Table 7) provided evidence that additional turns do not necessarily improve the performance significantly. Similarly, our results show that off-the-shelf LLMs' performance may not benefit from additional turns. In contrast, LLMs trained using **MAPoRL** exhibit improved performance as the number of collaboration turns increased, as shown in Figure 2.

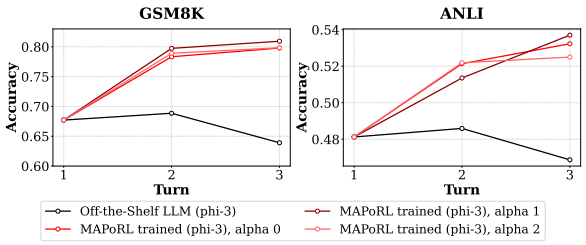

Figure 2: Performance comparison of different LLMs across tasks (left: GSM8k, right: ANLI) under various settings. We evaluate collaboration ability in five conditions: (1) off-the-shelf LLMs collaborating and (2) models trained using **MAPoRL** collaborating (with all incentive parameters (Section 4.4) $\alpha, \beta = 0, 1, 2$, respectively).

Remark 2 (Domain-Specific Knowledge Acquisition vs. Collaboration Ability Improvement)**.** One might question whether the performance gains observed in **MAPoRL**-trained models stem from acquiring domain-specific knowledge rather than improved collaboration ability. To address this, we compare off-the-shelf LLMs and **MAPoRL**-trained models by testing how well they perform on questions without any collaboration, providing **MAPoRL**-trained models only the original question – without interaction history – to check if their performance is solely due to domain knowledge learned during training. The results are as follows:

	Phi-3	MAPoRL T2	MAPoRL T3
GSM8k	0.609	0.604	0.611
ANLI	0.451	0.458	0.453

Here, we provide the same questions to the off-the-shelf Phi-3 model, the **MAPoRL**-trained turn-2 model, and the **MAPoRL**-trained turn-3 model. The similar performance across these models suggests that **MAPoRL** training did not enhance task-specific knowledge but rather improved the models' ability to collaborate effectively.

We also provide the changes in the fraction of responses that transition their correctness over multiple turns of **MAPoRL**. The fraction of Incorrect → Incorrect responses decreased, and the fraction of Correct → Incorrect responses also decreased, indicating that **MAPoRL** enhanced effective collaboration.

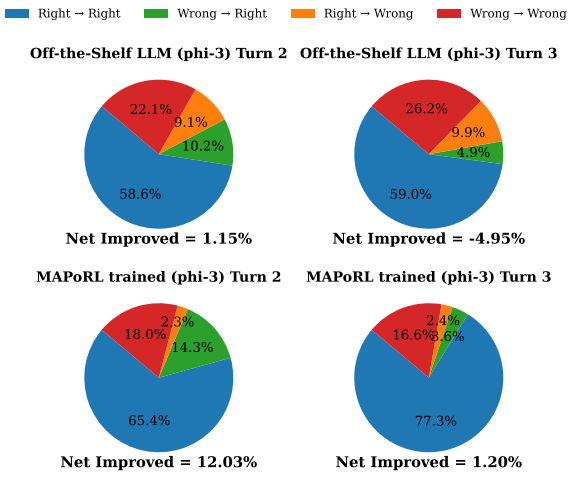

Figure 3: Changes in the fraction of responses that transition their correctness over multiple turns of **MAPoRL** on GSM8k.

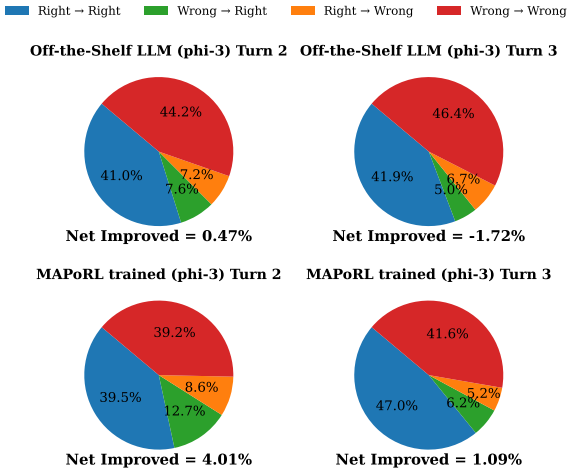

Figure 4: Changes in the fraction of responses that transition their correctness over multiple turns of **MAPoRL** on ANLI.

