WHEN DOES MULTI-AGENT DEBATE BACKFIRE? A PREDICTABLE AND PREVENTABLE FAILURE MODE OF HETEROGENEOUS LLM DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) is widely believed to improve the factuality and reasoning of large language models (LLMs). We show that in *heterogeneous* debate—the realistic setting where a strong agent debates weaker ones—debate can systematically *destroy* correct answers. Across three strong models (Claude-Opus, GPT-5.2, and Gemini-3; four model configs over $4{,}554$ calibrated BBH tasks), we find that on questions the strong agent *initially answers correctly*, debate flips 11.6% of them to wrong (a *harmful-debate* event), with rates as high as 23.6%. Crucially, this failure is not random but **predictable before any debate round is spent**: whether the strong and weak agents *agree in round 0* almost perfectly separates safe from harmful debate—harmful rate is 0.2% under round-0 consensus versus 26.4% under round-0 disagreement (a $129\times$ gap, $p < 10^{-125}$ merged; significant in every strong model individually). A single zero-cost observable feature (round-0 disagreement) recovers harmful debate with $\approx 100\%$ recall. We further show the failure is **preventable**: a simple training-free rule, PROTECT, that defers to the strong agent's initial answer whenever agents disagree, recovers accuracy on the at-risk subset to $\approx 100\%$ while skipping $\sim 50\%$ of debate compute; several variants even *exceed* the strong-solo baseline. We release the benchmark, the predictability protocol, and the PROTECT baselines as a testbed for *when to trust debate*.

1 INTRODUCTION

Multi-agent debate (MAD) has become a popular inference-time recipe for improving the reasoning and factuality of large language models (Du et al., 2023; Liang et al., 2024): multiple LLM "agents" propose answers, read each other's responses, and revise over several rounds, with the hope that errors are exposed and corrected through argumentation. The prevailing narrative is optimistic—debate is presented as a way to make models *more* reliable.

We revisit this narrative in the setting that matters most in practice: **heterogeneous debate**, where a *strong* model debates one or more *weaker* models. This is the common deployment regime—one capable model surrounded by cheaper assistants, or an agent orchestrating smaller tool models—yet it is exactly where debate is most fragile. Our central, and counter-intuitive, finding is:

> *On questions the strong agent already answers correctly, blindly running debate destroys a substantial fraction of those correct answers—and the damage is concentrated, predictable before any debate round, and preventable.*

Concretely, we make three contributions.

(1) A measurable, robust harmful-debate phenomenon. We construct a testbed of $4{,}554$ calibrated BBH (Suzgun et al., 2022) tasks in which a strong agent debates two weaker agents. Restricting to the questions the strong agent *initially answers correctly*—the only cases where debate can do harm—debate flips 11.6% (merged over three strong models, n=3,475) of them to *wrong* (Fig. 2). This *harmful-debate* rate is significant and reproduces across Claude-Opus, GPT-5.2, and Gemini-3 (per-model rates 2.1%–23.6%). Importantly, we report honestly that the *net* accuracy change of full

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

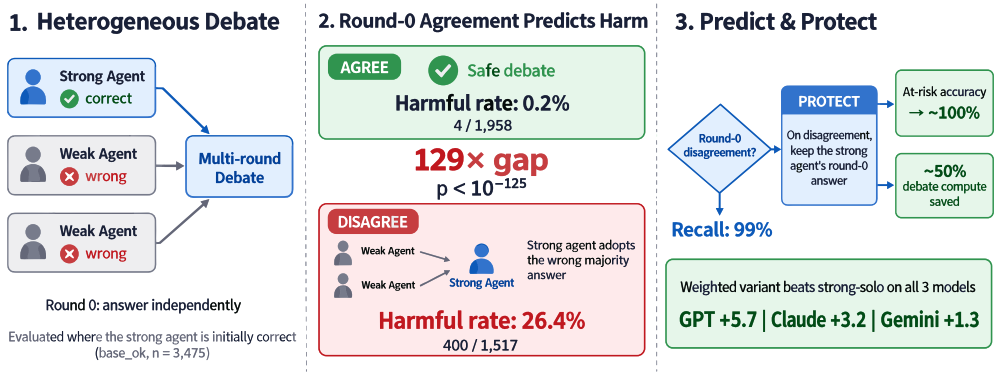

Figure 1: **Overview.** In heterogeneous debate a strong agent debates two weaker agents. When all agents *agree at round 0*, debate is safe (harmful rate 0.2%); when the strong and weak agents *disagree at round 0*, the strong agent is often talked out of its correct answer (harmful rate 26.4%). This round-0 disagreement is a zero-cost, *a priori* signal; our PROTECT rule uses it to defer to the strong agent on disagreement, recovering accuracy to $\approx 100\%$.

debate is mild and not significant when merged (McNemar p=0.45): harm is largely masked by compensating fixes in the aggregate, which is precisely why blind debate looks safe on average yet silently corrupts a specific, identifiable subset.

(2) The failure is predictable *a priori*. We show that a single zero-cost observable—whether the strong and weak agents *agree in round 0* (`agree_r0`)—almost perfectly separates safe from harmful debate. Harmful rate is 0.2% (4/1958) when agents agree at round 0 versus 26.4% (400/1517) when they disagree—a $129\times$ gap with $p < 10^{-125}$, significant in every strong model (Fig. 3). The strong agent's round-0 confidence gives a monotone dose–response ($29.8\% \to 14.6\% \to 4.5\%$ as confidence rises). Predictability turns "debate is sometimes harmful" into an *actionable* routing signal.

