

# SELF-IMPROVEMENT OF LANGUAGE MODELS BY POST-TRAINING ON MULTI-AGENT DEBATE

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Self-improvement, where models improve beyond their current performance without external supervision, remains a challenge. The core difficulty is sourcing a training signal stronger than what the model itself can currently produce. Majority voting has been shown to provide such a signal by aggregating over multiple samples, helping mitigate some of the inconsistencies in LM reasoning. In this work, we show that multi-agent debate—where models collaborate and exchange reasoning over multiple rounds—provides an even richer signal than single-round majority voting. We introduce Multi-Agent Consensus Alignment (MACA), which uses reinforcement learning (RL) to post-train models to effectively utilize multi-agent debate. We find that preference learning over full reasoning traces, learning to differentiate between majority and minority reasoning, is more effective than binary consensus rewards or SFT-based approaches for leveraging these debate signals. This produces three key improvements: models are (1) better at utilizing the multi-agent debate setting (+26.87% on MATH), (2) individually more accurate (+21.51% on MathQA), and (3) more self-consistent (+27.6% on GSM8K). We also see strong generalization to unseen benchmarks (+16.3% on GPQA, +11.6% on CommonsenseQA).

## 1 INTRODUCTION

Self-improvement in language models, where models train on self-generated signals without external supervision, is challenging. LMs suffer from inconsistent reasoning, often producing contradictory solutions to the same problem under exploratory sampling (Holtzman et al., 2020; Wang et al., 2023). While inference-time aggregation methods such as majority voting and multi-agent debate help improve the reliability of the final response (Wang et al., 2023; Li et al., 2024; Irving et al., 2018), they do not address the core problem: the internal reasoning instability that produces these inconsistencies in the first place. We argue that self-consistency, the ability to generate consistent outputs across diverse reasoning paths (Elazar et al., 2021; Wang et al., 2023), is a property of well-aligned reasoning systems. In the brain, consistency emerges through dedicated control mechanisms in prefrontal and anterior cingulate regions that resolve conflicts between competing neural activations, directing cognition toward coherent solutions (Miller and Cohen, 2001; Friston, 2010; Botvinick et al., 2004; Shenhav et al., 2013; Zhang et al., 2025). Language models, by contrast, lack mechanisms to align their diverse reasoning paths toward consensus.

Prior work shows that majority-vote signals can be an effective self-improvement signal (Wang et al., 2023; Zuo et al., 2025; Prasad et al., 2025). However, the models’ underlying poor self-consistency can reduce the efficacy of such reinforcement. When models sample low-quality, inconsistent reasoning traces, aggregation can be counterproductive: noisy arguments compound rather than cancel out (Radharapu et al., 2025). We show that this problem is particularly acute in small language models, where simple statistical aggregation can compound low-quality reasoning and collaboration can result in weak consensus.

To achieve effective self-improvement in language models, we introduce Multi-Agent Consensus Alignment (MACA), a reinforcement learning framework that post-trains models using debate-derived signals to strengthen self-consistency and collaborative reasoning. Through multi-agent debate, models observe which reasoning patterns converge across peers and learn to favor those patterns through reinforcement on debate-derived consensus signals. This creates a positive feedback

loop: improved self-consistency and ability to utilize peer context enables more effective debate, which generates higher-quality training signals for the next iteration (Fig. 1).

A key design question in MACA is how to translate the multi-agent debate into a training signal. We explore the use of majority/minority labels as rewards through majority-vote GRPO (Shao et al., 2024), as well as using them as preference labels for DPO (Rafailov et al., 2024) and KTO (Ethayarajh et al., 2024). We find that preference learning is more effective: while both use majority/minority labels, preference learning over full reasoning traces allows models to learn to differentiate between reasoning traces that survived through deliberation and those that did not, capturing patterns where agents successfully reference and build on each other’s reasoning. Crucially, this requires no ground-truth labels. Preference learning through relative comparison naturally grounds reasoning in the absence of external verification: when truth is ambiguous, judgments become robust through comparative assessment where majority reasoning provides a normative anchor while minority views introduce necessary variation (Festinger, 1954; Moscovici and Lage, 1976; Nemeth, 1986; Condorcet, 1972). We teach models to ground their reasoning through this comparison of debate-derived trajectories, learning stable reasoning through peer deliberation alone.

### Key contributions.

- **Method.** We introduce MACA, a reinforcement learning framework that post-trains models on consensus signals derived from multi-agent debate, teaching them to internalize collaborative reasoning patterns.
- **Insight.** We demonstrate that consensus signals from multi-agent debate, grounded in peer interaction, provide richer and more robust training signals for self-improvement than single-round majority voting. This creates a positive feedback loop where improved reasoning enables more effective debate, which generates higher-quality training signals.
- **Empirical validation.** We show that MACA improves multi-agent collaborative performance (+26.87% on MATH), individual reasoning (+21.51% on MathQA), and self-consistency (+27.6% on GSM8K), with strong generalization to unseen reasoning domains (+16.3% on GPQA, +11.6% on CommonsenseQA).

## 2 FORMALIZING SELF-CONSISTENCY

Given a prompt  $x$ , an LM with parameters  $\theta$  defines a distribution  $\pi_\theta(y|x) = \prod_{t=1}^{|y|} \pi_\theta(y_t|x, y_{<t})$  over reasoning trajectories  $y$ , from which answers  $a = A(y)$  are extracted. Under temperature sampling, the model samples from a modified distribution  $\pi_{\theta,\tau}(y|x)$  where token probabilities are adjusted by temperature  $\tau > 0$ . This induces an answer distribution  $P_{\theta,\tau}(a|x) = \sum_{y:A(y)=a} \pi_{\theta,\tau}(y|x)$ , which gives each answer’s probability by summing over all reasoning paths that lead to it. We denote the majority answer as  $a_{\theta,\tau}^*(x) = \arg \max_a P_{\theta,\tau}(a|x)$  with majority probability  $S_{\theta,\tau}^+(x) = P_{\theta,\tau}(a_{\theta,\tau}^*(x)|x)$ . This represents the total probability mass concentrated on the most likely answer, or the model’s internal consensus.

Temperature sampling enables exploration of diverse reasoning paths, but reduces the consistency of the final answer. While greedy decoding ( $\tau = 0$ ) trivially approaches perfect consistency, it eliminates exploration and often produces suboptimal solutions (Holtzman et al., 2020). Lower temperatures increase consistency but restrict reasoning diversity. A self-consistent model should maintain high  $S_{\theta,\tau}^+(x)$  even at high temperatures, allowing the model to access diverse reasoning trajectories while reliably converging on consistent answers. We measure self-consistency in two ways.

**Single-agent sampling consistency.** Computing  $S_{\theta,\tau}^+(x)$  directly requires summing probabilities over all trajectories that lead to the majority answer, which is untractable. Instead, we estimate it by sampling  $t$  independent trajectories with answers  $a_1, \dots, a_t$  and computing:  $s_t^{\theta,\tau}(x) = \frac{1}{t} \sum_{i=1}^t \mathbf{1}[a_i(x) = \hat{a}(x)]$ , where  $\hat{a}(x) = \text{Majority}\{a_1(x), \dots, a_t(x)\}$ . This measures the fraction of sampled trajectories that agree with the majority answer. As  $t \rightarrow \infty$ ,  $s_t^{\theta,\tau}(x) \rightarrow S_{\theta,\tau}^+(x)$ , providing a consistent estimate of the true majority probability.

**Multi-agent debate agreement.** When  $M$  agents produce answers  $a_1(x), \dots, a_M(x)$  through deliberation, we measure the fraction of agents converging on the majority answer:  $d_M^{\theta,\tau}(x) =$

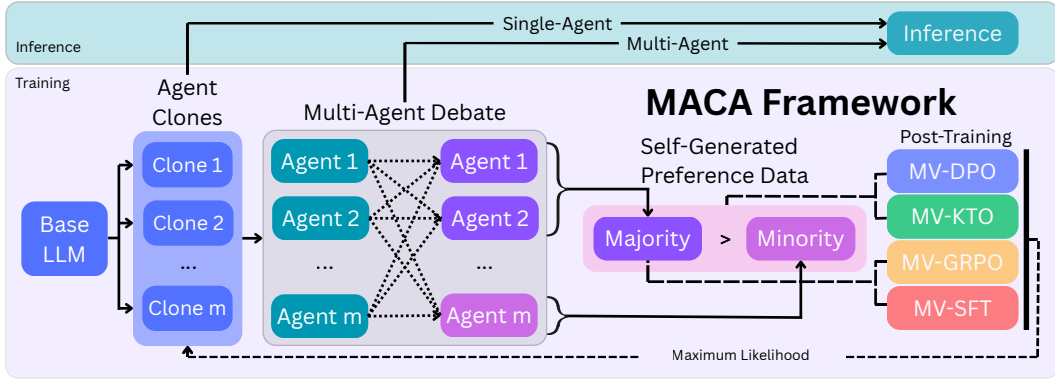


Figure 1: Multi-Agent Consensus Alignment framework: Multiple clones of a base LM engage in debate to generate majority and minority reasoning trajectories through multi-agent debate (collaborative exchange of reasoning). The framework splits responses based on alignment with majority consensus to create preference pairs. MV-GRPO compares online samples against majority signals, while MV-SFT imitates majority traces directly. In contrast, MV-DPO and MV-KTO utilize both positive (majority) and negative (minority) examples to learn to differentiate between these preference pairs. Updated agents can then be used for single-agent or multi-agent inference, or continue iterative training.

$\frac{1}{M} \sum_{m=1}^M \mathbf{1}[a_m(x) = \hat{a}(x)]$ , where  $\hat{a}(x) = \text{Majority}\{a_1(x), \dots, a_M(x)\}$ . Higher agreement indicates a stronger consensus; models with higher  $S_{\theta, \tau}^+(x)$  reach the same conclusion more frequently.

### 3 MACA: MULTI-AGENT CONSENSUS ALIGNMENT

Having formalized self-consistency, we now present a framework to improve it through post-training with self-generated signals from debate. In multi-agent debate,  $M$  copies of the same model engage in iterative discussion: each agent generates an initial response, then all agents see each other’s reasoning and update their answers over  $R - 1$  subsequent rounds of deliberation. Answers that persist indicate stronger reasoning. The framework requires no external supervision: agents supervise themselves by learning from their own debate dynamics. Specifically, for each prompt  $x$ , the debate produces final responses  $\mathcal{Y}(x) = \{y_1, \dots, y_M\}$  with extracted answers  $a_m = \mathcal{A}(y_m)$ . The majority consensus  $\hat{a}(x) = \text{Majority}\{a_1, \dots, a_M\}$  partitions  $\mathcal{Y}(x)$  into consensus-supporting  $\mathcal{G}^+(x) = \{y \in \mathcal{Y}(x) : \mathcal{A}(y) = \hat{a}(x)\}$  and dissenting  $\mathcal{G}^-(x) = \{y \in \mathcal{Y}(x) : \mathcal{A}(y) \neq \hat{a}(x)\}$  trajectories. This creates a fixed post-training dataset  $\mathcal{D}_{\text{post}} = \{(x, \hat{a}(x), \mathcal{G}^+(x), \mathcal{G}^-(x))\}_{x \in \mathcal{D}}$  where  $\mathcal{D}$  is the original set of prompts. Debate consensus, arising through deliberative exchange rather than statistical sampling, provides rich training signals. We adapt four post-training objectives to this self-generated data, treating consensus-supporting trajectories ( $\mathcal{G}^+$ ) as preferred and dissenting trajectories ( $\mathcal{G}^-$ ) as not preferred. By learning to separate these groups, the model internalizes the nuanced differences between stable consensus and dissenting reasoning. Details in Alg. 1 in App. E.

**Majority-Vote SFT (MV-SFT)** trains the model to mimic consensus-supporting trajectories:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^+ \in \mathcal{G}^+(x)} [\log \pi_{\theta}(y^+ | x)]. \quad (1)$$

**Majority-Vote GRPO (MV-GRPO)** uses online sampling with consensus-based rewards. For each prompt  $x$ , we sample multiple trajectories from the current policy and assign reward  $r_x(y) = \mathbf{1}[\mathcal{A}(y) = \hat{a}(x)]$  based on whether each sample’s answer matches the pre-computed consensus:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\theta}} \left[ \tilde{A}_x(y) \sum_t \log \pi_{\theta}(y_t | x, y_{<t}) \right] + \lambda \text{KL}(\pi_{\theta} \| \pi_{\text{ref}}), \quad (2)$$

where  $\tilde{A}_x(y) = r_x(y) - \bar{r}_x$  is the group-normalized advantage. We find that model inconsistency naturally yields both consensus and dissenting trajectories, allowing GRPO’s group normalization to contrast majority/minority outcomes within batches.

**Majority-Vote DPO (MV-DPO)** follows the standard DPO formulation with preference pairs constructed from our pre-generated debate outcomes:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{(y^+, y^-) \in \mathcal{G}^+(x) \times \mathcal{G}^-(x)} \left[ \log \sigma \left( \beta \left[ \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)} \right] \right) \right]. \quad (3)$$

By contrasting the model’s own consensus and dissenting trajectories, DPO’s log probability ratios capture differences across entire reasoning chains, not just final answers, allowing each token to contribute to the preference signal.

**Majority-Vote KTO (MV-KTO)** applies KTO’s unpaired formulation with debate-derived labels from our fixed dataset, with class-balancing weights  $\lambda_+$  and  $\lambda_-$ :

$$\begin{aligned} \mathcal{L}_{\text{KTO}}(\theta) = & -\lambda_+ \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^+ \in \mathcal{G}^+(x)} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} \right) \right] \\ & - \lambda_- \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^- \in \mathcal{G}^-(x)} \left[ \log \sigma \left( -\beta \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)} \right) \right]. \end{aligned} \quad (4)$$

KTO’s unpaired structure handles imbalanced outcomes where majority trajectories dominate.

## 4 RESULTS AND DISCUSSION

We evaluate MACA by post-training four instruction-tuned small LMs (Qwen-2B (Yang et al., 2024), Llama-3B (Grattafiori et al., 2024), Phi-4B (Abdin et al., 2024), Llama-8B (Grattafiori et al., 2024)) on six reasoning benchmarks (MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), MathQA (Amini et al., 2019), SVAMP (Patel et al., 2021), GPQA (Rein et al., 2023), CommonsenseQA (Talmor et al., 2019)). We use 4-bit quantization with QLoRA (Dettmers et al., 2023), limit responses to 256 tokens, and run multi-agent debate with  $M = 3$  clones and  $R = 2$  rounds—choices driven by the computational constraints of multi-agent debate RL on a single-node cluster. With temperature  $\tau = 1.0$ , this tests exploratory sampling under a budget adequate for efficient solvers while remaining trainable with multiple concurrent agents. Training efficiency analysis is in App. S. Crucially, improvements persist without token limits or quantization (App. K.2, App. P), indicating gains generalize beyond these settings. We compare (1) pre-trained models, (2) SFT baselines, and (3) MACA variants of GRPO, DPO, and KTO, training and evaluating on 1500/500 train/test splits per dataset independently unless otherwise specified. We report mean agent accuracy with standard deviation across three seeds. Debate prompts, training parameters, and other details are in App. F.

### 4.1 POST-TRAINING ON MAD IMPROVES DEBATE PERFORMANCE

We evaluate MACA in the multi-agent debate setting, examining how post-training improves both the final consensus accuracy and individual agent performance across debate rounds.

**Impact on multi-agent debate setting.** We evaluate MACA in the multi-agent setting using final-round majority-vote accuracy (the consensus after debate). Baselines are (1) the base model’s debate outcome (Du et al., 2023) and (2) MV-SFT on majority traces (Subramaniam et al., 2025). As shown in Table 1, post-training on debate-derived signals improves ensemble accuracy across all models and datasets. Preference learning variants (MV-DPO and MV-KTO) provide the largest and most consistent gains, up to +42.73 percentage points, by directly optimizing the log-probability gap between trajectories, with MV-KTO performing best on smaller models ( $\leq 3B$ ) and MV-DPO on larger ones (4 – 8B). Learning to differentiate between full reasoning trajectory pairs appears to better address credit assignment challenges in sparse final-answer supervision compared to both MV-GRPO’s scalar rewards and MV-SFT’s imitation learning, simultaneously improving answer accuracy while reinforcing higher-quality and more concise intermediate reasoning steps (App. N).

**Impact on debate dynamics.** Fig. 2 summarizes four metrics: initial round average single-agent accuracy, initial round ensemble majority vote (not used in the debate), final round average single-agent accuracy (conditioned on debate context/peer chains-of-thought), and final round ensemble majority vote (consensus). Post-training produces the largest gains where agents leverage peer feedback, since it directly teaches effective peer context utilization. We also observe that post-training

Table 1: Post-training multi-agent debate yields consistent accuracy improvements.

