

# 000 001 002 003 004 005 006 007 008 009 010 INTERNALIZING SELF-CONSISTENCY IN LANGUAGE MODELS: MULTI-AGENT CONSENSUS ALIGNMENT

005 **Anonymous authors**

006 Paper under double-blind review

## ABSTRACT

011 Language Models (LMs) are inconsistent reasoners, often generating contradic-  
012 tory responses to identical prompts. While inference-time methods can mitigate  
013 these inconsistencies, they fail to address the core problem: LMs struggle to reli-  
014 ably select reasoning pathways that lead to consistent outcomes under exploratory  
015 sampling. To address this, we formalize self-consistency as an intrinsic property of  
016 well-aligned reasoning models and introduce Multi-Agent Consensus Alignment  
017 (MACA), a reinforcement learning framework that post-trains models to favor rea-  
018 soning trajectories aligned with their internal consensus using majority/minority  
019 outcomes from multi-agent debate. These trajectories emerge from deliberative  
020 exchanges where agents ground reasoning in peer arguments, not just aggregation  
021 of independent attempts, creating richer consensus signals than single-round ma-  
022 jority voting. MACA enables agents to teach themselves to be more decisive and  
023 concise, and better leverage peer insights in multi-agent settings without external  
024 supervision, driving substantial improvements across self-consistency (+27.6%  
025 on GSM8K), single-agent reasoning (+23.7% on MATH), sampling-based infer-  
026 ence (+22.4% Pass@20 on MATH), and multi-agent ensemble decision-making  
027 (+42.7% on MathQA). These findings, coupled with strong generalization to un-  
028 seen benchmarks (+16.3% on GPQA, +11.6% on CommonsenseQA), demonstrate  
029 robust self-alignment that more reliably unlocks latent reasoning potential of lan-  
030 guage models.

## 1 INTRODUCTION

031 A fundamental trait of a reliable reasoning model is self-consistency: the intrinsic ability to produce  
032 stable outputs across various sampled reasoning paths (Elazar et al., 2021; Wang et al., 2022). In  
033 the human brain, this consistency emerges from the prefrontal and anterior cingulate cortices, which  
034 resolve conflicts between competing neural activations (Miller & Cohen, 2001) by balancing novelty  
035 and coherence (Friston, 2010; Botvinick et al., 2004; Shenhav et al., 2013; Zhang et al., 2025).  
036 This enables human reasoning to remain robust despite the inherent randomness of thought. In  
037 contrast, while probabilistic decoding in language models (LMs) gives access to diverse reasoning  
038 trajectories, it struggles to consistently select high-quality paths (Holtzman et al., 2020; Wang et al.,  
039 2022). Yet, current AI alignment research primarily focuses on human preferences and external  
040 values (Ouyang et al., 2022; Glaese et al., 2022), while overlooking the model’s self-alignment. The  
041 challenge remains: teaching models to sample diversely, i.e., exploring multiple valid reasoning  
042 paths like different theorem proofs or alternative chains of thought, while maintaining consistent  
043 quality and conclusions. Existing methods for mitigating sampling inconsistencies such as sampling  
044 multiple reasoning paths and aggregating via majority vote (Wang et al., 2022; Li et al., 2024) or  
045 using multi-agent debate (Irving et al., 2018) operate at inference time. While these reduce output  
046 variance, they do not improve the model’s internal reasoning stability. When models generate low-  
047 quality reasoning traces, aggregation can even be counterproductive: noisy arguments compound  
048 rather than cancel out, especially in ambiguous scenarios (Radharapu et al., 2025).

049 We formalize self-consistency as an intrinsic property of well-aligned reasoning models and intro-  
050 duce Multi-Agent Consensus Alignment (MACA), a reinforcement learning (RL) framework where  
051 multiple LM clones collaborate to solve problems through iterative debate. This debate serves as a  
052 simulation environment where agents explore solutions independently, then ground their reasoning  
053 through peer interaction and update their answers. Crucially, the reasoning paths exchanged during

054 this collaborative exploration, not just final majority answers, provide strong training signals for  
 055 teaching agents to recognize stable reasoning patterns (Fig. 1). These consensus-aligned trajectories  
 056 from debate contain richer signals than aggregating isolated reasoning attempts. We reinforce  
 057 majority-outcome traces where agents successfully converged through peer grounding. This teaches  
 058 models to internalize self-consistency: they learn from collaboratively refined reasoning and develop  
 059 an inductive bias toward consensus-forming trajectories even mid-generation, which resembles hu-  
 060 man intuition for sensing sound arguments before completing them (Kahneman, 2011). Learning  
 061 from these patterns teaches models to weigh multiple reasoning approaches, a skill that generalizes  
 062 beyond the original training tasks and makes them better ensemble participants.

063 Our experiments confirm that multi-agent debate produces more informative training signals than  
 064 single-round majority voting. We also observe that addressing consensus alignment through pre-  
 065 ference learning yields substantial improvements over scalar-reward RL and imitation learning. We  
 066 optimize the separation between majority and minority trajectories using majority vote variants of  
 067 DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024), outperforming GRPO (Shao et al.,  
 068 2024) and SFT (Subramaniam et al., 2025). This mirrors human preference formation through rel-  
 069 ative comparison (Festinger, 1957): when truth is ambiguous, judgments emerge through compara-  
 070 tive assessment where majority opinions provide normative pressure while minority views introduce  
 071 necessary variation (Moscovici, 1976; Nemeth, 1985; M. J. A. N. de Caritat, 1785). Training on  
 072 debate-derived preferences thus teaches models to ground reasoning in peer arguments, learning  
 073 efficient and stable reasoning through comparison rather than ground-truth labels.

074 **Key contributions.** Through extensive experiments on LMs across various reasoning benchmarks,  
 075 we empirically demonstrate that MACA achieves improvements on the following dimensions.

- 076 • **Self-consistency.** MACA shows improvements in answer consistency (up to +27.6% on  
 077 GSM8K) across different sampled reasoning paths.
- 078 • **Accuracy.** It also yields significant improvements in individual agent performance  
 079 (+23.7% on MATH), sampling-based inference (+22.4% Pass@20 on MATH), and multi-  
 080 agent performance (+42.7% on MathQA).
- 081 • **Generalization.** Training for self-consistency on mathematical reasoning transfers to all  
 082 evaluated tasks, including unseen domains (+11.3% on GPQA, +11.6% on Common-  
 083 senseQA), demonstrating that self-consistency is a foundational capability for general rea-  
 084 soning.

## 086 2 RELATED WORK

088 Existing approaches address sampling inconsistency primarily through inference-time techniques.  
 089 Self-consistency prompting (Wang et al., 2022; Li et al., 2024) samples multiple reasoning paths  
 090 and selects the majority-voted answer, with extensions for non-verifiable outputs (Chen et al., 2023)  
 091 and path pruning (Zhu et al., 2024a). Multi-agent debate frameworks (Du et al., 2023; Irving et al.,  
 092 2018) similarly utilize consensus across models to improve reliability, with recent work exploring  
 093 applications in scientific discovery (Gottweis et al., 2025). These methods, however, require  
 094 additional inference compute and do not internalize the self-consistency into the model. We in-  
 095 stead improve self-consistency through post-training that optimizes consensus signals via multi-  
 096 agent RL (Yang et al., 2021; Jiang & Lu, 2021; Zhu et al., 2024b; Zhan et al., 2025), strengthening  
 097 foundational reasoning abilities. Current training-time alternatives have limitations: relative log-  
 098 probability ranking (Huang et al., 2025) correlates weakly with accuracy compared to consensus  
 099 (App. I), while LLM-as-a-Judge approaches (Jiao et al., 2025) suffer from preference leakage (Li  
 100 et al., 2025) and bias under ambiguity (Radharapu et al., 2025). Majority vote RL methods such  
 101 as TTRL (Zuo et al., 2025) and ScDPO (Prasad et al., 2024) use GRPO and DPO, respectively, to  
 102 reinforce single-round majority vote, whereas our framework leverages multi-agent debate, while  
 103 supporting both preference learning and scalar-reward formulations. **Both TTRL and ScDPO are**  
 104 **specialized cases of the framework developed in this work, when reducing the multi-agent debate**  
 105 **parameters to a single-round majority vote.** Subramaniam et al. (2025) uses Supervised Fine-Tuning  
 106 (SFT) for multi-agent debate optimization. In this work we demonstrate that RL-based alternatives  
 107 achieve superior performance compared to SFT.

108 

### 3 FORMALIZING SELF-CONSISTENCY

109

110 Given a prompt  $x$ , an LM with parameters  $\theta$  defines a distribution  $\pi_\theta(y|x) = \prod_{t=1}^{|y|} \pi_\theta(y_t|x, y_{<t})$   
111 over reasoning trajectories  $y$ , from which answers  $a = A(y)$  are extracted. Under tempera-  
112 ture sampling, the model samples from a modified distribution  $\pi_{\theta,\tau}(y|x)$  where token proba-  
113 bilities are adjusted by temperature  $\tau > 0$ . This induces an answer distribution  $P_{\theta,\tau}(a|x) =$   
114  $\sum_{y:A(y)=a} \pi_{\theta,\tau}(y|x)$ , which gives each answer’s probability by summing over all reasoning paths  
115 that lead to it. We denote the majority answer as  $a_{\theta,\tau}^*(x) = \arg \max_a P_{\theta,\tau}(a|x)$  with majority  
116 probability  $S_{\theta,\tau}^+(x) = P_{\theta,\tau}(a_{\theta,\tau}^*(x)|x)$ . This represents the total probability mass concentrated on  
117 the most likely answer, or the model’s internal consensus.  
118

119 Temperature sampling enables exploration of diverse reasoning paths, but reduces the consistency  
120 of the final answer. While greedy decoding ( $\tau = 0$ ) trivially approaches perfect consistency, it elimi-  
121 nates exploration and often produces suboptimal solutions (Holtzman et al., 2020). Lower tempera-  
122 tures increase consistency but restrict reasoning diversity. A self-consistent model should maintain  
123 high  $S_{\theta,\tau}^+(x)$  even at high temperatures, allowing the model to access diverse reasoning trajectories  
124 while reliably converging on consistent answers. We measure self-consistency in two ways.  
125

126 **Single-agent sampling consistency.** Computing  $S_{\theta,\tau}^+(x)$  directly requires summing probabilities  
127 over all trajectories that lead to the majority answer, which is untractable. Instead, we estimate it by  
128 sampling  $t$  independent trajectories with answers  $a_1, \dots, a_t$  and computing:  
129

130 
$$s_t^{\theta,\tau}(x) = \frac{1}{t} \sum_{i=1}^t \mathbf{1}[a_i(x) = \hat{a}(x)], \quad \text{where } \hat{a}(x) = \text{Majority}\{a_1(x), \dots, a_t(x)\}. \quad (1)$$
131

132 This measures the fraction of sampled trajectories that agree with the majority answer. As  $t \rightarrow \infty$ ,  
133  $s_t^{\theta,\tau}(x) \rightarrow S_{\theta,\tau}^+(x)$ , providing a consistent estimate of the true majority probability.  
134

135 **Multi-agent debate agreement.** When  $M$  agents produce answers  $a_1(x), \dots, a_M(x)$  through de-  
136 liberation, we measure the fraction of agents converging on the majority answer:  
137

138 
$$d_M^{\theta,\tau}(x) = \frac{1}{M} \sum_{m=1}^M \mathbf{1}[a_m(x) = \hat{a}(x)], \quad \text{where } \hat{a}(x) = \text{Majority}\{a_1(x), \dots, a_M(x)\}. \quad (2)$$
139

140 Higher agreement indicates a stronger consensus. In other words, models with higher  $S_{\theta,\tau}^+(x)$  reach  
141 the same conclusion more frequently.  
142

143 

### 4 MACA: MULTI-AGENT CONSENSUS ALIGNMENT

144

145 Having formalized self-consistency, we now present a framework to improve it through post-training  
146 with self-generated signals from debate. In multi-agent debate,  $M$  copies of the same model engage  
147 in iterative discussion: each agent generates an initial response, then all agents see each other’s  
148 reasoning and update their answers over  $R - 1$  subsequent rounds of deliberation. Answers that  
149 persist indicate stronger reasoning. The framework requires no external supervision: agents su-  
150 pervise themselves by learning from their own debate dynamics. Specifically, for each prompt  $x$ ,  
151 the debate produces final responses  $\mathcal{Y}(x) = \{y_1, \dots, y_M\}$  with extracted answers  $a_m = A(y_m)$ .  
152 The majority consensus  $\hat{a}(x) = \text{Majority}\{a_1, \dots, a_M\}$  partitions  $\mathcal{Y}(x)$  into consensus-supporting  
153  $\mathcal{G}^+(x) = \{y \in \mathcal{Y}(x) : A(y) = \hat{a}(x)\}$  and dissenting  $\mathcal{G}^-(x) = \{y \in \mathcal{Y}(x) : A(y) \neq \hat{a}(x)\}$  tra-  
154 jectories. 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  
155  $\mathcal{D}$  is the original set of prompts. Debate consensus, arising through deliberative exchange rather  
156 than statistical sampling, provides rich training signals. We adapt four post-training objectives to  
157 this self-generated data, treating consensus-supporting trajectories ( $\mathcal{G}^+$ ) as preferred and dissenting  
158 trajectories ( $\mathcal{G}^-$ ) as not preferred. By learning to separate these groups, the model internalizes the  
159 nuanced differences between stable consensus and dissenting reasoning. See Alg. 1 in App. D for  
160 the complete iterative debate and post-training loop.  
161

162 **Majority-Vote SFT (MV-SFT)** trains the model to mimic consensus-supporting trajectories:  
163

164 
$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^+ \in \mathcal{G}^+(x)} [\log \pi_\theta(y^+|x)]. \quad (3)$$
165

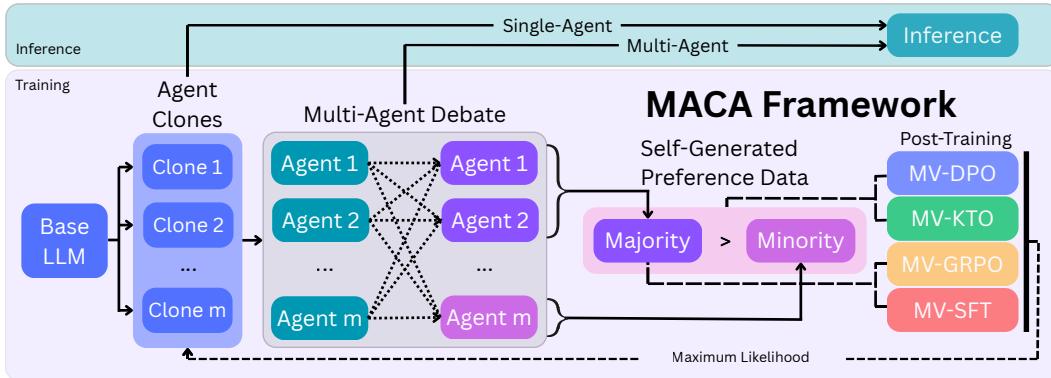


Figure 1: Multi-Agent Consensus Alignment framework: Multiple clones of a base LM engage in debate to generate majority and minority reasoning trajectories through multi-agent debate. The framework splits responses based on alignment with majority consensus to create preference pairs. MV-GRPO compares 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 relative separation between these preference pairs. Updated agents can then be used for single-agent or multi-agent inference, or continue iterative training.

