The Flip-Test: A Behavioral Diagnostic for Multi-Agent Debate Failures

Julia Hu **Alfred Shen** **Kumar Lakshmipathi**
Amazon Web Services
{juliahu, alfreshe, laksku}@amazon.com

Abstract

Multi-agent debate has been proposed to improve language model reasoning through adversarial interaction, yet empirical studies report that debate underperforms simpler baselines on open-ended reasoning (Choi et al., 2025). We propose a behavioral diagnostic—the *flip-test*—that instruments debate at the example level and measures how often, and in which direction, agents change their answers between initial and final rounds. On 2,417 MuSiQue examples with each of Llama 3.1 8B and Ministral 3 8B, flips are common (65–70% of debates). Our central result conditions on *opportunity*: a flip destroys an available correct answer $\sim$2.4$\times$ more often than it creates one (48.9% vs. 20.3% on Llama, 44.2% vs. 18.0% on Ministral; $p < 10^{-20}$). Unlike a raw directional split, this measure is invariant to how many correct answers were available, and it is statistically indistinguishable across the two model families. Three instruction-level interventions on a 100-example held-out set produce no improvement, suggesting simple prompt-level patches are insufficient. We separate this behavioral finding from mechanism: direct opponent-adoption is low ($\sim$10%), so the asymmetry is consistent with sycophancy but does not establish it. The signature is also scale-dependent—on three 70B-class models the flip rate roughly halves and debate turns net-positive—placing the harmful regime in the 8B capacity band rather than in debate itself. We frame the flip-test as a cheap behavioural probe for locating that regime.

1 Introduction

Multi-agent debate has attracted research interest as an inference-time mechanism for improving large language model (LLM) reasoning (Irving et al., 2018; Du et al., 2024; Liang et al., 2024). The motivating intuition is that agents producing diverse initial answers and critiquing one another should propagate evidence and converge on the correct answer more often than a single chain of reasoning. Recent empirical work complicates this picture: Choi et al. (2025) report that majority voting accounts for most of the gain attributed to debate, and adaptive routing (Eo et al., 2025) confirms debate is harmful on a substantial fraction of examples.

We address a specific gap: a *behavioral* description of where debate fails. We instrument the debate process and record, on each example, whether the endorsed answer moved toward correctness or away from it. We define a *flip* operationally—an agent's initial answer differs from the moderator's final answer—and, rather than scoring the raw direction of flips (which a difference in base rates can bias), we condition on whether a correct answer was available to lose, giving a low-cost diagnostic that assumes no particular causal mechanism.

Contributions. (1) The flip-test: a per-example diagnostic that decomposes debate behavior into a flip rate and an *opportunity-matched* destroy-vs-create comparison, the latter designed to avoid the base-rate bias that inflates or deflates raw directional counts. (2) On 2,417 MuSiQue examples per model, conditional on a flip debate destroys an available correct answer $\sim$2.4$\times$ more readily than it creates one ($p < 10^{-20}$ on both), an effect of equal magnitude across two independently developed model families (+28.6 vs. +26.2 pp). (3) A held-out study ($n = 100$) finds three instruction-level interventions all at or below baseline, indicating prompt-level patches are insufficient. (4) The signature reverses at 70B scale (§4.5), locating the harmful regime in the 8B capacity band. Direct opponent-adoption is low ($\sim$ 10%), so we read the asymmetry as consistent with sycophancy, not a proof of it.

Proceedings of the Seventh Workshop on Insights from Negative Results in NLP (Insights 2026), pages 70–75
October 29, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Multi-agent debate. Du et al. (2024) report that debate improves factual accuracy when initial answers are diverse and incorrect in different ways. Liang et al. (2024) introduce a divergence-encouraging variant after observing that LLMs "lock in" to a single solution. These works document conditions under which debate succeeds; we examine why it fails.

Voting and self-consistency. Wang et al. (2023) show that sampling reasoning paths and taking the majority recovers most prompting gains. Choi et al. (2025) extend this to debate: voting accounts for "most of the performance gains attributed to multi-agent debate." Our diagnostic complements this finding by showing that debate's marginal contribution is limited by frequent, accuracy-reducing answer changes—a pattern consistent with, but not proof of, sycophantic interaction.

Adaptive routing. Eo et al. (2025) activate debate only when a cheaper baseline is uncertain. Their finding that most examples do not benefit from debate aligns with our observation that flips are net-harmful; we explain *why* the harmful examples exist.