Model	Dataset	Pre / SFT (Baseline)		RL (Our Methods)			Best $\Delta$
		Debate	MV-SFT	MV-GRPO	MV-KTO	MV-DPO	
Qwen2B	MATH	32.40	37.07 $\pm$ 3.07	39.00 $\pm$ 1.74	<b>46.47 <math>\pm</math> 3.01</b>	42.60 $\pm$ 1.78	$\uparrow$ 14.07
	GSM8K	49.60	50.53 $\pm$ 1.36	54.13 $\pm$ 2.02	<b>63.07 <math>\pm</math> 0.64</b>	58.47 $\pm$ 1.62	$\uparrow$ 13.47
	MathQA	24.20	26.27 $\pm$ 0.58	29.93 $\pm$ 4.99	<b>32.60 <math>\pm</math> 0.72</b>	28.33 $\pm$ 0.31	$\uparrow$ 9.13
Llama3B	MATH	37.80	35.33 $\pm$ 1.62	48.33 $\pm$ 2.19	<b>52.93 <math>\pm</math> 0.99</b>	51.93 $\pm$ 1.67	$\uparrow$ 15.27
	GSM8K	65.60	64.60 $\pm$ 1.59	68.60 $\pm$ 1.00	<b>73.13 <math>\pm</math> 0.83</b>	71.67 $\pm$ 3.03	$\uparrow$ 7.80
	MathQA	21.60	40.07 $\pm$ 2.00	48.73 $\pm$ 1.60	<b>64.00 <math>\pm</math> 0.53</b>	63.13 $\pm$ 1.89	$\uparrow$ 42.73
Phi4B	MATH	44.40	45.53 $\pm$ 2.53	<b>49.93 <math>\pm</math> 1.33</b>	45.27 $\pm$ 0.70	46.73 $\pm$ 1.67	$\uparrow$ 5.53
	GSM8K	79.60	78.93 $\pm$ 0.61	82.67 $\pm$ 1.81	82.47 $\pm$ 1.14	<b>84.73 <math>\pm</math> 0.31</b>	$\uparrow$ 5.13
	MathQA	49.60	50.87 $\pm$ 1.62	<b>63.07 <math>\pm</math> 1.21</b>	51.53 $\pm$ 1.14	51.40 $\pm$ 1.44	$\uparrow$ 13.47
Llama8B	MATH	32.80	34.13 $\pm$ 0.70	45.93 $\pm$ 1.03	53.93 $\pm$ 1.80	<b>59.67 <math>\pm</math> 1.33</b>	$\uparrow$ 26.87
	GSM8K	74.00	66.27 $\pm$ 1.01	81.53 $\pm$ 2.81	81.00 $\pm$ 1.97	<b>81.93 <math>\pm</math> 1.51</b>	$\uparrow$ 8.60
	MathQA	44.60	44.13 $\pm$ 1.10	57.27 $\pm$ 0.61	62.00 $\pm$ 2.03	<b>69.27 <math>\pm</math> 1.55</b>	$\uparrow$ 24.67

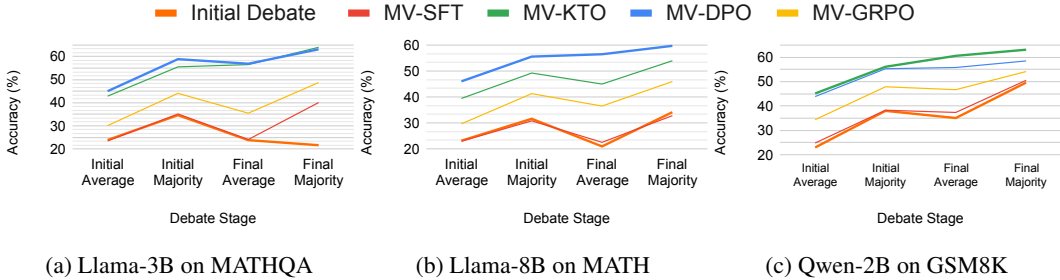


Figure 2: Debate-aware RL improves all stages of multi-agent debate. Incorporating debate context in RL teaches agents to leverage prior arguments, improving final-round consensus. Stages: initial round average, initial round majority vote, final round average, final round majority vote.

mitigates debate-driven performance degradation observed in MATHQA for Llama-3B. The contrastive signal allows models to learn from consensus patterns even when exposed to flawed reasoning trajectories, correcting rather than amplifying poor grounding behaviors.

**Impact on a single agent in zero-shot setting.** Post-training on debate also improves the performance of the individual agents in the initial round (see Initial Average in Fig. 2): across 12 model-dataset pairs, MV-RL methods consistently outperform the Base and MV-SFT baselines (Table 2) in single-agent zero-shot (single trajectory) settings. Self-guided preference learning (MV-DPO and MV-KTO) outperforms scalar rewards via MV-GRPO for all models except Phi-4B. MV-DPO is best in 7/12 cases, while MV-KTO is better in some cases for smaller LMs.

#### 4.2 POST-TRAINING ON MAD IMPROVES SELF-CONSISTENCY AND ACCURACY

We measure the effect of post-training on sampling consistency  $s_t^{\theta, \tau}(x)$ , the fraction of sampled trajectories that match the majority answer. As formalized in Sec. 2, we track the sampling consistency where  $s_t^{\theta, \tau}(x)$  converges to the modal probability  $S_{\theta, \tau}^+(x)$  as  $t \rightarrow \infty$ . For each model, we sample 20 trajectories on 500 held-out prompts and evaluate  $s_t^{\theta, \tau}(x)$  for  $t = 1$  to 20 (App. K.1). At  $t = 1$ , this metric primarily captures answer completeness, i.e., the percentage of responses that produce parseable answers within the token window, which post-training substantially improves. MV-DPO and MV-KTO achieve these gains through self-supervised preference learning alone, demonstrating that models can teach themselves more efficient reasoning without format rewards (App. N), which can otherwise be spurious (Huang et al., 2024; Srivastava et al., 2025). As sample size increases, the metric transitions to measuring true cross-sample agreement. The curves stabilize up to 27.6 percentage points above baseline, demonstrating that post-training increases answer concentration:

Table 2: Accuracy impact of MACA on single agent performance in zero-shot setting.

Model	Dataset	Base / SFT (Baseline)		RL (Our Methods)			Best $\Delta$
		Base	MV-SFT	MV-GRPO	MV-KTO	MV-DPO	
Qwen2B	MATH	7.67	11.51 $\pm$ 0.60	18.09 $\pm$ 0.71	20.18 $\pm$ 0.67	<b>23.49 <math>\pm</math> 2.30</b>	$\uparrow$ 15.82
	GSM8K	23.00	24.84 $\pm$ 0.87	34.40 $\pm$ 2.08	<b>45.13 <math>\pm</math> 1.80</b>	43.87 $\pm$ 1.92	$\uparrow$ 22.71
	MathQA	5.00	5.25 $\pm$ 0.33	17.27 $\pm$ 2.88	<b>22.16 <math>\pm</math> 1.14</b>	20.91 $\pm$ 0.47	$\uparrow$ 17.27
Llama3B	MATH	27.87	25.89 $\pm$ 0.56	35.22 $\pm$ 0.44	40.64 $\pm$ 1.25	<b>40.71 <math>\pm</math> 0.08</b>	$\uparrow$ 13.26
	GSM8K	57.33	55.98 $\pm$ 0.68	52.40 $\pm$ 2.84	<b>65.76 <math>\pm</math> 1.44</b>	64.98 $\pm$ 1.67	$\uparrow$ 8.80
	MathQA	23.87	23.44 $\pm$ 0.73	30.09 $\pm$ 1.98	42.84 $\pm$ 0.67	<b>45.00 <math>\pm</math> 2.23</b>	$\uparrow$ 21.13
Phi4B	MATH	34.60	34.60 $\pm$ 0.82	<b>37.42 <math>\pm</math> 0.16</b>	33.84 $\pm$ 0.78	34.62 $\pm$ 1.48	$\uparrow$ 2.82
	GSM8K	67.27	69.58 $\pm$ 0.76	67.13 $\pm$ 3.60	75.60 $\pm$ 1.80	<b>76.87 <math>\pm</math> 0.36</b>	$\uparrow$ 9.84
	MathQA	34.87	34.04 $\pm$ 0.58	<b>45.52 <math>\pm</math> 2.19</b>	33.91 $\pm$ 0.16	33.91 $\pm$ 0.50	$\uparrow$ 10.65
Llama8B	MATH	22.93	23.16 $\pm$ 0.14	29.66 $\pm$ 1.27	39.42 $\pm$ 0.44	<b>46.00 <math>\pm</math> 0.35</b>	$\uparrow$ 23.07
	GSM8K	57.93	42.09 $\pm$ 1.28	62.45 $\pm$ 6.01	72.36 $\pm$ 1.34	<b>77.36 <math>\pm</math> 0.27</b>	$\uparrow$ 19.43
	MathQA	29.67	30.84 $\pm$ 0.60	33.07 $\pm$ 1.11	38.42 $\pm$ 1.22	<b>51.18 <math>\pm</math> 0.24</b>	$\uparrow$ 21.51

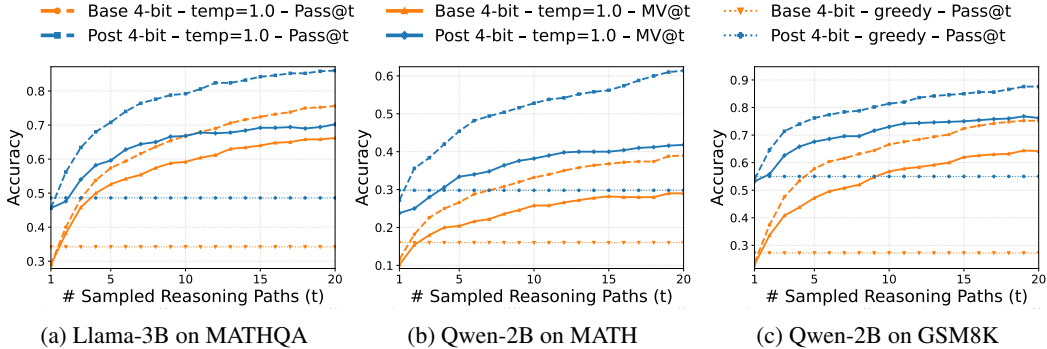


Figure 3: Post-training self-consistency improves sampling accuracy. Dashed: Pass@t (oracle upper bound), solid: MV@t (majority over  $t$  samples), dotted: greedy ( $\tau = 0$ ) accuracy. (Blue): post-trained model. (Orange): base model. Curves computed over 500 prompts.

models more consistently sample trajectories that converge despite high-temperature exploration. These improvements persist when tested without max token constraints (App. K.2).

In multi-agent debate, we evaluate the agreement metric  $d_M^{\theta, \tau}(x)$  defined in Sec. 2. Base small LMs initially struggle to reach meaningful consensus: for Qwen-2B on GSM8K, most of the “consensus” comes from random tie-breaking (1/3 agreement) or weak majorities (2/3), with only 13.4% reaching unanimity (Fig. 12, App. Q). Post-training with MACA thoroughly addresses this: non-parseable responses drop from 13.8% to 0.6%, no-agreement cases from 45.6% to 19.8%, while unanimous agreement triples from 13.4% to 43.4%. This confirms that MACA improves both individual reasoning quality and collaborative grounding, enabling genuine consensus rather than noisy aggregation (agreement distributions in App. Q).

**Impact on inference-time sampling.** We next examine how gains in self-consistency translate to inference-time performance under various sampling regimes. For each prompt, we draw  $t$  trajectories and report two metrics: Pass@t, the fraction of prompts for which at least one of the first  $t$  samples is correct (Chen et al., 2021), and MV@t, the fraction for which a majority vote over those  $t$  samples matches the ground truth (ties counted as incorrect) (Li et al., 2024). As shown in Fig. 3, post-training (blue) (i) lifts greedy ( $\tau=0$ ) accuracy, (ii) increases MV@t at each fixed  $t$ , and (iii) raises Pass@t, the empirical sampling upper bound, indicating higher achievable accuracy at any given sampling budget. When additional inference compute is available (i.e.,  $t > 1$ ), sampling techniques continue to see gains on top of the post-training improvements, showing MACA is complementary to inference-time sampling. The same pattern holds with full-precision, with im-

Table 3: Post-training self-consistency improves performance across general reasoning benchmarks. Models trained on datasets (columns) tested on benchmarks (rows). Bottom rows show generalization to unseen benchmarks: SVAMP (math), GPQA (science), CSQA (commonsense). All = joint training on combined datasets. Arrows show absolute gains over instruction-tuned model.

Test	Qwen2B (post-trained on)					Llama3B (post-trained on)				
	Base	MATH	GSM	MQA	All	Base	MATH	GSM	MQA	All
MATH	10.4	↑10.0	↑3.8	↑10.8	↑12.2	32.0	↑19.4	↑16.4	↑18.2	↑21.2
GSM	27.0	↑20.0	↑25.6	↑22.6	↑27.8	69.6	↑6.0	↑6.8	↑8.4	↑10.8
MQA	7.4	↑12.6	↑17.0	↑15.4	↑21.4	24.6	↑14.0	↑13.4	↑21.2	↑21.6
SVAMP	48.3	↑19.0	↑18.0	↑17.0	↑27.7	71.3	↑6.0	↑6.4	↑9.7	↑7.1
GPQA	0.5	↑6.0	↑5.3	↑12.8	↑16.3	0.7	↑5.4	↑6.3	↑9.8	↑10.7
CSQA	3.8	↑19.8	↑43.0	↑54.0	↑59.6	53.0	↑7.4	↑10.6	↑11.6	↑11.0

Table 4: Multi-agent: Post-training with debate (MV) is comparable to ground-truth (GT).

Llama-8B	Dataset	Debate	SFT		KTO		DPO		GRPO	
			GT	MV	GT	MV	GT	MV	GT	MV
Single-Agent	MATH	22.93	<b>23.73</b>	23.00	<b>41.20</b>	39.40	45.13	<b>46.40</b>	29.07	<b>31.13</b>
	GSM8K	57.93	40.20	<b>41.67</b>	<b>72.60</b>	70.87	76.33	<b>77.67</b>	61.27	<b>66.87</b>
Multi-Agent	MATH	32.80	<b>34.00</b>	33.40	<b>56.00</b>	55.80	<b>61.80</b>	60.80	<b>48.60</b>	44.80
	GSM8K	74.00	<b>65.20</b>	<b>65.20</b>	<b>81.20</b>	79.40	81.60	<b>83.0</b>	83.20	<b>84.20</b>

improvements comparable to their 4-bit counterparts (App. P). Finally, self-consistency gains strongly correlate with accuracy improvements ( $r > 0.86$  across all tested inference conditions; see App. K).

### Improving self-consistency on math datasets improves generalization to unseen benchmarks.

These consistency and accuracy gains extend beyond the training distribution: training to improve internal self-consistency on any mathematical dataset enhances performance across diverse reasoning tasks, including previously unseen domains. Building on recent work showing math training enhances general reasoning (DeepSeek-AI et al., 2026; Akter et al., 2025), we demonstrate that improving self-consistency—internal consensus strength—on math is essential for reliably unlocking these capabilities. Table 3 reports results for models trained on MATH, GSM8K, or MathQA individually, and on all three combined (All), using MV-DPO. Training on any single dataset improves performance across all reasoning tasks, including unseen math (SVAMP), science (GPQA), and commonsense reasoning (CSQA). Joint training achieves further improvements across nearly every benchmark, demonstrating that diverse training data amplifies self-consistency gains.

### 4.3 ABLATION STUDY

To understand MACA’s performance gains, we conduct ablation studies examining key components. We show that self-generated consensus signals outperform ground-truth supervision, including peer context improves relative grounding and debate utilization, and multi-round debate provides stronger signals than single-round majority vote. We also demonstrate in App. N that self-supervised preference learning serves as an effective implicit format reward, reinforcing more concise CoT construction, though most gains stem from reasoning improvements rather than formatting alone.

**RL with debate is comparable to RL with ground-truth.** Table 4 compares post-training with debate majority-vote labels, derived from the model’s own consensus without external supervision, to ground-truth labels on Llama-3B. Across post-training methods and in both single- and multi-agent settings, using Debate-MV is consistently comparable to ground-truth supervision, a trend that holds across other models as well (App. G). We additionally show in App. L that this comparable performance between Debate-MV and ground-truth supervision extends to general reasoning improvements under MACA across all proposed post-training methods and reasoning benchmarks tested, as observed in Table 3. Because Debate-MV scales naturally with sample size, its parity with ground-truth highlights self-supervised alignment as a promising direction.

Table 5: Single-agent: RL on debate majority-vote with peer context (MV-DPO) outperforms RL on single-round majority vote (TTRL, ScPO). Arrows show gains over instruction-tuned model. Statistics gathered over 3 seeds.

	MATH				GSM8K			
	Qwen-2B	Llama-3B	Phi-4B	Llama-8B	Qwen-2B	Llama-3B	Phi-4B	Llama-8B
TTRL	$\uparrow 18.0_{\pm 2.9}$	$\uparrow 5.3_{\pm 5.7}$	$\uparrow 6.1_{\pm 2.1}$	$\uparrow 7.5_{\pm 0.2}$	$\uparrow 24.5_{\pm 1.6}$	$\downarrow 22.3_{\pm 8.3}$	$\uparrow 4.8_{\pm 1.6}$	$\downarrow 13.5_{\pm 5.7}$
ScPO	$\uparrow 2.3_{\pm 1.1}$	$\uparrow 3.4_{\pm 0.2}$	$\uparrow 0.1_{\pm 0.5}$	$\uparrow 3.7_{\pm 0.6}$	$\uparrow 15.9_{\pm 3.4}$	$\uparrow 1.6_{\pm 0.8}$	$\uparrow 5.9_{\pm 0.5}$	$\uparrow 12.9_{\pm 1.2}$
MV-DPO	$\uparrow 16.7_{\pm 0.4}$	$\uparrow 12.5_{\pm 0.7}$	$\uparrow 6.9_{\pm 0.2}$	$\uparrow 17.1_{\pm 0.8}$	$\uparrow 23.8_{\pm 0.9}$	$\uparrow 3.7_{\pm 0.3}$	$\uparrow 7.3_{\pm 0.3}$	$\uparrow 18.2_{\pm 0.2}$

Table 6: Multi-agent: Post-training with debate context (peer CoTs) outperforms no context. Models: 3B = Llama-3B; 4B = Phi-4B; 8B = Llama-8B. Dataset: GSM8K

	Initial Round Avg		Final Round Avg		Final Round MV	
	NoCtx	Context	NoCtx	Context	NoCtx	Context
3B	65.87	<b>66.73</b>	67.87	<b>70.60</b>	70.60	<b>73.80</b>
4B	74.67	<b>76.60</b>	74.60	<b>77.80</b>	81.80	<b>84.40</b>
8B	75.80	<b>77.67</b>	78.00	<b>81.20</b>	81.60	<b>83.00</b>

**RL with debate outperforms RL with single-round majority vote.** We evaluate three training conditions in the single-agent setting to understand the impact of training on the debate versus initial round majority vote on the individual agents themselves, as shown in Table 5: (1) initial round majority vote with GRPO (akin to TTRL (Zuo et al., 2025)), (2) initial round majority vote with DPO (akin to ScPO without the weighted loss term (Prasad et al., 2025)), and (3) debate majority vote with DPO including peer context (MV-DPO from MACA). Training on the debate signal improves single-agent performance more than reinforcing the initial round majority vote. Additionally, Table 6 shows that including peer context during preference learning (MV-DPO) improves multi-agent debate performance (more models/datasets in App. H), with the largest gains in final-round individual accuracy, reflecting better use of peer chains-of-thought during deliberation (Fig. 2).

