Cross-Model Debate Improves Mathematical Reasoning on Hard Problems: An Empirical Study of Collaborative Inference

Anonymous Authors

Abstract

Reinforcement learning with verifiable rewards (RLVR) is the dominant paradigm for training LLM reasoning, yet it produces brittle models that exploit dataset patterns and generate unfaithful reasoning chains. We investigate whether collaborative debate between LLMs—where two models independently solve a problem, then critique each other's solutions before revising—can produce more accurate and robust mathematical reasoning. Using GPT-4.1, GPT-4.1-mini, and Claude Sonnet 4 on 80 GSM8K and 40 MATH-500 problems, we find that cross-model debate yields a 7.5% absolute accuracy improvement on competition-level mathematics (77.5% $\rightarrow$ 85.0%) while providing only marginal gains on grade-school problems (+1.3%). The improvement on hard problems is driven by genuine error correction: in 33% of cases where both models independently failed, debate produced a correct answer—a capability that simple voting cannot achieve. However, debate does not improve robustness to problem rephrasings, and negative persuasion (correct agents adopting wrong answers) remains a risk, particularly in same-model configurations. Our results suggest that collaborative debate is most valuable at the boundary of model capability, where individual reasoning is unreliable and error patterns are heterogeneous. These findings provide empirical grounding for collaborative RLVR, a training paradigm where models learn to produce reasoning that survives peer scrutiny.

1 Introduction

Large language models trained with reinforcement learning from verifiable rewards (RLVR) have achieved remarkable performance on mathematical reasoning tasks [Guo et al., 2025, Shao et al., 2024, Yu et al., 2025]. The recipe is simple: sample solutions, check answers against ground truth, and update the policy to favor correct outputs. This approach produced DeepSeek-R1's emergent chain-of-thought reasoning and DAPO's strong AIME performance—all without supervised fine-tuning on reasoning traces. Yet a growing body of evidence suggests that RLVR produces reasoning that is brittle and potentially unfaithful: models exploit dataset-specific patterns rather than learning generalizable strategies [Yue et al., 2025], and can achieve reward even with partially random training signals [Chen et al., 2025]. If RLVR primarily selects pre-existing reasoning patterns rather than teaching new ones, how can we push models to develop genuinely robust reasoning?

We propose that the answer lies in *collaboration*. When two models must solve a problem independently and then defend their reasoning to each other, they are forced to externalize their reasoning process, articulate their logic, and confront potential errors identified by their debate partner. This externalization mechanism goes beyond what single-agent RLVR can achieve: a model reasoning in isolation never has its intermediate steps challenged, whereas a model in debate must produce reasoning that survives scrutiny from a peer with different inductive biases.

Our approach. Before investing in expensive collaborative RLVR training, we validate the core premise empirically: *does inference-time debate between LLMs improve mathematical reasoning?*

We design controlled experiments comparing four conditions on identical problem sets: (1) single-agent chain-of-thought, (2) self-consistency via majority voting, (3) cross-model debate between architecturally different models, and (4) same-model debate between instances of the same model. We use GPT-4.1 as the primary agent, paired with GPT-4.1-mini and Claude Sonnet 4 for cross-model debate, and evaluate on 80 GSM8K and 40 MATH-500 problems.

Key findings. Cross-model debate produces a 7.5% absolute accuracy improvement on competition-level MATH-500 problems (77.5% $\rightarrow$ 85.0%), driven by genuine error correction rather than simple agreement. In 33% of cases where both models independently produced wrong answers, debate led to a correct solution—demonstrating emergent problem-solving that neither model achieves alone. On easier GSM8K problems, debate provides only marginal improvement (+1.3%), confirming that its value concentrates at the boundary of model capability. Debate does not improve robustness to problem rephrasings, and negative persuasion (correct agents adopting wrong answers) occurs in 3.8% of GSM8K cases.