Sycophancy. Sharma et al. (2024) show LLMs systematically defer to user assertions, including incorrect ones, and that RLHF reinforces this. Our flip asymmetry is one of several findings that would be consistent with sycophancy operating in agent-to-agent interaction; we discuss in §5.2 why the present data do not establish this directly.

Concurrent analyses of debate failure. A parallel line of recent work dissects debate breakdowns. Hao et al. (2026) decompose stance convergence and argue that not all answer changes reflect conformity, distinguishing principled revision from social pressure. Wynn et al. (2025) catalogue failure modes and show that inter-agent communication can actively degrade performance. Bertalanič and Fortuna (2026) find that isolated self-correction can outperform unguided homogeneous debate. Our contribution is complementary and methodological rather than taxonomic: where these works characterize *what* goes wrong, the flip-test supplies a cheap, per-example *instrument*—a flip rate plus an opportunity-matched destroy-vs-create rate that any practitioner can compute on their own workload without ground-truth-labeled failure categories, and which we use here to expose a scale-dependent reversal (§4.5) that a purely qualitative taxonomy would miss. We also contribute the negative result that instruction-level patches do not close the gap (§4), consistent with Bertalanič and Fortuna (2026)'s finding that guidance from within the debate is limited, and with evidence that LLMs struggle to self-correct reasoning without external signal (Huang et al., 2024; Madaan et al., 2023).

3 Methods

Datasets. We evaluate on the MuSiQue-Ans development set (Trivedi et al., 2022), treating all 2,417 multi-hop questions as our primary cell, and on GSM8K (Cobbe et al., 2021) (1,319 test items) as a consistency check. Correctness on MuSiQue is token F1 ≥ 0.5; on GSM8K it is exact numeric match.

Models. Two instruction-tuned 8B-class models served via Amazon Bedrock: Llama 3.1 8B Instruct (Meta) and Ministral 3 8B Instruct (Mistral AI). They differ in pretraining corpus, RLHF pipeline, and developer; they are not chosen as a scaling axis.

Protocols. **Greedy** (1 call): temp 0, 960 tokens. **Vote3** (3 calls): temp 0.7, 320 tokens each, majority vote. **Debate** (7 calls): two agents at temp 0.7 generate initial answers (192 tokens each), exchange a rebuttal round (128 each), exchange a closing round (64 each), and a temp-0 moderator (192 tokens) returns the final answer. All protocols share a 960-token ceiling; realized medians are $\sim$100 (greedy), $\sim$160 (vote3), $\sim$500 (debate).

Flip definition. For each agent $a \in \{A, B\}$, we compare the agent's initial round-1 answer $a_0(x)$ to the moderator's final answer $a_f(x)$ (lowercased, punctuation-stripped, whitespace-collapsed exact match). An example is *flipped* if at least one agent flips. Because the moderator's output is the protocol's actionable answer, this is an *endorsement-level* flip—the protocol stops endorsing an agent's initial answer—rather than a literal within-agent self-revision (§5.2). A flipped example is correct→wrong if at least one initial answer matched gold and the final is wrong, wrong→correct if no initial matched but the final is correct, and otherwise "stay-class." Directional percentages are over correct→wrong $\cup$

wrong→correct only.

Statistics. For each model we report the no-flip vs. flip accuracy gap with a bootstrap 95% CI (10,000 resamples, seed 42). Our primary directional statistic is *opportunity-matched* (below); we report the raw $50/50$ binomial only as a descriptive reference.

Why the raw $50/50$ null is the wrong test. A correct→wrong flip ("destroy") is possible only when ≥ 1 initial answer is correct; a wrong→correct flip ("create") only when neither is. These draw from different conditioning sets, so a $50/50$ null is one the design already violates—and the base rate runs *with* the asymmetry: for two agents at accuracy p, a correct answer is available with probability $1-(1-p)^2 > 0.5$ whenever $p > 29.3\%$, well below the observed greedy $40.3\%/55.6\%$. The raw majority could thus be a pure opportunity artifact. We instead compare the *destroy rate* $\Pr(\text{final wrong} \mid \geq 1 \text{ correct, flipped})$ to the *create rate* $\Pr(\text{final correct} \mid \text{none correct, flipped})$; their ratio and a two-proportion test are invariant to how many correct answers were available.