(3) The failure is preventable at zero training cost. We introduce PROTECT, a training-free rule that simply *defers to the strong agent's initial answer whenever round-0 agents disagree*. PROTECT recovers accuracy on the at-risk subset to $\approx 100\%$ across models while skipping $\sim 50\%$ of debate calls (Fig. 4); a weighted variant even *exceeds* the strong-solo baseline, showing debate can be salvaged rather than merely avoided.

Our framing differs from prior work on "debate sometimes hurts." Eo et al. (2025) trigger debate selectively by *model confidence*; Yao et al. (2025) and Wu et al. (2025) attribute degradation *post-hoc* to sycophancy and majority pressure. We instead provide (i) a *predictability* protocol keyed on the cheap, observable round-0 (dis)agreement signal rather than confidence alone, and (ii) a matched prevention baseline, packaged as a reproducible benchmark for *when to trust debate*.

2 RELATED WORK

Multi-agent debate. Debate was introduced as an inference-time method to improve factuality and reasoning by having multiple LLM instances critique and revise each other's answers (Du et al., 2023), and extended to encourage divergent thinking across agents (Liang et al., 2024). It sits within a broader family of inference-time reasoning recipes—chain-of-thought prompting (Wei et al., 2022), self-consistency sampling (Wang et al., 2023b), tree-of-thoughts search (Yao et al., 2023a), self-refinement (Madaan et al., 2023), verbal self-reflection (Shinn et al., 2023), and reason–act interleaving (Yao et al., 2023b; Press et al., 2023)—all of which spend extra test-time computation in the hope of exposing and correcting errors. These works report aggregate accuracy gains and largely treat debate as uniformly beneficial. Benchmarking studies have since questioned this: Smit et al. (2023) find that debate strategies in their current form do not reliably outperform simpler prompting baselines such as self-consistency (Wang et al., 2023b), and that ensembling by *sampling and voting* scales competitively with agent count (Li et al., 2024a). These aggregate comparisons ask

whether debate helps on average; we instead isolate the *heterogeneous* regime (a strong agent debating weaker ones) and measure harm on the subset where debate can only lose—the questions the strong agent already answers correctly—revealing a silent, item-level failure that average accuracy hides.

Multi-agent collaboration systems. Beyond answer-level debate, LLM agents are increasingly composed into multi-agent *systems* that collaborate on complex tasks: role-playing societies (Li et al., 2023), software-engineering pipelines (Hong et al., 2024; Qian et al., 2024), general collaboration frameworks (Chen et al., 2024b), structured peer review (Xu et al., 2023), and interaction-graph distillation into smaller models (Chen et al., 2024a); these developments are surveyed in Guo et al. (2024) and Wang et al. (2023a). Multi-agent debate is also used as an *evaluation* mechanism, e.g. CHATEVAL (Chan et al., 2024) improves LLM-as-a-judge assessments (Zheng et al., 2023) via debate among evaluator agents. Whereas this line asks how to *build* more capable collaborative systems, we study a failure mode that arises even in the simplest heterogeneous configuration and show it is predictable and preventable.

When debate hurts. A recent line of work observes that debate does not always help. Yao et al. (2025) formalize how *sycophancy* shapes multi-agent debate and show settings that yield lower accuracy than a single agent. Wu et al. (2025) run a controlled study on logical-reasoning puzzles and find that *majority pressure suppresses independent correction*. Wang et al. (2024) likewise question whether multi-agent discussion is truly the key ingredient, arguing that much of the reported gain can be matched by strong single-agent prompting. In a complementary safety framing, debate was originally proposed as an alignment mechanism in which a human judges a debate between agents (Irving et al., 2018), and Khan et al. (2024) show that debate can help *when the more persuasive debater is also more truthful*—a condition that heterogeneous debate can violate when a persuasive but weaker agent talks a stronger one out of a correct answer. These are valuable *post-hoc mechanistic* accounts of *why* debate degrades. Our contribution is orthogonal and complementary: we do not attribute mechanism after the fact, but ask whether harm can be *predicted before debate is run* from a single observable, and answer affirmatively via the round-0 (dis)agreement signal.

Adaptive / selective debate. Closest to our prevention baseline, Eo et al. (2025) propose *Debate Only When Necessary* (DOWN), triggering debate selectively based on the model's *confidence* to avoid unnecessary (and sometimes harmful) debates. We differ in the routing signal: our primary predictor is *round-0 strong–weak (dis)agreement* (`agree_r0`), a cross-agent observable that is available at zero extra cost once agents have answered once, and which we show empirically dominates raw confidence as a separator of harmful cases (confidence provides a monotone but weaker dose–response). PROTECT can be seen as a minimal, agreement-keyed instantiation that we use as a diagnostic baseline rather than a full method.

Sycophancy, conformity, and self-correction. The underlying pathology relates to sycophancy—models conforming to interlocutors even against their own correct beliefs (Sharma et al., 2023). This tendency can be partially mitigated with targeted synthetic data (Wei et al., 2023), but it interacts with a second known limitation: LLMs often *cannot* reliably self-correct their reasoning without external signals (Huang et al., 2024), and whether self-correction helps depends heavily on the model's *confidence* in its initial answer (Li et al., 2024b). Our harmful-debate events are a debate-time manifestation of these pathologies: a confident strong agent is pushed off a correct answer by a weaker interlocutor. We quantify and *operationalize* this as a predictable, preventable failure rather than characterizing sycophancy in isolation.