**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  $r_x(y) = 1[\mathcal{A}(y) = \hat{a}(x)]$  based on whether each sample’s answer matches the pre-computed consensus:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_\theta} \left[ \tilde{A}_x(y) \sum_t \log \pi_\theta(y_t | x, y_{<t}) \right] + \lambda \text{KL}(\pi_\theta \| \pi_{\text{ref}}), \quad (4)$$

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:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{(y^+, y^-) \in \mathcal{G}^+(x) \times \mathcal{G}^-(x)} \left[ \log \sigma \left( \beta \left[ \log \frac{\pi_\theta(y^+ | x)}{\pi_{\text{ref}}(y^+ | x)} - \log \frac{\pi_\theta(y^- | x)}{\pi_{\text{ref}}(y^- | x)} \right] \right) \right]. \quad (5)$$

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  $\lambda_-$ :

$$\begin{aligned} \mathcal{L}_{\text{KTO}}(\theta) = & -\lambda_+ \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^+ \in \mathcal{G}^+(x)} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y^+ | x)}{\pi_{\text{ref}}(y^+ | x)} \right) \right] \\ & -\lambda_- \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y^- \in \mathcal{G}^-(x)} \left[ \log \sigma \left( -\beta \log \frac{\pi_\theta(y^- | x)}{\pi_{\text{ref}}(y^- | x)} \right) \right]. \end{aligned} \quad (6)$$

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

**MACA as a generalization of self-guided majority-vote RL.** MACA subsumes prior majority-vote RL approaches such as TTRL (Zuo et al., 2025) and ScPO (Prasad et al., 2024), which use a single round majority vote from independently sampled responses. TTRL uses this to inform GRPO labels, while ScPO uses it to construct DPO pairs. By contrast, MACA introduces multi-agent, multi-round debate, including deliberation traces as conditioning contexts within the prompts during RL. This exposes the model to concrete examples of relative grounding, instances where consensus emerges through explicit engagement with peer arguments, allowing the model to learn not just to aggregate outputs, but to internalize the deliberative process by which consensus is reached.

---

216 

## 5 RESULTS AND DISCUSSION

217

218 We evaluate MACA by post-training four **instruction-tuned** small LMs (Qwen-2B (Yang et al.,  
219 Llama-3B (Grattafiori et al., 2024), Phi-4B (Abdin et al., 2024), and Llama-8B (Grattafiori  
220 et al., 2024)) on six reasoning benchmarks (MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al.,  
221 2021), MathQA (Amini et al., 2019), SVAMP (Patel et al., 2021), GPQA (Rein et al., 2023), and  
222 Commonsense-QA (Talmor et al., 2019)). We use 4-bit quantization with QLoRA (Dettmers et al.,  
223 2023) and limit responses to 256 tokens with temperature  $\tau = 1.0$ , which tests exploratory sampling  
224 under a budget adequate for efficient solvers on these benchmarks while reflecting realistic deploy-  
225 ment compute constraints (Tong et al., 2025; Husom et al., 2025). Improvements persist when tested  
226 with larger token limits (App. M). We instantiate multi-agent debate with  $M = 3$  clones and  $R = 2$   
227 rounds, and compare (1) pre-trained models, (2) SFT baselines, and (3) MACA variants of GRPO,  
228 DPO, and KTO. We train and evaluate models on 1500/500 train/test splits for each dataset inde-  
229 pendently, unless otherwise specified, isolating task-specific self-consistency improvements. We  
230 report mean agent accuracy with standard deviation across three seeds. Debate prompts, training  
231 parameters, multi-processing design, and other experimental details can be found in App. E.  
232

233 

### 5.1 POST-TRAINING IMPROVES SELF-CONSISTENCY

234 We measure the effect of post-training on sampling consistency  $s_t^{\theta, \tau}(x)$ , the fraction of sampled tra-  
235 jectories that match the majority answer. As formalized in Sec. 3, we track the sampling consistency  
236 where  $s_t^{\theta, \tau}(x)$  converges to the modal probability  $S_{\theta, \tau}^+(x)$  as  $t \rightarrow \infty$ . For each model, we sample  
237 20 trajectories on 500 held-out prompts and evaluate  $s_t^{\theta, \tau}(x)$  for  $t = 1$  to 20 (Fig. 2). At  $t = 1$ ,  
238 this metric primarily captures answer completeness, i.e., the percentage of responses that produce  
239 parseable answers within the token window, which post-training substantially improves. MV-DPO  
240 and MV-KTO achieve these gains through self-supervised preference learning alone, demon-  
241 strating that models can teach themselves more efficient reasoning without format rewards (App. M), which  
242 can otherwise be spurious (Huang et al., 2024; Srivastava et al., 2025). As sample size increases,  
243 the metric transitions to measuring true cross-sample agreement. The curves stabilize up to 27.6  
244 percentage points above baseline, demonstrating that post-training increases answer concentration:  
245 models more consistently sample trajectories that converge despite high-temperature exploration.  
246 These improvements persist when tested without max token constraints (App. J.1).  
247

248 In multi-agent debate, we evaluate the agreement metric  $d_M^{\theta, \tau}(x)$  defined in Sec. 3. Base small LMs  
249 initially struggle to reach meaningful consensus: for Qwen-2B on GSM8K, most of the “consensus”  
250 comes from random tie-breaking (1/3 agreement) or weak majorities (2/3), with only 13.4% reach-  
251 ing unanimity (Fig. 12, App. P). Post-training with MACA thoroughly addresses this: non-parseable  
252 responses drop from 13.8% to 0.6%, no-agreement cases from 45.6% to 19.8%, while unanimous  
253 agreement triples from 13.4% to 43.4%. This confirms that MACA improves both individual reason-  
254 ing quality and collaborative grounding, enabling genuine consensus rather than noisy aggregation  
255 (agreement distributions in App. P).  
256

257 

### 5.2 SELF-CONSISTENCY IMPROVES PROBLEM-SOLVING PERFORMANCE

258 **Impact on a single agent in zero-shot setting.** Across 12 model-dataset pairs, MV-RL methods  
259 consistently outperform the Base and MV-SFT baselines (Table 1) in single-agent zero-shot (single  
260 trajectory) settings. Self-guided preference learning (MV-DPO and MV-KTO) outperforms scalar  
261 rewards via MV-GRPO for all models except Phi-4B. MV-DPO is best in 7/12 cases, while MV-KTO  
262 is better in some cases for smaller LMs.  
263

264 **Impact on inference-time sampling.** We next examine how gains in self-consistency translate to  
265 inference-time performance under various sampling regimes. For each prompt, we draw  $t$  trajec-  
266 tories and report two metrics: Pass@t, the fraction of prompts for which at least one of the first  
267  $t$  samples is correct (Chen et al., 2021), and MV@t, the fraction for which a majority vote over  
268 those  $t$  samples matches the ground truth (ties counted as incorrect) (Li et al., 2024). As shown  
269 in Fig. 3, post-training (blue) (i) lifts greedy ( $\tau=0$ ) accuracy, (ii) increases MV@t at each fixed  $t$ ,  
270 and (iii) raises Pass@t, the empirical sampling upper bound, indicating higher achievable accuracy  
271 at any given sampling budget. When additional inference compute is available (i.e.,  $t > 1$ ), sam-  
272 pling techniques continue to see gains on top of the post-training improvements, showing MACA  
273

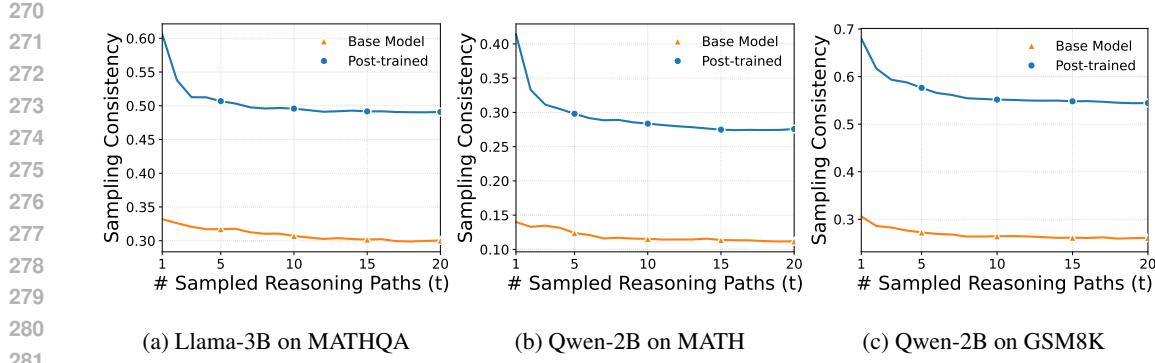


Figure 2: Consistency before and after MACA post-training. Pre-trained models (Orange) show low sampling consistency across sampled trajectories. Post-training with MACA (Blue) substantially improves sampling consistency. Averaged over 500 test prompts with 20 trajectories each.

Table 1: 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><math>23.49 \pm 2.30</math></b>	$\uparrow 15.82$
	GSM8K	23.00	$24.84 \pm 0.87$	$34.40 \pm 2.08$	<b><math>45.13 \pm 1.80</math></b>	$43.87 \pm 1.92$	$\uparrow 22.71$
	MathQA	5.00	$5.25 \pm 0.33$	$17.27 \pm 2.88$	<b><math>22.16 \pm 1.14</math></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><math>40.71 \pm 0.08</math></b>	$\uparrow 13.26$
	GSM8K	57.33	$55.98 \pm 0.68$	$52.40 \pm 2.84$	<b><math>65.76 \pm 1.44</math></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><math>45.00 \pm 2.23</math></b>	$\uparrow 21.13$
Phi4B	MATH	34.60	$34.60 \pm 0.82$	<b><math>37.42 \pm 0.16</math></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><math>76.87 \pm 0.36</math></b>	$\uparrow 9.84$
	MathQA	34.87	$34.04 \pm 0.58$	<b><math>45.52 \pm 2.19</math></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><math>46.00 \pm 0.35</math></b>	$\uparrow 23.07$
	GSM8K	57.93	$42.09 \pm 1.28$	$62.45 \pm 6.01$	$72.36 \pm 1.34$	<b><math>77.36 \pm 0.27</math></b>	$\uparrow 19.43$
	MathQA	29.67	$30.84 \pm 0.60$	$33.07 \pm 1.11$	$38.42 \pm 1.22$	<b><math>51.18 \pm 0.24</math></b>	$\uparrow 21.51$

is complementary to inference-time sampling. The same pattern holds with full-precision, with improvements comparable to their 4-bit counterparts (App. O). Finally, self-consistency gains strongly correlate with accuracy improvements ( $r > 0.86$  across all tested inference conditions; see App. J).

**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 2, 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 the relative separation 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. M).

**Impact on debate dynamics.** Fig. 4 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 mitigates debate-driven performance degradation observed in MATHQA for Llama-3B. The con-

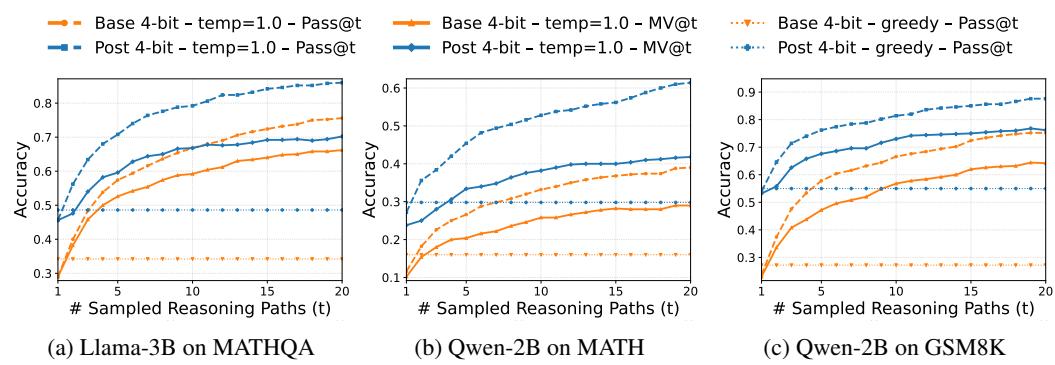


Figure 3: Post-training self-consistency improves sampling accuracy. Dashed: Pass@t (oracle upper bound), solid: MV@t (majority over  $t$  samples), dotted: greedy ( $\tau = 0$ ) accuracy. (Blue): post-trained model. (Orange): base model. Curves computed over 500 prompts.

Table 2: 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 3.01</math></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 0.64</math></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 0.72</math></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 0.99</math></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 0.83</math></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 0.53</math></b>	$63.13 \pm 1.89$	$\uparrow 42.73$
Phi4B	MATH	44.40	$45.53 \pm 2.53$	<b>49.93 <math>\pm 1.33</math></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 0.31</math></b>	$\uparrow 5.13$
	MathQA	49.60	$50.87 \pm 1.62$	<b>63.07 <math>\pm 1.21</math></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 1.33</math></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 1.51</math></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 1.55</math></b>	$\uparrow 24.67$

trastive signal allows models to learn from consensus patterns even when exposed to flawed reasoning trajectories, correcting rather than amplifying poor grounding behaviors.

**Improving self-consistency on math datasets improves general reasoning.** We demonstrate that training to improve internal self-consistency on any mathematical dataset enhances performance across diverse reasoning tasks, including previously unseen domains. Building on recent work

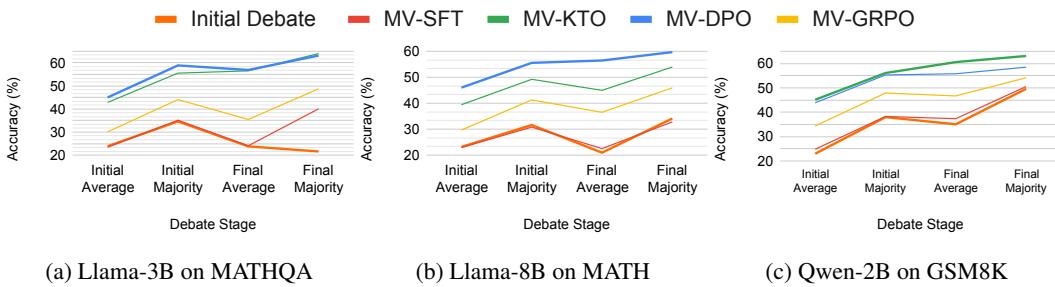


Figure 4: Debate-aware RL improves all stages of multi-agent debate. Incorporating debate context in RL teaches agents to leverage prior arguments, improving final-round consensus. Stages: initial round average, initial round majority vote, final round average, final round majority vote.

