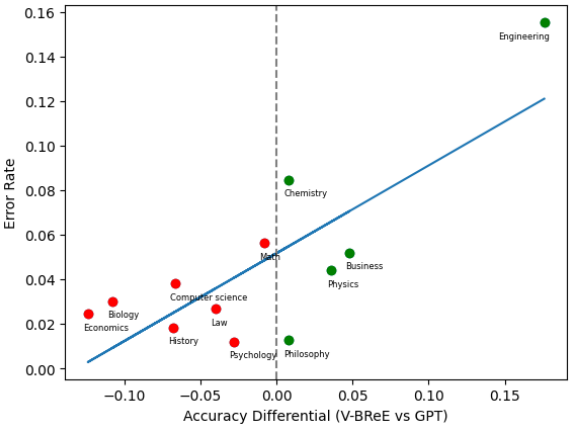

Figure 2: Correlation between domain-specific LaTeX formatting error rates and V-BReE accuracy gains ($r = 0.79, p = 0.0024$). High-friction technical environments (e.g., Engineering) prevent premature variance collapse, facilitating emergent reasoning gains that are absent in low-friction domains.

We also measured instances where only one ensemble member solved a question, to quantify potential gains. Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct contributed 267 unique correct answers (8.9% of total) missed by the 20B anchor (Table 1). These minority truths form the empirical basis for V-BReE; by enforcing structural context blinding, the protocol prevents the 20B model from acting as a sycophantic anchor, thereby preserving knowledge diversity and preventing convergence on shared, high-parameter errors.

Evaluating the V-BReE framework and exploring the deliberation deficit Aggregate V-BReE accuracy (64.6%) peaked 1.4% below the 20B anchor, despite a 61% minority recovery (MR) rate and 132 emergent recoveries (ER) (6.8% of total) (Table 2). This gap stems from a deliberation deficit: a tendency for the variance-based stopping rule to impede the deliberative process in high-consensus regimes. The presence of 132 ERs where the ensemble reached the truth despite zero baseline accuracy proves the protocol is a generative reasoning process rather than a simple consensus filter, yet these gains are offset by premature convergence on shared incorrect priors.

This deficit is non-uniform and mitigated by technical friction. We observed a strong correlation ($r = 0.79, p = 0.0024$) between a domain's LaTeX formatting error rate and its accuracy lead over the 20B anchor (Figure 2). In high-friction domains like Engineering (+17.6% lead), notation errors (15.5%) act as a stochastic reset, preventing inter-model variance (σ^2) from collapsing prematurely. This enforced variance sustains the refinement loop, whereas low-friction domains (Law, Economics) suffer the deliberation deficit, converging on shared misconceptions before the generative logic can engage. This performance zero-sum establishes the empirical necessity for extended variance threshold (EVT) testing done in Experiment 3.

Domain-specific accuracy-to-depth (t) profiles reveal three distinct mechanical archetypes (Figure 3). In Engineering, the ensemble acts as a distillation engine, with accuracy trending upward from $t = 8$ to 23, proving the protocol actively resolves incorrect chains into ground-truth. Biology exhibits latent recovery, where a "Trough-and-Spike" morphology demonstrates a deliberation deficit; here, late-cycle recovery ($t > 23$) suggests minority truths require extended cycles to overcome dominant incorrect priors. Conversely, Economics represents stagnant consensus, where flat accuracy indicates an entrenched shared prior that iteration cannot reset—a reflection of systemic training bias rather than protocol failure.

Testing the deliberation deficit hypothesis To validate the deliberation deficit hypothesis, we conducted a sensitivity analysis by manually tightening the variance threshold ($\tau_0 = 2.0$) and slowing the relaxation rate ($\lambda = 1.05$). This synthetic friction isolates the mechanical impact of processing depth from formatting-induced variance.

The transition to the tighter regime yielded a sig-

	GPT-OSS-20B	V-BReE Ensemble		
Domain	Accuracy	Accuracy	Minority Rec. (MR)	Emergent Rec. (ER)
Math	0.856	0.848	45	8
Physics	0.764	**0.800**	68	14
Chemistry	0.760	**0.768**	64	11
Business	0.716	**0.764**	34	19
Biology	0.860	0.752	21	6
Computer Science	0.752	0.685	55	7
Psychology	0.712	0.684	28	3
Economics	0.756	0.632	27	4
Engineering	0.384	**0.560**	52	28
Philosophy	0.508	**0.516**	33	11
History	0.540	0.472	20	8
Law	0.316	0.276	22	13
Overall	**0.660**	**0.646**	**469**	**132**

Table 2: V-BReE Performance and Recovery Metrics ($N = 3,001$). Bold values indicate performance exceeding the 20B anchor. MR (minority recovery) and ER (emergent recovery) represent 1/3 and 0/3 baseline accuracy recoveries, respectively.