**Debate improves consensus quality.** Finally, the debate process itself improves consensus signals. Base small LMs produce mostly random or weak majorities initially, but post-training increases unanimous agreement from 27.2% to 43.4% between initial and final rounds (Qwen-2B on GSM8K, App. Q). This iterative refinement provides more reliable training signals than independent sampling. While majority vote correlates strongly with accuracy (alternative ranking signals like log-probability are explored in App. J), post-training debate creates a self-reinforcing cycle: better debate generates higher-quality signals, producing models that debate more effectively. Iterative training yields continued improvements with diminishing returns, as shown in App. I.

## 5 CONCLUSION AND LIMITATIONS

We introduce MACA, a self-supervised framework that reinforces consensus signals from multi-agent debate to improve reasoning without external supervision. MACA yields substantial gains: +27.6% self-consistency on GSM8K, +21.51% accuracy on MathQA, +26.87% collaborative performance on MATH, and strong generalization to unseen benchmarks (+16.3% GPQA, +11.6% CommonsenseQA). By selecting debate trajectories where peer deliberation aligns with internal consensus, MACA provides natural supervision for stable reasoning, enabling models to self-improve consistency, efficiency, and accuracy without chain-of-thought supervision or external verification, while producing more concise reasoning and correcting cases where aggregation previously led to degeneration. MACA requires sufficient base model competence for meaningful consensus signals, may amplify existing biases, and lacks direct supervision of intermediate reasoning steps; future work could address these through alternative consensus methods, confidence-weighted voting, heterogeneous agents, and better use of minority traces. Nevertheless, gains on difficult unseen tasks (+10.0% on AMC, App. N) suggest consensus alignment unlocks latent capabilities beyond consistency alone.

## REFERENCES

- 432  
433  
434 March 2025. <https://huggingface.co/datasets/knoveleng/AMC-23>.
- 435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485
- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. (arXiv:2412.08905), December 2024. doi: 10.48550/arXiv.2412.08905. <http://arxiv.org/abs/2412.08905>. arXiv:2412.08905.
- Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-crossthink: Scaling self-learning beyond math reasoning. (arXiv:2504.13941), April 2025. doi: 10.48550/arXiv.2504.13941. <http://arxiv.org/abs/2504.13941>. arXiv:2504.13941.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North*, page 2357–2367, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. <http://aclweb.org/anthology/N19-1245>.
- Matthew M. Botvinick, Jonathan D. Cohen, and Cameron S. Carter. Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, 8(12):539–546, December 2004. ISSN 13646613. doi: 10.1016/j.tics.2004.10.003. <https://linkinghub.elsevier.com/retrieve/pii/S1364661304002657>.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. (arXiv:2107.03374), 2021. doi: 10.48550/arXiv.2107.03374. <http://arxiv.org/abs/2107.03374>. arXiv:2107.03374.
- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. (arXiv:2311.17311), November 2023. doi: 10.48550/arXiv.2311.17311. <http://arxiv.org/abs/2311.17311>. arXiv:2311.17311.
- 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:2110.14168), November 2021. doi: 10.48550/arXiv.2110.14168. <http://arxiv.org/abs/2110.14168>. arXiv:2110.14168.
- Jean-Antoine-Nicolas de Caritat marquis de Condorcet. *Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix*. Chelsea, New York, fac-sim. ed. edition, 1972. ISBN 9780828402521.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,

- 486 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang  
487 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai  
488 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,  
489 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,  
490 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,  
491 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,  
492 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghai Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng  
493 Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing  
494 Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanxia Zhao, Wen  
495 Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong  
496 Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu,  
497 Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xi-  
498 aosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia  
499 Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng  
500 Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong  
501 Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong,  
502 Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou,  
503 Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying  
504 Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda  
505 Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia  
506 Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu,  
507 Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via  
508 reinforcement learning. (arXiv:2501.12948), January 2026. doi: 10.48550/arXiv.2501.12948.  
<http://arxiv.org/abs/2501.12948>. arXiv:2501.12948.
- 509 Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning  
510 of quantized llms. (arXiv:2305.14314), May 2023. doi: 10.48550/arXiv.2305.14314. <http://arxiv.org/abs/2305.14314>. arXiv:2305.14314.
- 512 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving  
513 factuality and reasoning in language models through multiagent debate. (arXiv:2305.14325),  
514 May 2023. doi: 10.48550/arXiv.2305.14325. <http://arxiv.org/abs/2305.14325>.  
515 arXiv:2305.14325.
- 516 Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich  
517 Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language mod-  
518 els. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi:  
519 10.1162/tacl.a.00410. <https://aclanthology.org/2021.tacl-1.60/>.
- 520 Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model  
521 alignment as prospect theoretic optimization. (arXiv:2402.01306), November 2024. doi: 10.  
522 48550/arXiv.2402.01306. <http://arxiv.org/abs/2402.01306>. arXiv:2402.01306.
- 524 Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140,  
525 May 1954. ISSN 0018-7267, 1741-282X. doi: 10.1177/001872675400700202. <https://journals.sagepub.com/doi/10.1177/001872675400700202>.
- 527 Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11  
528 (2):127–138, February 2010. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn2787. <https://www.nature.com/articles/nrn2787>.
- 530 Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom  
531 Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici,  
532 Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat,  
533 Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan,  
534 Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary  
535 Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an  
536 ai co-scientist. (arXiv:2502.18864), February 2025. doi: 10.48550/arXiv.2502.18864. <http://arxiv.org/abs/2502.18864>. arXiv:2502.18864.
- 538 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad  
539 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,

540 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Ko-  
541 renev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava  
542 Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux,  
543 Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,  
544 Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,  
545 Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary,  
546 Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab  
547 AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco  
548 Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind That-  
549 tai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Kore-  
550 vaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,  
551 Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-  
552 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,  
553 Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jong-  
554 soo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala,  
555 Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid  
556 El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren  
557 Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin,  
558 Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi,  
559 Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew  
560 Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Ku-  
561 mar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoy-  
562 chev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan  
563 Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan,  
564 Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ra-  
565 mon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Ro-  
566 hit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan  
567 Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell,  
568 Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng  
569 Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer  
570 Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman,  
571 Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mi-  
572 haylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor  
573 Kerkez, Vincent Gouget, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei  
574 Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang  
575 Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Gold-  
576 schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning  
577 Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh,  
578 Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria,  
579 Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein,  
580 Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, An-  
581 drew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, An-  
582 nie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,  
583 Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leon-  
584 hardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu  
585 Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Mon-  
586 talvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao  
587 Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia  
588 Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide  
589 Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le,  
590 Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily  
591 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smoth-  
592 ers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni,  
593 Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia  
Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan,  
Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harri-  
son Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,  
Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James  
Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

- 594 nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang,  
595 Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Jun-  
596 jie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy  
597 Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang,  
598 Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,  
599 Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa,  
600 Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias  
601 Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L.  
602 Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike  
603 Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari,  
604 Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan  
605 Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong,  
606 Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth  
607 Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina  
608 Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez,  
609 Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Ray-  
610 mond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta,  
611 Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha  
612 Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy,  
613 Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil,  
614 Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe,  
615 Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan  
616 Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian,  
617 Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas  
618 Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Von-  
619 timitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla,  
620 Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang,  
621 Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang,  
622 Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin  
623 Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian,  
624 Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhi-  
625 wei Zhao, and Zhiyu Ma. The llama 3 herd of models. (arXiv:2407.21783), November 2024. doi:  
626 10.48550/arXiv.2407.21783. <http://arxiv.org/abs/2407.21783>. arXiv:2407.21783.
- 625 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn  
626 Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset.  
627 (arXiv:2103.03874), November 2021. doi: 10.48550/arXiv.2103.03874. <http://arxiv.org/abs/2103.03874>. arXiv:2103.03874.
- 628 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text  
629 degeneration. (arXiv:1904.09751), February 2020. doi: 10.48550/arXiv.1904.09751. <http://arxiv.org/abs/1904.09751>. arXiv:1904.09751.
- 630 Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song,  
631 and Denny Zhou. Large language models cannot self-correct reasoning yet. (arXiv:2310.01798),  
632 March 2024. doi: 10.48550/arXiv.2310.01798. <http://arxiv.org/abs/2310.01798>.  
633 arXiv:2310.01798.
- 634 Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. (arXiv:1805.00899),  
635 October 2018. doi: 10.48550/arXiv.1805.00899. <http://arxiv.org/abs/1805.00899>.  
636 arXiv:1805.00899.
- 637 Jiechuan Jiang and Zongqing Lu. Offline decentralized multi-agent reinforcement learning.  
638 (arXiv:2108.01832), 2023. doi: 10.48550/arXiv.2108.01832. <http://arxiv.org/abs/2108.01832>.  
639 arXiv:2108.01832.
- 640 Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F. Chen, Shafiq Joty, and Furu Wei. Preference  
641 optimization for reasoning with pseudo feedback. (arXiv:2411.16345), February 2025. doi: 10.  
642 48550/arXiv.2411.16345. <http://arxiv.org/abs/2411.16345>. arXiv:2411.16345.

- 648 Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang,  
649 Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge.  
650 (arXiv:2502.01534), May 2025. doi: 10.48550/arXiv.2502.01534. [http://arxiv.org/  
651 abs/2502.01534](http://arxiv.org/abs/2502.01534). arXiv:2502.01534.
- 652 Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. Self-prompting large language models  
653 for zero-shot open-domain qa. In *Proceedings of the 2024 Conference of the North American  
654 Chapter of the Association for Computational Linguistics: Human Language Technologies  
655 (Volume 1: Long Papers)*, page 296–310, Mexico City, Mexico, 2024. Association for Computa-  
656 tional Linguistics. doi: 10.18653/v1/2024.naacl-long.17. [https://aclanthology.org/  
657 2024.naacl-long.17](https://aclanthology.org/2024.naacl-long.17).
- 658 Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. *Annual  
659 Review of Neuroscience*, 24(1):167–202, March 2001. ISSN 0147-006X, 1545-4126. doi:  
660 10.1146/annurev.neuro.24.1.167. [https://www.annualreviews.org/doi/10.1146/  
661 annurev.neuro.24.1.167](https://www.annualreviews.org/doi/10.1146/annurev.neuro.24.1.167).
- 662 Serge Moscovici and Elisabeth Lage. Studies in social influence iii: Majority versus minority influ-  
663 ence in a group. *European Journal of Social Psychology*, 6(2):149–174, April 1976. ISSN 0046-  
664 2772, 1099-0992. doi: 10.1002/ejsp.2420060202. [https://onlinelibrary.wiley.  
665 com/doi/10.1002/ejsp.2420060202](https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2420060202).
- 666 Charlan J. Nemeth. Differential contributions of majority and minority influence. *Psychological  
667 Review*, 93(1):23–32, January 1986. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.93.  
668 1.23. <https://doi.apa.org/doi/10.1037/0033-295X.93.1.23>.
- 669 Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math  
670 word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the  
671 Association for Computational Linguistics: Human Language Technologies*, page 2080–2094,  
672 Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.  
673 168. <https://aclanthology.org/2021.naacl-main.168>.
- 674 Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit  
675 Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. Self-consistency preference optimiza-  
676 tion. (arXiv:2411.04109), 2025. doi: 10.48550/arXiv.2411.04109. [http://arxiv.org/  
677 abs/2411.04109](http://arxiv.org/abs/2411.04109). arXiv:2411.04109.
- 678 Bhaktipriya Radharapu, Manon Revel, Megan Ung, Sebastian Ruder, and Adina Williams. Ar-  
679 biters of ambivalence: Challenges of using llms in no-consensus tasks. (arXiv:2505.23820),  
680 May 2025. doi: 10.48550/arXiv.2505.23820. <http://arxiv.org/abs/2505.23820>.  
681 arXiv:2505.23820.
- 682 Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and  
683 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model.  
684 (arXiv:2305.18290), 2024. doi: 10.48550/arXiv.2305.18290. [http://arxiv.org/abs/  
685 2305.18290](http://arxiv.org/abs/2305.18290). arXiv:2305.18290.
- 686 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien  
687 Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a  
688 benchmark. (arXiv:2311.12022), November 2023. doi: 10.48550/arXiv.2311.12022. [http:  
689 //arxiv.org/abs/2311.12022](http://arxiv.org/abs/2311.12022). arXiv:2311.12022.
- 690 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
691 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits  
692 of mathematical reasoning in open language models. (arXiv:2402.03300), April 2024. doi:  
693 10.48550/arXiv.2402.03300. <http://arxiv.org/abs/2402.03300>. arXiv:2402.03300.
- 694 Amitai Shenhav, Matthew M. Botvinick, and Jonathan D. Cohen. The expected value of control:  
695 An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013. ISSN  
696 08966273. doi: 10.1016/j.neuron.2013.07.007. [https://linkinghub.elsevier.com/  
697 retrieve/pii/S0896627313006077](https://linkinghub.elsevier.com/retrieve/pii/S0896627313006077).

- 702 Pragma Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun  
703 Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuv  
704 Karthikeyan Shanmugam, and Doina Precup. Robust reward modeling via causal rubrics.  
705 (arXiv:2506.16507), 2025. doi: 10.48550/arXiv.2506.16507. [http://arxiv.org/abs/  
706 2506.16507](http://arxiv.org/abs/2506.16507). arXiv:2506.16507.
- 707 Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and  
708 Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains.  
709 (arXiv:2501.05707), March 2025. doi: 10.48550/arXiv.2501.05707. [http://arxiv.org/  
710 abs/2501.05707](http://arxiv.org/abs/2501.05707). arXiv:2501.05707.
- 711 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question  
712 answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference*  
713 *of the North*, page 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational  
714 Linguistics. doi: 10.18653/v1/N19-1421. <http://aclweb.org/anthology/N19-1421>.
- 715 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-  
716 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.  
717 (arXiv:2203.11171), March 2023. doi: 10.48550/arXiv.2203.11171. [http://arxiv.org/  
718 abs/2203.11171](http://arxiv.org/abs/2203.11171). arXiv:2203.11171.
- 719 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li,  
720 Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang,  
721 Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren  
722 Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang,  
723 Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin,  
724 Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong  
725 Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu,  
726 Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru  
727 Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. (arXiv:2407.10671), 2024. doi:  
728 10.48550/arXiv.2407.10671. <http://arxiv.org/abs/2407.10671>. arXiv:2407.10671.
- 729 Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and  
730 Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent  
731 reinforcement learning. (arXiv:2106.03400), October 2021. doi: 10.48550/arXiv.2106.03400.  
732 <http://arxiv.org/abs/2106.03400>. arXiv:2106.03400.
- 733 Wenhao Zhan, Scott Fujimoto, Zheqing Zhu, Jason D. Lee, Daniel R. Jiang, and Yonathan  
734 Efroni. Exploiting structure in offline multi-agent rl: The benefits of low interaction rank.  
735 (arXiv:2410.01101), October 2024. doi: 10.48550/arXiv.2410.01101. [http://arxiv.org/  
736 abs/2410.01101](http://arxiv.org/abs/2410.01101). arXiv:2410.01101.
- 737 Shuo Zhang, Yan Tian, Quanying Liu, and Haiyan Wu. The neural correlates of novelty and variabil-  
738 ity in human decision-making under an active inference framework. *eLife*, 13:RP92892, March  
739 2025. ISSN 2050-084X. doi: 10.7554/eLife.92892.4. [https://elifesciences.org/  
740 articles/92892](https://elifesciences.org/articles/92892).
- 741 Jiace Zhu, Yuanzhe Huang, Yingtao Shen, Jie Zhao, and An Zou. Path-consistency with prefix  
742 enhancement for efficient inference in llms. (arXiv:2409.01281), November 2025a. doi: 10.  
743 48550/arXiv.2409.01281. <http://arxiv.org/abs/2409.01281>. arXiv:2409.01281.
- 744 Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Er-  
745 mon, and Weinan Zhang. Madiff: Offline multi-agent learning with diffusion models.  
746 (arXiv:2305.17330), January 2025b. doi: 10.48550/arXiv.2305.17330. [http://arxiv.org/  
747 abs/2305.17330](http://arxiv.org/abs/2305.17330). arXiv:2305.17330.
- 748 Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen  
749 Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding,  
750 and Bowen Zhou. Ttrl: Test-time reinforcement learning. (arXiv:2504.16084), 2025. doi: 10.  
751 48550/arXiv.2504.16084. <http://arxiv.org/abs/2504.16084>. arXiv:2504.16084.

## A LLM USAGE DECLARATION

We declare that large language models (LLMs) were only used to assist with minor tasks, including grammar checking, language refinement, and basic coding support such as debugging. All research contributions, experimental design, data analysis, and scientific findings presented in this work are original to the authors.

## B RELATED WORK

Existing approaches address sampling inconsistency primarily through inference-time techniques. Self-consistency prompting (Wang et al., 2023; Li et al., 2024) samples multiple reasoning paths and selects the majority-voted answer, with extensions for non-verifiable outputs (Chen et al., 2023) and path pruning (Zhu et al., 2025a). Multi-agent debate frameworks (Du et al., 2023; Irving et al., 2018) similarly utilize consensus across models to improve reliability, with recent work exploring applications in scientific discovery (Gottweis et al., 2025). However, these methods require additional inference compute and do not internalize self-consistency into the model. To improve self-consistency and single rollout performance through post-training, current training-time alternatives have limitations: relative log-probability rankings as a proxy for relative confidence correlate less strongly with accuracy than majority vote signals, as we show (App. J), while LLM-as-a-Judge approaches (Jiao et al., 2025) suffer from preference leakage (Li et al., 2025) and bias under ambiguity (Radharapu et al., 2025). We improve self-consistency through post-training that optimizes consensus signals via multi-agent RL (Yang et al., 2021; Jiang and Lu, 2023; Zhu et al., 2025b; Zhan et al., 2024). Majority vote RL methods such as TTRL (Zuo et al., 2025) and ScPO (Prasad et al., 2025) use GRPO and DPO, respectively, to reinforce single-round majority vote from independently sampled responses. By contrast, MACA introduces multi-agent, multi-round debate, including deliberation traces as conditioning contexts within the prompts during RL. This exposes models to concrete examples of relative grounding where consensus emerges through explicit engagement with peer arguments, allowing them to learn not just to aggregate outputs but to internalize the deliberative process by which consensus is reached. Both TTRL and ScPO are specialized cases of the framework developed in this work, when reducing the multi-agent debate parameters to a single-round majority vote. For post-training the multi-agent debate setting specifically, Subramaniam et al. (2025) uses Supervised Fine-Tuning (SFT) for debate optimization, whereas we demonstrate that RL-based alternatives achieve superior performance.