378 Table 3: Post-training self-consistency improves performance across general reasoning benchmarks.  
 379 Models trained on datasets (columns) tested on benchmarks (rows). Bottom rows show generaliza-  
 380 tion to unseen benchmarks: SVAMP (math), GPQA (science), CSQA (commonsense). All = joint  
 381 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>

403 showing math training enhances general reasoning (DeepSeek-AI, 2025; Akter et al., 2025), we  
 404 demonstrate that improving self-consistency—internal consensus strength—on math is essential for  
 405 reliably unlocking these capabilities. Table 3 reports results for models trained on MATH, GSM8K,  
 406 or MathQA individually, and on all three combined (All), using MV-DPO. Training on any single  
 407 dataset improves performance across all reasoning tasks, including unseen math (SVAMP), science  
 408 (GPQA), and commonsense reasoning (CSQA). Joint training achieves further improvements across  
 409 nearly every benchmark, demonstrating that diverse training data amplifies self-consistency gains.

### 5.3 ABLATION STUDY

412 To understand the sources of performance gains from MACA, we conduct ablation studies exam-  
 413 ining key components. We show that self-generated consensus signals outperform ground-truth  
 414 supervision, including peer context during training improves relative grounding and debate utili-  
 415 zation, and multi-round debate provides stronger signals than single-round majority vote. We also  
 416 demonstrate in App. M that self-supervised preference learning serves as an effective implicit for-  
 417 mat reward by reinforcing more efficient and concise chain-of-thought construction, though most  
 418 gains stem from fundamental reasoning improvements rather than formatting alone.

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

428 **RL with debate outperforms RL with single-round majority vote.** We evaluate three training  
 429 conditions in the single-agent setting to understand the impact of training on the debate versus ini-  
 430 tial round majority vote on the individual agents themselves, as shown in Table 6: (1) initial round  
 431 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., 2024)), and (3) debate majority

432 Table 5: Multi-agent: Post-training with debate context (peer CoTs) outperforms no context.  
433

GSM8K	Initial Round Avg		Final Round Avg		Final Round MV	
	NoCtx	Context	NoCtx	Context	NoCtx	Context
Llama-3B	65.87	<b>66.73</b>	67.87	<b>70.60</b>	70.60	<b>73.80</b>
Phi-4B	74.67	<b>76.60</b>	74.60	<b>77.80</b>	81.80	<b>84.40</b>
Llama-8B	75.80	<b>77.67</b>	78.00	<b>81.20</b>	81.60	<b>83.00</b>

440 Table 6: Single-agent: RL on debate majority-vote with peer context (MV-DPO) outperforms RL  
441 on single-round majority vote (TTRL, ScPO). Arrows show gains over instruction-tuned model.  
442 Statistics gathered over 3 seeds.  
443

	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	$\uparrow 16.7 \pm 0.4$	$\uparrow 12.5 \pm 0.7$	$\uparrow 6.9 \pm 0.2$	$\uparrow 17.1 \pm 0.8$	$\uparrow 23.8 \pm 0.9$	$\uparrow 3.7 \pm 0.3$	$\uparrow 7.3 \pm 0.3$	$\uparrow 18.2 \pm 0.2$

451 vote with DPO including peer context (MV-DPO from MACA). Training on the debate signal im-  
452 proves single-agent performance more than reinforcing the initial round majority vote. Additionally,  
453 Table 5 shows that including peer context during preference learning (MV-DPO) improves multi-  
454 agent debate performance (more models/datasets in App. G), with the largest gains in final-round  
455 individual accuracy, reflecting better use of peer chains-of-thought during deliberation (Fig. 4).

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

## 465 6 CONCLUSION, LIMITATIONS, AND FUTURE WORK

466 We introduce self-consistency as an intrinsic property of well-aligned reasoning and present MACA,  
467 a self-supervised framework that teaches models to reliably sample coherent reasoning through re-  
468 inforcement of internal consensus signals. Without external supervision, MACA drives substantial  
469 improvements: +27.6% self-consistency on GSM8K, +22.4% Pass@20 on MATH, +23.7% zero-  
470 shot accuracy on MATH, strong generalization to unseen benchmarks (+16.3% GPQA, +11.6%  
471 CommonsenseQA), and enhanced multi-agent performance (+42.7% on MathQA). Through debate,  
472 agents attempt to ground their reasoning in peer context, and we select trajectories where such de-  
473 liberation aligns with internal consensus. These consensus-aligned examples provide natural super-  
474 vision for stable reasoning patterns, enabling models to self-improve their reasoning consistency, ef-  
475 ficiency, and accuracy without explicit chain-of-thought supervision or external answer verification.  
476 Through preference learning on debate signals, models produce more concise chains-of-thought  
477 without format rewards and correct cases where aggregation previously led to degeneration, signif-  
478 icantly improving the robustness of multi-agent debate. While MACA has limitations, including  
479 requiring sufficient base model competence to generate meaningful consensus signals, potentially  
480 amplifying existing biases in model outputs, and lacking direct supervision of intermediate rea-  
481 soning correctness, it demonstrates that consensus reinforcement has robust benefits. Future work  
482 could explore alternative consensus methods, confidence-weighted voting, heterogeneous agents,  
483 and better leveraging of minority traces. Sizeable gains on difficult unseen tasks (+10.0% on AMC,  
484 App. M) suggest consensus alignment unlocks latent capabilities beyond consistency itself. MACA  
485 demonstrates that language models can effectively harness internal deliberation for self-alignment,  
elevating their reasoning capabilities autonomously.

486 REFERENCES  
487

488 M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Java-  
489 heripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price,  
490 G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and  
491 Y. Zhang. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

492 S. N. Akter, S. Prabhumoye, M. Novikov, S. Han, Y. Lin, E. Bakhturina, E. Nyberg, Y. Choi, M. Pat-  
493 wary, M. Shoeybi, and B. Catanzaro. Nemotron-crossthink: Scaling self-learning beyond math  
494 reasoning. *arXiv preprint arXiv:2504.13941*, 2025.

495 A. Amini, S. R. Gabriel, A. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi. MathQA: To-  
496 wards interpretable math word problem solving with operation-based formalisms. In *Conference*  
497 *of the North American Chapter of the Association for Computational Linguistics: Human Lan-*  
498 *guage Technologies*, pp. 2357–2367, 2019.

499 M. M. Botvinick, J. D. Cohen, and C. S. Carter. Conflict monitoring and anterior cingulate cortex:  
500 An update. *Trends in Cognitive Sciences*, 8(12):539–546, 2004.

502 M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda,  
503 N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin,  
504 B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, Ł. Kaiser, M. Bavarian, C. Winter, P. Tillet,  
505 F. Petroski Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H.  
506 Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders,  
507 C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight,  
508 M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish,  
509 I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. *arXiv preprint*  
510 *arXiv:2107.03374*, 2021.

511 X. Chen, R. Aksitov, U. Alon, J. Ren, K. Xiao, P. Yin, S. Prakash, C. Sutton, X. Wang, and  
512 D. Zhou. Universal self-consistency for large language model generation. *arXiv preprint*  
513 *arXiv:2311.17311*, 2023.

514 K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, Ł. Kaiser, M. Plappert, J. Tworek, J. Hilton,  
515 R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv*  
516 *preprint arXiv:2110.14168*, 2021.

518 DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.  
519 *arXiv preprint arXiv:2501.12948*, 2025.

520 T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quan-  
521 tized LLMs. *arXiv preprint arXiv:2305.14314*, 2023.

523 Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in  
524 language models through multiagent debate. In *Proceedings of the 41st International Conference*  
525 *on Machine Learning*, 2023.

526 Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg. Mea-  
527 suring and improving consistency in pretrained language models. *Transactions of the Association*  
528 *for Computational Linguistics*, 9:1012–1031, 2021.

530 K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect  
531 theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*,  
532 pp. 12634–12651, 2024.

533 L. Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

535 K. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):  
536 127–138, 2010.

537 A. Glaese, N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger,  
538 M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu,  
539 F. Yang, A. See, S. S. Dathathri, R. Greig, C. Chen, D. Fritz, J. Sanchez Elias, R. Green,

540 S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hass-  
 541 abis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving. Improving alignment of dialogue agents  
 542 via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

543 J. Gottweis, W. H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissberger,  
 544 K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk,  
 545 A. Vahdat, P. Kohli, Y. Matias, A. Carroll, K. Kulkarni, N. Tomasev, Y. Guan, V. Dhillon, E. D. V.  
 546 Vaishnav, B. Lee, T. R. D. Costa, J. R. Penadés, G. Peltz, Y. Xu, A. Pawlosky, A. Karthikesalingam, and V. Natarajan. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

547 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur,  
 548 A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*,  
 549 2024.

550 D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Mea-  
 551 suring mathematical problem solving with the MATH dataset. In *Neural Information Processing  
 552 Systems Track on Datasets and Benchmarks*, 2021.

553 A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration.  
 554 In *International Conference on Learning Representations*, 2020.

555 A. Huang, A. Block, D. J. Foster, D. Rohatgi, C. Zhang, M. Simchowitz, J. T. Ash, and A. Krish-  
 556 namurthy. Self-improvement in language models: The sharpening mechanism. In *International  
 557 Conference on Learning Representations*, 2025.

558 J. Huang, X. Chen, S. Mishra, H. Zheng, A. W. Yu, X. Song, and D. Zhou. Large language mod-  
 559 els cannot self-correct reasoning yet. In *International Conference on Learning Representations*,  
 560 2024.

561 E. J. Husom, A. Goknil, M. Astekin, L. K. Shar, A. Kåsen, S. Sen, B. A. Mithassel, and A. Soylu.  
 562 Sustainable llm inference for edge ai: Evaluating quantized llms for energy efficiency, output  
 563 accuracy, and inference latency. *arXiv preprint arXiv:2504.03360*, 2025.

564 G. Irving, P. Christiano, and D. Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*,  
 565 2018.

566 J. Jiang and Z. Lu. Offline decentralized multi-agent reinforcement learning. *arXiv preprint  
 567 arXiv:2108.01832*, 2021.

568 F. Jiao, G. Guo, X. Zhang, N. F. Chen, S. Joty, and F. Wei. Preference optimization for reasoning  
 569 with pseudo feedback. In *International Conference on Learning Representations*, 2025.

570 D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

571 KnovelEng. AMC-23 dataset. Hugging Face Datasets, 2023. URL <https://huggingface.co/datasets/knoveleng/AMC-23>.

572 D. Li, R. Sun, Y. Huang, M. Zhong, B. Jiang, J. Han, X. Zhang, W. Wang, and H. Liu. Preference  
 573 leakage: A contamination problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025.

574 J. Li, J. Wang, Z. Zhang, and H. Zhao. Self-prompting large language models for zero-shot open-  
 575 domain QA. In *Conference of the North American Chapter of the Association for Computational  
 576 Linguistics*, 2024.

577 Marquis de Condorcet M. J. A. N. de Caritat. *Essai sur l'application de l'analyse à la probabilité  
 578 des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785.

579 E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of  
 580 Neuroscience*, 24:167–202, 2001.

581 S. Moscovici. Studies in social influence iii: Majority versus minority influence in a group. *Euro-  
 582 pean Journal of Social Psychology*, 6(2):149–202, 1976.

583 C. J. Nemeth. *The Influence of Minority Social Position on the Formation of Social Judgment*.  
 584 Psychology Press, 1985.

594 L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal,  
 595 X. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder,  
 596 P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with  
 597 human feedback. In *Advances in Neural Information Processing Systems*, 2022.

598 A. Patel, S. Bhattacharya, and N. Goyal. Are NLP models really able to solve simple math word  
 599 problems? In *Conference of the North American Chapter of the Association for Computational  
 600 Linguistics: Human Language Technologies*, pp. 2080–2094, 2021.

602 A. Prasad, W. Yuan, R. Y. Pang, J. Xu, M. Fazel-Zarandi, M. Bansal, S. Sukhbaatar, J. Weston, and  
 603 J. Yu. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*, 2024.

604 B. Radharapu, M. Revel, M. Ung, S. Ruder, and A. Williams. Arbiters of ambivalence: Challenges  
 605 of using LLMs in no-consensus tasks. *arXiv preprint arXiv:2505.23820*, 2025.

607 R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference  
 608 optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*,  
 609 2023.

610 D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman.  
 611 GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

612 Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu, and D. Guo. Deepseek-  
 613 math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint  
 614 arXiv:2402.03300*, 2024.

616 A. Shenhav, M. M. Botvinick, and J. D. Cohen. The expected value of control: An integrative theory  
 617 of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

618 P. Srivastava, H. Singh, R. Madhavan, G. Patil, S. Addepalli, A. Suggala, R. Aravamudhan,  
 619 S. Sharma, A. Laha, A. Raghav, K. Shanmugam, and D. Precup. Robust reward modeling  
 620 via causal rubrics. 2025.

622 V. Subramaniam, Y. Du, J. B. Tenenbaum, A. Torralba, S. Li, and I. Mordatch. Multiagent finetun-  
 623 ing: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.

625 A. Talmor, J. Herzig, N. Lourie, and J. Berant. CommonsenseQA: A question answering challenge  
 626 targeting commonsense knowledge. In *Conference of the North American Chapter of the Association  
 627 for Computational Linguistics: Human Language Technologies*, pp. 4149–4158, 2019.

628 Y. Tong, J. Yuan, and C. Hu. Enhancing quantization-aware training on edge devices via relative  
 629 entropy coresnet selection and cascaded layer correction. *arXiv preprint arXiv:2507.17768*, 2025.

631 X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou.  
 632 Self-consistency improves chain of thought reasoning in language models. *arXiv preprint  
 633 arXiv:2203.11171*, 2022.

634 A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei,  
 635 H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He,  
 636 J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng,  
 637 R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou,  
 638 X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu,  
 639 Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*,  
 640 2024.

641 Y. Yang, X. Ma, C. Li, Z. Zheng, Q. Zhang, G. Huang, J. Yang, and Q. Zhao. Believe what you  
 642 see: Implicit constraint approach for offline multi-agent reinforcement learning. In *Advances in  
 643 Neural Information Processing Systems*, pp. 10299–10312, 2021.

644 W. Zhan, S. Fujimoto, Z. Zhu, J. D. Lee, D. R. Jiang, and Y. Efroni. Exploiting structure in offline  
 645 multi-agent RL: The benefits of low interaction rank. In *International Conference on Learning  
 646 Representations*, 2025.

648 S. Zhang, Y. Tian, Q. Liu, and H. Wu. The neural correlates of novelty and variability in human  
649 decision-making under an active inference framework. *eLife*, 13:RP92892, 2025.  
650

651 J. Zhu, Y. Shen, J. Zhao, and A. Zou. Path-consistency: Prefix enhancement for efficient inference  
652 in LLM. *arXiv preprint arXiv:2409.01281*, 2024a.

