```
Solve the problem step by step and provide clear reasoning. Ensure that the reasoning is concise and directly relevant to solving the problem. Avoid adding commentary or unrelated content.

Present the final answer in the following format:

Answer: \\boxed{XX}"'}
```

ANLI.

```
{"role": "user", "content": f"'Premise: {sample["premise"]}

Hypothesis: {sample["hypothesis"]}


Please determine the relationship between the premise and the hypothesis. Choose one of the following: 'entailment,' 'neutral,' or 'contradiction.'
Start with concise reasoning for your choice and conclude with your final answer. You do not need to restate the premise and hypothesis. Present the final answer in the following format:

Answer: \\boxed{XX}}"'}
```

G.5 Post Turn 1 Prompt

```
{"role": "user", "content": f"' Question: {sample["question"]}

Solve the problem step by step and provide clear reasoning. Ensure that the reasoning is concise and directly relevant to solving the problem. Avoid adding commentary or unrelated content.

Present the final answer in the following format:

Answer: \\boxed{XX}"'}

{"role": "assistant", "contents": f"{agent_answer_for_turn_1}}

{"role": "user", "contents" f"Reward from a verifier of your answer: {score_value:.3f} out of 1.0, which means {feedback}" }

{"role": "user", "content": f"'
Agent {agent_num} solution: {agent_response}
Agent {agent_num} reward: {agent_response}


Agent {agent_num} solution: {agent_response}
Agent {agent_num} reward: {agent_response}

  .
  .
  .
```

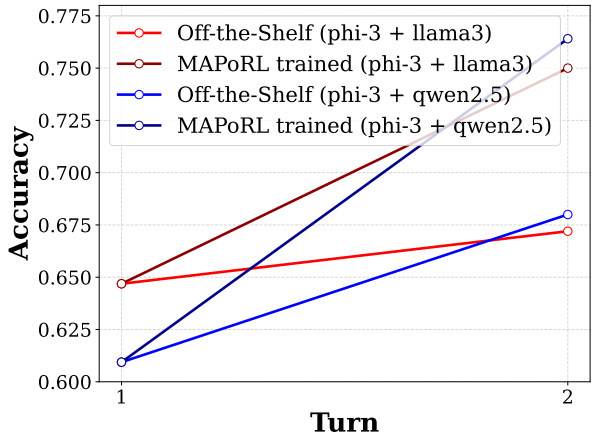

Figure 7: Performance comparison between off-the-shelf LLMs collaborations and **MAPoRL** trained LLM pairs. Off-the-shelf LLMs experiments were conducted with a 600-token limit, which is the double of the **MAPoRL** output token lengths.

Here, each reward represents the probability that a suggested answer is correct, as evaluated by a verifier. The reward value is between 0 and 1, with values closer to 1 indicating a higher likelihood of correctness. While these rewards offer useful context, they are not always perfect, though generally quite reliable.

Focus on providing a well–reasoned response that not only considers your own previous solution but also takes into account answers from other agents. If you believe your previous answer was incorrect, feel free to revise it. However, avoid repeating the same answer you or other agents have already provided. Also, internally think about the reward of your and other agents' answers. Ensure that your explanation clearly justifies your final answer. Please maintain your answer with very simple reasoning.

Once again, the question is: {question_for_input}'"}

.

.

.

(Stack these results by turn.)

G.6 Deferred Figure for Section 4.6

G.7 Ablation Study: Verifier Robustness

Since **MAPoRL** relies on a learned verifier to provide intermediate rewards, the robustness of this verifier is critical to the overall framework. To assess the verifier's influence, we conduct two ablation studies: (i) removing the verifier entirely, and (ii) varying the verifier's base model to evaluate the impact of architectural alignment.

Training Without Verifier Rewards. We first evaluate **MAPoRL** in a binary reward-only setting, where the reward at each episode is 1 if the final answer is correct, and 0 otherwise. No signal is provided for intermediate turns. Table 7 presents the results on GSM8K. Even in the absence of verifier-based shaping, **MAPoRL** shows improved performance over discussion turns, indicating that the collaborative training objective itself drives nontrivial gains.

Table 7: Performance with and without verifier rewards on GSM8K. Even without verifier shaping, multi-turn training yields improved outcomes.

Model	Turn 0	Turn 1	Turn 2
Off-the-shelf LLMs	0.677	0.689	0.639
MAPoRL(with verifier)	0.677	0.797	0.809
MAPoRL(w/o verifier)	0.677	0.734	0.746

While the absence of a verifier leads to somewhat lower performance, the continued improvement over turns suggests that multi-agent co-adaptation remains beneficial, even under sparse supervision.

Verifier-Model Architectural Alignment. We next examine the effect of mismatched architectures between the generation model and the verifier. Specifically, we train verifiers using different base models (e.g., Qwen, LLaMA, Gemma) and pair them with generators based on alternative architectures. We observe two consistent effects: (1) reward signals degrade when verifier and generation models are based on different families, and (2) during reinforcement learning, the generation model tends to stylistically drift toward the verifier's base model, often resulting in reduced accuracy. These findings emphasize the importance of architectural alignment between generator and verifier to ensure reward signal fidelity and prevent unintended distribution shifts.

Taken together, these studies empirically validate our design decision to co-train the verifier and generator on the same model base, providing stable and meaningful reward supervision during multi-turn training.

G.8 Ablation Study: Comparison to Single-Agent RL with Verifier

A natural question arises: why not apply RL to a single agent using the same verifier for supervision? While this is a valid and important consideration, it overlooks the broader objective of **MAPoRL**, which is not merely to improve task accuracy, but to enable the *emergence of collaboration* through multi-agent learning.

To empirically address this, we conduct an ablation study where a single LLM is trained using RL with verifier-provided rewards. The setting mirrors **MAPoRL**'s single-turn supervision, but without multi-agent interactions. As shown in Table 8, the single-agent RL model achieves a final accuracy of 0.732—higher than the off-the-shelf baseline but notably below **MAPoRL**-trained agents at later turns.

Table 8: Comparison of **MAPoRL**vs. single-agent RL with verifier rewards. **MAPoRL** demonstrates superior performance through collaborative refinement.

Model	Turn 0	Turn 1	Turn 2
Off-the-shelf LLMs	0.677	0.689	0.639
Single-Agent RL	0.732 (single turn)		
MAPoRL(Ours)	0.677	0.797	0.809