Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning

Lingkai Zu, Xiyue Peng, Liyu Yang, ShanghaiTech University

SI 2512 Reinforcement Learning

Spring, 2025

Background

Improving the reasoning capabilities of language models to better assist in mathematical reasoning, code generation, and other real-world applications has become a prominent research focus. Although large language models have demonstrated remarkable performance in these tasks, the more successful cases are all large-scale models supported by substantial computational resources. Aligning smaller-scale models offers a promising approach to investigating language model alignment when computational resources are limited. This raises the question:

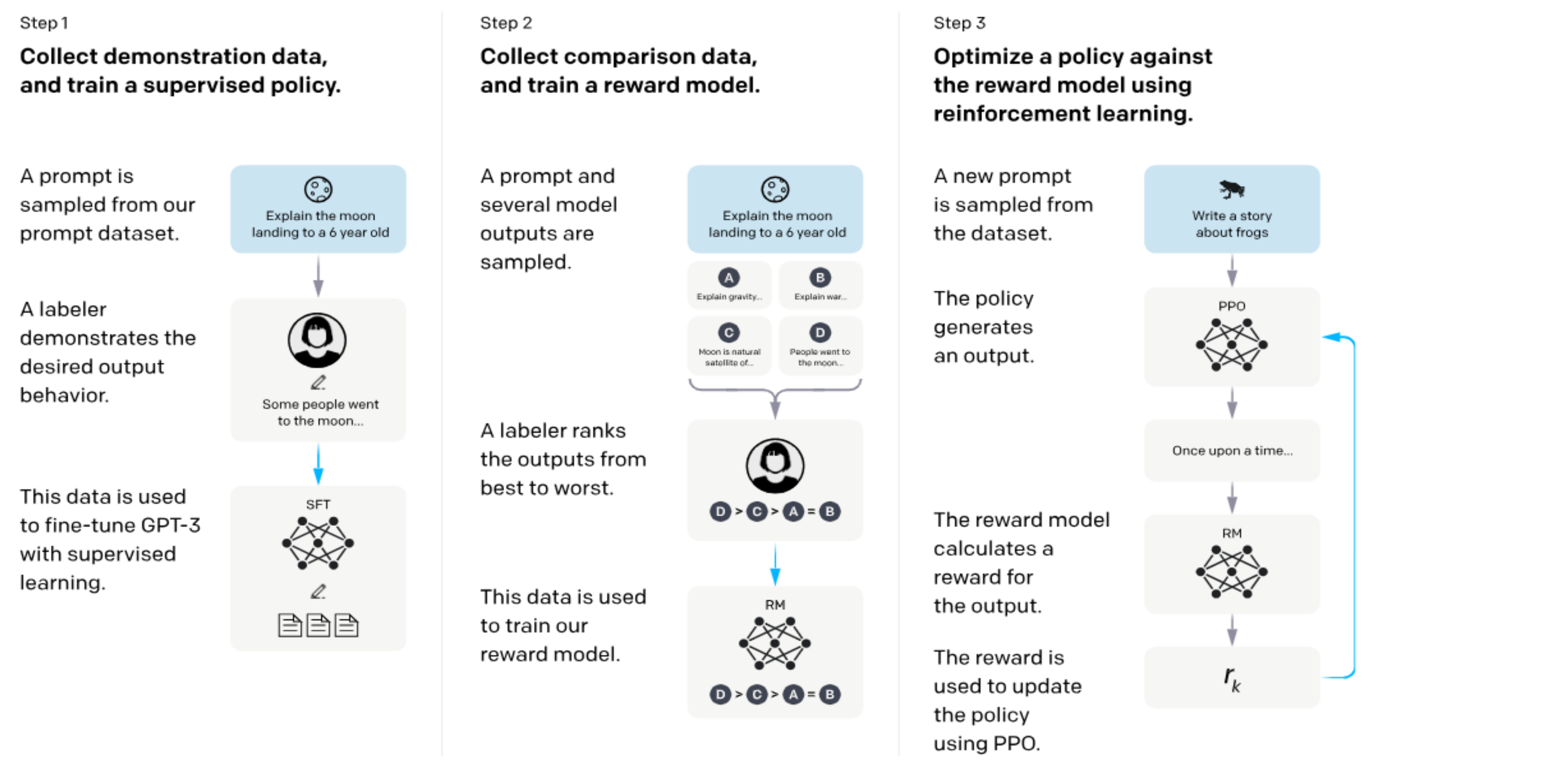
Can similar reasoning abilities emerge in smaller-scale models?