Caveat: no-flip ≡ initial agreement by construction. Under exact match, a no-flip example is one where neither initial answer differs from the final, which entails the two initials agree; so the flip rate is essentially one minus the initial-disagreement rate. (Empirically $816/840$ and $705/718$ rather than the definitional $840/840/718/718$; the few exceptions are normalization mismatches, not genuine agree-then-flip events.) Hence the no-flip/flip accuracy gap mostly re-describes that agreement correlates with easiness (Du et al., 2024); we treat it as descriptive. This is exactly why our headline conditions on *opportunity*, not flip status—the destroy-vs-create comparison lives *within* the flipped set, holding correct-answer availability fixed.

4 Results

4.1 Flips correlate with accuracy loss

Figure 1 reports the descriptive gap. On Llama, debate accuracy is 66.1% [62.8, 69.2] without a flip and 31.0% [28.8, 33.3] with a flip; Ministral: 75.3% [72.1, 78.4] vs. 33.7% [31.5, 36.0]. Because no-flip is initial agreement by construction (§3), we read this gap as *easy-vs-hard* rather than as a causal effect of suppressing flips, and we do not rest any mechanistic claim on it. Notably, most of the cross-model difference lives in the no-flip cell (75.3 vs. 66.1, +9.2 pp) while the flip-conditional accuracies are nearly equal (33.7 vs. 31.0), i.e. the two models differ mainly in baseline ability on easy items, not in how badly flips hurt. The causal headline is instead the opportunity-matched result in §4.2.

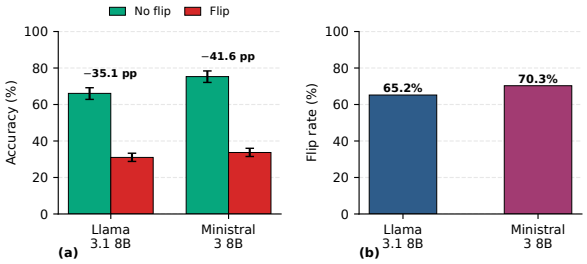

Figure 1: (a) Accuracy conditional on whether at least one agent flipped its answer during debate. Error bars are bootstrap 95% CIs over the two (disjoint) example groups. Because no-flip ≡ initial agreement (§3), this gap is descriptive—it contrasts easy agreed examples with harder disagreed ones—and is *not* our causal headline; see §4.2. (b) Flip rate on MuSiQue: 65.2% (Llama) and 70.3% (Ministral).

4.2 Conditional on a flip, debate destroys correct answers ∼2.4× more than it creates them

We compare the destroy and create rates defined in §3 (Figure 2). When a flip has a correct answer to lose, debate loses it $269/550 = 48.9\%$ of the time on Llama; when it has none to gain, it gains one only $208/1026 = 20.3\%$ of the time—a $2.41\times$ ratio. Ministral is nearly identical: 44.2% $(313/708)$ vs. 18.0% $(178/991)$, $2.46\times$. A two-proportion test separates the two rates decisively on both models ($z \approx 11.8$, $p < 10^{-20}$): *conditional on changing its endorsed answer, debate is ∼2.4× more likely to abandon a correct answer than to reach one.* The gap is the same size on both families (destroy−create = +28.6 vs. +26.2 pp), so it is a shared property, not a quirk of one model.

The non-directional majority. Directional flips are the minority: ∼70% of flips are *stay-class* (the endorsed answer changes string but not correctness). These carry no accuracy signal—hence their exclusion from the destroy/create rates—but their prevalence shows most endorsement changes are churn, not error correction, driving debate's to-

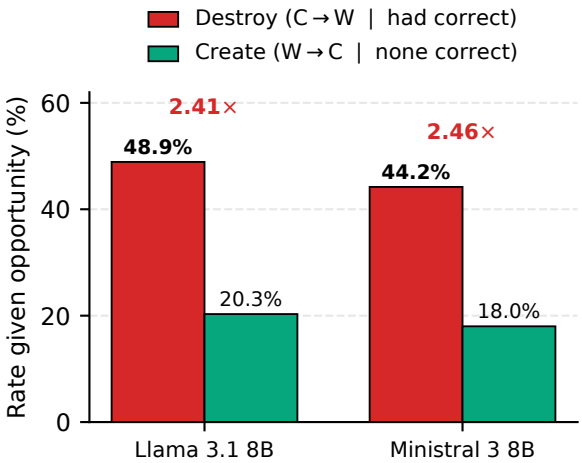

Figure 2: Opportunity-matched flip outcomes. *Destroy rate* = $\Pr(\text{final wrong} \mid \geq 1 \text{ initial correct})$; *create rate* = $\Pr(\text{final correct} \mid \text{no initial correct})$, among flipped examples. Debate destroys an available correct answer $\sim$2.4$\times$ more readily than it manufactures one on both models ($p < 10^{-20}$, two-proportion test). Once matched for opportunity the two models are statistically indistinguishable (+28.6 vs. +26.2 pp), i.e. Ministral's larger *raw* lean is just its having more correct answers to lose.

ken cost without accuracy return (§4.4).