## C EXAMPLES OF POST-TRAINED CONSISTENT REASONING IN SINGLE- AND MULTI- AGENT SETTINGS

This appendix presents examples from the debates, providing a qualitative signal of how post-training affects both single-agent reasoning and multi-agent collaboration. The evidence highlights visible improvements in mathematical reasoning, consensus formation, and systematic problem-solving.

### C.1 OVERVIEW OF KEY FINDINGS

Analysis of the transcripts reveals that post-training leads to more systematic and efficient reasoning in single-agent settings. Models demonstrate clearer step-by-step problem decomposition, make fewer calculation errors, and show improved conceptual understanding. Notably, responses become more concise and are less likely to be truncated due to inefficient reasoning.

In multi-agent debates, post-training enhances collaborative effects. Agents more frequently reference and build upon each other’s solutions, leading to more robust cross-verification and consensus formation. We observe that post-trained agents are more likely to resolve mistakes through discussion rather than propagate errors.

Our analysis shows that preference learning teaches models higher-quality and more concise reasoning without requiring explicit format rewards. The consensus mechanism itself becomes more effective at aggregating individual outputs, supporting both in-distribution and out-of-distribution transfer.

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

## C.2 CASE STUDY 1: SINGLE-AGENT MATHEMATICAL REASONING

### C.2.1 PROBLEM: CITIES A AND B RELATIVE SPEED

#### AMC Problem Statement

Cities  $A$  and  $B$  are 45 miles apart. Alicia lives in  $A$  and Beth lives in  $B$ . Alicia bikes towards  $B$  at 18 miles per hour. Leaving at the same time, Beth bikes toward  $A$  at 12 miles per hour. How many miles from City  $A$  will they be when they meet?

**Correct Solution:** Combined speed =  $18 + 12 = 30$  mph; Time to meet =  $45/30 = 1.5$  hours; Distance from  $A = 18 \times 1.5 = 27$  miles

**Ground Truth:** 27

### C.2.2 PRE-TRAINING RESPONSE: FUNDAMENTAL CONCEPTUAL CONFUSION

#### Base Model: Conceptual Breakdown

To solve this problem, we can use the concept of relative motion. Here's a step-by-step breakdown:

- Let's denote the time that Alicia travels as  $t$  hours. Since Beth travels in the opposite direction, we can say that the time she travels is also  $t$  hours.
- Since they are traveling towards each other, the distance between them will decrease at a rate of  $18 + 12 = 30$  miles per hour.
- The initial distance between them is 45 miles. After  $t$  hours, the total distance traveled by both Alicia and Beth will be the initial distance plus the distance traveled together, which is  $45 + 30t$  miles.
- We can set up an equation using the fact that their combined distance traveled equals the initial distance between them:  $45 + 30t = 45$ . This is because they meet when the sum of their distances is equal to the initial distance.
- Solving for  $t$ , we get  $t = 0$  hours. However, this doesn't make sense because it means they meet immediately after starting. Let's try again.
- Since we can't set up a simple equation with this method, let's think about it in terms of their relative speed: 12

[Response cuts off mid-sentence]

**Parsed Answer:** null

**Analysis: Multiple failure modes.** The pre-training response reveals several critical reasoning failures: incorrect problem setup, misapplication of distance-time relationships, and ultimately an illogical result. The model exhausts its token budget exploring unproductive solution paths before reaching a conclusion.

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

### C.2.3 POST-TRAINING RESPONSE: CORRECT SYSTEMATIC SOLUTION

#### Post-Training: Correct and Concise

To solve this problem, we can follow these steps:

- We know the distance between Cities A and B is 45 miles.
- Alicia and Beth are biking towards each other at a combined speed of  $18 + 12 = 30$  miles per hour.
- Since they are biking towards each other, we can use the concept of relative speed to find the time it takes for them to meet.
- We can calculate the time using the formula:  $\text{time} = \text{distance} / \text{speed}$ . Using this formula, the time they take to meet is  $45 \text{ miles} / 30 \text{ miles per hour}$ , which equals 1.5 hours.
- Now that we have the time, we can find the distance covered by Alicia, since we know her speed is 18 miles per hour. Distance covered by Alicia is  $18 \times 1.5$ , which equals 27 miles.

Therefore, Alicia and Beth will meet 27 miles from City A.

27

**Analysis: Systematic improvement.** Post-training produces a clear, systematic approach: correct conceptual application, logical step progression, and efficient completion within the token limit. This demonstrates improved reasoning quality and token efficiency achieved through preference learning without explicit format rewards.

### C.3 CASE STUDY 2: MULTI-AGENT MATHEMATICAL REASONING

#### C.3.1 PROBLEM: ROAD TRIP GAS MILEAGE CALCULATION

##### Problem Statement

Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Correct Solution:**  $100 \div 4 = 25 \text{ mpg}$ ;  $25 \times 12 = 300 \text{ miles}$

**Ground Truth:** 300

### C.4 COMPLETE MULTI-AGENT TRANSCRIPTS: PRE-TRAINING FAILURE (ITERATION 0)

#### C.4.1 ROUND 1: INITIAL AGENT RESPONSES

##### User Prompt

Solve the following math problem. Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

Provide a bullet point summary of your step-by-step reasoning. Your final answer should be a single numerical number, in the form `\boxed{answer}`, at the end of your response.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

### Agent 1: Mathematical Error

To solve this problem, I need to determine Sophia's car's fuel efficiency and then calculate how far she can drive on a full tank.

Step 1: Determine fuel efficiency from the given information.

- Sophia traveled 100 miles and used 4 gallons of gas
- Fuel efficiency = 100 miles  $\div$  4 gallons = 25 miles per gallon

Step 5: Convert the fuel efficiency from mpg to gallons per mile (GPM):

- Since 1 mpg means traveling 1 mile with each gallon, Sophia will use 4 gallons  $\times$  (1 mile/mpg) = 4 miles per tank.

Step 6: Calculate total distance for full tank:

- Total distance = 4 miles per tank  $\times$  12 gallons = 48 miles

48

### Agent 2: Incomplete Reasoning

If the distance covered in one tank of fuel is 100 miles, then the Miles Per Gallon (MPG) is calculated as follows...

[Response appears incomplete and truncated mid-calculation]

After reviewing the problem again, I need to be more careful with my calculations.

Let me reconsider: Sophia used 4 gallons to travel some distance, and her tank holds 12 gallons total.

I'm getting confused by the different approaches. Let me try again: 100 miles with 4 gallons means...

Final answer: 48

### Agent 3: Correct but Overruled

Let me solve this step by step:

Step 1: Calculate fuel efficiency from given data:

- Distance traveled: 100 miles
- Gas used: 4 gallons
- Fuel efficiency: 100  $\div$  4 = 25 miles per gallon

Step 2: Calculate range on full tank:

- Full tank capacity: 12 gallons
- Total range: 25 mpg  $\times$  12 gallons = 300 miles

Therefore, Sophia can drive 300 miles on a single tank of gas.

300

## C.4.2 ROUND 2: FAILED COLLABORATIVE DISCUSSION

### Agent 1: Doubling Down on Error

After reviewing all approaches, I maintain my calculation:

Based on the calculations, I'm confident that 48 miles is correct because we need to convert the efficiency properly.

48

**Pre-Training Result:** Consensus selected 48 miles despite Agent 3's correct reasoning.

## 972 C.4.3 COMPLETE MULTI-AGENT TRANSCRIPTS: POST-TRAINING SUCCESS (ITERATION 1)

973  
974 C.4.4 ROUND 1: PERFECT AGENT RESPONSES975 **Agent 1: Complete Correction**976  
977 To solve this problem, I need to determine Sophia’s fuel efficiency and calculate her range  
978 on a full tank.

979 Step 1: Calculate Sophia’s car’s fuel efficiency:

- 980
- Sophia traveled 100 miles since last fill-up
  - She used 4 gallons of gas
  - Fuel efficiency = 100 miles ÷ 4 gallons = 25 miles per gallon

984 Step 2: Calculate range on full tank:

- 985
- Tank capacity: 12 gallons
  - Range = 25 mpg × 12 gallons = 300 miles

987  
988 Therefore, Sophia can drive 300 miles on a single tank of gas.989 990  
991 **Unanimous Agreement**992  
993 **Agent 1:** “I see all agents used the same systematic approach. The calculation is clear: 25  
994 mpg × 12 gallons = 300 miles.”995 **Agent 2:** “Perfect agreement on methodology. The fuel efficiency calculation is the key  
996 step, then applying it to full tank capacity.”997 **Agent 3:** “Unanimous consensus on 300 miles. All reasoning aligns perfectly.”998  
999 **Post-Training Result:** Perfect consensus on 300 miles with all agents showing correct reasoning.

## 1000 C.4.5 ANALYSIS: TRAINING EFFECTS ON MULTI-AGENT COLLABORATION

1001  
1002 Post-training produces substantial improvements in both individual and collective reasoning. Be-  
1003 fore training, agents make calculation errors and the consensus mechanism selects incorrect answers  
1004 even when correct solutions are present, with two agents converging on the same mistake and over-  
1005 ruling the correct agent. After training, all agents apply correct methodology and reach unanimous  
1006 agreement on the right answer. These results demonstrate that preference learning encourages sys-  
1007 tematic step-by-step reasoning, reducing errors and their propagation during collaboration, while  
1008 multi-agent training enhances the consensus mechanism’s ability to reliably select correct solutions.1009  
1010 **D GROUNDING BEHAVIORS IN BASE-MODEL MULTI-AGENT DEBATES**1011  
1012 Having provided examples of the impact of post-training on single- and multi-agent reasoning, we  
1013 now highlight representative examples from some initial debates among base models demonstrating  
1014 spontaneous grounding behaviors that emerge during multi-agent reasoning. Note that each debate  
1015 consists of an initial round (three agents provide independent solutions) and a final round (agents  
1016 update after seeing peers’ responses).1017  
1018 **D.1 EXAMPLE 1: AIRPORT TIMING PROBLEM**1019 **Problem**1020  
1021 Amber and Josh are flying to France for vacation. Their flight leaves at 8:00 pm and as  
1022 an international flight, it’s recommended that they check in 2 hours early. It will take 45  
1023 minutes to drive to the airport and another 15 minutes to park their vehicle and make their  
1024 way to the terminal. What is the latest time they can leave their house to make it to the  
1025 airport in time?

1026 The correct solution requires working backwards from the 8:00 PM flight time: they must arrive by  
1027 6:00 PM (2 hours early), and with 1 hour total travel time, they must leave by 5:00 PM.

1028 **Initial round:**

1030 **Agent 1: Directional Error**

- 1031 • Flight leaves at 8:00 pm, need to check in 2 hours early = 6:00 pm arrival
- 1032 • Calculate travel time:  $45 + 15 = 60$  minutes = 1 hour
- 1033 • Add travel time to departure time:  $6:00 \text{ pm} + 1 \text{ hour} = 7:00 \text{ pm}$

1034 **Error: Adds travel time instead of subtracting it**  
1035 Answer: 7:00 pm

1037 **Agent 2: Fundamental Misunderstanding**

- 1038 • Find total travel time:  $45 + 15 = 60$  minutes = 1 hour
- 1039 • Add travel time to flight departure time:  $8:00 \text{ pm} + 1 \text{ hour} = 9:00 \text{ pm}$

1040 **Completely ignores 2-hour check-in requirement and adds travel time instead of subtracting it**  
1041 Answer: 9:00 pm

1042 **Agent 3: Correct Solution**

- 1043 • Flight at 8:00 pm requires arrival by 6:00 pm (2 hours early)
- 1044 • Travel time: 45 min driving + 15 min parking = 60 minutes
- 1045 • Must leave by:  $6:00 \text{ pm} - 1 \text{ hour} = 5:00 \text{ pm}$

1046 Answer: 5:00 pm

1047 **Final round (after seeing peers):**

1048 **Agent 1: Adopts Worse Error**

1049 “After reviewing other solutions, I update my approach:

- 1050 • Departure time from house: 8:00 pm
- 1051 • Total travel time:  $45 + 15 = 60$  minutes = 1 hour
- 1052 • Add travel time to departure:  $8:00 \text{ pm} + 1 \text{ hour} = 9:00 \text{ pm}$ ”

1053 **Abandons partial understanding to adopt Agent 2’s incorrect reasoning as well**  
1054 Answer: 9:00 pm

1055 **Agent 2: Persists with Error**

1056 Maintains original incorrect answer of 9:00 pm

1057 **Agent 3: Explicitly Identifies Error Pattern**

1058 “The other agents seem to be adding time to the flight departure rather than working back-  
1059 wards from the required arrival time. The correct approach:

- 1060 • Required arrival: 6:00 pm (2 hours before 8:00 pm flight)
- 1061 • Subtract travel time:  $6:00 \text{ pm} - 1 \text{ hour} = 5:00 \text{ pm}$  departure”

1062 **Recognizes error in other agents’ responses and upholds its own correct understanding**  
1063 Answer: 5:00 pm

1080 **Key observation.** Agent 1 had partial understanding (recognized the 6:00 PM arrival requirement)  
 1081 but abandoned this to adopt Agent 2’s more fundamental error. Agent 3 explicitly identifies the  
 1082 common error pattern of “adding time to flight departure” and maintains the correct backward cal-  
 1083 culation. This incidentally reflects exactly the case where statistical aggregation can be counter-  
 1084 productive: the inconsistencies in reasoning among the sampled trajectories outnumber the sensible  
 1085 agent, and this error gets propagated. Interestingly, we show that post-training is able to correct  
 1086 these cases despite the initial debate datasets containing examples of bad grounding.

## 1088 D.2 EXAMPLE 2: SELF-CORRECTION IN MATHEMATICAL REASONING

### 1090 Problem

1091 A store offers a 20% discount on all items. If Sarah buys a jacket originally priced at \$80  
 1092 and pants originally priced at \$50, how much does she pay in total after the discount?  
 1093

1094 The correct solution: Total original price = \$130, discount = 20% of \$130 = \$26, final price = \$104.

### 1096 Initial round:

#### 1098 Agent 1: Calculation Error

1099 Original total:  $\$80 + \$50 = \$130$   
 1100 Discount:  $20\% = 0.2 \times \$130 = \$26$   
 1101 Final price:  $\$130 + \$26 = \$156$   
 1102 **Error:** Adds discount instead of subtracting  
 1103 Answer: \$156  
 1104

#### 1106 Agent 2: Correct Approach

1107 Total: \$130  
 1108 After 20% discount:  $\$130 \times 0.8 = \$104$   
 1109 Answer: \$104  
 1110

#### 1112 Agent 3: Misapplies Discount

1113 Jacket discount: 20% of \$80 = \$16, price = \$64  
 1114 Pants discount: 20% of \$50 = \$10  
 1115 Total:  $\$64 + \$10 = \$74$   
 1116 **Error:** Adds discounted jacket price to pants discount amount  
 1117 Answer: \$74  
 1118

### 1119 Final round:

#### 1122 Agent 1: Self-Correction

1123 “I realize I added the discount instead of subtracting. Looking at Agent 2’s approach of  
 1124 multiplying by 0.8 is cleaner:  $\$130 \times 0.8 = \$104$ ”  
 1125 **Recognizes own error and adopts correct method**  
 1126 Answer: \$104  
 1127

#### 1128 Agent 3: Fixes Calculation

1129 “Jacket after discount:  $\$80 \times 0.8 = \$64$   
 1130 Pants after discount:  $\$50 \times 0.8 = \$40$   
 1131 Total:  $\$64 + \$40 = \$104$ ”  
 1132 Answer: \$104  
 1133

1134 **Key observation.** Agent 1 explicitly acknowledges their error (“I added the discount instead of  
 1135 subtracting”) and adopts Agent 2’s more efficient method. Agent 3 corrects their calculation error  
 1136 after seeing consistent peer answers.  
 1137

### 1138 D.3 EXAMPLE 3: GEOMETRIC PROBLEM - METHOD SIMPLIFICATION

#### 1140 Problem

1141 A rectangle has a perimeter of 24 units and an area of 32 square units. What are the dimen-  
 1142 sions of the rectangle?  
 1143

1144 The correct solution: If dimensions are length  $l$  and width  $w$ , then  $2(l + w) = 24$  so  $l + w = 12$ ,  
 1145 and  $lw = 32$ . Solving this system gives dimensions 8 and 4.  
 1146

#### 1147 Initial round:

#### 1149 Agent 1: Algebraic Approach

1150 Let length =  $l$ , width =  $w$   
 1151 From perimeter:  $2l + 2w = 24$ , so  $l + w = 12$   
 1152 From area:  $lw = 32$   
 1153 Setting up quadratic:  $w(12 - w) = 32$   
 1154  $12w - w^2 = 32$   
 1155  $w^2 - 12w + 32 = 0$   
 1156 Stops without solving the quadratic  
 1157

#### 1158 Agent 2: Direct Factoring

1159  $l + w = 12$  and  $lw = 32$   
 1160 Looking for factors of 32 that sum to 12: 8 and 4  
 1161 Dimensions:  $8 \times 4$   
 1162 Answer: 8 units by 4 units  
 1163  
 1164

#### 1165 Agent 3: Quadratic Solution

1166 From  $w^2 - 12w + 32 = 0$   
 1167 Using quadratic formula or factoring:  $(w - 8)(w - 4) = 0$   
 1168 So  $w = 8$  or  $w = 4$   
 1169 Dimensions:  $8 \times 4$   
 1170 Answer: 8 units by 4 units  
 1171  
 1172

#### 1173 Final round:

#### 1174 Agent 1: Adopts Simpler Method

1175 “Agent 2’s approach of finding factors of 32 that sum to 12 is much more direct than solving  
 1176 the quadratic. Since  $8 \times 4 = 32$  and  $8 + 4 = 12$ , the dimensions are 8 by 4.”  
 1177 Recognizes and adopts more efficient solution method  
 1178 Answer: 8 units by 4 units  
 1179  
 1180