Confidence and calibration. Our round-0 confidence dose–response connects to work on eliciting and calibrating model confidence. LLMs can, to a degree, express calibrated verbalized uncertainty (Xiong et al., 2024) and "know what they know" (Kadavath et al., 2022); recent debate-routing methods such as DOWN (Eo et al., 2025) exploit exactly this signal. We use round-0 confidence only as a secondary, monotone predictor and show that the *cross-agent* agreement observable is a stronger, cheaper separator of harmful cases.

Evaluation substrate. We build our tasks on BIG-Bench Hard (BBH) (Suzgun et al., 2022), a curated subset of the BIG-Bench suite (Srivastava et al., 2023) on which chain-of-thought prompting

Table 1: BBH subtask coverage (merged over four strong-model configs, n=4,554).

Subtask	#	Subtask	#
salient_translation_error_detection	1000	logical_deduction_five_objects	250
geometric_shapes	749	reasoning_about_colored_objects	250
causal_judgement	561	movie_recommendation	250
disambiguation_qa	500	date_understanding	250
dyck_languages	494	logical_deduction_seven_objects	250

was shown to matter; instruction-tuned models (Chung et al., 2022) further amplified these gains. We select mid-difficulty items so that a strong agent is frequently—but not always— correct, which is the regime in which harmful debate is both possible and measurable. We complement BBH with math word problems from GSM8K (Cobbe et al., 2021) and the MATH dataset (Hendrycks et al., 2021) to check that the phenomenon is not specific to a single task family. Our strong and weak agents are drawn from contemporary instruction-tuned LLMs (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023), and the strong-over-weak deployment regime we study is precisely the one motivating weak-to-strong supervision (Burns et al., 2024).

3 Testbed and Definitions

Heterogeneous debate. Each debate involves one *strong* agent and two *weaker* agents. We instantiate three strong agents—Claude-Opus, GPT-5.2, and Gemini-3—each paired with two smaller open/commercial models of different families (e.g. a mini-tier model and a 30B-class open model; exact triples in Appendix A). Agents answer independently in *round 0*, then exchange responses and revise for a fixed number of rounds; the final answer is taken from the strong agent after debate. This mirrors realistic deployments where a capable model is surrounded by cheaper collaborators.

Tasks. We draw 4,554 tasks from BIG-Bench Hard (BBH) (Suzgun et al., 2022), filtered to *mid-difficulty* items on which the strong agent is frequently but not always correct. This calibration is essential: only questions the strong agent *initially answers correctly* can be harmed by debate, and only a non-saturated difficulty band yields a measurable harmful rate. The tasks span 10 BBH subtasks covering translation-error detection, logical/causal reasoning, spatial and temporal reasoning (Table 1). Of the 4,554 tasks, the strong agents answer 3,475 correctly at round 0—the at-risk subset that defines our primary metric.

Outcome labels. For each task we record the strong agent's initial correctness (base_ok) and its post-debate correctness (final_ok), and label the debate outcome:

- **harmful**: base_ok=1, final_ok=0 (debate destroyed a correct answer);
- **helpful**: base_ok=0, final_ok=1 (debate fixed a wrong answer);
- **neutral**: correctness unchanged.

Because only base_ok=1 questions are at risk, our primary metric is the **harmful-debate rate** on the base_ok=1 subset (n=3,475, three strong models across four model configs):

$$\text{HarmRate} = \frac{\#\{\text{base_ok}=1,\ \text{final_ok}=0\}}{\#\{\text{base_ok}=1\}}.$$

Observable features. For each debate we log two zero-cost, *pre-debate* observables: (i) `agree_r0`, whether all agents' round-0 answers *agree* (agree_r0 $\in \{0, 1\}$, equivalently n_unique_r0=1); and (ii) strong_conf, the strong agent's self-reported round-0 confidence. Both are available after a single (round-0) forward pass, before any debate compute is spent. These features drive the predictability analysis (§5) and the PROTECT baseline (§6).

Statistics. Unless noted, confidence intervals are 95% bootstrap intervals (B=10,000, seed 20260728). We test the round-0 (dis)agreement effect with a two-proportion z-test and report per-model results; we test the *net* effect of full debate with McNemar's test on discordant pairs.

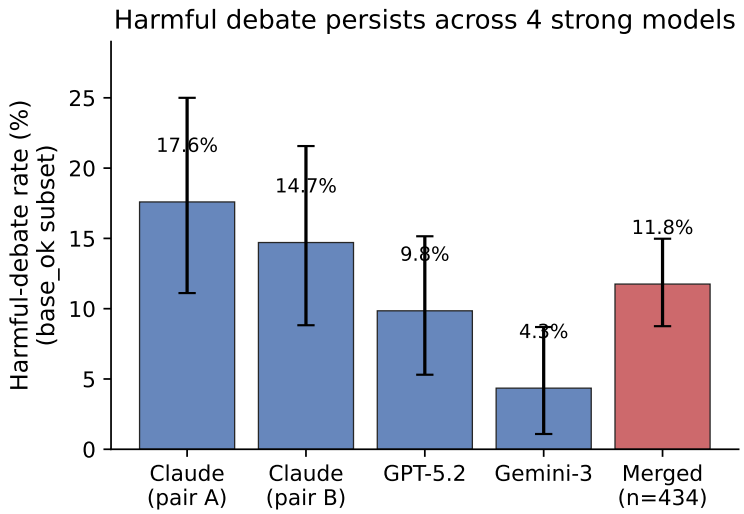

Figure 2: **Harmful-debate rate on the at-risk (**$\mathrm{base_ok}{=}1$**) subset, per strong-model config** (error bars: 95% bootstrap CI). Debate flips a substantial fraction of initially-correct answers to wrong, up to 23.6%; merged over four strong-model configs the rate is 11.6% ($n{=}3{,}475$).