4.3 Instruction-level interventions: a small held-out study

To probe whether prompt-level patches can reduce flipping, we ran a small held-out study on $n = 100$ MuSiQue examples with Llama 3.1 8B. This held-out subset has a baseline debate accuracy of 40%, slightly below the full-set 43.2% reported in Figure 3(a); the two figures refer to different sample sizes ($n = 100$ vs. $N = 2{,}417$) and should not be expected to coincide. All intervention deltas below are relative to the $n = 100$ baseline. **Anchoring** added: "Only change your answer if you identify a fundamental flaw in your original reasoning" (40% $\rightarrow$ 40%). **Justification-on-flip** required each agent to articulate the specific flaw before changing answers (40% $\rightarrow$ 39%). **Moderator review** asked the moderator to reject flips that lacked substantive justification (40% $\rightarrow$ 38%). None of the three reached baseline. With $n = 100$, individual differences are within sampling noise; the pattern that all three stayed at or below baseline is suggestive, not definitive, and a stronger conclusion would require broader prompt-engineering studies.

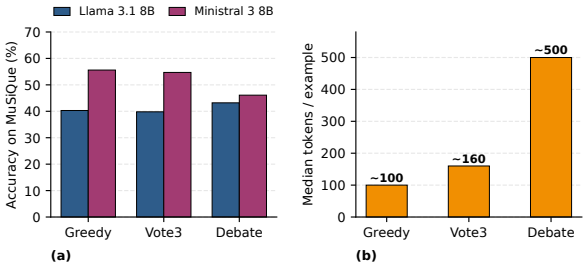

Figure 3: (a) Protocol accuracy on MuSiQue. Debate beats greedy on Llama by 2.9 pp but loses to greedy on Ministral by 9.5 pp; debate is not a uniformly best fixed protocol. (b) Median tokens generated per example. Debate consumes roughly 5$\times$ greedy and 3$\times$ vote3.

4.4 Protocol comparison and cost

Debate is dominant on Llama-MuSiQue (43.2% vs. greedy 40.3%) but is dominated by greedy on Ministral-MuSiQue (46.1% vs. 55.6%, Figure 3a). This dataset/model dependence rules out treating debate as a universal default. Crucially, debate's token cost is roughly 5$\times$ greedy and 3$\times$ vote3 (Figure 3b), so any small accuracy gain comes at substantial expense.

GSM8K consistency check. The flip-test reproduces qualitatively on GSM8K: flip-conditional accuracy is far below no-flip (Llama: 49.5% vs. 88.1%; Ministral: 57.6% vs. 93.4%) and directional flips lean correct$\rightarrow$wrong on both models ($\sim$77%). We do not foreground GSM8K because arithmetic and multi-hop QA differ in normalization and ambiguity, but the pattern is consistent.

4.5 The flip signature reverses at 70B scale

The most common question about an 8B-only result is whether it is an artifact of capacity (§5.1). To probe this directly, we ran the identical MuSiQue protocol on three 70B-class models—Llama 3.1 70B, Llama 3.3 70B, and Mistral Large—and applied the same flip-test.

Two things change together at scale (Table 1). The flip rate roughly halves (65–70% $\rightarrow$ 29–40%), and the sign of debate's net effect flips: debate *beats* both baselines on all three 70B models (+6.9, +2.5, +1.5 pp over greedy), whereas at 8B it hurt Ministral by 9.5 pp. This is direct evidence that the harmful-flip regime is a property of the 8B capacity band, not of debate in general, favoring the non-monotonic account of §5.1. We do *not* over-claim a predictive rule from five cells: flip rate vs. the debate$-$greedy margin is $r = -0.62$ ($p = 0.26$, $n = 5$), so flip rate is a weak screen,