1181 **Key observation.** Agent 1 explicitly acknowledges that Agent 2’s factor-finding approach is “much  
 1182 more direct” than the algebraic method, demonstrating recognition of solution efficiency. Post-  
 1183 training reinforces these behaviors, enabling the model to learn patterns of concise and efficient  
 1184 reasoning.  
 1185

### 1186 D.4 SUMMARY

1187 These examples demonstrate observable grounding behaviors in multi-agent debates:

**Algorithm 1** Multi-Agent Consensus Alignment Via Debate-RL

---

```

1188 1: input: agents  $\{\pi_{\theta_m}\}_{m=1}^M$ , rounds  $R$ , iters  $L$ , batch size  $B$ , method in MV-SFT, MV-GRPO, MV-DPO, or
1189 MV-KTO
1190 2: for  $\ell = 1, \dots, L$  do
1191 3:   sample batch  $\{x^{(\ell,i)}\}_{i=1}^B$  of prompts from  $\mathcal{D}$ 
1192 4:   for  $i = 1, \dots, B$  do ▷ Generate debate trajectories
1193 5:      $y_{m,1}^{(\ell,i)} \sim \pi_{\theta_m}(\cdot | x^{(\ell,i)}) \quad \forall m$ 
1194 6:     for  $r = 2, \dots, R$  do ▷ Condition on peers' previous round outputs
1195 7:        $x_{m,r}^{(\ell,i)} = [x^{(\ell,i)}; \{y_{j,r-1}^{(\ell,i)}\}_{j \neq m}]$ 
1196 8:        $y_{m,r}^{(\ell,i)} \sim \pi_{\theta_m}(\cdot | x_{m,r}^{(\ell,i)}) \quad \forall m$ 
1197 9:     end for
1198 10:    Parse final answers  $a_m^{(\ell,i)} = \mathcal{A}(y_{m,R}^{(\ell,i)}) \quad \forall m$ 
1199 11:    Aggregate consensus  $\hat{a}(x^{(\ell,i)}) = \text{Majority}\{a_1^{(\ell,i)}, \dots, a_M^{(\ell,i)}\}$ 
1200 12:    Extract final contexts  $x_m^{(\ell,i)} \leftarrow x_{m,R}^{(\ell,i)}$  for all  $m$ 
1201 13:    Partition responses by consensus alignment:
1202      $\mathcal{G}_+^{(\ell,i)} = \{y_{m,R}^{(\ell,i)} : a_m^{(\ell,i)} = \hat{a}(x^{(\ell,i)})\}$ 
1203      $\mathcal{G}_-^{(\ell,i)} = \{y_{m,R}^{(\ell,i)} : a_m^{(\ell,i)} \neq \hat{a}(x^{(\ell,i)})\}$ 
1204 14:   end for
1205 15:   build dataset:  $\mathcal{D}_{\text{post}} = \{(x^{(\ell,i)}, \hat{a}(x^{(\ell,i)}), \mathcal{G}_+^{(\ell,i)}, \mathcal{G}_-^{(\ell,i)})\}_{i=1}^B$ 
1206 16:   training data by method:
1207     MV-SFT: for each  $m$ , use  $\{(x_m^{(\ell,i)}, y_{m,R}^{(\ell,i)}) : y_{m,R}^{(\ell,i)} \in \mathcal{G}_+^{(\ell,i)}\}_i$ .
1208     MV-GRPO: for each  $m$ , store  $\{(x_m^{(\ell,i)}, \hat{a}(x^{(\ell,i)}))\}_i$  for reward computation on new samples.
1209     MV-DPO/MV-KTO: pool  $\{(x_m^{(\ell,i)}, y_{m,R}^{(\ell,i)})\}_{i,m}$  partitioned by  $\mathcal{G}_+, \mathcal{G}_-$  across agents.
1210 17:   update policies: apply majority-vote objective to refine each  $\pi_{\theta_m}$ 
1211 18: end for

```

---

**Error propagation:** Agents sometimes abandon partially correct reasoning to adopt peer errors (Example 1, Agent 1).

**Explicit error recognition:** Agents identify specific errors in their own reasoning (“I added the discount instead of subtracting”) and in peers’ approaches (“adding time to flight departure”).

**Method adoption:** Agents recognize more efficient solution methods from peers and explicitly acknowledge this (“Agent 2’s approach... is much more direct”).

**Self-correction through peer comparison:** Seeing consistent peer answers prompts agents to re-visit and correct their calculations.

These behaviors generate training signals beyond simple majority voting. The debate format captures not just correct answers but also common error patterns, self-correction mechanisms, and explicit reasoning about solution quality, providing rich supervision for teaching models to recognize and adopt sound reasoning strategies.

## E MACA ALGORITHM: ADDITIONAL IMPLEMENTATION DETAILS

### E.1 AGENT HETEROGENEITY AND SPECIALIZATION

In our current experiments, we use homogeneous agents (clones) that update independently over a single training iteration, enabling evaluation of both individual and averaged performance to measure the benefits of divergence. We also explore the impact of training over multiple such iterations of debate and post-training in App. I. For multi-iteration runs, agent checkpoints can be managed in two ways: agents may either remain synchronized by resetting checkpoints to the best performing one after every iteration, or diverge by updating independently from distinct trajectory data, allowing specialization and diversity through differential learning. Additionally, the MACA framework also supports heterogeneous agents, allowing different language models to be independently optimized through this framework, where agents represent distinct models or architectures from the outset, each bringing inherently different capabilities or perspectives.

## F EXPERIMENTAL SETUP

### F.1 TRAINING PARAMETERS

We conducted hyperparameter sweeps across learning rates from  $1e-7$  to  $1e-5$ , finding that  $1e-5$  consistently performed best across all methods. For preference-based methods (MV-KTO, MV-DPO), we used  $\beta = 0.1$  throughout all experiments. LoRA ranks and alpha values were adjusted based on model size and computational constraints, with rank/alpha of 64-128 for MV-DPO and MV-KTO, 32-64 for MV-GRPO, and 128 for MV-SFT. Tables 7–10 provide complete parameter specifications.

Table 7: Training parameters.

Parameter	SFT	GRPO	DPO	KTO
Learning rate	$1 \times 10^{-5-7}$	$1 \times 10^{-5-7}$	$1 \times 10^{-5-7}$	$1 \times 10^{-5-7}$
Weight decay	$1 \times 10^{-2}$	$1 \times 10^{-2}$	$1 \times 10^{-2}$	$1 \times 10^{-2}$
Batch size	1–8	8	1–8	1–8
Epochs	1–3	1–3	1–3	1–3
Gradient accumulation steps	2–4	2–4	2–4	2–4
Optimizer	AdamW	AdamW	AdamW	AdamW
LoRA rank ( $r$ )	8–128	8–64	8–128	8–128
LoRA alpha	8–128	8–64	8–128	8–128
LoRA dropout	0.1	0.1	0.1	0.1
LoRA target modules		q_proj, k_proj, v_proj, o_proj		
Entropy coefficient	-	0.01	-	-
Beta (regularization)	-	-	0.1–0.3	0.1–0.3
Number of generations per step	-	8	-	-
Maximum sequence length		Model-dependent (2048–4096)		
Mixed precision		bf16		
Quantization		4-bit (BitsAndBytesConfig)		

Table 8: Generation parameters.

Parameter	Value
Temperature	1.0
Top-p sampling	0.9
Maximum new tokens	256
Do sample	True
Pad token ID	EOS token ID

Table 9: Multi-agent debate parameters.

Parameter	Value
Number of agents	3
Debate rounds	2
Batch debate size	8–24
Use majority vote	True
Include debate context	True
Use async debate	True

Table 10: Model and dataset configuration.

Parameter	Value
Base models	Phi-4B, Qwen-2B, Llama-3B
Training datasets	GSM8K, MATH, MathQA
Test datasets	GSM8K, MATH, MathQA, SVAMP, GPQA, CSQA, AMC
Training size	1500 samples
Test size	500 samples
Validation size	500 samples
Use quantization	4-bit BitsAndBytesConfig
Use adapter mode	True (LoRA)
Use scheduler	True

GSM8K, MATH, and MathQA used 1500/500/500 train/valid/test splits. SVAMP, GPQA, CSQA, and AMC were used only for testing, with 300, 448, 500, and 40 test examples, respectively.

## F.2 DATASETS

We evaluated model mathematical reasoning using seven publicly available datasets spanning a range of difficulty and subject areas: MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), MathQA (Amini et al., 2019), SVAMP (Patel et al., 2021), GPQA (Rein et al., 2023), AMC 23 (kno (2025)), and CommonsenseQA/CSQA (Talmor et al., 2019).

**MATH** (Hendrycks et al., 2021): The MATH dataset contains 12,500 high school mathematics problems from algebra, geometry, combinatorics, and number theory, each requiring multi-step reasoning and precise mathematical solutions.

**GSM8K** (Cobbe et al., 2021): GSM8K is composed of 8,500 grade-school-level word problems emphasizing arithmetic and logical reasoning, with step-by-step annotated solutions for each problem.

**MathQA** (Amini et al., 2019): MathQA features over 37,000 question-answer pairs based on quantitative reasoning, requiring models to convert natural language problems into mathematical expressions and perform multi-step computation.

**SVAMP** (Patel et al., 2021): SVAMP consists of carefully reworded arithmetic word problems designed to probe model robustness and prevent reliance on annotation artifacts, focusing on genuine multi-step arithmetic reasoning.

**GPQA** (Rein et al., 2023): GPQA provides 448 expert-curated, multiple-choice questions covering graduate-level biology, physics, and chemistry, emphasizing deep, multi-stage reasoning and robust factual understanding.

**AMC 23** (kno, 2025): This dataset includes recent problems drawn from the 2023 American Mathematics Competitions (AMC), spanning algebra, combinatorics, geometry, and number theory, and is useful for assessing model performance on expert-constructed math tasks

**CommonsenseQA** (Talmor et al., 2019): CommonsenseQA is a challenging multiple-choice question-answering dataset requiring models to apply commonsense reasoning over diverse everyday scenarios. The dataset comprises 12,247 questions, each designed to probe deeper, non-trivial conceptual knowledge beyond factual recall, making it a benchmark for evaluating commonsense understanding in language models.

These datasets collectively enable a thorough evaluation of analytical reasoning capabilities in large language models.

### 1350 F.3 MULTI-AGENT DEBATE INFRASTRUCTURE

1351

1352 Our multi-agent debate system addresses the computational challenges of training and deploying  
 1353 multiple agents efficiently through specialized infrastructure for quantized training isolation and  
 1354 scalable debate inference.

1355

1356

#### 1357 F.3.1 QUANTIZED MARL TRAINING

1358

1359 Current post-training libraries lack support for multi-gpu distributed training with 4-bit quanti-  
 1360 zation and PEFT adapters. We implement training isolation where each agent trains on a single  
 1361 GPU with explicit resource assignment, using gradient checkpointing and paged optimizers  
 1362 (paged\_adamw\_8bit) for memory efficiency. This enables stable multi-agent RL with QLoRA while  
 1363 preserving quantization benefits.

1364

1365

#### 1366 F.3.2 SCALABLE DEBATE INFERENCE FRAMEWORK

1367

1368 The inference system enables efficient multi-agent debates through adapter hot-swapping and dy-  
 1369 namic resource management, designed to scale seamlessly with available compute resources. A cen-  
 1370 tralized base model manager maintains a pool of quantized base models shared across agents, imple-  
 1371 menting lazy loading and caching for lightweight LoRA adapters that enable rapid agent switching  
 1372 without full model reloading. An adapter job scheduler coordinates debate workloads across avail-  
 1373 able GPUs through adaptive load balancing, priority queues for debate tasks ordered by round index,  
 1374 and automatic cluster adaptation that detects available hardware and distributes workloads accord-  
 1375 ingly. This architecture enables concurrent multi-agent inference on hardware that would typically  
 1376 support only single full-precision models, with debate throughput scaling linearly with compute  
 1377 resources while maintaining memory efficiency through quantization and adapter sharing.

1378

### 1379 F.4 DEBATE PROMPT TEMPLATES

1380

1381 This appendix details the prompts used in our multi-agent debate system.

1382

1383

#### 1384 F.4.1 INITIAL QUESTION PROMPT

1385

1386 Used in the first round of debate.

1387

1388

#### 1389 For GPQA, MathQA, CommonsenseQA

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

#### Initial Prompt: Multiple Choice

Answer the following multiple choice question as accurately  
 as possible.  
 {question}

Provide a bullet point summary of your step-by-step reasoning

.  
 Your final answer should be a single choice letter in the  
 form

\boxed{answer}, at the end of your response.

#### For GSM8K, SVAMP, AMC, MATH

Table 11: Single-agent accuracy: Unsupervised post-training using debate-majority-vote-derived supervision (DMV) performs comparably to or outperforms supervised training using ground truth (GT) labels. Bold indicates the better score in each pair.

Model	Dataset	Debate	SFT		KTO		DPO		GRPO	
			GT	DMV	GT	DMV	GT	DMV	GT	DMV
Qwen2B	MATH	7.67	12.13	<b>12.20</b>	12.67	<b>19.67</b>	17.13	<b>23.00</b>	<b>21.33</b>	17.33
	GSM8K	23.00	24.20	<b>24.80</b>	45.40	<b>47.20</b>	<b>46.47</b>	44.20	<b>39.73</b>	32.07
Llama3B	MATH	27.87	<b>26.80</b>	25.60	12.53	<b>39.20</b>	38.40	<b>40.67</b>	<b>36.80</b>	35.00
	GSM8K	57.33	54.13	<b>56.73</b>	27.27	<b>66.27</b>	56.87	<b>66.73</b>	<b>54.93</b>	52.20
Phi4B	MATH	34.60	33.07	<b>35.20</b>	<b>33.73</b>	33.00	<b>37.00</b>	36.33	<b>37.80</b>	37.33
	GSM8K	67.27	<b>71.67</b>	70.20	<b>76.13</b>	75.47	75.80	<b>76.60</b>	<b>74.80</b>	70.40
Llama8B	MATH	22.93	<b>23.73</b>	23.00	<b>41.20</b>	39.40	45.13	<b>46.40</b>	29.07	<b>31.13</b>
	GSM8K	57.93	40.20	<b>41.67</b>	<b>72.60</b>	70.87	76.33	<b>77.67</b>	61.27	<b>66.87</b>

#### Initial Prompt: Math Problem

Solve the following math problem. {question}

Provide a bullet point summary of your step-by-step reasoning

Your final answer should be a single numerical number, in the form

`\boxed{answer}`, at the end of your response.

#### F.4.2 DEBATE PHASE PROMPT

Used in subsequent rounds of debate to incorporate peer context.

#### Debate Phase Prompt

Here are solutions from other agents:  
{agent\_responses}

Using each response as additional advice, give an updated bullet point summary of your step-by-step reasoning to the question: {question}

Make sure your final answer is in the form `\boxed{answer}`, at the end of your response.

## G ABLATION: GROUND TRUTH VS DEBATE MAJORITY VOTE

Tables 11 and 12 report ablations comparing post-training on oracle ground-truth (GT) labels versus (DMV) labels across MATH and GSM8K benchmarks.

In the single-agent setting (Table 11), DMV supervision proves consistently competitive with, and often superior to, GT supervision across Qwen2B, Llama-3B, Phi-4B, and Llama-8B. Gains are most pronounced for preference-based objectives such as KTO and DPO, where DMV provides a more stable learning signal. By contrast, the oracle GT signal sometimes produces degraded performance, particularly in KTO. Notably, DMV provides a robust alternative that avoids these pitfalls, often leading to stronger outcomes without requiring external supervision.

Table 12: Multi-agent accuracy: Unsupervised post-training using debate-majority-vote-derived supervision (*DMV*) performs comparably to or outperforms supervised training using ground truth (*GT*) labels. Bold indicates the better score in each pair.

Model	Dataset	Init	SFT		KTO		DPO		GRPO	
			<i>GT</i>	<i>DMV</i>	<i>GT</i>	<i>DMV</i>	<i>GT</i>	<i>DMV</i>	<i>GT</i>	<i>DMV</i>
Qwen2B	MATH	32.4	38.6	<b>40.6</b>	40.4	<b>46.2</b>	39.4	<b>41.2</b>	<b>43.6</b>	40.2
	GSM8K	49.6	50.0	<b>51.6</b>	<b>66.2</b>	63.8	<b>61.6</b>	57.0	<b>60.4</b>	51.8
Llama3B	MATH	37.8	<b>35.6</b>	33.6	15.8	<b>51.8</b>	50.8	<b>51.4</b>	49.2	<b>50.8</b>
	GSM8K	65.6	<b>65.8</b>	64.0	29.4	<b>73.8</b>	67.0	<b>73.8</b>	64.4	<b>69.6</b>
Phi4B	MATH	44.4	44.0	<b>46.0</b>	<b>46.8</b>	44.6	<b>50.2</b>	48.6	<b>51.4</b>	50.8
	GSM8K	79.6	<b>79.8</b>	78.4	<b>82.6</b>	81.2	83.8	<b>84.4</b>	81.6	<b>83.4</b>
Llama8B	MATH	32.8	<b>34.0</b>	33.4	<b>56.0</b>	55.8	<b>61.8</b>	60.8	<b>48.6</b>	44.8
	GSM8K	74.0	<b>65.2</b>	<b>65.2</b>	<b>81.2</b>	79.4	81.6	<b>83.0</b>	83.2	<b>84.2</b>

Table 13: Multi-agent comparison of Context vs No Context runs across Initial Debate, SFT, KTO, DPO, and GRPO. Bold indicates which setting performed better in each pair.