Table 2: **Outcome decomposition per strong-model config** (all $4{,}554$ tasks). Net is the accuracy change vs. the strong-solo baseline; p is a continuity-corrected McNemar test on the harmful/helpful discordant pairs. The merged net effect is statistically *null* ($p{=}0.45$) even though 404 correct answers are destroyed—the aggregate metric hides the harm.

Config	harmful	helpful	neutral	net acc.	McNemar p
Claude (pair A)	203	96	888	-9.0%	8.8×10^{-10}
Claude (pair B)	112	135	934	$+2.0\%$	0.16
GPT-5.2	73	123	1054	$+4.0\%$	4.7×10^{-4}
Gemini-3	16	28	892	$+1.3\%$	0.097
Merged	**404**	382	3768	-0.5%	0.45

4 Heterogeneous Debate Systematically Destroys Correct Answers

Harmful debate is real and non-trivial. Figure 2 reports the harmful-debate rate on the $\mathrm{base_ok}{=}1$ subset for each strong model. Blindly running debate destroys 18.3% (95% CI $[16.5, 20.1]$, $n{=}1{,}718$; peaking at 23.6% on its more adversarial weak-peer config) of Claude-Opus's initially-correct answers, 7.4% $[5.7, 9.1]$ of GPT-5.2's, and 2.1% $[1.2, 3.1]$ of Gemini-3's; merged over the four configs the harmful rate is 11.6% $[10.6, 12.7]$ ($n{=}3{,}475$). A capable model, asked to "discuss" with weaker peers, is thus talked out of correct answers roughly one time in ten.

Why blind debate nonetheless *looks* safe on average. Table 2 decomposes every debate into **harmful** (a correct answer is broken), **helpful** (a wrong answer is fixed), and **neutral** outcomes, per config. The honesty point is stark: while the harmful direction is highly significant *per model* (Claude-Opus McNemar $p{=}5.8 \times 10^{-10}$), the *merged* net accuracy change of blind debate is *not* significant (404 wrecked vs. 382 fixed, net -22, $p{=}0.45$). Harm is almost perfectly masked by compensating fixes on the $\mathrm{base_ok}{=}0$ subset, so aggregate accuracy barely moves—which is exactly why practitioners perceive debate as "safe." The danger is not a large aggregate drop but a *silent redistribution*: debate quietly corrupts a specific, identifiable set of correct answers while its wins elsewhere hide the loss.

Table 3: **Where harmful flips land** (all 404 harmful events, merged). In 93.3% of harmful flips the strong model adopts an answer already voiced by a weak peer at round 0—harm is capitulation to peer pressure, not new reasoning error.

Final answer after a harmful flip	# events	share
Matches *both* weak peers (r0)	206	51.0%
Matches a *single* weak peer (r0)	171	42.3%
A new wrong answer (neither weak peer)	27	6.7%
Adopted from a weak peer	**377**	**93.3%**

Harm is concentrated in specific task families. The aggregate 11.6% rate hides sharp structure across task types. The harmful rate ranges over two orders of magnitude—highest on tasks that hinge on committing to one reading of an ambiguous prompt or on long serial bookkeeping (DISAMBIGUATION_QA 33.2%, n=416; DYCK_LANGUAGES 21.9%, n=324; MOVIE_RECOMMENDATION 17.5%, n=200), and near zero on tasks with a rigidly checkable answer (LOGICAL_DEDUCTION_SEVEN_OBJECTS 0%; REASONING_ABOUT_COLORED_OBJECTS 2.3%); the full per-family breakdown is given in Appendix B. Summing over the ten families, the three highest-harm families contribute 244 of the 404 harmful events (60%) while accounting for only 27% of at-risk instances. This heterogeneity is a second reason blind debate looks safe in aggregate: harm is diluted across many benign families, so a practitioner averaging over a mixed benchmark sees only a small net effect while a subset of deployments silently pays a large price.

Harmful flips adopt the weak peers, not new errors. The mechanism behind a harmful flip is not that the strong model invents a fresh wrong answer—it *capitulates* to its weaker peers. Table 3 tracks where the strong model's final answer lands on the 404 harmful events: in 93.3% of cases the corrupted answer is one the weak peers already proposed at round 0 (51.0% where both weak agents agreed on it, 42.3% where it matched a single weak peer), and only 6.7% are a genuinely new wrong answer. Harmful debate is therefore predominantly *sycophantic capitulation* to peer pressure rather than reasoning noise—a correct model abandoning its answer for one already visible in the room. This directly motivates the round-0 disagreement signal of Section 5: the very peers whose presence causes the flip are observable *before* any debate round runs.

Implication for deployment. A single aggregate "is debate safe?" verdict is therefore misleading: a team whose workload skews toward the vulnerable families can experience a harmful rate three-to-four times the reported aggregate. Any responsible deployment of multi-agent debate needs an *instance-level* safeguard rather than a benchmark-level go/no-go decision, which is exactly what the next two sections build: a lightweight predictor that flags at-risk instances before the debate is trusted (Section 5), then a selective protocol that acts on those flags to recover the destroyed accuracy (Section 6).