Model	Greedy	Vote3	Debate	Flip rate
8B-class (primary)				
Llama 3.1 8B	40.3	39.8	**43.2**	65.2%
Ministral 3 8B	**55.6**	54.7	46.1	70.3%
70B-class				
Llama 3.1 70B	57.2	56.7	**64.1**	39.0%
Llama 3.3 70B	60.8	61.7	**63.3**	39.9%
Mistral Large	46.7	46.6	**48.2**	29.2%

Table 1: Accuracy (%) and flip rate on MuSiQue ($N = 2{,}417$); best protocol per row in bold. Vote3 is included because Choi et al. (2025) attribute most of debate's gain to voting and debate costs $\sim$3$\times$ Vote3's tokens—yet debate never beats Vote3 by more than $\sim$2.5 pp, and at 8B Vote3 essentially ties Greedy. At 70B scale debate beats both baselines on all three models and the flip rate roughly halves, reversing the harmful 8B pattern.

not a decision procedure. (Accuracy and flip rate here reproduce exactly against the 8B cells; a harmonized directional breakdown at 70B is deferred to avoid mixing correctness conventions.)

5 Discussion

5.1 Are flips just because the models are small?

One explanation is the *weak-reasoning hypothesis*: 8B models are too weak to tell a correct answer from a plausible wrong one, so they oscillate at random. Random oscillation predicts equal destroy and create rates; the $\sim$2.4$\times$ asymmetry (§4.2) rules that out. What remains open is a *non-monotonic* capacity story: mid-capacity models may sense a shaky correct answer without being able to defend it, while stronger models leave the regime entirely. The 70B results (§4.5) fit that exit—flips grow rare and net-positive—but two 8B points cannot separate a genuine mid-capacity peak from a smooth capacity trend, so we do not claim one.

5.2 What the asymmetry does and does not show

A natural interpretation is sycophantic capitulation (Sharma et al., 2024): agents abandon correct answers under confident wrong opposition. To probe this we measured the *adoption rate*—among initially-disagreeing examples, how often the final wrong answer normalize-equals the opponent's initial wrong answer. It is only $\sim 10\%$ on both models; the remaining correct$\rightarrow$wrong finals are paraphrases or new wrong formulations. So the asymmetry is real but does not establish literal answer-adoption; the mechanism may instead be moderator synthesis bias, or initially-correct positions being less defensible on hard instances.

Endorsement-level, and what it does not isolate. A flip is defined at the *endorsement* level (initial answer differs from the moderator's final output), not as a within-agent revision. This mixes three processes the present logs cannot separate: (i) peer persuasion, (ii) moderator synthesis bias, and (iii) spontaneous instability a model would show under *any* re-prompting. Two absent controls would disentangle them—a *self-reflection* control (one agent reconsiders, no peer) bounding (iii), and a *moderator-free* variant (agents self-finalize) separating (i) from (ii). Lacking both, our rates are *upper bounds* on peer-induced change; adding these ablations is the highest-value next experiment.

5.3 Limitations and recommendation

Our study covers five models across two capacity tiers, one debate topology, and primarily MuSiQue; the intervention study is small (n=100); and the causal mechanism behind the asymmetry remains open (§5.2). We therefore position the flip-test as a behavioural instrument, not a protocol-selection rule: with only five cells the flip-rate/margin correlation is not significant, so we make no prescriptive claim. What the diagnostic does deliver is a cheap, opportunity-matched readout of whether debate is in its harmful regime on a given model, and evidence that this regime is real at 8B and recedes by 70B.

References

Blaž Bertalanič and Carolina Fortuna. 2026. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. *arXiv preprint arXiv:2605.00914*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving

factuality and reasoning in language models through multiagent debate. In *ICML*.

Sugyeong Eo, Hyeonseok Moon, Evelyn Hayoon Zi, Chanjun Park, and Heuiseok Lim. 2025. Debate only when necessary: Adaptive multiagent collaboration for efficient LLM reasoning. *arXiv preprint arXiv:2504.05047*.

Xiqi Hao, Zengqing Wu, Yu-Xuan Qiu, Chuan Xiao, Ruiqi Xu, Shuyuan Zheng, and Jianbin Qin. 2026. Not all flips are conformity: Decomposing stance convergence in multi-agent LLM debate. *arXiv preprint arXiv:2606.00820*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *ICLR*.

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

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. In *EMNLP*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *ICLR*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. In *TACL*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

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