Model	Dataset	Debate	SFT		KTO		DPO		GRPO	
			<i>Context</i>	<i>NoCtx</i>	<i>Context</i>	<i>NoCtx</i>	<i>Context</i>	<i>NoCtx</i>	<i>Context</i>	<i>NoCtx</i>
Qwen2B	MATH	32.4	<b>40.6</b>	38.0	46.2	<b>46.6</b>	41.2	<b>43.4</b>	<b>40.2</b>	39.8
	GSM8K	49.6	<b>51.6</b>	<b>51.6</b>	<b>63.8</b>	59.2	57.0	<b>57.8</b>	51.8	<b>57.2</b>
Llama3B	MATH	37.8	33.6	<b>37.4</b>	<b>51.8</b>	27.0	<b>51.4</b>	51.0	<b>50.8</b>	47.0
	GSM8K	65.6	64.0	<b>65.4</b>	<b>73.8</b>	59.0	<b>73.8</b>	70.6	<b>69.6</b>	50.0
Phi4B	MATH	44.4	<b>46.0</b>	45.0	44.6	<b>46.0</b>	<b>48.6</b>	48.4	50.8	<b>52.8</b>
	GSM8K	79.6	78.4	<b>80.0</b>	81.2	<b>82.0</b>	<b>84.4</b>	81.8	83.4	<b>82.0</b>
Llama8B	MATH	32.8	<b>33.4</b>	31.4	<b>55.8</b>	49.2	<b>60.8</b>	57.2	<b>44.8</b>	44.6
	GSM8K	74.0	65.2	<b>67.4</b>	79.4	<b>80.6</b>	<b>83.0</b>	81.6	<b>84.2</b>	71.0

In the multi-agent setting (Table 12), the advantages of DMV supervision become even clearer. Across nearly all models and methods, DMV either matches or exceeds GT labels. The gains are especially consistent under preference-learning formulations (KTO, DPO), where DMV supervision yields more reliable improvements to consensus-based performance. While GT labels retain competitive strength in certain cases (e.g., Phi-4B on MATH tasks), DMV repeatedly delivers higher or more stable final-round accuracies.

Together, these results show that debate-derived majority-vote supervision provides an effective, scalable alternative to oracle ground truth. DMV not only mitigates the instability observed when GT is used in preference-learning objectives, but also enhances both single- and multi-agent training. These findings underscore the efficacy of unsupervised alignment signals, leveraging a model’s own consensus dynamics, as a robust substitute for human-labeled supervision.

## H ABLATION: EFFECT OF PEER CONTEXT IN MULTI-AGENT DEBATE

Tables 13 and 14 examine whether conditioning on peer responses during training improves debate performance.

Training with peer context substantially improves multi-agent debate performance (Table 14). While initial round (single-agent) performance shows modest benefits from context training (5 out of 8 cases improve), the advantages become pronounced in final-round multi-agent settings. Context training improves final-round individual accuracy in 7 out of 8 cases, with particularly large gains for models like L8B-MATH (57.80 vs. 51.47) and L3B-MATH (45.80 vs. 41.13). Most importantly, final-round majority voting benefits from context in 6 out of 8 cases, with GSM8K tasks showing consistent improvements across all model sizes (1.4-3.2 percentage points) and larger models

Table 14: Post-training with peer context teaches agents to utilize other agents’ responses in the debate format for more effective final round ensemble reasoning (MV-DPO).

Model-Data	Initial Round Avg		Final Round Avg		Final Round MV	
	NoCtx	Context	NoCtx	Context	NoCtx	Context
Qwen2B-MATH	22.33	<b>23.00</b>	37.13	<b>38.33</b>	<b>43.40</b>	41.20
Qwen2B-GSM8K	<b>46.53</b>	44.20	52.53	<b>54.73</b>	<b>57.80</b>	57.00
Llama3B-MATH	40.27	<b>40.67</b>	41.13	<b>45.80</b>	51.00	<b>51.40</b>
Llama3B-GSM8K	65.87	<b>66.73</b>	67.87	<b>70.60</b>	70.60	<b>73.80</b>
Phi4B-MATH	<b>39.47</b>	36.33	<b>39.53</b>	34.73	48.40	<b>48.60</b>
Phi4B-GSM8K	74.67	<b>76.60</b>	74.60	<b>77.80</b>	81.80	<b>84.40</b>
Llama8B-MATH	<b>49.87</b>	46.40	51.47	<b>57.80</b>	57.20	<b>60.80</b>
Llama8B-GSM8K	75.80	<b>77.67</b>	78.00	<b>81.20</b>	81.60	<b>83.00</b>

Table 15: Iterative alternation between debate generation and post-training across four iterations shows substantial initial gains (It-0 to It-1) followed by diminishing returns. Model abbreviations: Q2B=Qwen-2B, L3B=Llama-3B, L8B=Llama-8B, P4B=Phi-4B.

Model-Data	Initial Round Avg				Final Round Avg				Final Round MV			
	It-0	It-1	It-2	It-3	It-0	It-1	It-2	It-3	It-0	It-1	It-2	It-3
Q2B-MATH	7.67	17.40	18.00	<b>19.33</b>	21.47	43.13	40.73	<b>43.93</b>	32.40	47.40	41.20	<b>48.60</b>
Q2B-GSM8K	23.00	44.20	44.67	<b>44.73</b>	35.07	58.67	<b>59.93</b>	59.47	49.60	60.80	62.20	<b>62.60</b>
L3B-MATH	27.87	40.93	39.33	<b>41.93</b>	24.27	48.27	46.93	<b>48.73</b>	37.80	<b>55.00</b>	54.00	53.60
L3B-GSM8K	57.33	64.60	<b>67.07</b>	65.20	49.20	68.73	<b>71.13</b>	69.07	65.60	72.00	<b>74.60</b>	70.80
P4B-MATH	34.60	43.27	43.67	<b>43.80</b>	34.37	48.00	48.93	<b>50.20</b>	44.40	55.00	55.80	<b>57.40</b>
P4B-GSM8K	67.27	75.73	<b>76.20</b>	75.47	68.53	77.00	<b>79.20</b>	77.60	79.60	81.40	<b>84.40</b>	83.00
L8B-MATH	22.93	44.53	44.67	<b>44.87</b>	22.53	55.73	<b>57.60</b>	56.07	32.80	58.20	<b>60.40</b>	59.80
L8B-GSM8K	57.93	77.00	77.87	<b>78.80</b>	56.53	80.80	80.73	<b>82.93</b>	74.00	82.80	82.60	<b>85.60</b>

achieving substantial gains on MATH tasks (e.g., L8B-MATH improving from 57.20 to 60.80). These results demonstrate that context-aware training teaches agents to effectively leverage peer feedback during deliberation, leading to stronger consensus outcomes in multi-agent debate.

## I ITERATIVE IMPROVEMENT

Table 15 examines whether iterative training beyond the first iteration yields continued improvements. While It-1 produces the substantial gains reported in our main results, iterations 2 and 3 demonstrate continued modest improvements: in 23 of 24 evaluation settings, either It-2 or It-3 achieves the best performance. For example, Phi-4B on MATH majority vote increases from 55.00 (It-1) to 57.40 (It-3), and Llama-8B on GSM8K improves from 82.80 to 85.60.

These gains show clear diminishing returns compared to the It-0→It-1 jump, with typical improvements of 1-3 percentage points between iterations. Performance occasionally dips between adjacent iterations (e.g., L3B-GSM8K majority vote: 72.00→74.60→70.80) but generally trends upward. This pattern suggests that iterative training continues to extract useful signal from debate-generated data, though with decreasing marginal benefit after the initial iteration.

## J DPO PAIR SELECTION STRATEGY ANALYSIS

We compared two strategies for creating preference pairs from multi-agent debate data: majority/minority partitioning versus confidence-based selection using model log-probabilities. Analysis covered 2,226 agent responses from 742 problems across multiple models and datasets. Majority/minority partitioning substantially outperforms confidence-based selection across all metrics. Majority responses achieve 68.0% accuracy versus 28.1% for minority responses (39.8 percentage point gap), while high-confidence responses achieve only 51.5% accuracy versus 33.2% for low-

confidence (18.3 percentage point gap). The majority strategy yields an effect size of Cohen’s  $d = 1.832$  compared to 0.281 for confidence-based selection, a 6.5 $\times$  difference in discriminative power. Additionally, majority voting provides usable preference signals in 70.5% of examples versus 35.2% for confidence-based selection. These results validate using majority vote consensus for DPO pair selection, demonstrating that collective agreement provides more reliable quality signals than individual model confidence for mathematical reasoning tasks.

## K SELF-CONSISTENCY AND ACCURACY CORRELATION ANALYSIS

We analyze the correlation between self-consistency and accuracy improvements across three experimental conditions to assess the robustness of our findings under different token generation limits and quantization settings. We test three conditions: (1) Token Capped + Quantized (256 tokens, 4-bit), our standard experimental setup matching computational constraints; (2) Token Uncapped + Non-Quantized (2048 tokens with no observed truncation, full precision), representing maximum generation quality without computational constraints; and (3) Token Uncapped + Quantized (2048 tokens with no observed truncation, 4-bit), a balanced approach removing token truncation while maintaining efficiency.

Table 16: Self-consistency vs accuracy correlation across experimental conditions. “Capped” refers to 256 token limit, which is used throughout the work due to computational constraints in multi-agent RL settings, and “Uncapped” refers to 2048 token limit where no response truncation occurred.

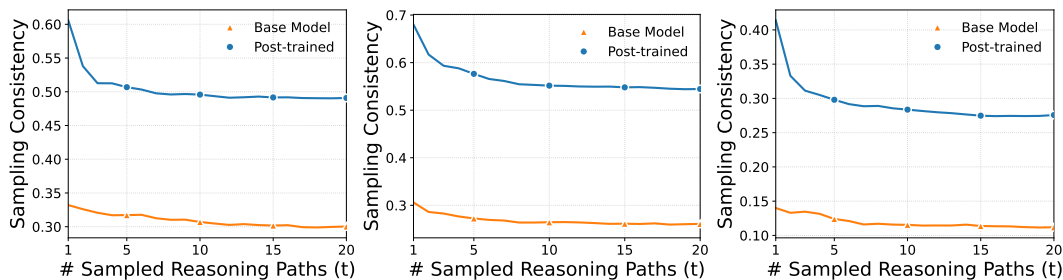
Condition	Model-Dataset	Self-Consistency (%)		Accuracy (%)		Correlation ( $r$ )
		Base	Post	Base	Post	
Capped +Quant	Llama3B-MathQA	31.2	<b>49.4</b>	29.1	<b>45.8</b>	0.954
	Qwen2B-Math	12.0	<b>27.8</b>	11.6	<b>26.8</b>	0.989
	Qwen2B-GSM8K	26.4	<b>54.0</b>	25.9	<b>52.7</b>	0.982
Uncapped +Full	Llama3B-MathQA	49.5	<b>68.5</b>	45.1	<b>60.9</b>	0.869
	Qwen2B-Math	50.4	<b>57.2</b>	48.2	<b>54.4</b>	0.944
	Qwen2B-GSM8K	51.4	<b>76.4</b>	50.8	<b>75.5</b>	0.933
Uncapped +Quant	Llama3B-MathQA	49.8	<b>67.3</b>	45.8	<b>60.8</b>	0.899
	Qwen2B-Math	42.8	<b>49.7</b>	38.9	<b>45.2</b>	0.935
	Qwen2B-GSM8K	50.6	<b>73.5</b>	48.9	<b>71.3</b>	0.927

All conditions show strong positive correlations ( $r > 0.86$ ) between self-consistency and accuracy, validating self-consistency as a robust proxy for model performance across computational settings. Interestingly, capped conditions show slightly higher correlations (mean  $r = 0.975$ ) compared to uncapped conditions (mean  $r = 0.915$  for both quantized and full precision). This tighter coupling likely reflects how post-training teaches more efficient reasoning: models learn to better utilize limited token budgets, producing fewer truncated responses while achieving higher accuracy within constraints. In uncapped settings, post-trained models generate 22-36% shorter responses while still being more accurate, demonstrating that this efficiency persists without token limits. Four-bit quantization minimally impacts correlation strength, with quantized uncapped (mean  $r = 0.920$ ) closely matching full precision uncapped (mean  $r = 0.915$ ), demonstrating that computational efficiency can be achieved without degrading the consistency-accuracy relationship. Post-training consistently improves both metrics across all conditions, with self-consistency gains ranging from 6.9 to 27.6 percentage points and accuracy gains from 6.3 to 24.7 percentage points, confirming that our approach remains effective under varying computational constraints.

### K.1 SELF-CONSISTENCY CURVES

Figure 4 shows sampling consistency before and after MACA post-training. Pre-trained models show low sampling consistency across sampled trajectories, while post-training substantially improves consistency, with curves stabilizing up to 27.6 percentage points above baseline.

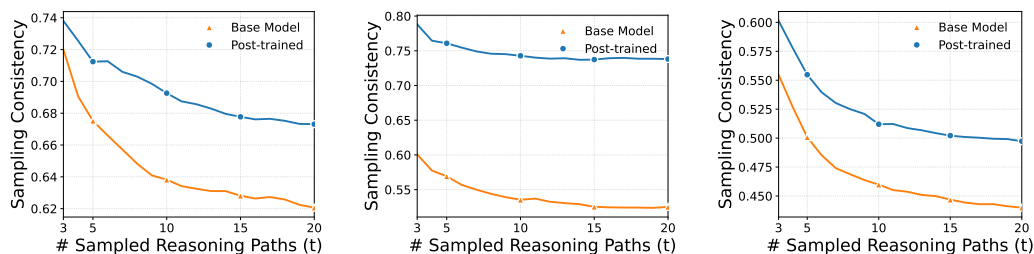
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631



(a) Llama-3B on MATHQA (b) Qwen-2B on GSM8K (c) Qwen-2B on MATH

1632 Figure 4: Consistency before and after MACA post-training with token constraints (maximum of  
1633 256 tokens). Pre-trained models (Orange) show low sampling consistency across sampled trajec-  
1634 tories. Post-training with MACA (Blue) substantially improves sampling consistency. Averaged over  
1635 500 test prompts with 20 trajectories each.

1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647



(a) Llama-3B on MathQA (b) Qwen-2B on GSM8K (c) Qwen-2B on MATH

1648 Figure 5: Consistency before and after MACA post-training without token constraints. Models  
1649 trained with 256-token debates still show gains when tested with full-length responses, though with  
1650 reduced effect sizes due to the weaker training signal compared to testing conditions. Colors: Blue:  
1651 post-trained model, Orange: base model.

1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664

## K.2 SELF-CONSISTENCY IMPROVEMENTS WITHOUT TOKEN CONSTRAINTS

Figure 5 shows self-consistency curves for the aforementioned model-dataset pairs without token constraints. Improvements persist across all configurations, demonstrating that MACA’s benefits are not merely artifacts of addressing truncation. The effect sizes are slightly smaller than in our main results due to train-test mismatch: these models were trained on debate signals from 256-token responses but tested without constraints. Training on full-length debates would likely yield larger improvements, as the consensus signals would be stronger and better aligned with test conditions.

## L IMPACT OF MACA ON GENERAL REASONING

1665 We demonstrate that debate-derived majority vote supervision achieves comparable performance  
1666 to ground truth supervision while enabling effective generalization to unseen reasoning domains.  
1667 Tables 17 and 18 present direct comparisons between these supervision approaches across mathe-  
1668 matical training domains and out-of-distribution tasks.

1669 Table 17 shows that debate-derived supervision performs comparably to ground truth labels on train-  
1670 ing domains, with methods trading wins across model-dataset combinations. Table 18 reveals that  
1671 both supervision approaches generalize effectively to unseen reasoning tasks, including mathemat-  
1672 ical word problems (SVAMP), science reasoning (GPQA), and commonsense reasoning (CSQA).  
1673 Both methods show substantial improvements over base performance across all domains, confirm-  
ing that MACA develops transferable reasoning capabilities.

Table 17: Within-distribution performance comparison: Debate-derived majority vote supervision vs. ground truth supervision on training domains. Models are trained on MATH, GSM8K, and MathQA datasets. Bold indicates the better method for each model-dataset pair, demonstrating that debate-derived supervision achieves comparable performance to ground truth labels.

Model	Method	MATH		GSM8K		MathQA	
		Debate-MV	GT	Debate-MV	GT	Debate-MV	GT
Qwen2B	Base	10.4	10.4	27.0	27.0	7.4	7.4
	SFT	<b>10.8</b>	10.4	25.6	<b>26.4</b>	8.2	<b>8.8</b>
	GRPO	19.4	<b>21.0</b>	45.2	<b>48.6</b>	18.6	<b>19.6</b>
	KTO	22.6	<b>23.2</b>	<b>54.8</b>	54.6	<b>28.8</b>	28.6
	DPO	<b>24.8</b>	24.2	51.4	<b>52.0</b>	<b>24.2</b>	24.0
Llama3B	Base	32.0	32.0	69.6	69.6	24.6	24.6
	SFT	<b>33.2</b>	32.4	64.2	64.2	<b>26.4</b>	25.2
	GRPO	45.8	<b>46.4</b>	<b>75.8</b>	74.8	<b>36.2</b>	31.8
	KTO	<b>48.0</b>	47.8	76.0	<b>76.8</b>	<b>41.4</b>	40.6
	DPO	53.2	<b>53.6</b>	<b>80.4</b>	77.8	<b>46.2</b>	45.4

Table 18: Cross-domain generalization: Debate-derived majority vote supervision vs. ground truth supervision on unseen reasoning tasks. Models trained on mathematical datasets (MATH, GSM8K, MathQA) generalize effectively to diverse reasoning domains. Bold indicates the better method for each model-dataset pair, showing that both supervision approaches transfer well to out-of-distribution tasks.

Model	Method	SVAMP		GPQA		CSQA	
		Debate-MV	GT	Debate-MV	GT	Debate-MV	GT
Qwen2B	Base	48.30	48.30	0.45	0.45	3.80	3.80
	SFT	<b>53.30</b>	53.00	<b>17.90</b>	0.89	16.80	<b>18.80</b>
	GRPO	<b>60.30</b>	58.33	<b>8.70</b>	7.81	<b>10.80</b>	9.60
	KTO	76.00	<b>76.33</b>	16.70	<b>16.96</b>	<b>63.40</b>	60.80
	DPO	<b>65.00</b>	64.67	19.64	<b>20.98</b>	<b>62.2</b>	60.80
Llama3B	Base	71.30	71.30	0.67	0.67	53.00	53.00
	SFT	68.33	<b>72.0</b>	<b>2.23</b>	2.23	57.40	<b>57.80</b>
	GRPO	75.00	<b>79.33</b>	<b>6.92</b>	5.13	<b>63.20</b>	59.40
	KTO	76.70	<b>78.67</b>	8.93	<b>9.82</b>	<b>62.20</b>	61.80
	DPO	78.40	<b>80.67</b>	<b>11.40</b>	11.60	<b>64.00</b>	62.40

These results demonstrate that debate-generated consensus signals provide an effective unsupervised alternative to ground truth supervision, achieving comparable performance without human annotation. This approach offers significant advantages for scaling reasoning improvements to new domains or large datasets where expert labels are unavailable or prohibitively expensive.