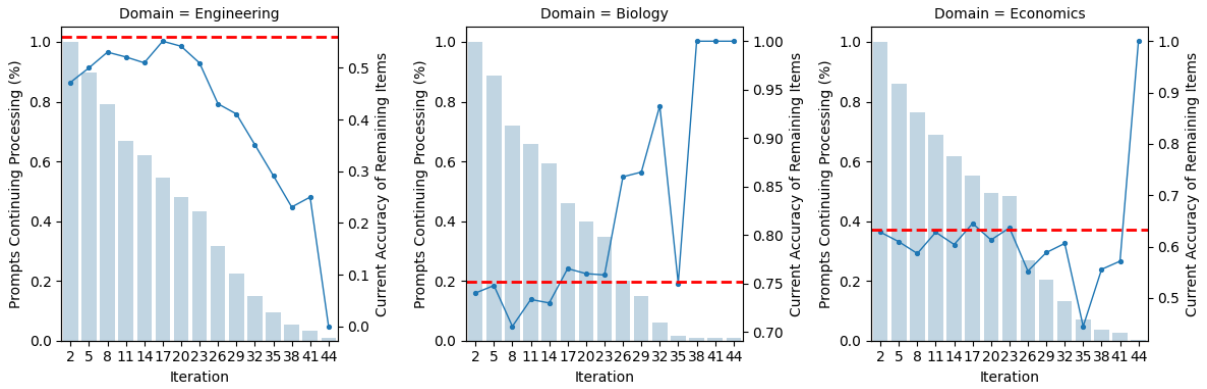

Figure 3: Archetypes of Ensemble Survival. (Left) **Distillation** (Engineering): Deliberation actively manufactures accuracy over time. (Center) **Latent Recovery** (Biology): Logical recovery emerges only after extended refinement ($t > 23$). (Right) **Stagnant Consensus** (Economics): Accuracy plateaus at an unrecoverable floor, indicating entrenched shared priors.

nificant accuracy shift from 0.500 to 0.572 (+7.2%). This confirms that simulating technical friction can rescue complex reasoning chains otherwise lost to premature consensus. The shift to 0.572 was driven by an asymmetric exchange between successful recovery (incorrect $\rightarrow$ correct) and stochastic decay (correct $\rightarrow$ incorrect). As shown in Table 3, the protocol facilitated 43 recoveries against 17 instances of decay. A McNemar's test confirms this asymmetry is highly significant ($p = 0.0011$), rejecting the null hypothesis that extended deliberation merely introduces random noise. The resulting 2.5:1 recovery-to-decay ratio demonstrates that, in general, the deliberation deficit poses a significantly greater threat to accuracy than the risk of "overthinking" (stochastic decay).

The effectiveness of the extended variance threshold (EVT) is highly domain-dependent (Figure 4), falling into three distinct tiers: high utility (Biology, STEM), where accuracy gains of $+10\%$ to $+17\%$ confirm the protocol successfully bridges the deliberation deficit in complex, objective fields; zero utility (Economics), where performance remains fixed at the 0.500 baseline, validating the sys-

	EVT Outcome	
Baseline Outcome	**Correct**	**Incorrect**
Correct	163 (Stable)	**17 (Decay)**
Incorrect	**43 (Recovery)**	137 (Gap)

Table 3: Contingency Table for the $N = 360$ Reasoning Stress Test. The asymmetric shift from Baseline Incorrect to EVT Correct (43 items) versus the shift from Baseline Correct to EVT Incorrect (17 items) represents a significant net deliberation gain (McNemar's $p = 0.0011$).

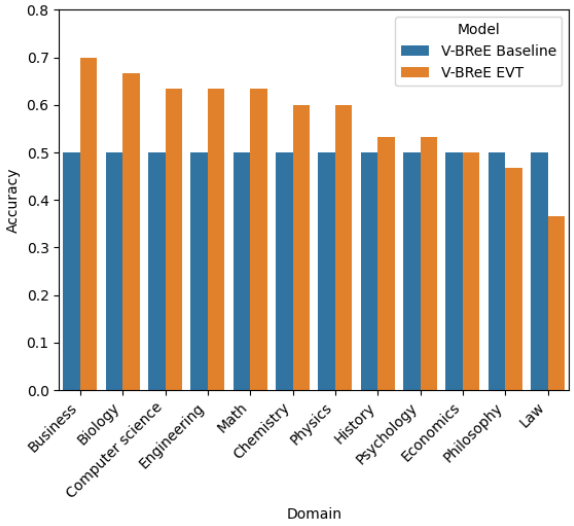

Figure 4: Performance delta under EVT ($n = 360$) relative to the 0.500 baseline. Results illustrate the transition from catalytic recovery in objective STEM domains to stochastic decay in subjective fields.