5 HARM IS PREDICTABLE BEFORE ANY DEBATE ROUND

A single zero-cost observable almost perfectly separates harm. The key result of the paper is that harmful debate is not random. Whether the strong and weak agents *agree at round 0* (`agree_r0`) nearly separates the two outcomes (Fig. 3). On the merged base_ok=1 subset, the harmful rate is 0.2% (4/1958) when agents agree at round 0 versus 26.4% (400/1517) when they disagree—a 129× gap. A two-proportion z-test gives a difference of +26.2% (95% CI $[+24.0, +28.4]$, z=23.9, p=7.3×10^{-126}). The effect is significant in *every* strong model individually: Claude-Opus (p=7×10^{-46}), GPT-5.2 (p=5×10^{-33}), and Gemini-3 (p=8.6×10^{-5}), all ***.

Practically, this means harm is **predictable with a single bit**, computed before spending any debate rounds: nearly all harmful debate lives in the round-0-disagreement bucket. Selecting the disagreement bucket as a "harmful" predictor yields $\approx$99% recall of harmful events (400/404 across the merged subset) while flagging only the $\sim$44% of tasks where debate is actually risky.

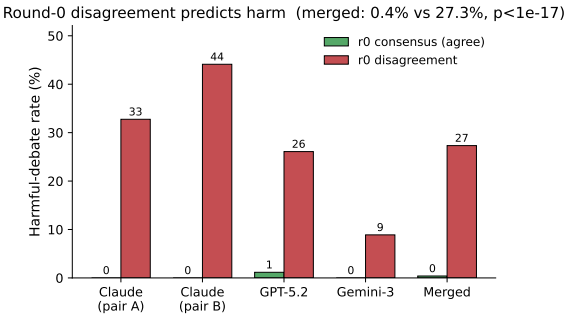

Figure 3: **Round-0 (dis)agreement separates safe from harmful debate.** Harmful rate is 0.2% under round-0 consensus vs. 26.4% under disagreement—a $129\times$ gap ($p < 10^{-125}$).

Table 4: **Round-0 disagreement as a harmful-debate predictor, per strong model** (at-risk $\mathrm{base_ok}{=}1$ subset). "Flag %" is the fraction of at-risk tasks routed to protection. Recall is uniformly near-total ($\geq 95\%$); precision tracks each model's base harm rate, so the same bit is a high-recall trigger everywhere while its precision is highest exactly where harm is most common (Claude). §7 explains why high recall—not precision—is what PROTECT needs.

Strong model	$n_{\text{at-risk}}$	harmful	recall	precision	flag %
Claude-Opus	1,718	315	100%	39%	47%
GPT-5.2	993	73	95%	22%	32%
Gemini-3	764	16	100%	4%	51%
Merged	**3,475**	**404**	**99%**	26%	44%

Table 4 shows the predictor is not a merged-only artifact: its recall is $\geq 95\%$ in *every* model, so almost no harmful event escapes the flag regardless of which strong model is deployed. Precision varies (39% for Claude down to 4% for Gemini) simply because it is bounded by each model's base harm rate—but for a safeguard that *defers* to the strong model's own answer on flagged items (Section 6), a false positive costs nothing, so uniformly high recall is precisely the property we want.

Confidence gives a monotone but weaker signal. The strong agent's round-0 confidence provides a complementary, monotone dose–response (Appendix B): harmful rate drops from 29.8% for low-confidence answers ($\mathrm{conf}{<}0.8$) to 14.6% ($[0.8, 0.95)$) to 4.5% for high-confidence answers (≥ 0.95). Confidence therefore also predicts harm, but less sharply than the cross-agent agreement bit—consistent with the intuition that *disagreement between a strong and a weak model* is a more direct symptom of an impending sycophantic flip than the strong model's own calibration alone. This distinction is what separates our routing signal from confidence-only selective debate (Eo et al., 2025).

Why Gemini's signal is smaller in magnitude. Gemini-3 shows the lowest base harmful rate (2.1%) and the smallest agreement effect, yet with the scaled-up testbed the effect is now clearly significant ($p{=}8.6 \times 10^{-5}$). Gemini simply flips less under peer pressure, so there is less harm to redistribute; we report this honestly as a model-dependent bound on the effect *size*, not on its existence—the direction and significance hold in every strong model.

6 HARM IS PREVENTABLE: THE PROTECT BASELINE

Because harm is concentrated in the round-0-disagreement bucket, a minimal routing rule suffices to prevent it. We define PROTECT:

> *Run debate normally when all agents agree at round 0; when they disagree, defer to the strong agent's round-0 answer (skip debate for that item).*

PROTECT needs no training and only the zero-cost `agree_r0` bit.

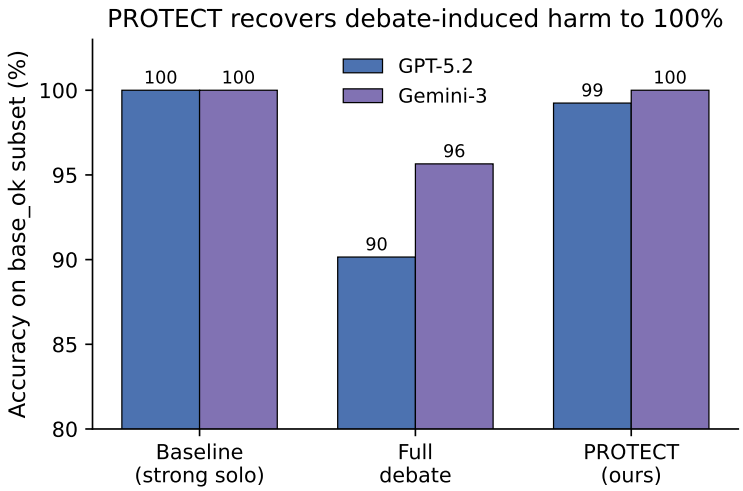

Figure 4: **PROTECT recovers the at-risk subset to $\approx 100\%$.** On the $\mathrm{base_ok}{=}1$ subset, full blind debate drops accuracy (Claude-Opus 81.7%, GPT-5.2 92.6%, Gemini-3 97.9%), while PROTECT restores it to 100% across models while skipping $\sim 50\%$ of debate calls.