653 Z. Zhu, M. Liu, L. Mao, B. Kang, M. Xu, Y. Yu, S. Ermon, and W. Zhang. MADiff: Offline multi-  
654 agent learning with diffusion models. In *Advances in Neural Information Processing Systems*,  
655 2024b.

656

657 Y. Zuo, K. Zhang, L. Sheng, S. Qu, G. Cui, X. Zhu, H. Li, Y. Zhang, X. Long, E. Hua, B. Qi, Y. Sun,  
658 Z. Ma, L. Yuan, N. Ding, and B. Zhou. Ttrl: Test-time reinforcement learning. *arXiv preprint*  
659 *arXiv:2504.16084*, 2025.

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 **A LLM USAGE DECLARATION**  
703704 We declare that large language models (LLMs) were only used to assist with minor tasks, including  
705 grammar checking, language refinement, and basic coding support such as debugging. All research  
706 contributions, experimental design, data analysis, and scientific findings presented in this work are  
707 original to the authors.  
708709 **B EXAMPLES OF POST-TRAINED CONSISTENT REASONING IN SINGLE- AND**  
710 **MULTI- AGENT SETTINGS**  
711712 This appendix presents examples from the debates, providing a qualitative signal of how post-  
713 training affects both single-agent reasoning and multi-agent collaboration. The evidence highlights  
714 visible improvements in mathematical reasoning, consensus formation, and systematic problem-  
715 solving.  
716717 **B.1 OVERVIEW OF KEY FINDINGS**  
718719 Analysis of the transcripts reveals that post-training leads to more systematic and efficient reasoning  
720 in single-agent settings. Models demonstrate clearer step-by-step problem decomposition, make  
721 fewer calculation errors, and show improved conceptual understanding. Notably, responses become  
722 more concise and are less likely to be truncated due to inefficient reasoning.  
723724 In multi-agent debates, post-training enhances collaborative effects. Agents more frequently refer-  
725 ence and build upon each other’s solutions, leading to more robust cross-verification and consensus  
726 formation. We observe that post-trained agents are more likely to resolve mistakes through discus-  
727 sion rather than propagate errors.  
728729 Our analysis shows that preference learning teaches models higher-quality and more concise rea-  
730 soning without requiring explicit format rewards. The consensus mechanism itself becomes more  
731 effective at aggregating individual outputs, supporting both in-distribution and out-of-distribution  
732 transfer.  
733734 **B.2 CASE STUDY 1: SINGLE-AGENT MATHEMATICAL REASONING**  
735736 **B.2.1 PROBLEM: CITIES A AND B RELATIVE SPEED**737 **AMC Problem Statement**738 Cities  $A$  and  $B$  are 45 miles apart. Alicia lives in  $A$  and Beth lives in  $B$ . Alicia bikes towards  
739  $B$  at 18 miles per hour. Leaving at the same time, Beth bikes toward  $A$  at 12 miles per hour.  
740 How many miles from City  $A$  will they be when they meet?  
741742 **Correct Solution:** Combined speed =  $18 + 12 = 30$  mph; Time to meet =  $45/30 = 1.5$  hours;  
743 Distance from  $A$  =  $18 \times 1.5 = 27$  miles  
744745 **Ground Truth:** 27  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

756 B.2.2 PRE-TRAINING RESPONSE: FUNDAMENTAL CONCEPTUAL CONFUSION  
757758 **Base Model: Conceptual Breakdown**  
759760 To solve this problem, we can use the concept of relative motion. Here's a step-by-step  
761 breakdown:  
762763 

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

778 [Response cuts off mid-sentence]

779 **Parsed Answer:** null780 **Analysis: Multiple failure modes.** The pre-training response reveals several critical reasoning  
781 failures: incorrect problem setup, misapplication of distance-time relationships, and ultimately an  
782 illogical result. The model exhausts its token budget exploring unproductive solution paths before  
783 reaching a conclusion.784 B.2.3 POST-TRAINING RESPONSE: CORRECT SYSTEMATIC SOLUTION  
785786 **Post-Training: Correct and Concise**  
787788 To solve this problem, we can follow these steps:  
789790 

- 791 • We know the distance between Cities A and B is 45 miles.
- 792 • Alicia and Beth are biking towards each other at a combined speed of  $18 + 12 = 30$   
793 miles per hour.
- 794 • Since they are biking towards each other, we can use the concept of relative speed  
795 to find the time it takes for them to meet.
- 796 • We can calculate the time using the formula: time = distance / speed. Using this  
797 formula, the time they take to meet is  $45$  miles /  $30$  miles per hour, which equals  
798  $1.5$  hours.
- 799 • Now that we have the time, we can find the distance covered by Alicia, since we  
800 know her speed is 18 miles per hour. Distance covered by Alicia is  $18 \times 1.5$ , which  
801 equals 27 miles.

802 Therefore, Alicia and Beth will meet 27 miles from City A.  
803

27

804 **Analysis: Systematic improvement.** Post-training produces a clear, systematic approach: correct  
805 conceptual application, logical step progression, and efficient completion within the token limit.  
806 This demonstrates improved reasoning quality and token efficiency achieved through preference  
807 learning without explicit format rewards.808  
809

810 B.3 CASE STUDY 2: MULTI-AGENT MATHEMATICAL REASONING  
811812 B.3.1 PROBLEM: ROAD TRIP GAS MILEAGE CALCULATION  
813814 **Problem Statement**815 Sophia is thinking of taking a road trip in her car, and would like to know how far she can  
816 drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she  
817 needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that  
818 her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?  
819820 **Correct Solution:**  $100 \div 4 = 25 \text{ mpg}$ ;  $25 \times 12 = 300 \text{ miles}$ 821 **Ground Truth:** 300822 B.4 COMPLETE MULTI-AGENT TRANSCRIPTS: PRE-TRAINING FAILURE (ITERATION 0)  
823824 B.4.1 ROUND 1: INITIAL AGENT RESPONSES  
825826 **User Prompt**827 Solve the following math problem. Sophia is thinking of taking a road trip in her car, and  
828 would like to know how far she can drive on a single tank of gas. She has traveled 100 miles  
829 since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The  
830 owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can  
831 Sophia drive on a single tank of gas?832 Provide a bullet point summary of your step-by-step reasoning. Your final answer should be  
833 a single numerical number, in the form \boxed{answer}, at the end of your response.  
834835 **Agent 1: Mathematical Error**836 To solve this problem, I need to determine Sophia's car's fuel efficiency and then calculate  
837 how far she can drive on a full tank.

838 Step 1: Determine fuel efficiency from the given information.

839

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

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

841

- Since 1 mpg means traveling 1 mile with each gallon, Sophia will use  $4 \text{ gallons} \times$   
842  $(1 \text{ mile/mpg}) = 4 \text{ miles per tank.}$

843 Step 6: Calculate total distance for full tank:

844

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

845 **48**846 **Agent 2: Incomplete Reasoning**847 If the distance covered in one tank of fuel is 100 miles, then the Miles Per Gallon (MPG) is  
848 calculated as follows...

849 [Response appears incomplete and truncated mid-calculation]

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

851 Let me reconsider: Sophia used 4 gallons to travel some distance, and her tank holds 12  
852 gallons total.853 I'm getting confused by the different approaches. Let me try again: 100 miles with 4 gallons  
854 means...855 Final answer: **48**

856

857

858

859

860

861

862

863

864  
865**Agent 3: Correct but Overruled**

866 Let me solve this step by step:

867 Step 1: Calculate fuel efficiency from given data:

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

871 Step 2: Calculate range on full tank:

- 872 • Full tank capacity: 12 gallons
- 873 • Total range:  $25 \text{ mpg} \times 12 \text{ gallons} = 300 \text{ miles}$

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

875  
876  
877  
300

878

879  
880**B.4.2 ROUND 2: FAILED COLLABORATIVE DISCUSSION**881  
882**Agent 1: Doubling Down on Error**

883 After reviewing all approaches, I maintain my calculation:

884 Based on the calculations, I'm confident that 48 miles is correct because we need to convert  
885 the efficiency properly.886  
887  
48888  
889**Pre-Training Result:** Consensus selected 48 miles despite Agent 3's correct reasoning.890  
891  
892**B.4.3 COMPLETE MULTI-AGENT TRANSCRIPTS: POST-TRAINING SUCCESS (ITERATION 1)**893  
894**B.4.4 ROUND 1: PERFECT AGENT RESPONSES**895  
896**Agent 1: Complete Correction**897 To solve this problem, I need to determine Sophia's fuel efficiency and calculate her range  
898 on a full tank.

899 Step 1: Calculate Sophia's car's fuel efficiency:

- 900 • Sophia traveled 100 miles since last fill-up
- 901 • She used 4 gallons of gas
- 902 • Fuel efficiency =  $100 \text{ miles} \div 4 \text{ gallons} = 25 \text{ miles per gallon}$

903 Step 2: Calculate range on full tank:

- 904 • Tank capacity: 12 gallons
- 905 • Range =  $25 \text{ mpg} \times 12 \text{ gallons} = 300 \text{ miles}$

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

907  
908  
300909  
910  
911**Unanimous Agreement**912 **Agent 1:** "I see all agents used the same systematic approach. The calculation is clear:  $25 \text{ mpg} \times 12 \text{ gallons} = 300 \text{ miles}$ ."913 **Agent 2:** "Perfect agreement on methodology. The fuel efficiency calculation is the key  
914 step, then applying it to full tank capacity."915 **Agent 3:** "Unanimous consensus on 300 miles. All reasoning aligns perfectly."916  
917**Post-Training Result:** Perfect consensus on 300 miles with all agents showing correct reasoning.

918 B.4.5 ANALYSIS: TRAINING EFFECTS ON MULTI-AGENT COLLABORATION  
919

920 Post-training produces substantial improvements in both individual and collective reasoning. Be-  
921 fore training, agents make calculation errors and the consensus mechanism selects incorrect answers  
922 even when correct solutions are present, with two agents converging on the same mistake and over-  
923 ruling the correct agent. After training, all agents apply correct methodology and reach unanimous  
924 agreement on the right answer. These results demonstrate that preference learning encourages sys-  
925 tematic step-by-step reasoning, reducing errors and their propagation during collaboration, while  
926 multi-agent training enhances the consensus mechanism’s ability to reliably select correct solutions.  
927

928 C GROUNDING BEHAVIORS IN BASE-MODEL MULTI-AGENT DEBATES  
929

930 Having provided examples of the impact of post-training on single- and multi-agent reasoning, we  
931 now highlight representative examples from some initial debates among base models demonstrating  
932 spontaneous grounding behaviors that emerge during multi-agent reasoning. Note that each debate  
933 consists of an initial round (three agents provide independent solutions) and a final round (agents  
934 update after seeing peers’ responses).

935 C.1 EXAMPLE 1: AIRPORT TIMING PROBLEM  
936937 **Problem**

938 Amber and Josh are flying to France for vacation. Their flight leaves at 8:00 pm and as  
939 an international flight, it’s recommended that they check in 2 hours early. It will take 45  
940 minutes to drive to the airport and another 15 minutes to park their vehicle and make their  
941 way to the terminal. What is the latest time they can leave their house to make it to the  
942 airport in time?  
943

944 The correct solution requires working backwards from the 8:00 PM flight time: they must arrive by  
945 6:00 PM (2 hours early), and with 1 hour total travel time, they must leave by 5:00 PM.  
946

947 **Initial round:**948 **Agent 1: Directional Error**

949

- 950 • Flight leaves at 8:00 pm, need to check in 2 hours early = 6:00 pm arrival
- 951 • Calculate travel time:  $45 + 15 = 60$  minutes = 1 hour
- 952 • Add travel time to departure time: 6:00 pm + 1 hour = 7:00 pm

953 **Error: Adds travel time instead of subtracting it**

954 Answer: 7:00 pm

955 **Agent 2: Fundamental Misunderstanding**

956

- 957 • Find total travel time:  $45 + 15 = 60$  minutes = 1 hour
- 958 • Add travel time to flight departure time: 8:00 pm + 1 hour = 9:00 pm

959 **Completely ignores 2-hour check-in requirement and adds travel time instead of subtracting  
960 it**

961 Answer: 9:00 pm

962 **Agent 3: Correct Solution**

963

- 964 • Flight at 8:00 pm requires arrival by 6:00 pm (2 hours early)
- 965 • Travel time: 45 min driving + 15 min parking = 60 minutes
- 966 • Must leave by: 6:00 pm - 1 hour = 5:00 pm

967 Answer: 5:00 pm

972  
973**Final round (after seeing peers):**

974

**Agent 1: Adopts Worse Error**

975

“After reviewing other solutions, I update my approach:

977

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

980

**Abandons partial understanding to adopt Agent 2’s incorrect reasoning as well**

981

Answer: 9:00 pm

983

984

**Agent 2: Persists with Error**

985

Maintains original incorrect answer of 9:00 pm

987

988

**Agent 3: Explicitly Identifies Error Pattern**

989

“The other agents seem to be adding time to the flight departure rather than working backwards from the required arrival time. The correct approach:

991

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

994

**Recognizes error in other agents’ responses and upholds its own correct understanding**

995

Answer: 5:00 pm

996

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

1004

1005

**C.2 EXAMPLE 2: SELF-CORRECTION IN MATHEMATICAL REASONING**

1006

1007

**Problem**

1008

1009

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

1010

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

1012

**Initial round:**

1013

1014

**Agent 1: Calculation Error**

1015

Original total:  $\$80 + \$50 = \$130$

1016

Discount:  $20\% = 0.2 \times \$130 = \$26$

1017

Final price:  $\$130 + \$26 = \$156$

1018

**Error: Adds discount instead of subtracting**

1019

Answer: \$156

1020

1021

**Agent 2: Correct Approach**

1022

1023

Total: \$130

1024

After 20% discount:  $\$130 \times 0.8 = \$104$

1025

Answer: \$104

1026

1027

1028 Jacket discount: 20% of \$80 = \$16, price = \$64

1029 Pants discount: 20% of \$50 = \$10

1030 Total: \$64 + \$10 = \$74

1031 **Error: Adds discounted jacket price to pants discount amount**

1032 Answer: \$74

1033

1034

**Final round:**

1035

1036

1037 "I realize I added the discount instead of subtracting. Looking at Agent 2's approach of  
1038 multiplying by 0.8 is cleaner:  $\$130 \times 0.8 = \$104$ "1039 **Recognizes own error and adopts correct method**

1040 Answer: \$104

1041

1042

1043

1044 "Jacket after discount:  $\$80 \times 0.8 = \$64$ 1045 Pants after discount:  $\$50 \times 0.8 = \$40$ 1046 Total:  $\$64 + \$40 = \$104$ "

1047 Answer: \$104

1048

1049

1050 **Key observation.** Agent 1 explicitly acknowledges their error ("I added the discount instead of  
subtracting") and adopts Agent 2's more efficient method. Agent 3 corrects their calculation error  
after seeing consistent peer answers.