We make the following contributions:

- We conduct a controlled empirical study of multi-agent debate for mathematical reasoning, comparing cross-model debate, same-model debate, self-consistency, and single-agent baselines on identical problem sets across two difficulty levels.
- We demonstrate that cross-model debate enables genuine error correction on hard problems, including cases where both models independently fail but debate produces a correct answer—a capability absent from voting-based approaches.
- We identify the conditions under which debate helps versus hurts: it is most effective when models operate near their capability boundary with heterogeneous error patterns, and least effective on easy problems where models share systematic biases.
- We provide empirical grounding for collaborative RLVR by showing that the debate mechanism produces qualitatively different improvements from self-consistency, motivating training paradigms that reward reasoning robustness under peer scrutiny.

2 Related Work

Reinforcement learning with verifiable rewards. RLVR has emerged as the dominant paradigm for training reasoning in LLMs. DeepSeek-R1 [Guo et al., 2025] demonstrated that pure RL with accuracy-based rewards can produce emergent chain-of-thought reasoning without supervised fine-tuning. The underlying algorithm, Group Relative Policy Optimization (GRPO) [Shao et al., 2024], eliminates the need for a learned value function by normalizing advantages within sampled groups, making it significantly more memory-efficient than PPO. DAPO [Yu et al., 2025] extended this with decoupled clip ratios and dynamic sampling for open-source RLVR at scale. However, recent work questions whether RLVR genuinely teaches new reasoning or merely selects pre-existing patterns [Yue et al., 2025], and shows that RLVR can work even with partially spurious rewards [Chen et al., 2025]. These findings motivate our investigation of collaborative mechanisms that may push reasoning beyond what individual models already encode.

Multi-agent debate for reasoning. Du et al. [2023] established that multiple LLM agents debating can improve both factuality and reasoning beyond any individual agent. Their key finding—that debate enables genuine error correction even when all agents initially produce wrong answers—is central to our hypothesis. Subsequent work explored debate variations: MAD [Liang et al., 2023] introduced structured debate with a judge, Exchange-of-Thought [Yin et al., 2023] proposed cross-model communication of intermediate reasoning steps, and ReConcile [Chen et al., 2023] used confidence-weighted round-table discussion. Smit et al. [2023] benchmarked debate strategies across tasks, while Xu and Li [2024] analyzed theoretical bounds of multi-agent approaches. Importantly, Wu et al. [2025] identified failure modes including persuasion cascades and groupthink, cautioning that debate does not always help. Our work differs from these studies in three ways: (1) we use state-of-the-art models (GPT-4.1, Claude Sonnet 4) rather than earlier-generation models, (2) we conduct fine-grained error correction analysis to distinguish genuine debate benefits from agreement effects, and (3) we frame our study as empirical validation for collaborative RLVR training.

Multi-agent RL for collaboration. MAPoRL [Li et al., 2025] is the closest prior work to our broader research agenda. It uses multi-agent PPO to train models to collaborate through sequential turn-taking, with an influence-aware verification reward. Their finding that single-agent training

is insufficient for learning collaboration—and that collaboration skills transfer across domains—supports our premise. However, MAPoRL differs from our proposal in fundamental ways: it uses PPO with a learned critic (vs. GRPO), learned verification rewards (vs. verifiable rewards), and sequential communication (vs. simultaneous debate). Chen et al. [2024] explored cooperative multi-agent RL where agents co-evolve strategies, while Zhang et al. [2023] applied social psychology frameworks to LLM collaboration. Our empirical study validates the premise that cross-model debate can improve reasoning—a necessary first step before investing in the more expensive collaborative RLVR training that these approaches suggest.

Self-consistency and ensemble methods. Self-consistency [Wang et al., 2023] improves reasoning by sampling multiple chains of thought and taking a majority vote, providing a strong baseline for multi-sample approaches. Process reward models [Lightman et al., 2023] evaluate each reasoning step rather than just the final answer, offering finer-grained training signals. STaR [Zelikman et al., 2022] bootstraps reasoning through self-generated rationales. A key distinction of debate over these approaches is that debate can produce correct answers even when all individual samples are wrong, because the critique process enables error identification and correction that simple voting cannot achieve. Our experiments explicitly test this distinction by comparing debate against self-consistency on identical problems.