## M POST-TRAINING METHOD IMPACT ON LOG-PROBABILITY DISTRIBUTION

Figure 6 shows how different post-training methods affected log-probability distributions for Qwen2B on MATH. Preference-based methods (DPO, KTO) increased the density of majority distributions, with KTO showing particularly pronounced effects: higher peaks and tighter tails indicating more concentrated probability mass around consensus responses. SFT left the majority distribution largely unchanged while substantially reducing minority distribution density. GRPO created the most dramatic separation between distributions, shifting both leftward (lower log-probabilities overall) but with the minority distribution shifting much more substantially than the majority. While these different patterns (probability concentration for preference methods versus selective penalization for GRPO) all corresponded with performance improvements, further research is needed to better understand the relationship between these specific distributional changes and the impact on consistency.

1728  
 1729  
 1730  
 1731  
 1732  
 1733  
 1734  
 1735  
 1736  
 1737  
 1738  
 1739  
 1740  
 1741  
 1742  
 1743  
 1744  
 1745  
 1746  
 1747  
 1748  
 1749  
 1750  
 1751  
 1752  
 1753  
 1754  
 1755  
 1756  
 1757  
 1758  
 1759  
 1760  
 1761  
 1762  
 1763  
 1764  
 1765  
 1766  
 1767  
 1768  
 1769  
 1770  
 1771  
 1772  
 1773  
 1774  
 1775  
 1776  
 1777  
 1778  
 1779  
 1780  
 1781

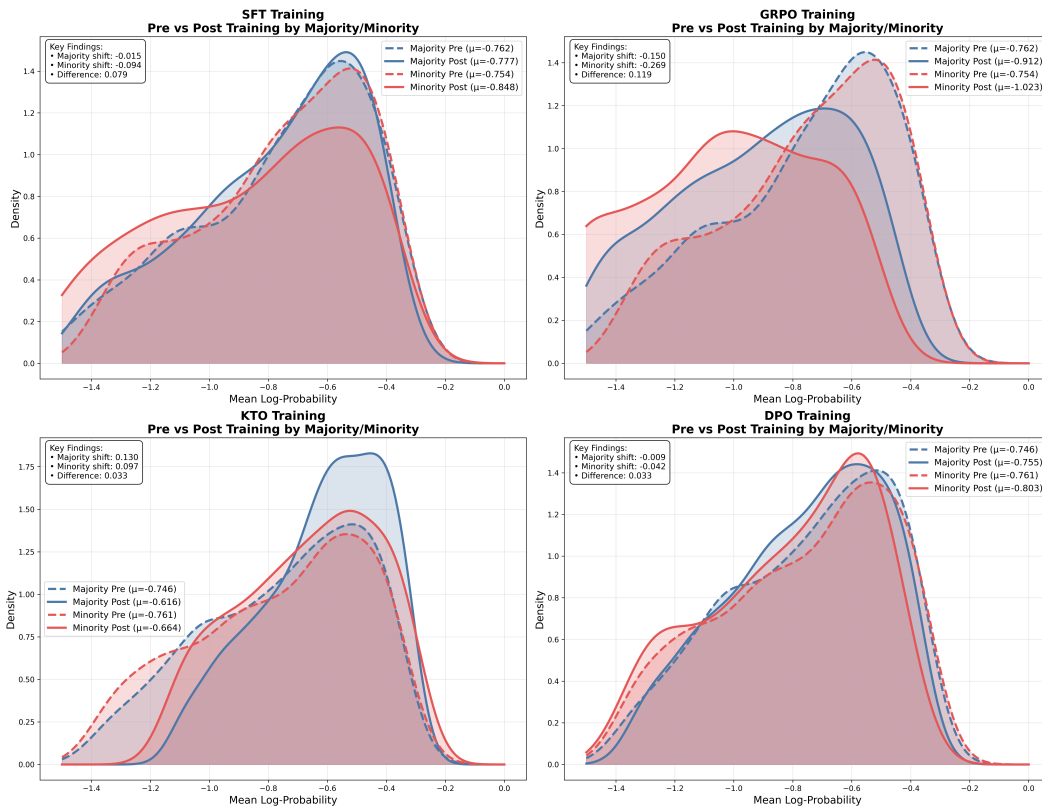


Figure 6: Log probability distributions for majority and minority answers before and after post-training (Qwen2B, Math).

## N PREFERENCE LEARNING AS AN IMPLICIT FORMAT REWARD

This section investigates how preference learning through consensus signals acts as an implicit format reward, teaching models to produce more efficient and accurate reasoning without explicit formatting supervision. We examine three key aspects: the baseline capability requirements for effective consensus formation, how token constraints interact with reasoning improvements, and the decomposition of performance gains into format improvements versus problem-solving accuracy improvements. Crucially, we demonstrate that while preference learning does function effectively as an implicit format reward, most of the performance gains observed are attributed to fundamental improvements in problem-solving accuracy rather than mere formatting compliance.

### N.1 BASELINE CAPABILITY REQUIREMENTS FOR CONSENSUS FORMATION

Our experiments reveal that MACA requires sufficient baseline model capability to generate meaningful consensus signals. When models lack foundational problem-solving ability, they fail to produce the correct responses necessary for consensus-based reinforcement. Table 19 illustrates this limitation: under a strict 256-token limit, Qwen2B produces no correct responses on AMC, resulting in no useful consensus to reinforce. Increasing the token limit to 512 partially mitigates this by allowing more reasoning space, though inefficient reasoning patterns can still cause truncation.

Table 19: Impact of token limits and baseline capability on AMC performance. Shows accuracy percentages for base models and post-trained models (“All”) under different token constraints. When models lack sufficient baseline capability (Qwen2B at 256 tokens), consensus formation fails as no correct responses are generated for reinforcement.

	max_new_tokens = 256		max_new_tokens = 512	
	Base	All	Base	All
Qwen2B	0.0	0.0	5.0	<b>12.5</b>
Llama3B	7.5	<b>10.0</b>	10.0	<b>20.0</b>

This baseline capability requirement has important implications for applying MACA: models must possess some initial problem-solving ability on the target domain to benefit from consensus-based training. However, once this threshold is met, we observe that improvements on easier datasets can generalize to more challenging tasks, suggesting that enhanced self-consistency helps overcome some limitations of consensus formation on difficult problems.

### N.2 TOKEN EFFICIENCY AND REASONING QUALITY IMPROVEMENTS

Post-training simultaneously addresses token efficiency and reasoning quality. Models trained with debate under constrained token limits generate more parseable answers within the budget while achieving higher accuracy on those answers. To verify that improvements extend beyond mere format optimization, we evaluated models post-trained with 256-token debates using 512-token test budgets.

Table 20 demonstrates that performance gains persist under increased token limits, indicating genuine reasoning enhancement rather than just format improvements. We also analyze the sequence lengths of the trajectories generated for the analysis in App. K, showing that uncapped post-trained models generate responses that are 22-36% shorter than their base models. This evidence suggests that self-guided preference learning functions as an implicit format reward, teaching models to produce more concise and effective reasoning patterns that generalize across computational budgets and task difficulties.

### N.3 DECOMPOSING THE SOURCES OF IMPROVEMENT

To understand whether our gains stem from improved reasoning or merely better formatting, we decompose performance improvements into their constituent components. Our analysis separates gains from better completion (avoiding truncation) versus fundamental reasoning improvements.

Table 20: Improvements from post-training with `max_new_tokens = 256` translate when tested with larger token counts (512). “All” represents a model post-trained on Math, MathQA, and GSM8K; its row reports improvement deltas over the base model performance.

Model	Trained on	Tested on ( <code>max_new_tokens = 512</code> )					
		MATH	GSM8K	MathQA	SVAMP	AMC	GPQA
Qwen2B	Base	32.8	47.4	18.8	53.7	7.5	12.5
	All	↑15.6	↑24.6	↑22.2	↑20.0	↑2.5	↑11.2
Llama3B	Base	60.8	76.8	43.0	78.0	10.0	9.2
	All	↑6.4	↓0.2	↑11.2	↑3.0	↑10.0	↑14.5

Table 21: Decomposition of performance improvements into completion gains versus reasoning gains. “From Completion” represents accuracy gained if all truncated base model responses had been allowed to complete. “From Better Reasoning” represents the remaining improvement attributable to fundamental problem-solving enhancement. Percentages show the relative contribution of each component to total gains.

Model	Dataset	Before	After	Total Gain	From Completion	From Better Reasoning
Qwen2B	CSQA	3.8	63.4	↑59.6	↑18.6 (31%)	↑41.0 (69%)
	GPQA	0.4	16.7	↑16.3	↑0.7 (4%)	↑15.6 (96%)
	GSM8K	24.6	54.8	↑30.2	↑0.0 (0%)	↑30.2 (100%)
	MATH	10.4	22.6	↑12.2	↑0.0 (0%)	↑12.2 (100%)
	MathQA	7.4	28.8	↑21.4	↑2.0 (9%)	↑19.4 (91%)
	SVAMP	48.3	76.0	↑27.7	↑0.4 (1%)	↑27.3 (99%)
Llama3B	CSQA	53.0	64.0	↑11.0	↑1.2 (11%)	↑9.8 (89%)
	GPQA	0.7	11.4	↑10.7	↑0.9 (8%)	↑9.8 (92%)
	GSM8K	69.6	80.4	↑10.8	↑0.2 (2%)	↑10.6 (98%)
	MATH	41.2	53.2	↑12.0	↑0.2 (2%)	↑11.8 (98%)
	MathQA	29.2	46.2	↑17.0	↓0.2 (-1%)	↑17.2 (101%)
	SVAMP	71.3	75.7	↑4.3	↑0.0 (0%)	↑4.3 (100%)

The results reveal that 69-100% of improvements stem from better reasoning rather than just avoiding truncation. While base models exhibit high truncation rates (e.g., 74.8% for Qwen2B on CSQA, 82.8% on MATH), post-training reduces these dramatically. However, even after accounting for completion improvements, substantial reasoning gains remain, confirming that our method teaches fundamentally better problem-solving, not merely more efficient token usage.

#### N.4 FORMATTING IMPROVEMENTS AND REMAINING REASONING GAINS

To further isolate reasoning improvements from formatting effects, we analyze how post-training affects answer formatting compliance. Table 22 shows the percentage of responses with correct reasoning but incorrect formatting, and calculates reasoning gains that persist beyond all formatting improvements. Post-training dramatically reduces formatting losses (e.g., from 38.8% to 0.6% on CSQA for Qwen2B). However, the “Reasoning Gain Beyond Formatting” column reveals that substantial improvements remain even after perfect formatting is assumed. For instance, Qwen2B’s 59.6 percentage point improvement on CSQA includes 38.2 points from better formatting; the remaining 21.4 points represent accuracy improvements distinct from formatting.

#### N.5 PARSER IMPLEMENTATION DETAILS

To quantify formatting improvements versus reasoning gains, we employed two parsers with different strictness levels. Our standard parser requires answers in `\boxed{}` format and returns None for any deviation. The relaxed parser, used for impact analysis in Table 22, accepts common natural

Table 22: Analysis of formatting improvements versus reasoning gains. “Accuracy Lost to Format Errors” shows the percentage of responses with correct reasoning but incorrect formatting (e.g., writing “The answer is A” instead of  $\boxed{A}$ ). “Reasoning Gain Beyond Formatting” shows improvements that persist even after accounting for all formatting fixes, calculated as: Total Gain - (Format Loss Before - Format Loss After).

Model	Dataset	Accuracy Lost to Format Errors (%)		Reasoning Gain
		Before Training	After Training	Beyond Formatting (%)
Qwen2B	CSQA	38.8	0.6	↑21.4
	GPQA	6.2	1.6	↑11.7
	MathQA	5.8	0.6	↑16.2
Llama3B	CSQA	7.0	1.4	↑5.4
	GPQA	7.6	3.8	↑6.9
	MathQA	5.8	2.8	↑14.0

language patterns like “The answer is A” or “Answer: 42” by searching the entire response for valid answer formats.

This relaxed parser first attempts the strict extraction, then falls back to regex patterns that capture answers expressed naturally in text. For multiple choice, it accepts patterns like “[Tt]he answer is ([A-E])” or “Answer: ([A-E])”. For numerical answers, it extracts from patterns like “= 42” at line endings or “The final answer is 42”. When multiple patterns match, it takes the last occurrence, mimicking how humans identify the final answer in a reasoning chain.

The gap between strict and relaxed parser accuracies precisely measures the “Accuracy Lost to Format Errors” in Table 22. Strict formatting compliance is a key measure of a model’s ability to follow instructions and is critical for downstream applications that rely on reliable parsing of LM outputs, while the relaxed parser accounts for human-interpretable correctness. This dual evaluation reveals that preference learning through consensus not only teaches proper formatting as an implicit reward but fundamentally improves problem-solving capabilities, with reasoning gains persisting even when formatting constraints are removed.

## O TRAINING CURVES

The training dynamics of our post-training methods provide insight into how models learn to refine responses towards multi-agent consensus preferences. Figures 7a–8b show example training curves across key metrics during post-training with MV-DPO, MV-KTO, MV-GRPO, and MV-SFT.

Across all methods, the reward margins between chosen (consensus) and rejected (non-consensus) responses increase consistently, indicating effective preference learning. MV-DPO and MV-KTO reveal this pattern strongly: margins start near zero and grow steadily as the models optimize towards favoring majority-preferred outputs. MV-KTO achieves similar reward improvements despite not requiring strict paired comparisons, demonstrating its robust learning dynamics.

Training accuracy converges for all methods, reaching high classification levels between consensus and non-consensus responses, showing that each approach successfully reinforces desired sampling behaviors. Correspondingly, losses decrease smoothly without signs of instability or collapse, indicating stable training processes.

Log probabilities of rejected responses decline across post-training methods, reflecting the models’ increasing tendency to assign lower likelihood to outputs outside the consensus. This is most pronounced in MV-DPO and MV-GRPO, where rejected rewards fall more steeply, delineating a clear separation between preferred and discouraged responses. MV-SFT, operating via imitation learning, exhibits strong improvements in token accuracy and loss while reinforcing consensus-aligned responses effectively.

Overall, the asymmetric reward trajectories, where chosen response rewards remain relatively stable or increase slightly while rejected response rewards decline sharply, suggest that post-training

1944 primarily discourages generation of minority or outlier outputs. This mechanism is a core driver  
1945 behind the improved sampling consistency and reasoning quality observed in our experiments.

1946 These training curves collectively support our hypothesis that post-training with majority vote pref-  
1947 erences enables models to internalize collective agreement notions and reproduce responses better  
1948 aligned with multi-agent consensus.  
1949

## 1950 P POST-TRAINING SELF-CONSISTENCY IMPROVEMENTS TRANSLATE FROM 1951 4-BIT QUANTIZED MODEL TO FULL MODEL 1952 1953

1954 As shown in Figures 9-11, post-training improvements on 4-bit quantized models transfer to full-  
1955 precision models, and the improvement margin persists with larger numbers of trajectories sampled  
1956 ( $t = 50$ ).  
1957

## 1958 Q IMPACT OF POST-TRAINING ON DEBATE AGREEMENT RATES 1959

1960 Figure 12 shows how agent agreement patterns evolve through debate, comparing base models (Iter-  
1961 ation 0) against consensus-trained models (Iteration 1). These distributions directly measure whether  
1962 our training successfully shifts probability mass toward consensus trajectories  $\mathcal{G}_+(x)$ .  
1963

1964 Base models show relatively uniform agreement distributions, with only 13.4% of examples achiev-  
1965 ing full consensus (3/3 agreement) in the final round. After consensus post-training with MV-DPO,  
1966 this increases over three-fold to 43.4%, with systematic improvements across all agreement lev-  
1967 els: increased mass at 2/3 and 3/3 agreement, decreased mass at 1/3 agreement. This redistribution  
1968 confirms that training drives the policy toward consensus-supporting trajectories.

1969 Additionally, unparseable responses drop from 11% to 0.6% without explicit format rewards. Since  
1970 our training favors consensus completions, which must finish within token limits to be compar-  
1971 able, the method implicitly rewards efficient, complete reasoning patterns. This suggests consensus  
1972 alignment naturally encourages concise and coherent reasoning as a prerequisite for measurable  
1973 agreement.

## 1974 R PERFORMANCE COMPARISON OF RL ON DEBATE VS. SINGLE-ROUND 1975 MAJORITY VOTE RL IN SINGLE-AGENT SETTING 1976 1977

1978 Table 23 reports the full results referenced in Table 5, including standard deviation over 3 seeds.  
1979 We find that MACA with final round majority vote with debate context outperforms single-round  
1980 majority-vote RL with DPO (akin to ScPO (Prasad et al., 2025)) in all 8 configurations tested, and  
1981 RL with GRPO (akin to TTRL (Zuo et al., 2025)) in 6 out of 8 configurations tested, with the other  
1982 2 configs being within the standard deviation.