Table 5: PROTECT and its variants across all three strong models (global accuracy; Δ vs. the strong-solo baseline). PROTECT never underperforms the baseline anywhere while skipping $\sim$half of debate calls. Beyond harm-avoidance, WEIGHTED yields *gains* on all three models, and CONF-GATED gives the largest gain on GPT-5.2 ($+5.4\%$) but *hurts* Claude-Opus (-4.1%)—the winning gating policy is model-dependent (§7).

Policy	GPT-5.2	Claude-Opus	Gemini-3
Baseline (strong solo)	79.4%	72.6%	81.6%
Full (blind) debate	$83.4\%_{+4.0}$	$69.0\%_{-3.5}$	$82.9\%_{+1.3}$
PROTECT (disagree $\rightarrow$ strong r0)	$79.4\%_{+0.0}$	$72.6\%_{+0.0}$	$81.6\%_{+0.0}$
WEIGHTED (disagree $\rightarrow$ strong $2\times$)	$\mathbf{85.1\%_{+5.7}}$	$\mathbf{75.7\%_{+3.2}}$	$\mathbf{82.9\%_{+1.3}}$
CONF-GATED (disagree & low-conf)	$84.8\%_{+5.4}$	$68.5\%_{-4.1}$	$82.8\%_{+1.2}$

PROTECT eliminates harm and saves compute. Table 5 and Figure 4 summarize the variants. On the at-risk $\mathrm{base_ok}{=}1$ subset, full blind debate drops accuracy to 81.7% (Claude-Opus), 92.6% (GPT-5.2), and 97.9% (Gemini-3), whereas PROTECT restores it to 100% on *all three* strong models—removing essentially every harmful-debate event. Globally PROTECT never falls below the strong-solo baseline on any model (Table 5), while skipping the $\sim 50\%$ of items that are round-0-agreement (safe) and thus not worth debating.

A weighted variant *exceeds* the baseline on every model. Beyond merely avoiding harm, the WEIGHTED variant (give the strong agent a double vote on disagreement) improves *global* accuracy over the strong-solo baseline on *all three* strong models: $+5.7\%$ on GPT-5.2, $+3.2\%$ on Claude-Opus, and $+1.3\%$ on Gemini-3 (Table 5). This is important for framing: the paper is not merely "debate is bad"—there exists a training-free routing policy that *beats* both blind debate and the strong-solo baseline, turning a documented failure mode into a usable gain. A *confidence-gated* variant (protect only when agents disagree *and* the strong agent is low-confidence) gives the largest gain on GPT-5.2 ($+5.4\%$) but *hurts* Claude-Opus (-4.1%), so the best gating rule is model-dependent, a dependence we analyze in §7.

7 DISCUSSION AND LIMITATIONS

Net effect vs. concentrated harm. We deliberately report that the *net* accuracy change of full debate is mild and not significant when merged (McNemar $p{=}0.45$, §4), even though the harmful

direction is highly significant per model. We view this as a feature of the finding, not a weakness: the harm is real and predictable on an identifiable subset, but is masked in aggregate by compensating fixes. A benchmark that only tracked aggregate accuracy would miss it entirely; our subset-and-predictability framing is what surfaces the failure.

Predictor precision has an upper bound. The round-0-disagreement predictor achieves $\approx 100\%$ *recall* of harmful events but only 26% *precision* (many disagreements are still resolved correctly by debate). Crucially, the PROTECT application does not require high precision: flagging a safe disagreement item merely reverts it to the strong agent's (correct) round-0 answer at no accuracy cost, so a conservative, high-recall predictor is sufficient for prevention. Improving precision (to know *when disagreement is safe to debate*) is an open direction: a grid search over training-free multi-feature conjunctions cannot exceed 26% precision while retaining 100% recall (Appendix A, Table 5)—since every harmful event is preceded by a round-0 disagreement, any stricter rule sacrifices guaranteed prevention—so higher precision likely requires learned predictors that trade recall for precision along the frontier we characterise.

The winning policy is model-dependent. The confidence-gated variant is the best policy on GPT-5.2 ($+5.4\%$) but was *worse* than baseline on Claude-Opus (-4.1%), where low-confidence disagreements are cases the strong agent also gets wrong. Thus the optimal routing policy depends on the strong model's confidence calibration—itself an interesting open axis. We report both the win and this dependence honestly.

Gemini's weak signal. Gemini-3 exhibits both the lowest harmful rate and the smallest agreement effect. With the scaled-up testbed the effect is nonetheless clearly significant (p=8.6×10^{-5}); qualitatively Gemini conforms less under peer pressure, leaving less harm to predict. This bounds the effect size, not its existence or significance.

Statistical power and event counts. The at-risk subset yields 404 harmful events over $3{,}475$ tasks. This is ample to establish the round-0 (dis)agreement effect (which reaches p=7.3×10^{-126} merged and is significant in every strong model, including Gemini-3 with 16 harmful events at p=8.6×10^{-5}). The direction and significance are robust across all four model configs; the bootstrap CIs we report make the per-model effect sizes explicit.