temic expiration hypothesis regarding entrenched shared priors; and negative utility (Law, Philosophy), where net regressions (e.g., Law falling to 0.367) represent an "overthinking tax." In these linguistically fluid domains, extended deliberation can destabilize correct initial consensus, triggering stochastic decay. Ultimately, deliberation acts as a catalyst for truth in technical reasoning, but a solvent for consensus in subjective reasoning.

Examining the activity rate and BERTScore The validation set ($n = 70$) achieved a 42.4% activity rate (25.7% corrective; 0% regression), while the test set ($n = 3001$) yielded 37.7% (19.7% corrective; 4.4% regression). When gpt-oss-20b provided the initial response, the activity rate dropped to 15.6% (6.9% gains vs. 6.4% regressions), confirming the 20B anchor as the most rigid initial state. Semantic analysis reveals that logical discernment scales with model accuracy rather than raw size: mean BERT F1 scores for updates were 0.95 (Llama-3.1-8B), 0.93 (Qwen2.5-7B), and 0.90 (gpt-oss-20b). The lower F1 score for the 7B model relative to the 8B indicates it performs more significant logical overhauls, positioning it as a more effective driver of distillation despite a smaller parameter footprint.

7 Discussion

Minority utility and the anti-sycophancy mandate Reasoning is not a monolithic function of parameter scale; the 7B and 8B models provided 8.9% unique correctness on instances where the 20B anchor failed. In traditional ensembles, these "islands of competence" are often lost to sycophancy. Standard consensus methods such as majority voting effectively discard minority insights by smoothing outputs toward the most frequent common denominator (Wynn et al., 2025). This structural silencing precludes emergence: an architecture designed solely for consensus cannot synthesize solutions absent in the initial baseline. V-BReE's context blinding acts as a mechanical mandate to preserve these minority truths against the authoritative pull of high-parameter priors.

Blinded refinement and emergent consensus Unlike debate-based frameworks that intensify sycophantic pressure, V-BReE enforces structural context blinding. By stripping downstream models of MCQ selections, evaluation scores, and agent identities, the protocol forces the ensemble to judge reasoning chains solely on logical merit. Here, agreement is not a negotiated consensus but a stable attraction point in the reasoning space. This explains the 132 emergent recoveries in Experiment 2: the truth was not voted into existence, but distilled through blinded refinement. This generative process enables the ensemble to bypass the reasoning ceilings of its constituents, reaching solutions natively inaccessible to any individual agent.

The friction paradox The $+17.6\%$ lead in Engineering reveals a friction paradox: formatting noise acting as a functional stochastic reset. While low-friction domains often fall into a consensus trap and settle on shared incorrect priors, the syntactic complexity of Engineering prevented rolling variance from "cooling" prematurely. These formatting hurdles effectively reset the reasoning loop, forcing the deeper, multi-turn deliberation required for logical recovery. Consequently, noise served as a safeguard against premature agreement. This implies that for technical reasoning, the computational cost of iterative depth is outweighed by gains in accuracy.

The subjectivity threshold While the friction paradox benefits objective domains, Experiment 3 reveals a subjectivity threshold where the utility of iterative refinement depends on whether a task is convergent or divergent. In objective STEM domains ground truth acts as a logical "basin of attraction", focusing the ensemble reasoning over itera-

tions. Conversely, in subjective domains like Law or Philosophy, the lack of rigid logical guardrails causes the refinement window to act as a corrosive solvent. In these divergent spaces, extended deliberation destabilizes stable truths as models over-analyze semantic nuances until correct consensus dissolves. This identifies an "overthinking tax": a maximum effective depth where additional computation actively degrades accuracy.

Systemic expiration The failure of both natural and synthetic friction to improve Economics results suggests a state of systemic expiration. We posit that when models share an identical training-induced bias deliberation becomes a closed loop. If every agent begins with the same entrenched misconception, blinded review cannot generate the variance required to break that consensus. This identifies a critical boundary for V-BReE: while the framework successfully resolves reasoning failures and uncovers emergent solutions, it cannot independently overcome knowledge deficits where the ground truth is absent from the entire model population.