1051

1052

**C.3 EXAMPLE 3: GEOMETRIC PROBLEM - METHOD SIMPLIFICATION**

1053

1054

**Problem**

1055

1056

A rectangle has a perimeter of 24 units and an area of 32 square units. What are the dimensions of the rectangle?

1057

1058

1059

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

1060

1061

**Initial round:**

1062

1063

**Agent 1: Algebraic Approach**

1064

Let length =  $l$ , width =  $w$ 

1065

From perimeter:  $2l + 2w = 24$ , so  $l + w = 12$ 

1066

From area:  $lw = 32$ 

1067

Setting up quadratic:  $w(12 - w) = 32$ 

1068

 $12w - w^2 = 32$ 

1069

 $w^2 - 12w + 32 = 0$ 

1070

**Stops without solving the quadratic**

1071

1072

1073

1074

1075

1076

1077

1078

1079

**Agent 2: Direct Factoring** $l + w = 12$  and  $lw = 32$ 

Looking for factors of 32 that sum to 12: 8 and 4

Dimensions:  $8 \times 4$ 

Answer: 8 units by 4 units

1080

## Agent 3: Quadratic Solution

1081

From  $w^2 - 12w + 32 = 0$ 

1082

Using quadratic formula or factoring:  $(w - 8)(w - 4) = 0$ 

1083

So  $w = 8$  or  $w = 4$ 

1084

Dimensions:  $8 \times 4$ 

1085

Answer: 8 units by 4 units

1086

1087

## Final round:

1088

## Agent 1: Adopts Simpler Method

1089

1090

“Agent 2’s approach of finding factors of 32 that sum to 12 is much more direct than solving the quadratic. Since  $8 \times 4 = 32$  and  $8 + 4 = 12$ , the dimensions are 8 by 4.”

1091

Recognizes and adopts more efficient solution method

1092

Answer: 8 units by 4 units

1093

1094

1095

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

1100

1101

## C.4 SUMMARY

1102

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

1103

1104

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

1105

1106

**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”).

1107

1108

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

1109

1110

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

1111

1112

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.

1113

1114

## D MACA ALGORITHM: ADDITIONAL IMPLEMENTATION DETAILS

1115

1116

## D.1 AGENT HETEROGENEITY AND SPECIALIZATION

1117

1118

1119

1120

1121

1122

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. H. 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.

1131

1132

1133

---

1134 **Algorithm 1** Multi-Agent Consensus Alignment Via Debate-RL

---

1135

1136 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  
1137 MV-KTO

1138 2: **for**  $\ell = 1, \dots, L$  **do**

1139 3:   sample batch  $\{x^{(\ell,i)}\}_{i=1}^B$  of prompts from  $\mathcal{D}$

1140 4:   **for**  $i = 1, \dots, B$  **do** ▷ Generate debate trajectories

1141 5:      $y_{m,1}^{(\ell,i)} \sim \pi_{\theta_m}(\cdot | x^{(\ell,i)}) \quad \forall m$

1142 6:     **for**  $r = 2, \dots, R$  **do** ▷ Condition on peers' previous round outputs

1143 7:        $x_{m,r}^{(\ell,i)} = [x^{(\ell,i)}; \{y_{j,r-1}^{(\ell,i)}\}_{j \neq m}]$

1144 8:        $y_{m,r}^{(\ell,i)} \sim \pi_{\theta_m}(\cdot | x_{m,r}^{(\ell,i)}) \quad \forall m$

1145 9:     **end for**

1146 10:   Parse final answers  $a_m^{(\ell,i)} = \mathcal{A}(y_{m,R}^{(\ell,i)}) \quad \forall m$

1147 11:   Aggregate consensus  $\hat{a}(x^{(\ell,i)}) = \text{Majority}\{a_1^{(\ell,i)}, \dots, a_M^{(\ell,i)}\}$

1148 12:   Extract final contexts  $x_m^{(\ell,i)} \leftarrow x_{m,R}^{(\ell,i)}$  for all  $m$

1149 13:   Partition responses by consensus alignment:  
1150      $\mathcal{G}_+^{(\ell,i)} = \{y_{m,R}^{(\ell,i)} : a_m^{(\ell,i)} = \hat{a}(x^{(\ell,i)})\}$   
1151      $\mathcal{G}_-^{(\ell,i)} = \{y_{m,R}^{(\ell,i)} : a_m^{(\ell,i)} \neq \hat{a}(x^{(\ell,i)})\}$

1152 14: **end for**

1153 15: **build dataset:**  $\mathcal{D}_{\text{post}} = \{(x^{(\ell,i)}, \hat{a}(x^{(\ell,i)}), \mathcal{G}_+^{(\ell,i)}, \mathcal{G}_-^{(\ell,i)})\}_{i=1}^B$

1154 16: **training data by method:**

1155     MV-SFT: for each  $m$ , use  $\{(x_m^{(\ell,i)}, y_{m,R}^{(\ell,i)}) : y_{m,R}^{(\ell,i)} \in \mathcal{G}_+^{(\ell,i)}\}_i$ .

1156     MV-GRPO: for each  $m$ , store  $\{(x_m^{(\ell,i)}, \hat{a}(x^{(\ell,i)}))\}_i$  for reward computation on new samples.

1157     MV-DPO/MV-KTO: pool  $\{(x_m^{(\ell,i)}, y_{m,R}^{(\ell,i)})\}_{i,m}$  partitioned by  $\mathcal{G}_+, \mathcal{G}_-$  across agents.

1158 17: **update policies:** apply majority-vote objective to refine each  $\pi_{\theta_m}$

1159 18: **end for**

---

## E EXPERIMENTAL SETUP

### E.1 TRAINING PARAMETERS

1164 We conducted hyperparameter sweeps across learning rates from 1e-7 to 1e-5, finding that 1e-5  
1165 consistently performed best across all methods. For preference-based methods (MV-KTO, MV-  
1166 DPO), we used  $\beta = 0.1$  throughout all experiments. LoRA ranks and alpha values were adjusted  
1167 based on model size and computational constraints, with rank/alpha of 64-128 for MV-DPO and  
1168 MV-KTO, 32-64 for MV-GRPO, and 128 for MV-SFT. Tables 7–10 provide complete parameter  
1169 specifications.

1170 GSM8K, MATH, and MathQA used 1500/500/500 train/valid/test splits. SVAMP, GPQA, CSQA,  
1171 and AMC were used only for testing, with 300, 448, 500, and 40 test examples, respectively.

### E.2 DATASETS

1174 We evaluated model mathematical reasoning using seven publicly available datasets spanning a  
1175 range of difficulty and subject areas: MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021),  
1176 MathQA (Amini et al., 2019), SVAMP (Patel et al., 2021), GPQA (Rein et al., 2023), AMC 23 Knov-  
1177 elEng (2023), and CommonsenseQA/CSQA (Talmor et al., 2019).

1179 **MATH** (Hendrycks et al., 2021): The MATH dataset contains 12,500 high school mathematics  
1180 problems from algebra, geometry, combinatorics, and number theory, each requiring multi-step reasoning  
1181 and precise mathematical solutions.

1182 **GSM8K** (Cobbe et al., 2021): GSM8K is composed of 8,500 grade-school-level word problems  
1183 emphasizing arithmetic and logical reasoning, with step-by-step annotated solutions for each prob-  
1184 lem.

1185 **MathQA** (Amini et al., 2019): MathQA features over 37,000 question-answer pairs based on quan-  
1186 titative reasoning, requiring models to convert natural language problems into mathematical expres-  
1187 sions and perform multi-step computation.

1188  
1189

Table 7: Training parameters.

1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208

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)		

1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219

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

1220  
1221  
1222

**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.

1223  
1224  
1225  
1226

**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.

1227  
1228  
1229

**AMC 23** (KnovelEng, 2023): 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

1230  
1231  
1232  
1233  
1234

**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.

1235  
1236  
1237

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

1238  
1239

### E.3 MULTI-AGENT DEBATE INFRASTRUCTURE

1240  
1241

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

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

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

### E.3.1 QUANTIZED MARL TRAINING

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

### E.3.2 SCALABLE DEBATE INFERENCE FRAMEWORK

The inference system enables efficient multi-agent debates through adapter hot-swapping and dynamic resource management, designed to scale seamlessly with available compute resources. A centralized base model manager maintains a pool of quantized base models shared across agents, implementing lazy loading and caching for lightweight LoRA adapters that enable rapid agent switching without full model reloading. An adapter job scheduler coordinates debate workloads across available GPUs through adaptive load balancing, priority queues for debate tasks ordered by round index, and automatic cluster adaptation that detects available hardware and distributes workloads accordingly. This architecture enables concurrent multi-agent inference on hardware that would typically support only single full-precision models, with debate throughput scaling linearly with compute resources while maintaining memory efficiency through quantization and adapter sharing.

### E.4 DEBATE PROMPT TEMPLATES

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

#### E.4.1 INITIAL QUESTION PROMPT

Used in the first round of debate.

#### For GPQA, MathQA, CommonsenseQA

1296 Initial Prompt: Multiple Choice  
1297  
1298  
1299 Answer the following multiple choice question as accurately  
1300 as possible.  
1301 {question}  
1302  
1303 Provide a bullet point summary of your step-by-step reasoning  
1304 .  
1305 Your final answer should be a single choice letter in the  
1306 form  
1307 \boxed{answer}, at the end of your response.  
1308

## For GSM8K, SVAMP, AMC, MATH

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.

#### E.4.2 DEBATE PHASE PROMPT

1324 Used in subsequent rounds of debate to incorporate peer context.

1325  
1326 Debate Phase Prompt  
1327  
1328 Here are solutions from other agents:  
1329 {agent\_responses}  
1330  
1331 Using each response as additional advice, give an updated  
1332 bullet point summary of your step-by-step reasoning to  
1333 the question: {question}  
1334  
1335 Make sure your final answer is in the form \boxed{answer}, at  
1336 the end of your response.  
1337

## F 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.

1343 In the single-agent setting (Table 11), DMV supervision proves consistently competitive with, and  
1344 often superior to, GT supervision across Qwen2B, Llama-3B, Phi-4B, and Llama-8B. Gains are  
1345 most pronounced for preference-based objectives such as KTO and DPO, where DMV provides  
1346 a more stable learning signal. By contrast, the oracle GT signal sometimes produces degraded  
1347 performance, particularly in KTO. Notably, DMV provides a robust alternative that avoids these  
1348 pitfalls, often leading to stronger outcomes without requiring external supervision.

1350 Table 11: Single-agent accuracy: Unsupervised post-training using debate-majority-vote-derived  
 1351 supervision (DMV) performs comparably to or outperforms supervised training using ground truth  
 1352 (GT) labels. Bold indicates the better score in each pair.

1354 Model	1355 Dataset	1356 Debate	1357 SFT		1358 KTO		1359 DPO		1360 GRPO	
			1361 <i>GT</i>	1362 <i>DMV</i>	1363 <i>GT</i>	1364 <i>DMV</i>	1365 <i>GT</i>	1366 <i>DMV</i>	1367 <i>GT</i>	1368 <i>DMV</i>
1357 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
1359 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
1361 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
1363 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>

1365 Table 12: Multi-agent accuracy: Unsupervised post-training using debate-majority-vote-derived  
 1366 supervision (DMV) performs comparably to or outperforms supervised training using ground truth  
 1367 (GT) labels. Bold indicates the better score in each pair.

1369 Model	1370 Dataset	1371 Init	1372 SFT		1373 KTO		1374 DPO		1375 GRPO	
			1376 <i>GT</i>	1377 <i>DMV</i>	1378 <i>GT</i>	1379 <i>DMV</i>	1380 <i>GT</i>	1381 <i>DMV</i>	1382 <i>GT</i>	1383 <i>DMV</i>
1372 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
1374 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>
1376 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>
1378 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>

1382 In the multi-agent setting (Table 12), the advantages of DMV supervision become even clearer.  
 1383 Across nearly all models and methods, DMV either matches or exceeds GT labels. The gains are  
 1384 especially consistent under preference-learning formulations (KTO, DPO), where DMV supervision  
 1385 yields more reliable improvements to consensus-based performance. While GT labels retain  
 1386 competitive strength in certain cases (e.g., Phi-4B on MATH tasks), DMV repeatedly delivers higher or  
 1387 more stable final-round accuracies.

1388 Together, these results show that debate-derived majority-vote supervision provides an effective,  
 1389 scalable alternative to oracle ground truth. DMV not only mitigates the instability observed when  
 1390 GT is used in preference-learning objectives, but also enhances both single- and multi-agent training.  
 1391 These findings underscore the efficacy of unsupervised alignment signals, leveraging a model’s own  
 1392 consensus dynamics, as a robust substitute for human-labeled supervision.

## 1394 G ABLATION: EFFECT OF PEER CONTEXT IN MULTI-AGENT DEBATE

1396 Tables 13 and 14 examine whether conditioning on peer responses during training improves debate  
 1397 performance.

1398 Training with peer context substantially improves multi-agent debate performance (Table 14). While  
 1399 initial round (single-agent) performance shows modest benefits from context training (5 out of 8  
 1400 cases improve), the advantages become pronounced in final-round multi-agent settings. Context  
 1401 training improves final-round individual accuracy in 7 out of 8 cases, with particularly large gains  
 1402 for models like L8B-MATH (57.80 vs. 51.47) and L3B-MATH (45.80 vs. 41.13). Most importantly,  
 1403 final-round majority voting benefits from context in 6 out of 8 cases, with GSM8K tasks show-

1404 Table 13: Multi-agent comparison of Context vs No Context runs across Initial Debate, SFT, KTO,  
 1405 DPO, and GRPO. Bold indicates which setting performed better in each pair.

Model	Dataset	Debate	SFT		KTO		DPO		GRPO	
			Context	NoCtx	Context	NoCtx	Context	NoCtx	Context	NoCtx
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

1418 Table 14: Post-training with peer context teaches agents to utilize other agents’ responses in the  
 1419 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>

1433 ing consistent improvements across all model sizes (1.4-3.2 percentage points) and larger models  
 1434 achieving substantial gains on MATH tasks (e.g., L8B-MATH improving from 57.20 to 60.80).  
 1435 These results demonstrate that context-aware training teaches agents to effectively leverage peer  
 1436 feedback during deliberation, leading to stronger consensus outcomes in multi-agent debate.

## H ITERATIVE IMPROVEMENT

1440 Table 15 examines whether iterative training beyond the first iteration yields continued improve-  
 1441 ments. While It-1 produces the substantial gains reported in our main results, iterations 2 and 3  
 1442 demonstrate continued modest improvements: in 23 of 24 evaluation settings, either It-2 or It-3  
 1443 achieves the best performance. For example, Phi-4B on MATH majority vote increases from 55.00  
 1444 (It-1) to 57.40 (It-3), and Llama-8B on GSM8K improves from 82.80 to 85.60.