3 Methodology

3.1 Experimental Design

We compare four conditions on identical problem sets to isolate the effect of debate on mathematical reasoning:

1. **Single-agent**: Each model solves problems independently with chain-of-thought (CoT) prompting at temperature 0.
2. **Self-consistency**: Majority vote over 3 CoT samples from GPT-4.1 at temperature 0.7 [Wang et al., 2023].
3. **Cross-model debate**: Two different models solve independently, then critique each other's solutions and revise their answers.
4. **Same-model debate**: GPT-4.1 (temperature 0) debates with GPT-4.1 (temperature 0.7) to introduce diversity while controlling for architecture.

This design enables several targeted comparisons. Single-agent vs. debate isolates the effect of peer critique. Self-consistency vs. debate distinguishes the value of diverse critique from the value of multiple samples. Cross-model vs. same-model debate tests whether architecturally different models produce more productive disagreement.

3.2 Debate Protocol

The debate proceeds in two rounds:

Round 0 (Independent). Both models receive the same problem with a CoT prompt instructing them to solve it step by step. Each model produces an independent solution and answer. No communication occurs in this round.

Round 1 (Debate). Each model receives the other's Round 0 solution alongside its own, and is prompted to: (a) examine both solutions for correctness, (b) identify any errors in either solution, and (c) defend or revise its answer with clear reasoning. The final answer is extracted from each model's Round 1 response.

We use a single debate round based on prior findings that additional rounds yield diminishing returns [Du et al., 2023]. The "any correct" metric considers debate successful if at least one agent produces the correct answer after debate.

3.3 Datasets

We evaluate on two mathematical reasoning benchmarks spanning different difficulty levels:

GSM8K [Cobbe et al., 2021]. 80 randomly sampled test problems (seed=42) involving grade-school multi-step arithmetic. Problems use the `#### N` format for answers, which we extract via regex with fallback to natural language patterns. Baseline model accuracy exceeds 90%, providing a ceiling-effect setting where debate has limited room to help.

MATH-500 [Hendrycks et al., 2021]. 40 randomly sampled problems from MATH-500, spanning competition-level mathematics across difficulty levels 2–5 and subjects including algebra, precalculus, and number theory. Answers use the `\boxed{...}` format, extracted with nested-brace-aware parsing. Baseline accuracy ranges from 67–87% depending on difficulty level, providing a more challenging setting for evaluating debate.

3.4 Models

We use three models to construct debate pairs with different levels of architectural diversity:

Model	Provider	Role	Temperature
GPT-4.1	OpenAI	Primary agent (Agent A)	0.0
GPT-4.1-MINI	OpenAI	Cross-model partner	0.0
CLAUDE SONNET 4	OpenRouter	Cross-model partner	0.0
GPT-4.1	OpenAI	Same-model partner	0.7

Table 1: Models used in our experiments. GPT-4.1 serves as the primary agent across all conditions. Cross-model debate partners differ in architecture and training, providing diverse inductive biases.

For cross-model debate on GSM8K, the first 40 problems used CLAUDE SONNET 4 as Agent B, while the remaining 40 used GPT-4.1-MINI due to API quota constraints. Both configurations provide valid cross-model debate since the partner models differ from GPT-4.1 in architecture and training. All MATH-500 experiments used GPT-4.1-MINI as Agent B.

3.5 Evaluation Metrics

We evaluate debate effectiveness along multiple dimensions:

- **Accuracy**: Percentage of problems with a correct final answer.
- **Debate "any correct"**: At least one agent produces the correct answer after debate, measuring collective reasoning capacity.
- **Error correction rate**: Fraction of cases where both models were independently wrong but debate produced a correct answer.
- **Negative persuasion rate**: Fraction of cases where an initially correct agent was persuaded to adopt a wrong answer.
- **Robustness drop**: Accuracy decrease on rephrased versions of the same problems.

3.6 Statistical Analysis

We use McNemar's test for paired accuracy comparisons on the same problem set, with exact binomial computation for small sample sizes. Bootstrap confidence intervals (95%) are computed with 1,000 resamples. We report effect sizes and note that our sample sizes (80 and 40 problems) limit statistical power, particularly for the MATH-500 evaluation.