Scope. Our study covers three strong models (four model configs) and $4{,}554$ mid-difficulty BBH tasks with a fixed heterogeneous configuration (one strong, two weak). Extending to more model families, task domains beyond BBH, larger debates, and longer horizons is future work. We release the testbed, the predictability protocol, and the PROTECT baselines to enable it.

8 CONCLUSION

We showed that heterogeneous multi-agent debate systematically destroys correct answers—flipping 11.6% of a strong agent's initially-correct answers to wrong—yet this harm is *predictable before any debate round* from a single zero-cost observable (round-0 strong–weak disagreement; harmful rate 0.2% vs. 26.4%, $p < 10^{-125}$) and *preventable* by a training-free routing rule (PROTECT) that recovers the at-risk subset to 100% and, in a weighted variant, exceeds the strong-solo baseline. We release the benchmark, predictability protocol, and baselines as a testbed for *when to trust debate*, arguing that debate should be routed, not applied blindly.

ETHICS STATEMENT

This work studies a failure mode of multi-agent debate and proposes a training-free mitigation; it uses only public BIG-Bench Hard tasks and involves no human subjects, private data, or personally identifiable information. Our central finding—that blind heterogeneous debate can silently destroy correct answers—is a safety-relevant observation rather than a capability that could be misused: the accompanying PROTECT baseline is a defensive measure intended to make debate-based systems more reliable. We report negative and non-significant results honestly (e.g. the merged net effect is not statistically significant) to avoid overstating the phenomenon. We see no foreseeable harmful application of the released testbed beyond its intended use as a diagnostic for when to trust debate.

REPRODUCIBILITY STATEMENT

All results in this paper are computed from released per-task records and analysis scripts. The exact strong/weak agent triples for all four model configs, the debate protocol, and the answer-normalisation and statistical procedures are specified in Appendix A. Every routing variant (PROTECT, WEIGHTED, CONF-GATED) is obtained by *offline replay* on the same logged transcripts, so no additional model calls are needed to reproduce Table 5. Bootstrap confidence intervals use B=10,000 resamples with a fixed seed (20260728); the round-0 (dis)agreement effect uses a two-proportion z-test and the net effect of debate uses McNemar's test on discordant pairs. We release the testbed (task set, per-task logs with round-0/final answers, features, and labels), the predictability protocol, and the PROTECT baselines as anonymous supplementary material to enable end-to-end reproduction of every table and figure.

USE OF LARGE LANGUAGE MODELS

Large language models are the object of study in this paper: the debate agents, predictors, and PROTECT routing all operate over the outputs of the GPT, Claude, and Gemini families described in Appendix A, and all such usage is documented as part of the experimental protocol. Separately, the authors used an LLM-based assistant for writing aid, LaTeX formatting, code refactoring, and verification. The authors reviewed all assisted changes; the assistant did not generate experimental data or determine the paper's conclusions. The authors take full responsibility for all content of the paper.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning (ICML)*, 2024.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*, 2024.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *International Conference on Machine Learning (ICML)*, 2024a.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong

Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*, 2024b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

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

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

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

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *International Conference on Machine Learning (ICML)*, 2024.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research (TMLR)*, 2024a.

Loka Li, Guangyi Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*, 2024b.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

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. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

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. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*, 2023.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023a.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

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

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

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

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

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*, 2023.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

A REPRODUCIBILITY DETAILS

Debate configurations. Table 6 lists the exact strong/weak agent triples behind the four model configs. Each config pairs one strong agent with two cheaper models of different families, so that the "heterogeneous" setting is not an artefact of a single weak-model choice. Claude-Opus is evaluated under two distinct weak-peer pairs (pair A and pair B), which is why three strong models yield four configs; the two Claude configs bracket the per-config harmful rate (23.6% and 13.1%, merged 18.3%).

Table 6: Agent triples per config. Agent 0 (bold) is the *strong* agent whose post-debate answer is scored; the other two are weaker peers.

Config	Strong (agent 0)	Weak peer 1	Weak peer 2
Claude (pair A)	**claude-opus-4-8**	gpt-4.1-mini	qwen3-30b-a3b
Claude (pair B)	**claude-opus-4-8**	gemini-2.5-flash-lite	doubao-seed-2-0-lite
GPT	**gpt-5.2**	gpt-4.1-mini	qwen3-30b-a3b
Gemini	**gemini-3-flash**	gpt-4.1-nano	doubao-seed-2-0-lite

Debate protocol. Every debate runs the same fixed schedule. In *round 0* each of the three agents answers the task independently, with no visibility into the others' responses; we record each agent's answer and (for the strong agent) a self-reported confidence in $[0, 1]$. Agents then exchange all round-0 responses and revise for a fixed number of rounds; the debated answer reported for a task is the strong agent's final-round answer. The strong-solo baseline uses the strong agent's round-0 answer directly (no exchange). All routing variants in the main paper (Section 6, PROTECT) are computed by *offline replay* on the same logged transcripts—no additional API calls—so the comparison between full debate, PROTECT, WEIGHTED, and CONF-GATED is exactly matched per task.