1445 These gains show clear diminishing returns compared to the It-0→It-1 jump, with typical improve-  
 1446 ments of 1-3 percentage points between iterations. Performance occasionally dips between adjacent  
 1447 iterations (e.g., L3B-GSM8K majority vote: 72.00→74.60→70.80) but generally trends upward.  
 1448 This pattern suggests that iterative training continues to extract useful signal from debate-generated  
 1449 data, though with decreasing marginal benefit after the initial iteration.

## I DPO PAIR SELECTION STRATEGY ANALYSIS

1453 We compared two strategies for creating preference pairs from multi-agent debate data: majori-  
 1454 ty/minority partitioning versus confidence-based selection using model log-probabilities. Analysis  
 1455 covered 2,226 agent responses from 742 problems across multiple models and datasets. Majori-  
 1456 ty/minority partitioning substantially outperforms confidence-based selection across all metrics.  
 1457 Majority responses achieve 68.0% accuracy versus 28.1% for minority responses (39.8 percentage  
 1458 point gap), while high-confidence responses achieve only 51.5% accuracy versus 33.2% for low-

1458 Table 15: Iterative alternation between debate generation and post-training across four iterations  
 1459 shows substantial initial gains (It-0 to It-1) followed by diminishing returns. Model abbreviations:  
 1460 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>

1474 confidence (18.3 percentage point gap). The majority strategy yields an effect size of Cohen’s  $d =$   
 1475 1.832 compared to 0.281 for confidence-based selection, a  $6.5\times$  difference in discriminative power.  
 1476 Additionally, majority voting provides usable preference signals in 70.5% of examples versus 35.2%  
 1477 for confidence-based selection. These results validate using majority vote consensus for DPO pair  
 1478 selection, demonstrating that collective agreement provides more reliable quality signals than individual  
 1479 model confidence for mathematical reasoning tasks.

## J SELF-CONSISTENCY AND ACCURACY CORRELATION ANALYSIS

1483 We analyze the correlation between self-consistency and accuracy improvements across three exper-  
 1484 imental conditions to assess the robustness of our findings under different token generation limits  
 1485 and quantization settings. We test three conditions: (1) Token Capped + Quantized (256 tokens,  
 1486 4-bit), our standard experimental setup matching computational constraints; (2) Token Uncapped +  
 1487 Non-Quantized (2048 tokens with no observed truncation, full precision), representing maximum  
 1488 generation quality without computational constraints; and (3) Token Uncapped + Quantized (2048  
 1489 tokens with no observed truncation, 4-bit), a balanced approach removing token truncation while  
 1490 maintaining efficiency.

1491 Table 16: Self-consistency vs accuracy correlation across experimental conditions. “Capped” refers  
 1492 to 256 token limit, which is used throughout the work due to computational constraints in multi-  
 1493 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

1508 All conditions show strong positive correlations ( $r > 0.86$ ) between self-consistency and accuracy,  
 1509 validating self-consistency as a robust proxy for model performance across computational settings.  
 1510 Interestingly, capped conditions show slightly higher correlations (mean  $r = 0.975$ ) compared to  
 1511 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

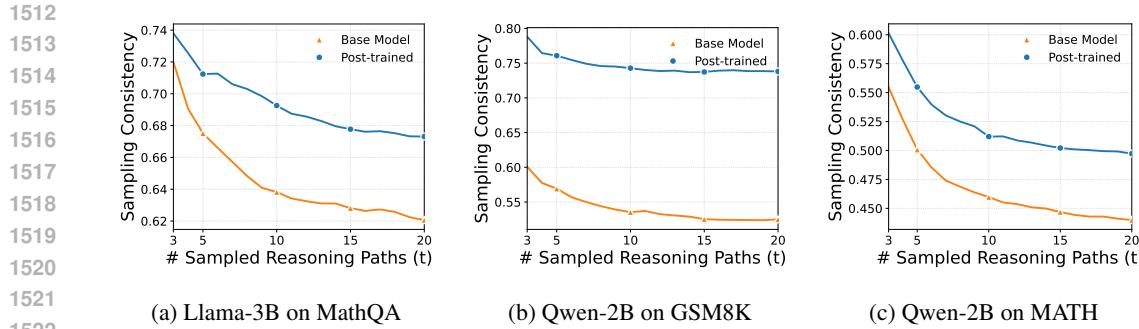


Figure 5: Self-consistency improvements persist without token constraints. Models trained with 256-token debates still show gains when tested with full-length responses, though with reduced effect sizes due to the weaker training signal compared to testing conditions. Colors: **Blue**: post-trained model, **Orange**: base model.

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.

### J.1 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.

## K IMPACT OF MACA ON GENERAL REASONING

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

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

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.

1566 Table 17: Within-distribution performance comparison: Debate-derived majority vote supervision  
 1567 vs. ground truth supervision on training domains. Models are trained on MATH, GSM8K, and  
 1568 MathQA datasets. Bold indicates the better method for each model-dataset pair, demonstrating that  
 1569 debate-derived supervision achieves comparable performance to ground truth labels.

1570

1571 Model	1572 Method	1573 MATH		1574 GSM8K		1575 MathQA	
		1576 Debate-MV	1577 GT	1578 Debate-MV	1579 GT	1580 Debate-MV	1581 GT
1574 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
1578 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

1582

1583 Table 18: Cross-domain generalization: Debate-derived majority vote supervision vs. ground truth  
 1584 supervision on unseen reasoning tasks. Models trained on mathematical datasets (MATH, GSM8K,  
 1585 MathQA) generalize effectively to diverse reasoning domains. Bold indicates the better method  
 1586 for each model-dataset pair, showing that both supervision approaches transfer well to out-of-  
 1587 distribution tasks.

1588

1589 Model	1590 Method	1591 SVAMP		1592 GPQA		1593 CSQA	
		1594 Debate-MV	1595 GT	1596 Debate-MV	1597 GT	1598 Debate-MV	1599 GT
1592 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
1596 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

1601

## 1602 L POST-TRAINING METHOD IMPACT ON LOG-PROBABILITY DISTRIBUTION

1603

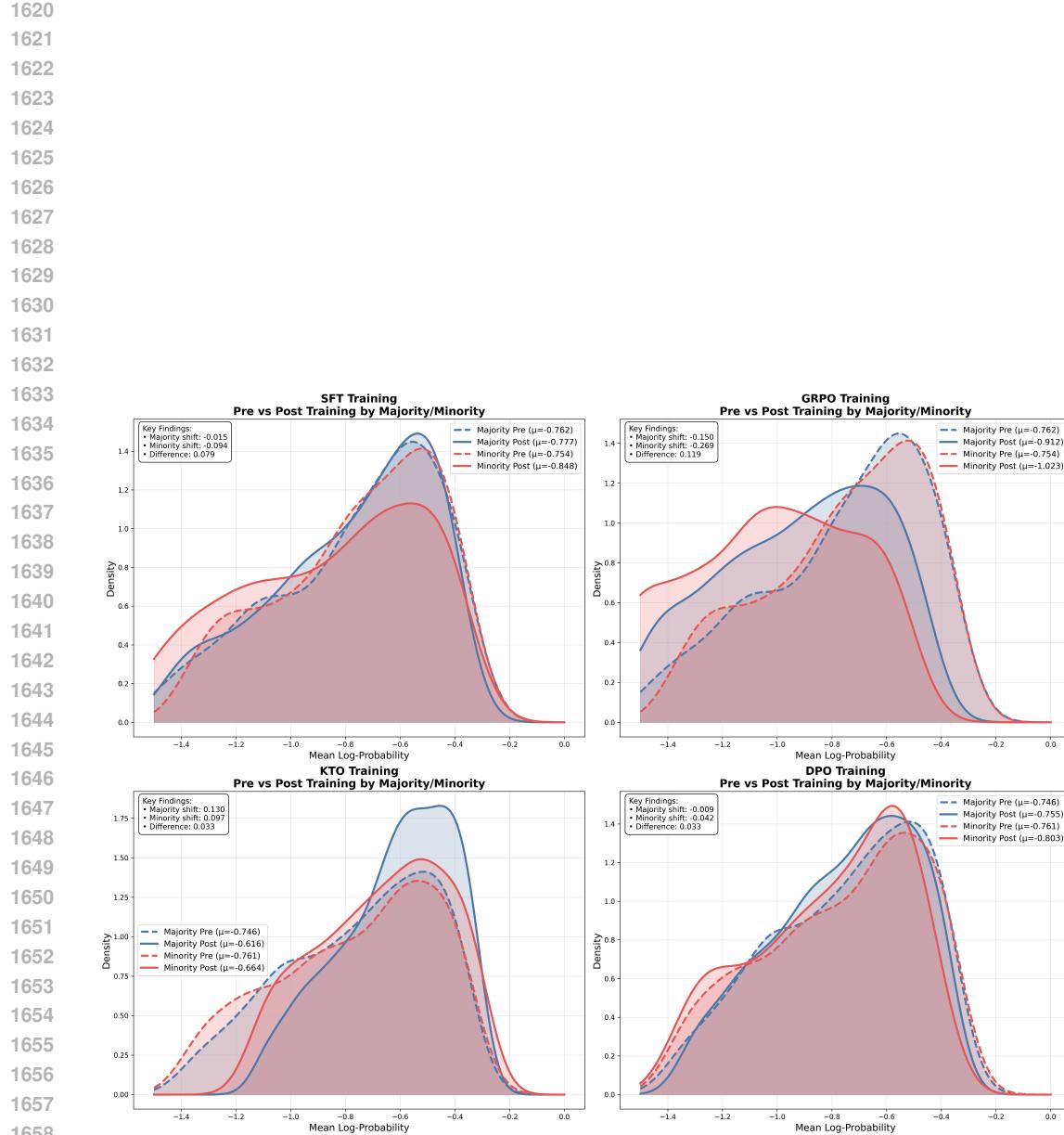
1604 Figure 6 shows how different post-training methods affected log-probability distributions for Qwen-  
 1605 2B on MATH. Preference-based methods (DPO, KTO) increased the density of majority distribu-  
 1606 tions, with KTO showing particularly pronounced effects: higher peaks and tighter tails indicating  
 1607 more concentrated probability mass around consensus responses. SFT left the majority distribu-  
 1608 tion largely unchanged while substantially reducing minority distribution density. GRPO created  
 1609 the most dramatic separation between distributions, shifting both leftward (lower log-probabilities  
 1610 overall) but with the minority distribution shifting much more substantially than the majority. While  
 1611 these different patterns (probability concentration for preference methods versus selective penal-  
 1612 ization for GRPO) all corresponded with performance improvements, further research is needed to  
 1613 better understand the relationship between these specific distributional changes and the impact on  
 1614 consistency.

1615

## 1616 M PREFERENCE LEARNING AS AN IMPLICIT FORMAT REWARD

1617

1618 This section investigates how preference learning through consensus signals acts as an implicit for-  
 1619 mat reward, teaching models to produce more efficient and accurate reasoning without explicit for-  
 1620 matting supervision. We examine three key aspects: the baseline capability requirements for ef-



1646  
 1647  
 1648  
 1649  
 1650  
 1651  
 1652  
 1653  
 1654  
 1655  
 1656  
 1657  
 1658  
 1659  
 1660  
 1661  
 1662  
 1663  
 1664  
 1665  
 1666  
 1667  
 1668  
 1669  
 1670  
 1671  
 1672  
 1673