4.4 Experiment 2: Reward Shaping with Collaboration Incentives

In addition to the multi-agent independent PPO framework, we then investigate the auxiliary incentive mechanism designed to enhance collaborative interactions. To analyze the impact of the incentive parameters (α and β, Section 3.3), we simplify our experimental setup by limiting the total number of debate turns to 2 and analyze the following

cases. Here, α_0 and α_1 correspond to incentives for an agent's own revision, capturing *critical reasoning* (extracting useful information from incorrect answers) and *persuadability* (accepting correct information), respectively. Meanwhile, β_0 and β_1 correspond to incentives for influencing others, where β_0 encourages providing incorrect but useful responses, and β_1 reflects an agent's ability to *persuade others* with correct answers.

To analyze the impact of the incentive parameters (α and β, Section 3.3), we simplify our experimental setup by limiting the total number of debate turns to 2 and analyze the following cases. Here, α_0 and α_1 correspond to the incentives related to an agent's own revision of the answer, while β_0 and β_1 correspond to the incentives related to the agent's influence on other agents' answers.

(α_0, α_1)	RWR	RWW	WRW	WRR	Δ_0	Δ_1
(0, 0)	0.0529	0.0563	0.1244	0.2286	0.1757	0.0661
(0, 2)	0.0270	0.0521	0.1259	0.2194	0.1924	0.0738
(2, 0)	0.0500	0.0563	0.1241	0.2272	0.1772	0.0678

Table 3: Analysis of answer revision patterns under different α parameters. The columns RWR through WRR show the proportion of each transition type, where the three letters indicate Answer(t), Answer(t+1), and Majority(t) respectively. R and W stand for right and wrong answer. Δ_0 measures the difference in transitions from wrong to right answers when the majority is wrong (WRW − RWW) which is related to α_0, while Δ_1 measures transitions when the majority is right (WRR − RWR) which is related to α_1.

Analysis of α_0 and α_1. We compare baseline $(\alpha_0, \alpha_1) = (0, 0)$ against two configurations: $(0, 2)$ and $(2, 0)$. When α_1 was increased to 2, we observe a 9.5% improvement in Δ_1, indicating that incentivizing agents to follow correct majority opinions effectively improved performance. When α_0 was increased to 2, we observed a smaller (2.57%) improvement in Δ_0, suggesting that rewarding agents for deviating from incorrect majority opinions had a positive but limited effect.

(β_0, β_1)	RWR	RWW	WRW	WRR	Δ_0	Δ_1
(0, 0)	0.0070	0.0453	0.0226	0.0221	0.0151	-0.0227
(0, 2)	0.0686	0.0461	0.0231	0.0230	0.0161	-0.0230
(2, 0)	0.0011	0.0360	0.0161	0.0188	0.0177	-0.0199

Table 4: Analysis of majority opinion influence under different β parameters. Meaning of the column is the same as Table 3.

Analysis of β_0 and β_1. We compare baseline $(\beta_0, \beta_1) = (0, 0)$ against configurations $(0, 2)$ and $(2, 0)$. Increasing β_1 to 2 resulted in a slight decrease in Δ_1 (-1.32%), indicating that incentivizing agents based on their influence when correct did not improve outcomes. However, increasing β_0 to 2 lead to a substantial improvement in Δ_0 (17.2%), suggesting that rewarding agents for constructive influence even when wrong (providing useful incorrect answers that lead to better future responses) significantly enhanced collaborative performance.

For a total debating turns of 3, we also plot the collaboration performance using models trained with $\alpha_i = \beta_i = 0, 1, 2$ for $i = 1, 2$ on the GSM8K and ANLI tasks (Figure 2). The results showed some performance improvement, though the gain was relatively modest.

4.5 Experiment 3: Collaboration Ability Acquired by MAPoRL Is Transferable

Here, we investigate the transferability of collaboration abilities acquired through **MAPoRL** across different datasets not used during training. We evaluate LLMs trained with **MAPoRL** on one dataset when applied to tasks from other datasets. For instance, we assesse models trained on ANLI when solving tasks from GSM8k, along with other dataset combinations. The results, presented in Table 5, demonstrate that collaboration abilities learned through **MAPoRL** are indeed transferable across datasets. This suggests that the models acquire a *meta*-capability for effective collaboration, even when encountering novel, unseen tasks.

Training → Evaluation	Model	Turn 1	Turn 2	Turn 3
ANLI → GSM8K	Off-the-shelf	0.677	0.688	0.640
	Trained	0.677	0.712	0.720
GSM8K → ANLI	Off-the-shelf	0.482	0.486	0.468
	Trained	0.482	0.499	0.507

Table 5: Performance comparison (Accuracy) of 3-agent collaboration using *off-the-shelf* vs. *trained* LLMs. For each dataset pair (rows in bold), the first row shows the off-the-shelf performance and the second row shows the trained model performance, across Turns 1–3.