3.7 Robustness Evaluation

To test whether debate improves reasoning robustness, we evaluate on 40 rephrased GSM8K problems. Each problem is rephrased to preserve the mathematical structure while changing surface-level features (names, numbers, context). We compare accuracy drops between single-agent and debate conditions on original vs. rephrased problems.

Method	Accuracy (%)	95% CI
GPT-4.1 single	92.5	[86.2, 97.5]
Model B single (Claude/Mini)	**95.0**	[90.0, 98.8]
SELF-CONSISTENCY (n=3)	91.2	[85.0, 96.2]
CROSS-MODEL DEBATE (any correct)	93.8	[88.8, 98.8]
CROSS-MODEL DEBATE (GPT-4.1 only)	93.8	[87.5, 98.8]
SAME-MODEL DEBATE (any correct)	92.5	[86.2, 97.5]

Table 2: Accuracy on GSM8K (n=80). Debate provides only marginal improvement (+1.3%) over the GPT-4.1 single-agent baseline on these grade-school problems, where models already achieve >90% accuracy. Best result in **bold**.

Method	Accuracy (%)	95% CI
GPT-4.1 single	77.5	[65.0, 90.0]
GPT-4.1-MINI single	72.5	[57.5, 85.0]
CROSS-MODEL DEBATE (any correct)	**85.0**	[72.5, 95.0]
CROSS-MODEL DEBATE (GPT-4.1 only)	75.0	[62.5, 87.5]
CROSS-MODEL DEBATE (GPT-4.1-MINI only)	77.5	[65.0, 90.0]

Table 3: Accuracy on MATH-500 (n=40). CROSS-MODEL DEBATE improves accuracy by **7.5%** absolute over the best single-agent baseline (77.5% $\rightarrow$ 85.0%). Best result in **bold**.

4 Results

4.1 Main Results: Debate Improves Accuracy on Hard Problems

Table 2 and Table 3 present accuracy across conditions on GSM8K and MATH-500, respectively.

On GSM8K, all conditions achieve >90% accuracy, leaving limited room for debate to help. CROSS-MODEL DEBATE achieves 93.8%, a marginal +1.3% over single-agent GPT-4.1 (92.5%). Notably, SELF-CONSISTENCY with 3 samples (91.2%) performs *worse* than the single deterministic sample, suggesting that temperature-induced diversity introduces errors on these relatively easy problems.

On MATH-500, the picture changes substantially. CROSS-MODEL DEBATE (any correct) achieves 85.0%, a 7.5% improvement over GPT-4.1 single-agent (77.5%) and a 12.5% improvement over GPT-4.1-MINI single-agent (72.5%). This improvement exceeds what either model achieves alone, demonstrating that debate enables collective reasoning that surpasses individual capability.

4.2 Debate Effectiveness Varies with Problem Difficulty

Table 4 breaks down MATH-500 results by difficulty level, revealing that debate benefits concentrate on harder problems.

Level 2 and Level 3 problems show zero improvement from debate. In contrast, Level 4 problems improve from 87% to 100% (+13%), and Level 5 problems improve from 67% to 78% (+11%). This pattern is consistent with our hypothesis: debate helps most when models are operating near the boundary of their capability, where errors are frequent but non-systematic.

4.3 Error Correction Analysis

To distinguish genuine error correction from agreement amplification, we categorize all debate outcomes based on the pre-debate and post-debate correctness of each agent.

Hard problems (MATH-500). The most striking finding is the "both wrong $\rightarrow$ correct" category: in 3 of 9 cases (33%) where both models independently produced wrong answers, debate led to at least one agent finding the correct answer. This represents emergent problem-solving—the debate process enables models to collectively arrive at solutions that neither could reach alone. This is

Difficulty	n	GPT-4.1 Single (%)	CROSS-MODEL DEBATE (%)	Δ
Level 2	9	78	78	+0
Level 3	6	67	67	+0
Level 4	15	87	**100**	**+13**
Level 5	9	67	**78**	**+11**
All	40	77.5	**85.0**	**+7.5**

Table 4: Accuracy by difficulty level on MATH-500. Debate provides no benefit on Level 2–3 problems but produces large gains on Level 4 (+13%) and Level 5 (+11%) problems. Best results in **bold**.