Metrics and statistics. Correctness uses exact-match after answer normalisation (case- and whitespace-insensitive, option-letter canonicalisation). The harmful-debate rate is computed on the $\mathrm{base_ok}{=}1$ subset only. Bootstrap CIs use $B{=}10{,}000$ resamples with seed 20260728; the round-0 (dis)agreement effect is tested with a two-proportion z-test (per model and merged), and the net effect of full debate with McNemar's test on discordant pairs. All per-task records (round-0 and final answers, features, labels) and the analysis scripts are released with the testbed.

Multi-feature predictor: no free lunch under full recall. We test whether combining zero-cost, training-free features observable *before* debate—round-0 (dis)agreement, number of distinct round-0 answers (n_{uniq}), and the strong agent's round-0 confidence—can raise predictor precision beyond the single round-0-disagreement rule (precision 26.4%, recall 99%, flags 43.7% of the at-risk pool). A grid search over all conjunctions finds that *no multi-feature refinement beats the single rule while preserving* 100% *recall*: because every harmful event is preceded by a round-0 disagreement, any stricter AND-condition necessarily drops recall. Precision is therefore capped at 26.4% under the recall-100% constraint. Relaxing recall trades it for precision along a clean frontier (Table 7): e.g. the conjunction "disagree $\wedge$ strong_conf < 0.95" reaches precision 38.6% at recall 80% (best F1 $= 0.52$), and precision $\approx 37.8\%$ is attainable at recall 90%. This confirms that improving precision requires *giving up guaranteed prevention*, motivating richer (learned) predictors as future work.

Predictor (training-free)	Precision	Recall	F1	%flagged
round-0 disagree (single rule)	26.4%	99.0%	0.42	43.7%
strong_conf < 0.95	18.7%	80.9%	0.30	50.4%
$n_{\text{uniq}} \geq 3$	40.4%	22.8%	0.29	6.6%
disagree $\wedge$ strong_conf < 0.95	38.6%	80.0%	**0.52**	24.1%

Table 7: Zero-cost multi-feature predictors on the at-risk subset ($\mathrm{base_ok}{=}1$, $n{=}3475$, 404 harmful). No conjunction improves precision above 26.4% while keeping 100% recall; precision only rises by trading recall away.

B ADDITIONAL BREAKDOWNS

Per-task-family harmful rate. Figure 5 gives the full breakdown of the harmful-debate rate across all ten BBH task families on the at-risk (base_ok=1) subset, complementing the three-family summary in the main paper. The rate spans two orders of magnitude—from 33.2% on DISAMBIGUATION_QA down to 0% on LOGICAL_DEDUCTION_SEVEN_OBJECTS—confirming that debate harm is concentrated in a specific, identifiable class of tasks (those requiring a committed interpretation or a long serial trace) rather than being a uniform noise floor.

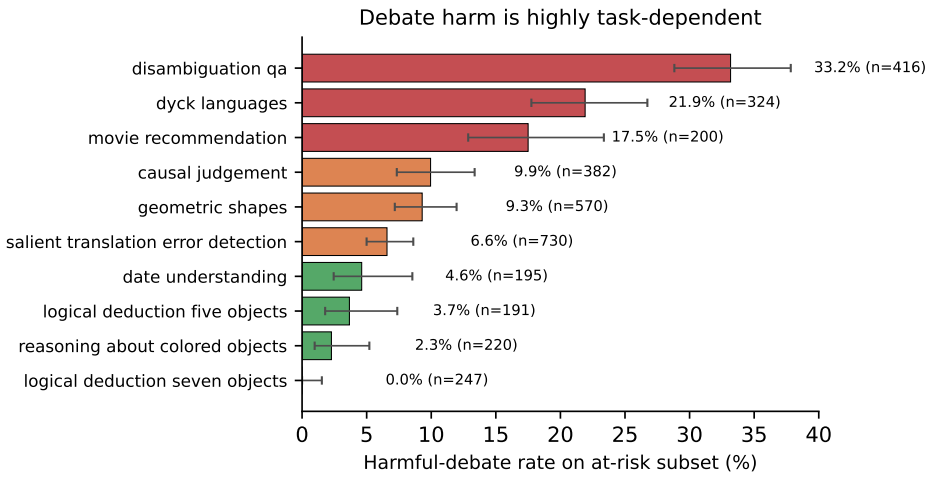

Figure 5: Harmful-debate rate broken down by BBH task family (at-risk subset, 95% Wilson CI).

Round-0 confidence dose–response. Figure 6 shows the monotone dose–response between the strong agent's round-0 confidence and the harmful rate, referenced from main-paper Section 5. The harmful rate falls from 29.8% for low-confidence answers (conf<0.8) to 14.6% ([0.8, 0.95)) to 4.5% for high-confidence answers ($\geq$ 0.95)—confirming that confidence is a complementary but weaker signal than the cross-agent round-0 agreement bit.

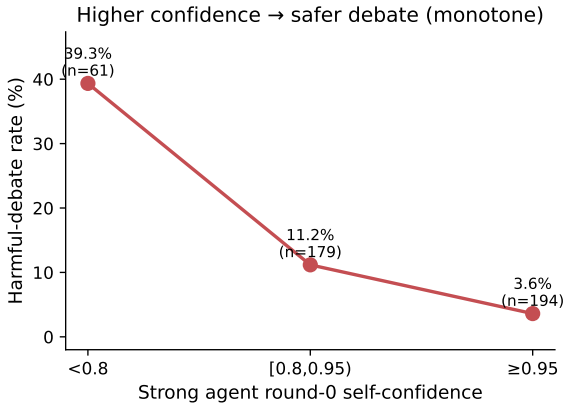

Figure 6: Monotone dose–response in strong-agent round-0 confidence vs. harmful rate.