8 Limitations

While V-BReE demonstrates significant generative potential, it is subject to several key constraints. First, the parameterization trade-off presents a binary optimization challenge where an extended threshold prevents premature STEM exits but imposes an "overthinking tax" on subjective domains. Second, as a logic-distillation engine rather than a retrieval system, the framework cannot overcome a shared knowledge deficit. No amount of refinement can generate novel information if it is absent from the entire model pool. Third, while MCQ benchmarks allowed for quantifiable tracking, they do not fully capture the framework's generative scope, which is natively designed for the free-form rationales not explicitly tested here. Finally, the reasoning ceiling remains tethered to pool diversity; future research must determine if these emergent mechanics persist across wider parameter gaps (e.g., 7B vs. 700B) where smaller models may lose the ability to effectively intervene in high-parameter reasoning chains.

9 Conclusion

This study introduces the Variance-Thresholded Blinded Refinement Ensemble (V-BReE), a framework that counters model sycophancy through a structural blinding protocol. By shifting the ensemble objective from consensus-seeking to logic distillation, we demonstrate that an iterative pool of 7B, 8B, and 20B models can significantly outperform individual baselines and facilitate emergent reasoning, particularly in high-complexity domains. Our findings establish two critical phenomena:

- **The friction paradox:** Technical noise and formatting errors act as a functional stochastic reset, preventing the deliberation deficit and driving a $+17.6\%$ accuracy lead in Engineering.
- **Judgment blinding:** By stripping authorship and evaluative history, V-BReE eliminates sycophantic pressure, yielding a 37.7% activity rate—a key indicator of the ensemble's willingness to pivot. This mechanical fluidity enabled 132 instances of emergent recovery ($0/3$ baseline accuracy), proving that when models are blinded to the majority's identity, they prioritize logical consistency over social conformity.

Ultimately, V-BReE demonstrates that the "wisdom of the crowd" in LLMs is a generative process rather than a statistical average. Robust reasoning emerges not through smoothed agreement, but when models are forced to survive the friction of their own iterative logic.

10 Future work

Limitations identified in this study suggest three trajectories for high-fidelity ensemble research. First, we propose **adaptive variance tuning** to mitigate the "overthinking tax," utilizing dynamic thresholds (τ) that adjust via real-time entropy detection to balance subjective fluidity against technical rigor. Second, **cross-scale benchmarking** is required to evaluate if recovery capacity scales as the parameter gap widens between frontier (70B+) and specialized expert models. Finally, we propose a hybrid **RAG-V-BReE** architecture to overcome systemic expiration; by injecting external facts into the silent revision loop, ensembles could potentially resolve logic-based reasoning failures and fact-based knowledge deficits simultaneously.

References

Milad Ebrahimi Abyazandi, Sherwin Darryl D'souza, Jeremy Greenwood, and Jason Ives. 2026. V-BReE. https://https://github.com/JasonIves/V-BReE.

Mahak Agarwal and Divyam Khanna. 2025. When Persuasion Overrides Truth in Multi-Agent LLM Debates: Introducing a Confidence-Weighted Persuasion Override Rate (CW-POR). *Preprint*, arXiv:2504.00374.

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.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *Preprint*, arXiv:2305.14325.

Andrew Estornell and Yang. 2024. Multi-LLM Debate: Framework, Principals, and Interventions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Preprint*, arXiv:2303.17760.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving Multi-Agent Debate with Sparse Communication Topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *Preprint*, arXiv:2305.19118.

Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2024. Refining LLMs Outputs with Iterative Consensus Ensemble (ICE). *medRxiv*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains. *Preprint*, arXiv:2501.05707.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024a. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? *Preprint*, arXiv:2402.18272.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as We Do, Not as You Think: the Conformity of Large Language Models. *Preprint*, arXiv:2501.13381.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *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.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. Peacemaker or Troublemaker: How Sycophancy Shapes Multi-Agent Debate. *Preprint*, arXiv:2509.23055.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. Conformity in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3854–3872, Vienna, Austria. Association for Computational Linguistics.