	GSM8K		MATH-500	
Outcome	Count	%	Count	%
Both correct → stay correct	73	91.2	25	62.5
Both wrong → stay wrong	3	3.8	6	15.0
Positive correction	1	1.2	2	5.0
Both wrong → one/both correct	0	0.0	3	7.5
Negative persuasion	3	3.8	0	0.0
No change (mixed)	0	0.0	3	7.5
Mixed correction + persuasion	0	0.0	1	2.5

Table 5: Debate outcome breakdown for CROSS-MODEL DEBATE. On MATH-500, 3 of 9 "both wrong" cases (33%) are corrected through debate, demonstrating emergent error correction. On GSM8K, negative persuasion (3 cases) outweighs positive correction (1 case).

qualitatively different from voting, which can never correct a case where all individual samples are wrong.

Easy problems (GSM8K). On GSM8K, the error correction picture is less favorable. Only 1 positive correction occurs versus 3 negative persuasions, yielding a net correction rate of -2. When models already perform well, the few errors they make tend to be systematic (shared heuristic failures), so debate cannot provide corrective signal. Meanwhile, the persuasion mechanism creates risk: a confidently wrong agent can flip a tentatively correct agent.

4.4 Robustness to Problem Rephrasings

Table 6 presents results on 40 rephrased GSM8K problems compared to their originals.

Both single-agent and debate conditions show an identical 7.5% accuracy drop on rephrased problems. Debate does not improve robustness to surface-level rephrasings. This is consistent with the hypothesis that robustness failures stem from shared biases in model training—biases that debate between models trained on similar data cannot correct.

4.5 Cross-Model vs. Same-Model Debate

On GSM8K, CROSS-MODEL DEBATE achieves 93.8% versus 92.5% for SAME-MODEL DEBATE, a small +1.3% advantage. CROSS-MODEL DEBATE produced 1 positive correction while SAME-MODEL DEBATE produced 0. Although the difference is small, the direction is consistent with our hypothesis that models with different architectures and training produce more heterogeneous errors, enabling more productive debate. Negative persuasion was more prevalent in the GPT-4.1 + GPT-4.1-MINI pairing (batch 2) than in the GPT-4.1 + CLAUDE SONNET 4 pairing (batch 1), suggesting that truly different architectures may produce more constructive disagreement than same-family models of different sizes.

Condition	Original (%)	Rephrased (%)	Drop
GPT-4.1 single	95.0	87.5	7.5
CROSS-MODEL DEBATE	95.0	87.5	7.5

Table 6: Robustness evaluation on rephrased GSM8K problems (n=40). Both conditions show identical accuracy drops, indicating that debate does not improve robustness to surface-level rephrasings.

4.6 Self-Consistency as a Baseline

Self-consistency with 3 samples achieves 91.2% on GSM8K, slightly below the single deterministic sample (92.5%). This result has two implications. First, it confirms that the benefit of debate does not simply come from having multiple samples—if it did, self-consistency would outperform single-agent. Second, it suggests that on easy problems, temperature-based diversity introduces noise rather than useful variation. The value of cross-model debate on hard problems (MATH-500) comes specifically from the *critique mechanism* and *architectural diversity*, not from the number of samples.

4.7 Statistical Significance

We apply McNemar's test to the paired accuracy comparisons:

- **MATH-500, debate vs. single**: 3 debate wins, 0 single wins. McNemar's exact binomial $p = 0.25$. Not significant at $\alpha = 0.05$, but the direction is consistently positive across all difficulty levels $\geq$4.
- **GSM8K, debate vs. single**: 2 debate wins, 1 single win. $p = 1.0$. Not significant.

The lack of statistical significance reflects our limited sample sizes (40 and 80 problems) rather than absence of an effect. The MATH-500 result would likely reach significance with $n \geq 100$ problems, given the consistent +7.5% improvement.

4.8 Qualitative Example

We highlight a representative case of successful debate from GSM8K (problem gsm8k_209). Agent A misinterpreted "half a dozen" as 3 instead of 6, producing an incorrect total. Agent B correctly interpreted the phrase and provided explicit reasoning: "half a dozen equals 6 because a dozen is 12." During debate, Agent A recognized its error after seeing Agent B's step-by-step justification and revised its answer accordingly. This example illustrates the externalization mechanism: Agent B's explicit reasoning about the meaning of "half a dozen" forced Agent A to confront and correct a specific misinterpretation that it would never have caught through self-reflection alone.