These findings demonstrate that models trained through **MAPoRL** on one task can effectively generalize their collaborative capabilities to different, unrelated tasks. This generalization ability suggests that **MAPoRL** develops fundamental collaborative skills that transcend specific task domains.

4.6 Experiment 4: MAPoRL with Heterogeneous LLMs Can Help

In this experiment, we investigate collaborative learning between different foundation models, specifically examining co-training between (Phi3 3.4B and Qwen2.5 3B) and (Phi3 3.4B and Llama3-8B) pairs. In single-model evaluations, both Phi3 and Qwen2.5 3B demonstrate stronger performance compared to Llama3-8B. Due to GPU memory constraints necessitating simultaneous loading of two base models, we conduct experiments in a two-agent, two-turn environment. This setup enables us to explore whether models with heterogeneous capabilities could effectively collaborate to enhance the overall performance (Figure 7). The synergistic effects are particularly evident when models with different strengths worked together, suggesting that diverse model partnerships can yield better outcomes than individual model performance alone when we have **MAPoRL**.

4.7 Experiment 5: Naïve Supervised Fine-Tuning Using High-Quality Collaboration Samples May Not Induce Collaborative Behaviors

In this experiment, we investigate whether models could learn collaborative behavior through SFT on high-quality debate trajectories. We generated 12,800 trajectories using the multi-agent system (Figure 1) with off-the-shelf LLMs to match the training sample size used in **MAPoRL** for GSM8K. To provide favorable conditions for SFT, we allow a maximum of 600 tokens per response, which exceeded the token limit used in our **MAPoRL** experiments. We selected the top 10% of trajectories using the following criteria: 1) excluding trajectories without well-formatted answers, 2) filtering out trajectories where the final majority voting result was incorrect, and 3) selecting 1,280 trajectories based on the verifier's score of the final answer, which evaluates both correctness and reasoning quality. Interestingly, the results indicate that SFT not only failed to enhance collaborative behaviors, but also led to a decline in performance compared to the off-the-shelf model. Specifically, for turn-2, accuracy dropped to 0.578 ($\Delta = -0.111$), and for turn-3, it further decreased to 0.525 ($\Delta = -0.114$)[3]. This suggests that either substantially more training data would be required to learn effective collaborative behaviors, or that SFT might not be an effective approach for inducing such behaviors. Contemporaneously, Subramaniam et al. (2025) and Zhao et al. (2025) enhance multi-agent performance by incorporating new techniques into iterative SFT with their own data augmentation to generate effective collaboration examples, demonstrating its potential when combined with additional refinements. In contrast, our approach does not leverage data augmentation, but uses RL.

[3]Initially, these unexpected results led us to validate our findings through multiple experiments with varying temperatures for language generation. The consistent performance degradation across turns was observed in all the cases. This pattern suggests fundamental challenges in using SFT to maintain collaborative performance across multiple debate turns.

5 Concluding Remarks, Limitations, and Potential Risks

In this paper, we have introduced **MAPoRL**, a new post-training paradigm that leverages multi-agent RL to explicitly foster the collaboration among multiple LLMs. Unlike methods that rely solely on prompting or single-agent fine-tuning, **MAPoRL** focuses on *co-training* multiple LLMs, ensuring that each agent adapts its policy not just to immediate feedback, but also to the strategic behaviors of other agents over multiple interactive turns. By incorporating a verifier network for reward shaping with incentives, the framework guides each agent's responses that account for both short-term correctness and long-term collaborative potential, thus promoting collaborative discussions that lead to more accurate final answers.

Through an extensive set of experiments on reasoning-intensive tasks – such as GSM8K for mathematical problem-solving and ANLI for logical natural language inference – our results demonstrate that off-the-shelf LLMs often do not improve the overall performance with additional debate turns. In contrast, **MAPoRL**-trained agents show significant improvements with accuracy increasing as collaboration progresses. Crucially, these collaborative abilities are shown transferable across tasks, suggesting that once LLMs learn to collaborate, they can retain a generalizable "collaboration skill" applicable to different domains. Furthermore, our experiments with heterogeneous LLMs highlight that **MAPoRL** can also foster collaborative synergy even among models of varying capabilities.

Limitations

Since we use instruction prompts as inputs to the LLMs, the output can vary significantly depending