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

1674  
 1675  
 1676  
 1677  
 1678  
 1679  
 1680  
 1681  
 1682  
 1683  
 1684  
 1685  
 1686  
 1687  
 1688  
 1689  
 1690  
 1691  
 1692  
 1693  
 1694  
 1695  
 1696  
 1697  
 1698  
 1699  
 1700  
 1701  
 1702  
 1703  
 1704  
 1705  
 1706  
 1707  
 1708  
 1709  
 1710  
 1711  
 1712  
 1713  
 1714  
 1715  
 1716  
 1717  
 1718  
 1719  
 1720  
 1721  
 1722  
 1723  
 1724  
 1725  
 1726  
 1727  
 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  
 1782  
 1783  
 1784  
 1785  
 1786  
 1787  
 1788  
 1789  
 1790  
 1791  
 1792  
 1793  
 1794  
 1795  
 1796  
 1797  
 1798  
 1799  
 1800  
 1801  
 1802  
 1803  
 1804  
 1805  
 1806  
 1807  
 1808  
 1809  
 1810  
 1811  
 1812  
 1813  
 1814  
 1815  
 1816  
 1817  
 1818  
 1819  
 1820  
 1821  
 1822  
 1823  
 1824  
 1825  
 1826  
 1827  
 1828  
 1829  
 1830  
 1831  
 1832  
 1833  
 1834  
 1835  
 1836  
 1837  
 1838  
 1839  
 1840  
 1841  
 1842  
 1843  
 1844  
 1845  
 1846  
 1847  
 1848  
 1849  
 1850  
 1851  
 1852  
 1853  
 1854  
 1855  
 1856  
 1857  
 1858  
 1859  
 1860  
 1861  
 1862  
 1863  
 1864  
 1865  
 1866  
 1867  
 1868  
 1869  
 1870  
 1871  
 1872  
 1873  
 1874  
 1875  
 1876  
 1877  
 1878  
 1879  
 1880  
 1881  
 1882  
 1883  
 1884  
 1885  
 1886  
 1887  
 1888  
 1889  
 1890  
 1891  
 1892  
 1893  
 1894  
 1895  
 1896  
 1897  
 1898  
 1899  
 1900  
 1901  
 1902  
 1903  
 1904  
 1905  
 1906  
 1907  
 1908  
 1909  
 1910  
 1911  
 1912  
 1913  
 1914  
 1915  
 1916  
 1917  
 1918  
 1919  
 1920  
 1921  
 1922  
 1923  
 1924  
 1925  
 1926  
 1927  
 1928  
 1929  
 1930  
 1931  
 1932  
 1933  
 1934  
 1935  
 1936  
 1937  
 1938  
 1939  
 1940  
 1941  
 1942  
 1943  
 1944  
 1945  
 1946  
 1947  
 1948  
 1949  
 1950  
 1951  
 1952  
 1953  
 1954  
 1955  
 1956  
 1957  
 1958  
 1959  
 1960  
 1961  
 1962  
 1963  
 1964  
 1965  
 1966  
 1967  
 1968  
 1969  
 1970  
 1971  
 1972  
 1973  
 1974  
 1975  
 1976  
 1977  
 1978  
 1979  
 1980  
 1981  
 1982  
 1983  
 1984  
 1985  
 1986  
 1987  
 1988  
 1989  
 1990  
 1991  
 1992  
 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  
 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  
 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  
 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  
 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  
 2268  
 2269  
 2270  
 2271  
 2272  
 2273  
 2274  
 2275  
 2276  
 2277  
 2278  
 2279  
 2280  
 2281  
 2282  
 2283  
 2284  
 2285  
 2286  
 2287  
 2288  
 2289  
 2290  
 2291  
 2292  
 2293  
 2294  
 2295  
 2296  
 2297  
 2298  
 2299  
 2300  
 2301  
 2302  
 2303  
 2304  
 2305  
 2306  
 2307  
 2308  
 2309  
 2310  
 2311  
 2312  
 2313  
 2314  
 2315  
 2316  
 2317  
 2318  
 2319  
 2320  
 2321  
 2322  
 2323  
 2324  
 2325  
 2326  
 2327  
 2328  
 2329  
 2330  
 2331  
 2332  
 2333  
 2334  
 2335  
 2336  
 2337  
 2338  
 2339  
 2340  
 2341  
 2342  
 2343  
 2344  
 2345  
 2346  
 2347  
 2348  
 2349  
 2350  
 2351  
 2352  
 2353  
 2354  
 2355  
 2356  
 2357  
 2358  
 2359  
 2360  
 2361  
 2362  
 2363  
 2364  
 2365  
 2366  
 2367  
 2368  
 2369  
 2370  
 2371  
 2372  
 2373  
 2374  
 2375  
 2376  
 2377  
 2378  
 2379  
 2380  
 2381  
 2382  
 2383  
 2384  
 2385  
 2386  
 2387  
 2388  
 2389  
 2390  
 2391  
 2392  
 2393  
 2394  
 2395  
 2396  
 2397  
 2398  
 2399  
 2400  
 2401  
 2402  
 2403  
 2404  
 2405  
 2406  
 2407  
 2408  
 2409  
 2410  
 2411  
 2412  
 2413  
 2414  
 2415  
 2416  
 2417  
 2418  
 2419  
 2420  
 2421  
 2422  
 2423  
 2424  
 2425  
 2426  
 2427  
 2428  
 2429  
 2430  
 2431  
 2432  
 2433  
 2434  
 2435  
 2436  
 2437  
 2438  
 2439  
 2440  
 2441  
 2442  
 2443  
 2444  
 2445  
 2446  
 2447  
 2448  
 2449  
 2450  
 2451  
 2452  
 2453  
 2454  
 2455  
 2456  
 2457  
 2458  
 2459  
 2460  
 2461  
 2462  
 2463  
 2464  
 2465  
 2466  
 2467  
 2468  
 2469  
 2470  
 2471  
 2472  
 2473  
 2474  
 2475  
 2476  
 2477  
 2478  
 2479  
 2480  
 2481  
 2482  
 2483  
 2484  
 2485  
 2486  
 2487  
 2488  
 2489  
 2490  
 2491  
 2492  
 2493  
 2494  
 2495  
 2496  
 2497  
 2498  
 2499  
 2500  
 2501  
 2502  
 2503  
 2504  
 2505  
 2506  
 2507  
 2508  
 2509  
 2510  
 2511  
 2512  
 2513  
 2514  
 2515  
 2516  
 2517  
 2518  
 2519  
 2520  
 2521  
 2522  
 2523  
 2524  
 2525  
 2526  
 2527  
 2528  
 2529  
 2530  
 2531  
 2532  
 2533  
 2534  
 2535  
 2536  
 2537  
 2538  
 2539  
 2540  
 2541  
 2542  
 2543  
 2544  
 2545  
 2546  
 2547  
 2548  
 2549  
 2550  
 2551  
 2552  
 2553  
 2554  
 2555  
 2556  
 2557  
 2558  
 2559  
 2560  
 2561  
 2562  
 2563  
 2564  
 2565  
 2566  
 2567  
 2568  
 2569  
 2570  
 2571  
 2572  
 2573  
 2574  
 2575  
 2576  
 2577  
 2578  
 2579  
 2580  
 2581  
 2582  
 2583  
 2584  
 2585  
 2586  
 2587  
 2588  
 2589  
 2590  
 2591  
 2592  
 2593  
 2594  
 2595  
 2596  
 2597  
 2598  
 2599  
 2600  
 2601  
 2602  
 2603  
 2604  
 2605  
 2606  
 2607  
 2608  
 2609  
 2610  
 2611  
 2612  
 2613  
 2614  
 2615  
 2616  
 2617  
 2618  
 2619  
 2620  
 2621  
 2622  
 2623  
 2624  
 2625  
 2626  
 2627  
 2628  
 2629  
 2630  
 2631  
 2632  
 2633  
 2634  
 2635  
 2636  
 2637  
 2638  
 2639  
 2640  
 2641  
 2642  
 2643  
 2644  
 2645  
 2646  
 2647  
 2648  
 2649  
 2650  
 2651  
 2652  
 2653  
 2654  
 2655  
 2656  
 2657  
 2658  
 2659  
 2660  
 2661  
 2662  
 2663  
 2664  
 2665  
 2666  
 2667  
 2668  
 2669  
 2670  
 2671  
 2672  
 2673  
 2674  
 2675  
 2676  
 2677  
 2678  
 2679  
 2680  
 2681  
 2682  
 2683  
 2684  
 2685  
 2686  
 2687  
 2688  
 2689  
 2690  
 2691  
 2692  
 2693  
 2694  
 2695  
 2696  
 2697  
 2698  
 2699  
 2700  
 2701  
 2702  
 2703  
 2704  
 2705  
 2706  
 2707  
 2708  
 2709  
 2710  
 2711  
 2712  
 2713  
 2714  
 2715  
 2716  
 2717  
 2718  
 2719  
 2720  
 2721  
 2722  
 2723  
 2724  
 2725  
 2726  
 2727  
 2728  
 2729  
 2730  
 2731  
 2732  
 2733  
 2734  
 2735  
 2736  
 2737  
 2738  
 2739  
 2740  
 2741  
 2742  
 2743  
 2744  
 2745  
 2746  
 2747  
 2748  
 2749  
 2750  
 2751  
 2752  
 2753  
 2754  
 2755  
 2756  
 2757  
 2758  
 2759  
 2760  
 2761  
 2762  
 2763  
 2764  
 2765  
 2766  
 2767  
 2768  
 2769  
 2770  
 2771  
 2772  
 2773  
 2774  
 2775  
 2776  
 2777  
 2778  
 2779  
 2780  
 2781  
 2782  
 2783  
 2784  
 2785  
 2786  
 2787  
 2788  
 2789  
 2790  
 2791  
 2792  
 2793  
 2794  
 2795  
 2796  
 2797  
 2798  
 2799  
 2800  
 2801  
 2802  
 2803  
 2804  
 2805  
 2806  
 2807  
 2808  
 2809  
 2810  
 2811  
 2812  
 2813  
 2814  
 2815  
 2816  
 2817  
 2818  
 2819  
 2820  
 2821  
 2822  
 2823  
 2824  
 2825  
 2826  
 2827  
 2828  
 2829  
 2830  
 2831  
 2832  
 2833  
 2834  
 2835  
 2836  
 2837  
 2838  
 2839  
 2840  
 2841  
 2842  
 2843  
 2844  
 2845  
 2846  
 2847  
 2848  
 2849  
 2850  
 2851  
 2852  
 2853  
 2854  
 2855  
 2856  
 2857  
 2858  
 2859  
 2860  
 2861  
 2862  
 2863  
 2864  
 2865  
 2866  
 2867  
 2868  
 2869  
 2870  
 2871  
 2872  
 2873  
 2874  
 2875  
 2876  
 2877  
 2878  
 2879  
 2880  
 2881  
 2882  
 2883  
 2884  
 2885  
 2886  
 2887  
 2888  
 2889  
 2890  
 2891  
 2892  
 2893  
 2894  
 2895  
 2896  
 2897  
 2898  
 2899  
 2900  
 2901  
 2902  
 2903  
 2904  
 2905  
 2906  
 2907  
 2908  
 2909  
 2910  
 2911  
 2912  
 2913  
 2914  
 2915  
 2916  
 2917  
 2918  
 2919  
 2920  
 2921  
 2922  
 2923  
 2924  
 2925  
 2926  
 2927  
 2928  
 2929  
 2930  
 2931  
 2932  
 2933  
 2934  
 2935  
 2936  
 2937  
 2938  
 2939  
 2940  
 2941  
 2942  
 2943  
 2944  
 2945  
 2946  
 2947  
 2948  
 2949  
 2950  
 2951  
 2952  
 2953  
 2954  
 2955  
 2956  
 2957  
 2958  
 2959  
 2960  
 2961  
 2962  
 2963  
 2964  
 2965  
 2966  
 2967  
 2968  
 2969  
 2970  
 2971  
 2972  
 2973  
 2974  
 2975  
 2976  
 2977  
 2978  
 2979  
 2980  
 2981  
 2982  
 2983  
 2984  
 2985  
 2986  
 2987  
 2988  
 2989  
 2990  
 2991  
 2992  
 2993  
 2994  
 2995  
 2996  
 2997  
 2998  
 2999  
 3000  
 3001  
 3002  
 3003  
 3004  
 3005  
 3006  
 3007  
 3008  
 3009  
 3010  
 3011  
 3012  
 3013  
 3014  
 3015  
 3016  
 3017  
 3018  
 3019  
 3020  
 3021  
 3022  
 3023  
 3024  
 3025  
 3026  
 3027  
 3028  
 3029  
 3030  
 3031  
 3032  
 3033  
 3034  
 3035  
 3036  
 3037  
 3038  
 3039  
 3040  
 3041  
 3042  
 3043  
 3044  
 3045  
 3046  
 3047  
 3048  
 3049  
 3050  
 3051  
 3052  
 3053  
 3054  
 3055  
 3056  
 3057  
 3058  
 3059  
 3060  
 3061  
 3062  
 3063  
 3064  
 3065  
 3066  
 3067  
 3068  
 3069  
 3070  
 3071  
 3072  
 3073  
 3074  
 3075  
 3076  
 3077  
 3078  
 3079  
 3080  
 3081  
 3082  
 3083  
 3084  
 3085  
 3086  
 3087  
 3088  
 3

1728  
1729 Table 20: Improvements from post-training with `max_new_tokens` = 256 translate when tested with  
1730 larger token counts (512). “All” represents a model post-trained on Math, MathQA, and GSM8K;  
1731 its row reports improvement deltas over the base model performance.

Model	Trained on	Tested on ( <code>max_new_tokens</code> = 512)					
		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

1740  
1741 Table 21: Decomposition of performance improvements into completion gains versus reasoning  
1742 gains. “From Completion” represents accuracy gained if all truncated base model responses had  
1743 been allowed to complete. “From Better Reasoning” represents the remaining improvement at-  
1744 tributable to fundamental problem-solving enhancement. Percentages show the relative contribution  
1745 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%)

#### 1760 M.4 FORMATTING IMPROVEMENTS AND REMAINING REASONING GAINS

1761  
1762 To further isolate reasoning improvements from formatting effects, we analyze how post-training  
1763 affects answer formatting compliance. Table 22 shows the percentage of responses with correct  
1764 reasoning but incorrect formatting, and calculates reasoning gains that persist beyond all formatting  
1765 improvements. Post-training dramatically reduces formatting losses (e.g., from 38.8% to 0.6% on  
1766 CSQA for Qwen2B). However, the “Reasoning Gain Beyond Formatting” column reveals that sub-  
1767 stantial improvements remain even after perfect formatting is assumed. For instance, Qwen2B’s 59.6  
1768 percentage point improvement on CSQA includes 38.2 points from better formatting; the remaining  
1769 21.4 points represent accuracy improvements distinct from formatting.

#### 1770 M.5 PARSER IMPLEMENTATION DETAILS

1771  
1772 To quantify formatting improvements versus reasoning gains, we employed two parsers with differ-  
1773 ent strictness levels. Our standard parser requires answers in `\boxed{}` format and returns `None`  
1774 for any deviation. The relaxed parser, used for impact analysis in Table 22, accepts common natural  
1775 language patterns like “The answer is A” or “Answer: 42” by searching the entire response for valid  
1776 answer formats.

1777  
1778 This relaxed parser first attempts the strict extraction, then falls back to regex patterns that capture  
1779 answers expressed naturally in text. For multiple choice, it accepts patterns like “[Tt]he answer is  
1780 ([A-E])” or “Answer: ([A-E])”. For numerical answers, it extracts from patterns like “= 42” at line  
1781 endings or “The final answer is 42”. When multiple patterns match, it takes the last occurrence,  
1782 mimicking how humans identify the final answer in a reasoning chain.

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 (%) Before Training	Accuracy Lost to Format Errors (%) After Training	Reasoning Gain 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

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.

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

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

1836 Table 23: Single-agent: RL with debate outperforms RL with single-round majority vote. Bench-  
 1837 mark against TTRL (single-round R0 MV-GRPO), ScPO (single-round R0 MV-DPO), MACA (de-  
 1838 bate final round R1 MV-DPO with debate context after collaborative exchange). Accuracy averaged  
 1839 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
R0 MV-GRPO	<b>27.4 ± 2.9</b>	45.5 ± 5.7	45.9 ± 2.1	43.5 ± 0.2
R0 MV-DPO	11.7 ± 1.1	43.6 ± 0.2	39.9 ± 0.5	39.7 ± 0.6
R1 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
R0 MV-GRPO	<b>50.9 ± 1.6</b>	50.5 ± 8.3	81.2 ± 1.6	54.3 ± 5.7
R0 MV-DPO	42.3 ± 3.4	74.4 ± 0.8	82.3 ± 0.5	80.7 ± 1.2
R1 MV-DPO	50.2 ± 0.9	<b>76.5 ± 0.3</b>	<b>83.7 ± 0.3</b>	<b>86.0 ± 0.2</b>

## O POST-TRAINING SELF-CONSISTENCY IMPROVEMENTS TRANSLATE FROM 4-BIT QUANTIZED MODEL TO FULL MODEL

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

## P IMPACT OF POST-TRAINING ON DEBATE AGREEMENT RATES

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

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

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

## Q PERFORMANCE COMPARISON OF RL ON DEBATE VS. SINGLE-ROUND MAJORITY VOTE RL IN SINGLE-AGENT SETTING

Table 23 reports the full results referenced in Table 6, including standard deviation over 3 seeds. We find that MACA with final round majority vote with debate context outperforms single-round majority-vote RL with DPO (akin to ScPO (Prasad et al., 2024)) in all 8 configurations tested, and RL with GRPO (akin to TTRL (Zuo et al., 2025)) in 6 out of 8 configurations tested, with the other 2 configs being within the standard deviation.

## R EFFICIENCY

Here we detail the compute performance tradeoffs of MACA’s MV-DPO with debate context against other single-agent benchmark methods. ScPO requires 0.075-0.58 GPU hours achieving -0.6%

1890 to +13.6% gains, TTTRL requires 2.2-7.7 GPU hours achieving -26.4% to +18% changes, while  
1891 MACA (MV-DPO with debate) requires 0.73-1.68 GPU hours achieving consistent positive gains.  
1892 Our efficiency plots demonstrate that MV-DPO on debate provides the best performance-efficiency  
1893 tradeoff—higher performance than ScPO/TTTRL with comparable computational cost to ScPO and  
1894 significantly lower cost than TTTRL.