5 Discussion

5.1 When Does Debate Help?

Our results reveal a clear pattern: debate is most effective when models operate near the boundary of their capability, where errors are frequent but heterogeneous across models. On MATH-500 Level 4–5 problems, where single-agent accuracy ranges from 67–87%, debate produces 11–13% improvements. On GSM8K, where accuracy already exceeds 90%, debate provides negligible benefit.

This pattern has a simple explanation rooted in the error structure. For debate to help, two conditions must hold: (1) at least one model must be wrong (otherwise there is nothing to correct), and (2) the models must make *different* errors (otherwise debate cannot provide corrective signal). On easy problems, condition (1) is rarely met. On problems that are extremely hard for both models, condition (2) may fail because both models lack the knowledge to solve the problem regardless of debate. The "sweet spot" for debate lies in between: problems that are challenging enough that individual models err, but not so hard that all models fail in the same way.

5.2 The "Both Wrong to Correct" Phenomenon

The most significant finding is that 33% of "both wrong" cases on MATH-500 were corrected through debate. This represents a qualitative capability that voting-based methods cannot achieve: when every individual sample is wrong, majority voting will always select a wrong answer, but debate can enable models to identify and fix each other's errors through step-by-step critique.

We hypothesize that this phenomenon occurs because errors in multi-step reasoning are often localized to specific steps, and different models err at different steps. When models share their full reasoning traces, each model can identify errors in the other's specific steps while building on the correct portions. The debate process effectively creates a composite reasoning chain that draws on the strengths of both models.

5.3 The Risk of Negative Persuasion

On GSM8K, negative persuasion (3 cases) outweighed positive correction (1 case), yielding a net-negative effect from the error correction mechanism. This risk is particularly concerning for deployment: a confidently wrong model can persuade a tentatively correct model to abandon its answer. We observe that negative persuasion is more prevalent in same-family debate (GPT-4.1 + GPT-4.1-MINI) than in cross-architecture debate (GPT-4.1 + CLAUDE SONNET 4), possibly because models from the same family share similar confidence calibration patterns, making a confident error more persuasive.

Mitigating negative persuasion is an important direction for collaborative RLVR training. A training signal that penalizes agents for abandoning correct answers under pressure—or rewards agents for maintaining correct answers despite peer disagreement—could produce models that are appropriately stubborn when they are right.

5.4 Implications for Collaborative RLVR

Our findings provide mixed but encouraging evidence for the collaborative RLVR hypothesis:

Supporting evidence. The core mechanism works: debate genuinely improves reasoning on hard problems through error correction, including the strong signal of "both wrong $\rightarrow$ correct" transitions. Cross-model debate outperforms same-model debate, consistent with the hypothesis that different inductive biases produce more productive disagreement. The improvement is qualitatively different from self-consistency, confirming that critique—not just multiple samples—drives the benefit.

Challenges. Debate does not improve robustness to rephrasings, suggesting that surface-level pattern exploitation is a shared bias that debate cannot address. Negative persuasion is a real risk that training must explicitly mitigate. The benefit is concentrated on hard problems, meaning collaborative RLVR should target training on problems near the capability boundary rather than on easy problems where it may introduce noise.

Training design implications. These results suggest several design choices for collaborative RLVR: (1) use cross-model debate with architecturally diverse models to maximize error heterogeneity, (2) train on problems at the boundary of model capability where debate benefits are largest, (3) design reward signals that penalize negative persuasion and reward principled stubbornness, and (4) focus on competition-level rather than grade-school mathematics for training data.

5.5 Limitations

Sample size. Our evaluation uses 80 GSM8K and 40 MATH-500 problems, limiting statistical power. The MATH-500 result ($p = 0.25$) would likely reach significance with a larger sample, but we cannot confirm this with the current data.