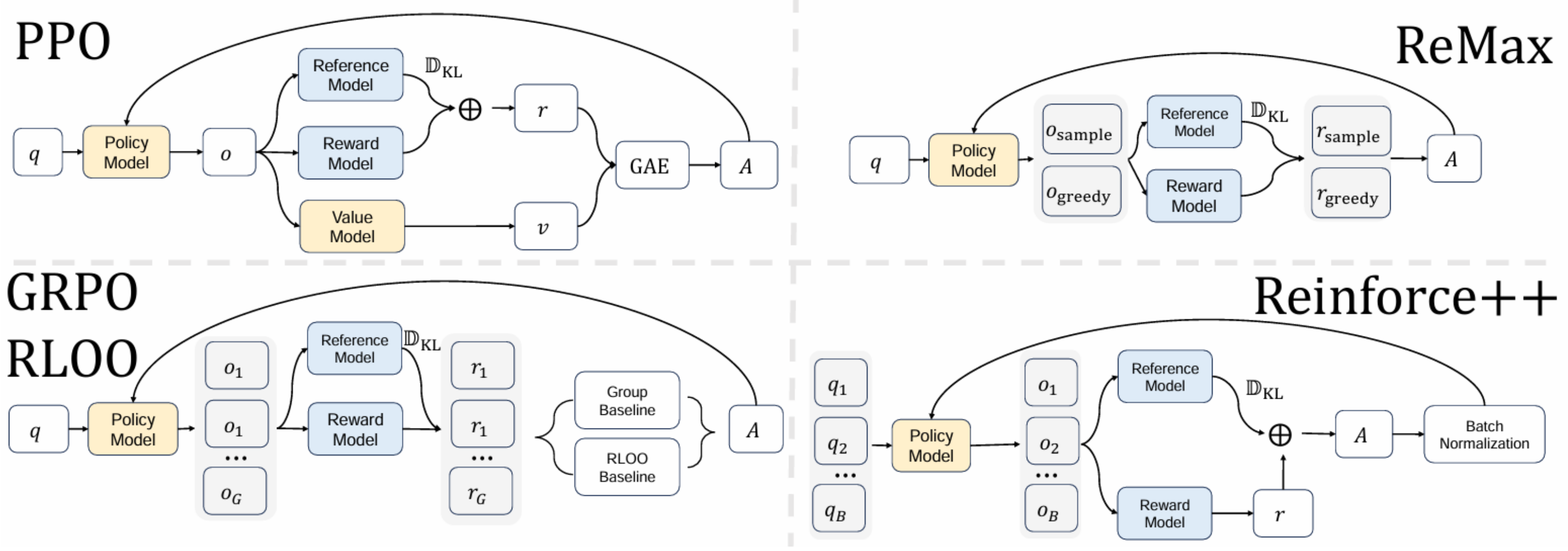
Synthetic Logic Puzzles Framework

An example of a Knights and Knaves (K&K) puzzle	
Problem	Knights always tell the truth, and knaves always lie. You meet 2 inhabitants: Zoey, and Oliver. Zoey: "*Oliver is not a knight*". Oliver: "*Oliver is a knight if and only if Zoey is a knave*". So who is a knight and who is a knave?
Solution	(1)Zoey is a knave; (2)Oliver is a knight.