A Comparative accuracy of V-BReE and constituent models by domain

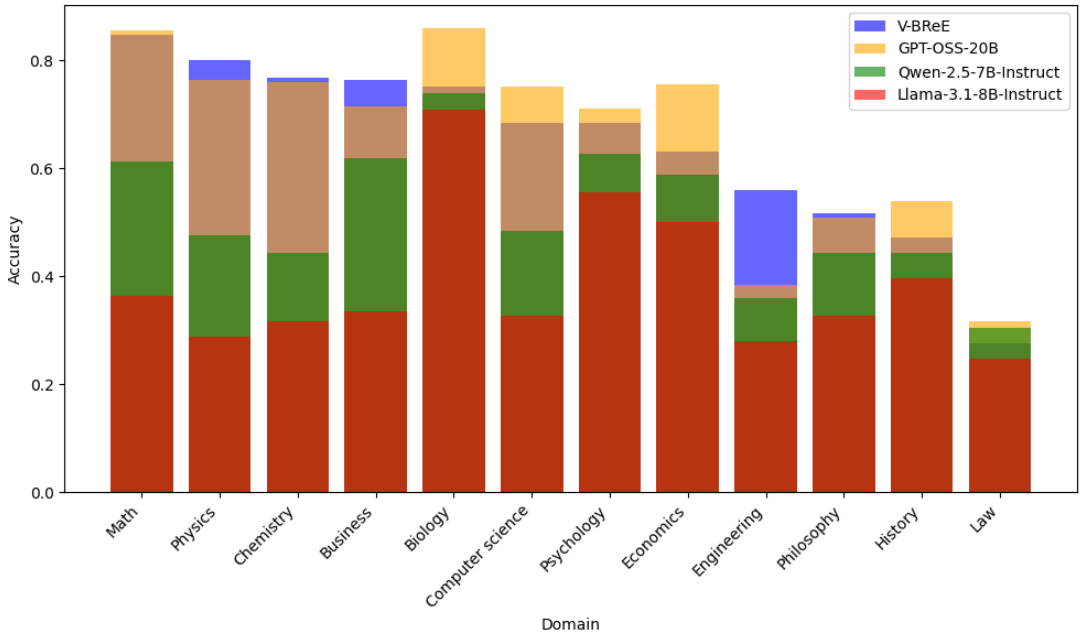

Figure 5: Full-sample *(n = 3001)* comparative accuracy by domain, including the V-BReE framework and the constituent models used for testing.

B Error rates and accuracy differentials by domain

Domain	LaTeX Error Rate	Δ Accuracy (V-BReE vs. GPT)
Engineering	0.155	+0.176
Chemistry	0.085	+0.008
Math	0.056	-0.008
Business	0.052	+0.048
Physics	0.044	+0.036
Computer Science	0.038	-0.067
Biology	0.030	-0.108
Law	0.027	-0.040
Economics	0.025	-0.124
History	0.018	-0.068
Philosophy	0.013	+0.008
Psychology	0.012	-0.028

Table 4: Correlation Analysis of Formatting Friction and Performance Delta. Pearson $r = 0.79, p = 0.0024$. Error rates represent the frequency of invalid LaTeX strings per domain, while Δ Accuracy represents the performance shift of the V-BReE ensemble relative to the GPT-OSS-20B anchor.

C V-BReE accuracy under extended variance threshold (EVT) conditions

Domain	EVT Accuracy	Δ (vs. 0.500)
Business	0.700	+0.200
Biology	0.667	+0.167
Engineering	0.633	+0.133
Computer Science	0.633	+0.133
Math	0.633	+0.133
Chemistry	0.600	+0.100
Physics	0.600	+0.100
History	0.533	+0.033
Psychology	0.533	+0.033
Economics	0.500	0.000
Philosophy	0.467	-0.033
Law	0.367	-0.133

Table 5: Domain-Specific Reasoning Performance under Extended Variance Threshold (EVT). Results represent the accuracy on a balanced $N = 360$ subsample (Baseline = 0.500). Note the clear performance divergence between objective STEM/Life Sciences and subjective/interpretive domains.

D Example input format

```
Question: When small pieces of material are assembled into a larger piece, the combined surface area
Existing Answer: When small pieces of material are assembled into a larger piece, the combined
    surface area slightly decreases because the surfaces that were exposed on the individual pieces
    become internal and are no longer exposed.

Possible Choices:
A. greatly increases
B. slightly increases
C. is halved
D. becomes zero
E. greatly decreases
F. is quadrupled
G. is doubled
H. is unchanged
I. slightly decreases
J. decreases

Task:
1. Rate the answer from 0 to 100 by aggregating scores for clarity (0-30 points), completeness (0-30
    points), and accuracy (0-40 points). Empty responses should receive the minimum score.
2. Refine the answer for maximum clarity, completeness, and accuracy. Remove filler and omit feedback
     or references to the original version. Restrict your response to 600 tokens. If no improvements
     are possible, provide the text verbatim.
3. Based on your updated response, identify the single letter from the possible choices provided
    corresponding to the best answer.
```

E Example response format


```
{
  "score": 65,
  "response": When small pieces of material are joined to form a larger piece, the surfaces that were
       exposed on the individual pieces become internal and are no longer exposed. Consequently, the
       total exposed surface area of the assembled object is slightly less than the sum of the
      surface areas of the separate pieces.",
  "letter": "I"
}
```