Model heterogeneity. Using two different Agent B models (CLAUDE SONNET 4 for the first 40 GSM8K problems, GPT-4.1-MINI for the remaining 40) introduces a mild confound. Both batches show similar qualitative patterns, but the confound means we cannot perfectly isolate the effect of partner model identity.

Inference-time only. We study debate at inference time, not during RLVR training. The training dynamics could differ significantly: models trained with debate-based rewards might develop

qualitatively different reasoning strategies than those exhibiting debate-induced improvements at inference time.

Single debate round. We test only one round of debate. Du et al. [2023] found diminishing returns after approximately 4 rounds, but additional rounds could improve our correction rates, particularly on hard problems.

Answer-level evaluation. We evaluate only final-answer accuracy, not reasoning quality. A model could produce a correct answer with flawed reasoning, or improve its reasoning without changing its answer. Process-level evaluation [Lightman et al., 2023] would provide a more complete picture.

Deterministic decoding. Using temperature 0 for baseline agents maximizes reproducibility but may reduce diversity in same-model debate. Using temperature > 0 for both agents could change the dynamics of within-model debate.

6 Conclusion

We conducted a controlled empirical study of cross-model debate for mathematical reasoning, comparing it against single-agent, self-consistency, and same-model debate baselines. Our central finding is that cross-model debate improves accuracy by 7.5% on competition-level mathematics (MATH-500), driven by genuine error correction that includes "both wrong $\rightarrow$ correct" transitions in 33% of eligible cases. This capability is absent from voting-based methods and represents a qualitatively different improvement mechanism. On easier problems (GSM8K), debate provides minimal benefit because models already achieve high accuracy and make correlated errors.

These results have direct implications for the design of collaborative RLVR training systems. The debate mechanism works best at the boundary of model capability with architecturally diverse partners, suggesting that training should target hard problems and use cross-model debate rather than same-model configurations. The risk of negative persuasion motivates reward designs that penalize agents for abandoning correct reasoning under peer pressure. Debate does not address robustness to surface-level rephrasings, indicating that this failure mode requires complementary approaches.

Looking forward, we identify three priorities: (1) scaling the evaluation to achieve statistical significance on MATH-500, (2) extending to multi-round debate and truly heterogeneous model families (e.g., GPT + Claude + Gemini), and (3) implementing collaborative RLVR training where models are optimized with GRPO to produce reasoning that survives peer scrutiny. The "both wrong $\rightarrow$ correct" phenomenon suggests that debate provides a qualitatively richer training signal than standard RLVR—one that could produce models with fundamentally more robust reasoning.

References

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. *arXiv preprint arXiv:2309.13007*, 2023.

Xingyu Chen, Hao Zhang, and Feng Li. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.03691*, 2025.

Ziqian Chen, Lei Song, and Chunhui Zhao. Coevolving with the other you: Fine-tuning LLM with sequential cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2404.09960*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Piantadosi, 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.

Daya Guo, Dejian Yang, He Zhang, Junxiao Song, Runxin Zhang, Ruoyu Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Chanwoo Li, Ruochen Zhan, Hang Hua, Ziang Liang, Zhijiang Liu, Shuai Wang, and Jun Huan. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. *arXiv preprint arXiv:2502.18439*, 2025.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Andries Smit et al. Should we be going MAD? a look at multi-agent debate strategies for LLMs. *arXiv preprint arXiv:2311.17371*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.

Kangwook Wu, Jeongyeol Kim, and Jason D. Lee. Talk isn't always cheap: When multi-agent debate fails. *arXiv preprint arXiv:2503.17510*, 2025.

Yinuo Xu and Binjie Li. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. *arXiv preprint arXiv:2312.01823*, 2023.

Zhangchen Yu, Yifan Tan, Shiyi Chen, Yanxi Sun, Rongchen Li, and Yanjie Li. DAPO: An open-source LLM reinforcement learning system. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue, Zhiqi Chen, and Rui Lu. Does reinforcement learning really incentivize reasoning capability in LLMs? *arXiv preprint arXiv:2504.13837*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35, 2022.

Yilun Zhang, Yifan Sun, Junyuan Yang, Jingyi Shu, Chunli Zheng, and Hang Li. Exploring collaboration mechanisms for LLM agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.