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

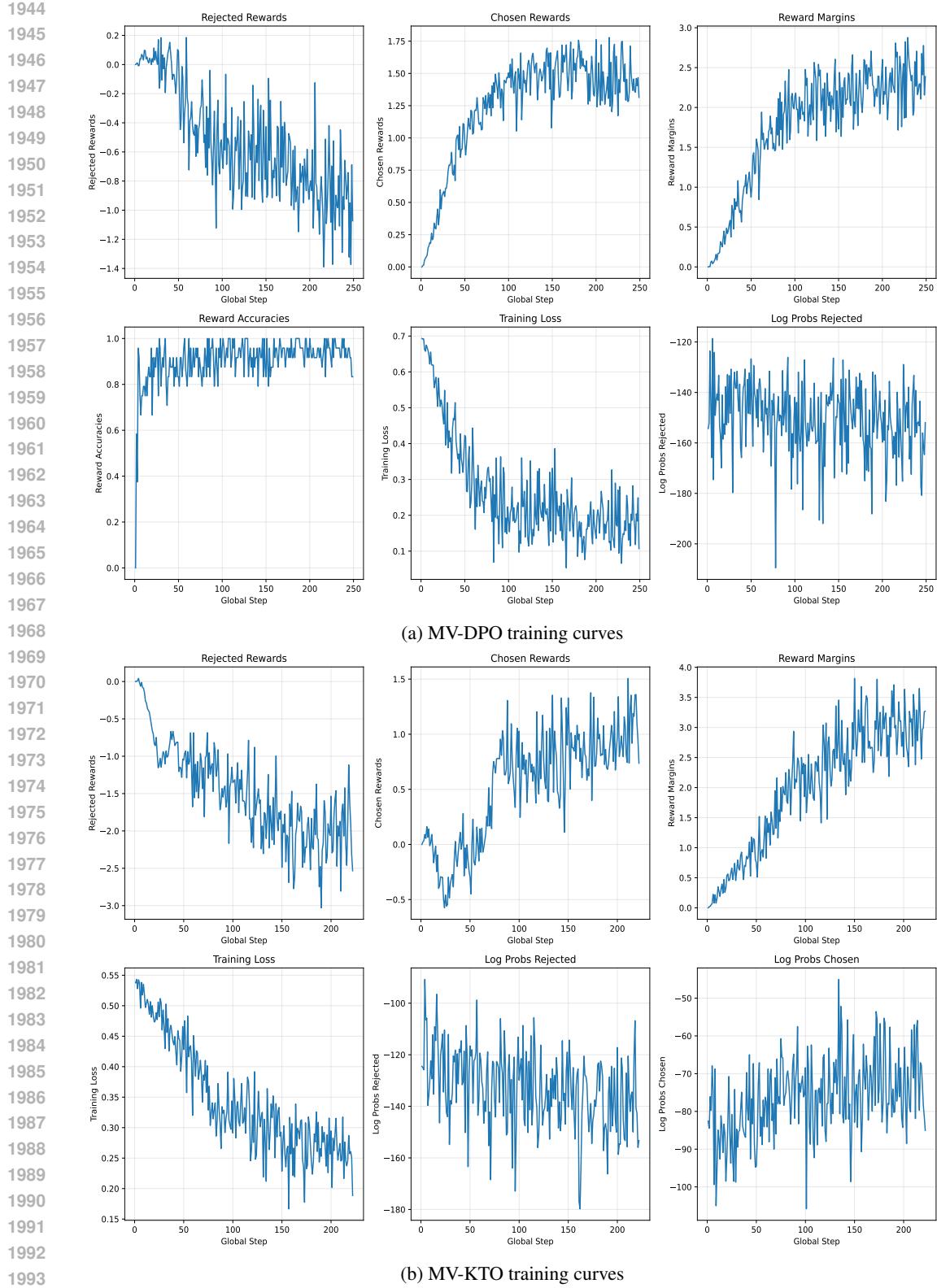
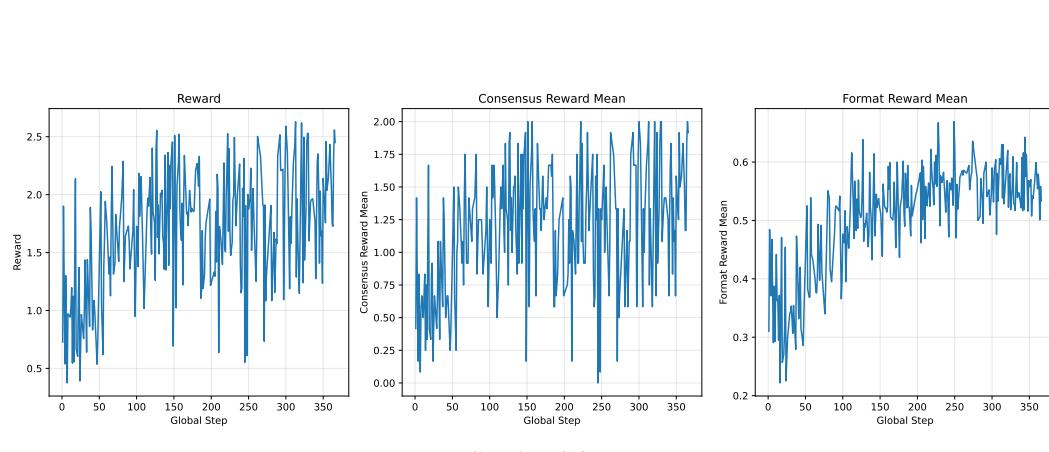
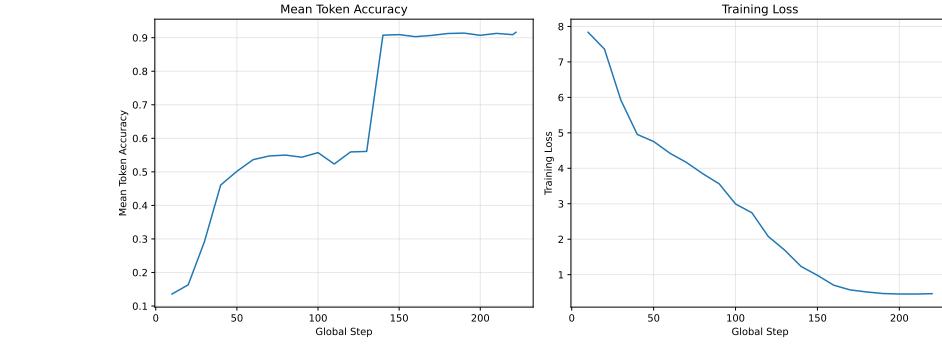


Figure 7: Training curves for MV-DPO and MV-KTO. (a) MV-DPO: increasing reward margin between majority (chosen) and minority (rejected) responses, with declining rejected log probabilities. (b) MV-KTO: similar reward margin increase using unpaired examples, with rejected log probabilities decreasing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

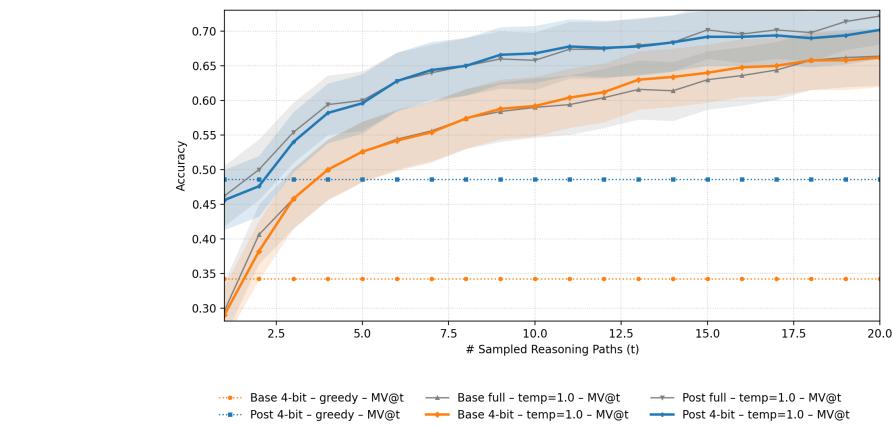


(a) MV-GRPO training curves



(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.

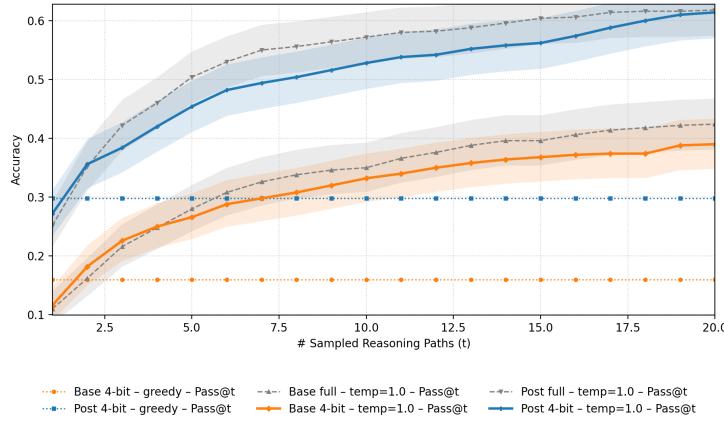


Figure 10: Qwen-2B on MATH (Pass@t, t=20). **Blue**: post-trained 4-bit model, **Orange**: base 4-bit model, **Grey**: full-precision model.

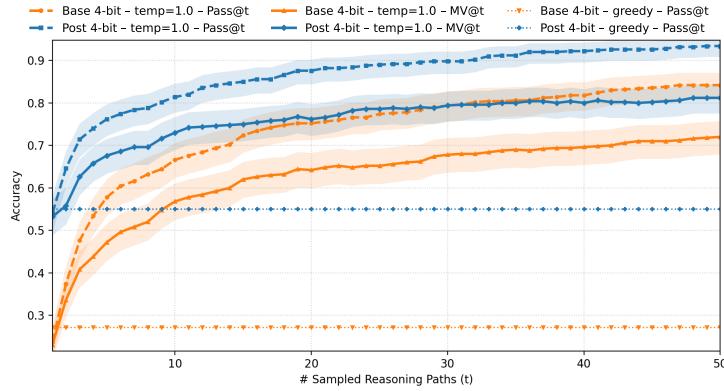


Figure 11: Qwen-2B on GSM8K (Pass@t, MV@t; t=50). **Blue**: post-trained 4-bit model, **Orange**: base 4-bit model.

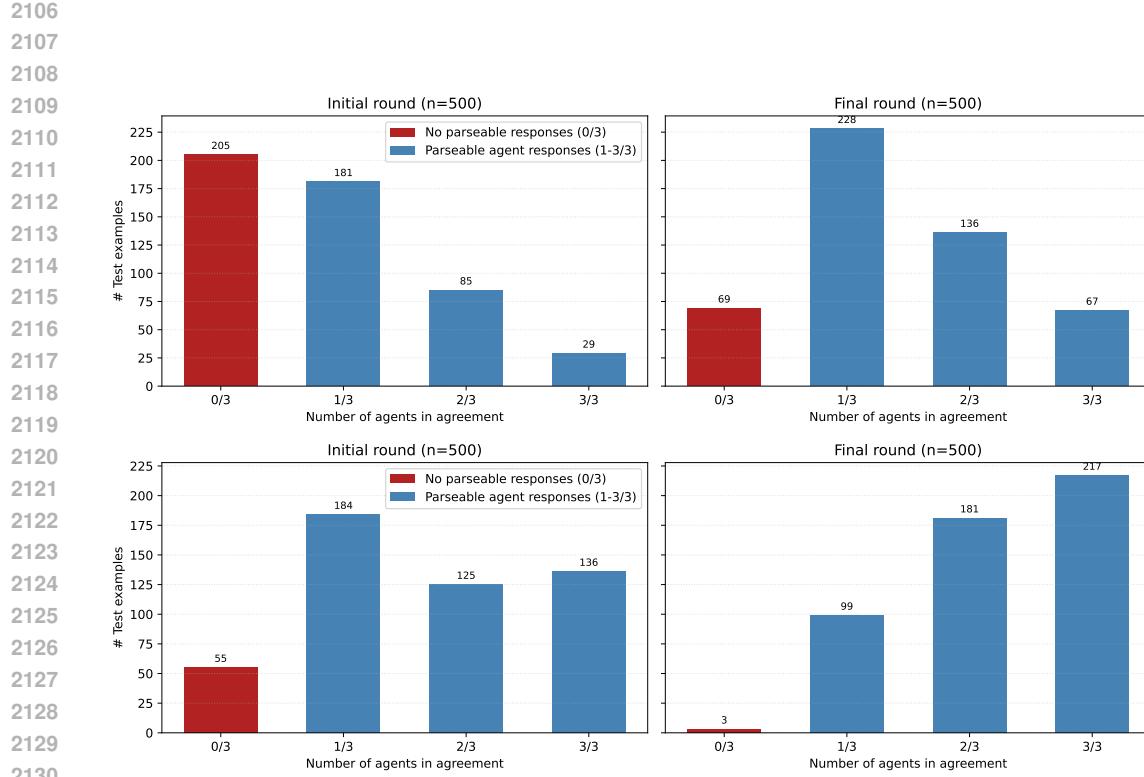


Figure 12: MACA drives meaningful improvements in both answer completeness and agent agreement, verifying the probability mass reallocation to the consensus set of reasoning trajectories (Ex: Qwen2B on GSM8K). Top: base model debate; bottom: post-trained model debate; left: initial debate round; right: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

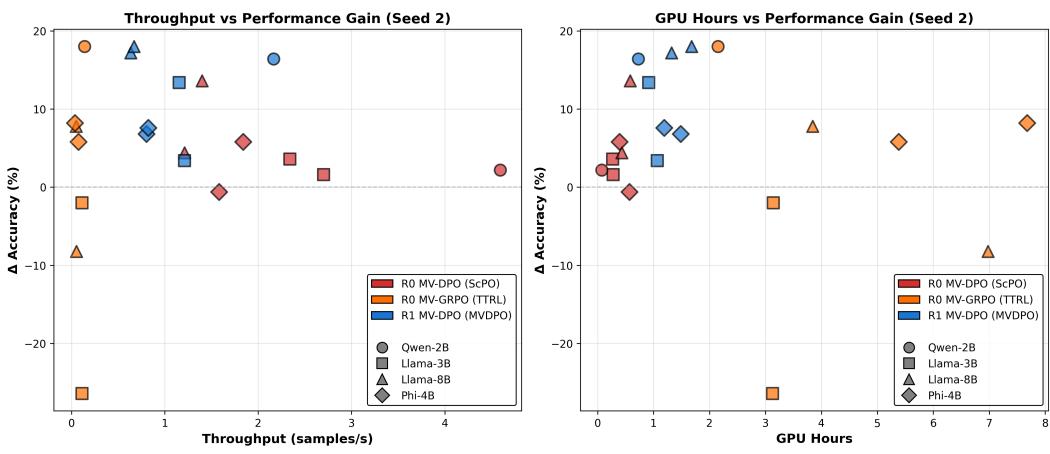


Figure 13: GPU throughput and training runtime compared with performance lift for single-agent post-training methods