## 1983 S EFFICIENCY 1984 1985

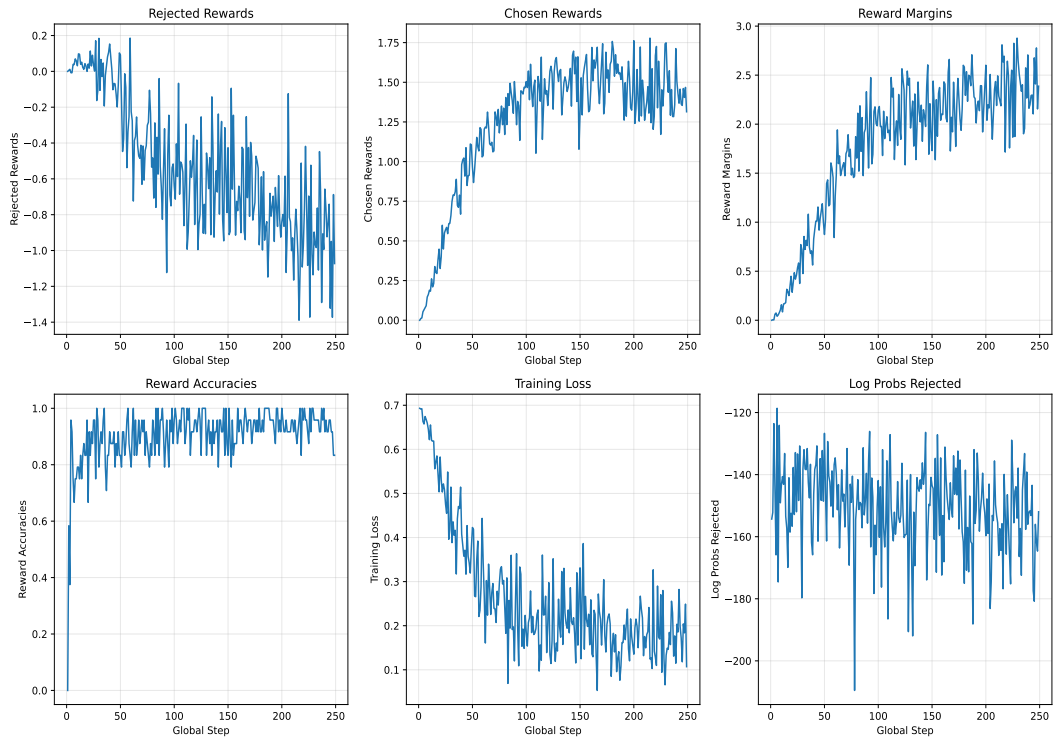
1986 Here we detail the compute performance tradeoffs of MACA’s MV-DPO with debate context against  
1987 other single-agent benchmark methods. ScPO requires 0.075-0.58 GPU hours achieving -0.6%  
1988 to +13.6% gains, TTRL requires 2.2-7.7 GPU hours achieving -26.4% to +18% changes, while  
1989 MACA (MV-DPO with debate) requires 0.73-1.68 GPU hours achieving consistent positive gains.  
1990 Our efficiency plots demonstrate that MV-DPO on debate provides the best performance-efficiency  
1991 tradeoff—higher performance than ScPO/TTRL with comparable computational cost to ScPO and  
1992 significantly lower cost than TTRL.  
1993  
1994  
1995  
1996  
1997

1998  
 1999  
 2000  
 2001  
 2002  
 2003  
 2004  
 2005  
 2006  
 2007  
 2008  
 2009  
 2010  
 2011  
 2012  
 2013  
 2014  
 2015  
 2016  
 2017  
 2018  
 2019  
 2020  
 2021  
 2022  
 2023  
 2024  
 2025  
 2026  
 2027  
 2028  
 2029  
 2030  
 2031  
 2032  
 2033  
 2034  
 2035  
 2036  
 2037  
 2038  
 2039  
 2040  
 2041  
 2042  
 2043  
 2044  
 2045  
 2046  
 2047  
 2048  
 2049  
 2050  
 2051

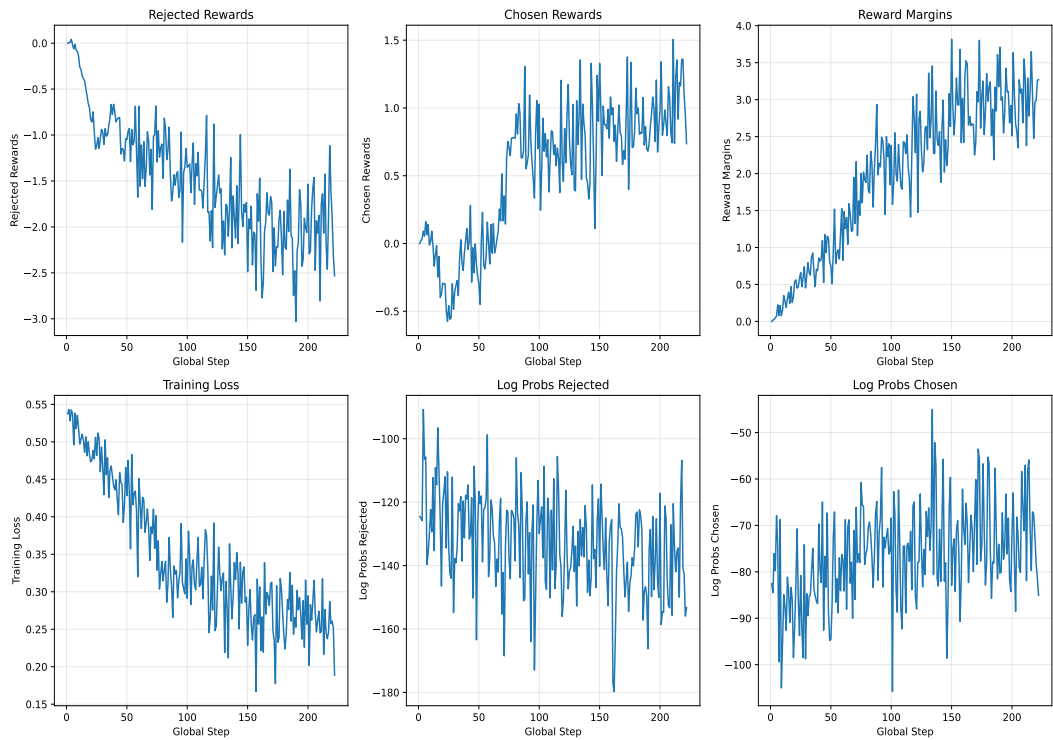
Table 23: Single-agent: RL with debate outperforms RL with single-round majority vote. Benchmark against TTRL (single-round R0 MV-GRPO), ScPO (single-round R0 MV-DPO), MV-DPO (MACA: debate final round R1 MV-DPO with debate context after collaborative exchange). Accuracy averaged over 3 seeds with standard deviation reported in comparison to base(instruction-tuned) model.

	Qwen-2B	Llama-3B	Phi-4B	Llama-8B
<b>MATH</b>				
Base	9.4	40.2	39.8	36.0
TTRL	<b>27.4 ± 2.9</b>	45.5 ± 5.7	45.9 ± 2.1	43.5 ± 0.2
ScPO	11.7 ± 1.1	43.6 ± 0.2	39.9 ± 0.5	39.7 ± 0.6
MV-DPO	26.1 ± 0.4	<b>52.7 ± 0.7</b>	<b>46.7 ± 0.2</b>	<b>53.1 ± 0.8</b>
<b>GSM8K</b>				
Base	26.4	72.8	76.4	67.8
TTRL	<b>50.9 ± 1.6</b>	50.5 ± 8.3	81.2 ± 1.6	54.3 ± 5.7
ScPO	42.3 ± 3.4	74.4 ± 0.8	82.3 ± 0.5	80.7 ± 1.2
MV-DPO	50.2 ± 0.9	<b>76.5 ± 0.3</b>	<b>83.7 ± 0.3</b>	<b>86.0 ± 0.2</b>

2052  
 2053  
 2054  
 2055  
 2056  
 2057  
 2058  
 2059  
 2060  
 2061  
 2062  
 2063  
 2064  
 2065  
 2066  
 2067  
 2068  
 2069  
 2070  
 2071  
 2072  
 2073  
 2074  
 2075  
 2076  
 2077  
 2078  
 2079  
 2080  
 2081  
 2082  
 2083  
 2084  
 2085  
 2086  
 2087  
 2088  
 2089  
 2090  
 2091  
 2092  
 2093  
 2094  
 2095  
 2096  
 2097  
 2098  
 2099  
 2100  
 2101  
 2102  
 2103  
 2104  
 2105



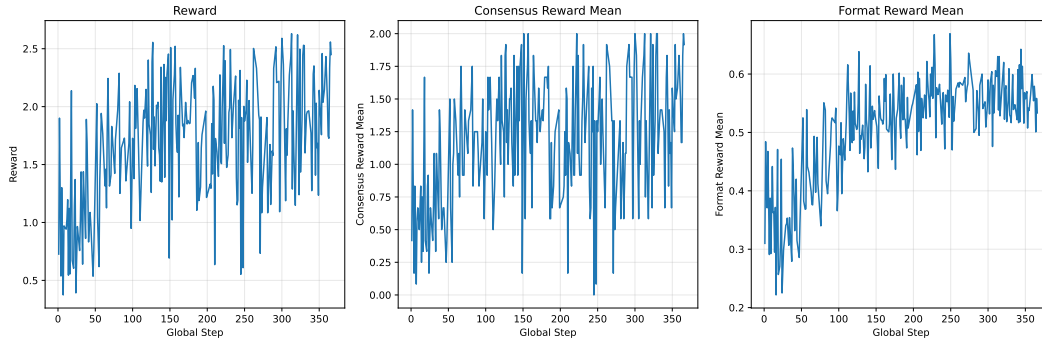
(a) MV-DPO training curves



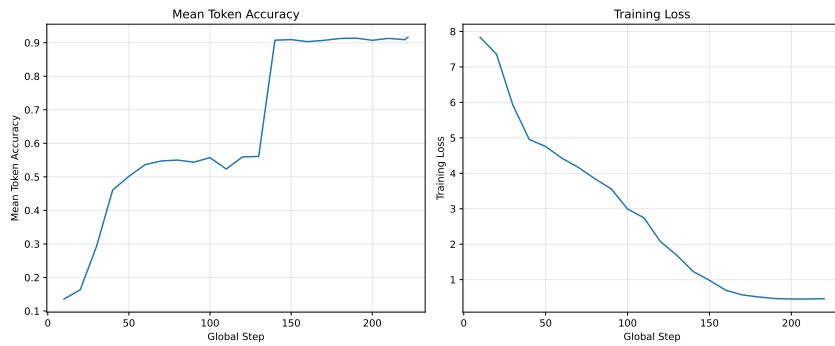
(b) MV-KTO training curves

Figure 7: Training curves for MV-DPO and MV-KTO. (a) MV-DPO: increasing reward margin between majority (chosen) and minority (rejected) responses, with declining rejected log probabilities. (b) MV-KTO: similar reward margin increase using unpaired examples, with rejected log probabilities decreasing and chosen increasing.

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159



(a) MV-GRPO training curves



(b) MV-SFT training curves

Figure 8: Training curves for MV-GRPO and MV-SFT. (a) MV-GRPO: consensus and format rewards both increase. (b) MV-SFT: token accuracy increases while loss decreases.

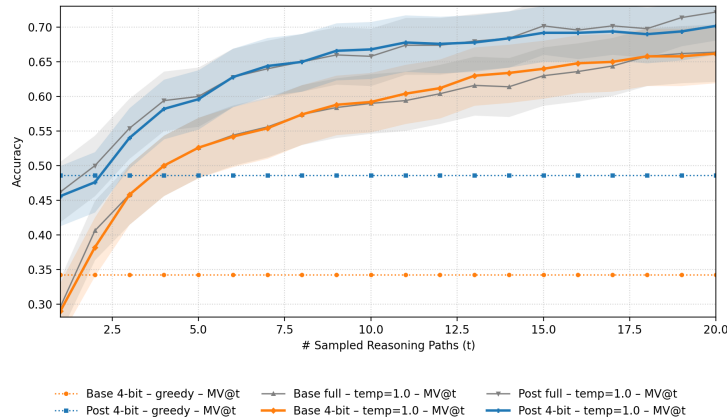


Figure 9: Llama-3B on MathQA (MV@t, t=20). Blue: post-trained 4-bit model, Orange: base 4-bit model, Grey: full-precision model.

2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213

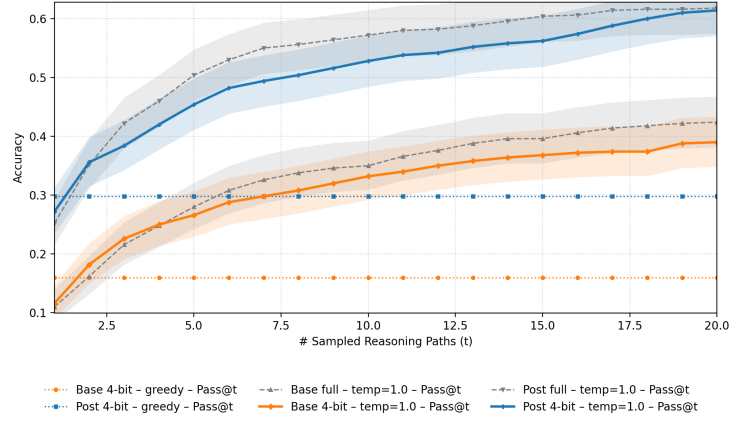


Figure 10: Qwen-2B on MATH (Pass@t, t=20). **Blue**: post-trained 4-bit model, **Orange**: base 4-bit model, **Grey**: full-precision model.

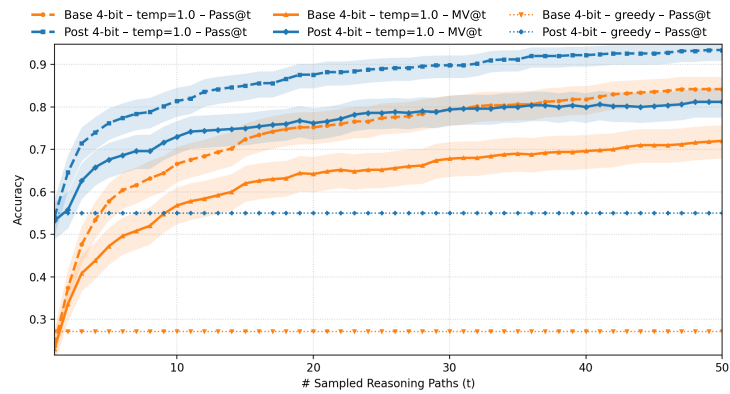


Figure 11: Qwen-2B on GSM8K (Pass@t, MV@t; t=50). **Blue**: post-trained 4-bit model, **Orange**: base 4-bit model.

2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267

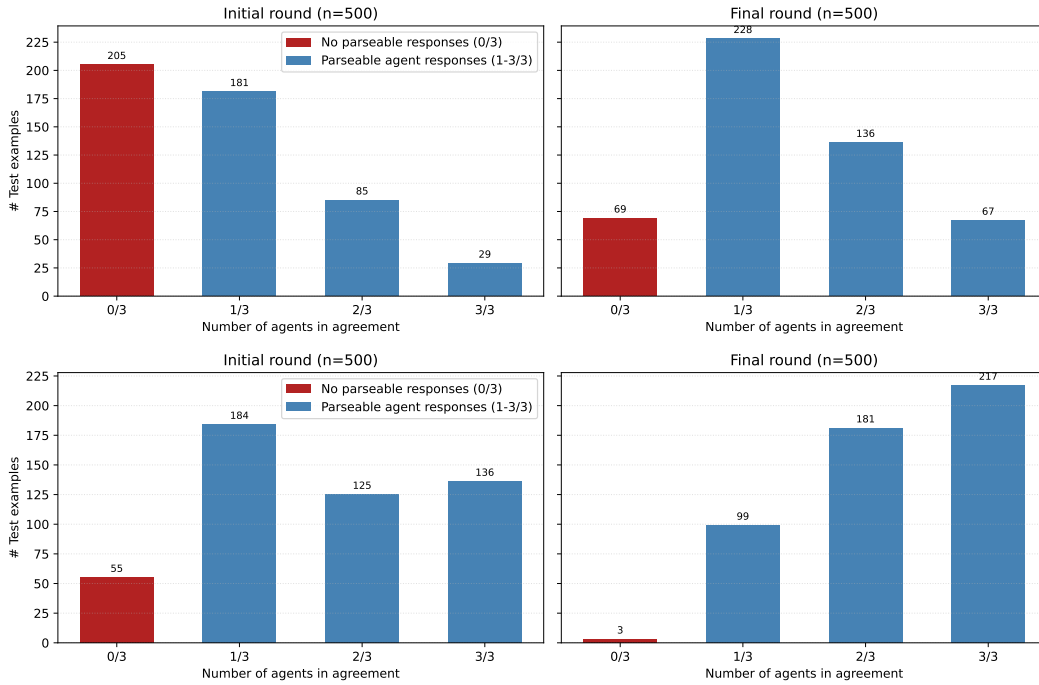


Figure 12: MACA drives meaningful improvements in both answer completeness and agent agreement, verifying the probability mass reallocation to the consensus set of reasoning trajectories (Ex: Qwen2B on GSM8K). Top: base model debate; bottom: post-trained model debate; left: initial debate round; right: final debate round.

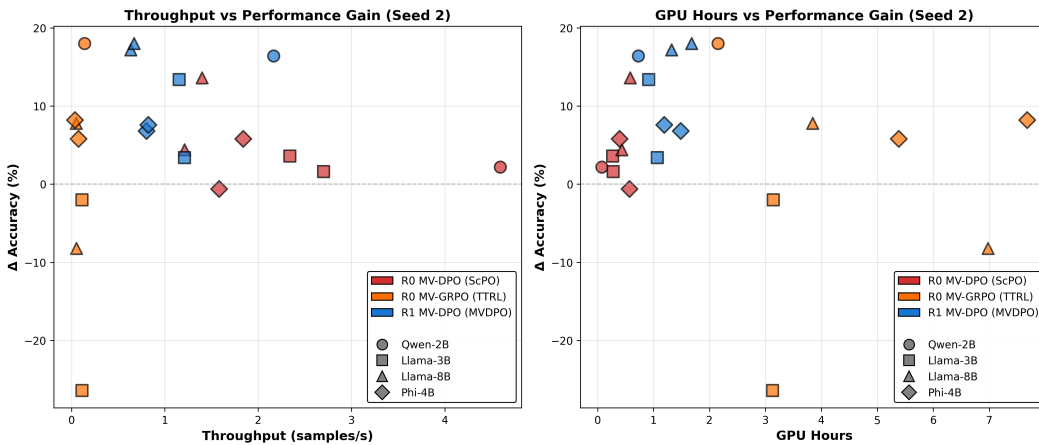


Figure 13: GPU throughput and training runtime compared with performance lift for single-agent post-training methods