Reward Type	Reward Value	Condition
s_{answer}	2	The answer fully matches the ground truth
	-1.5	The answer partially matches the ground truth
	-2	The answer cannot be parsed or missing
s_{format}	1	The format is correct
	-1	The format is incorrect

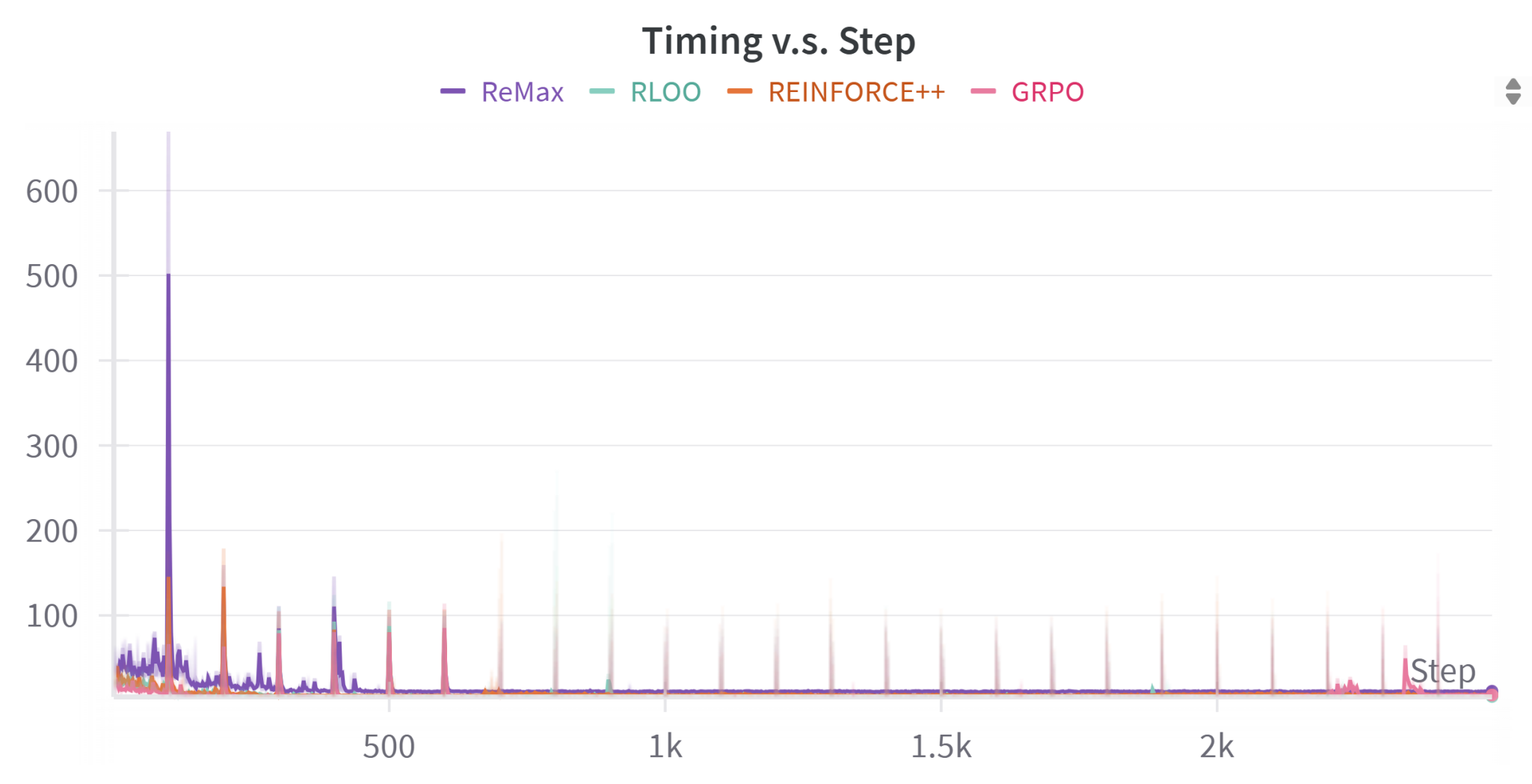
Post Training Method Overview

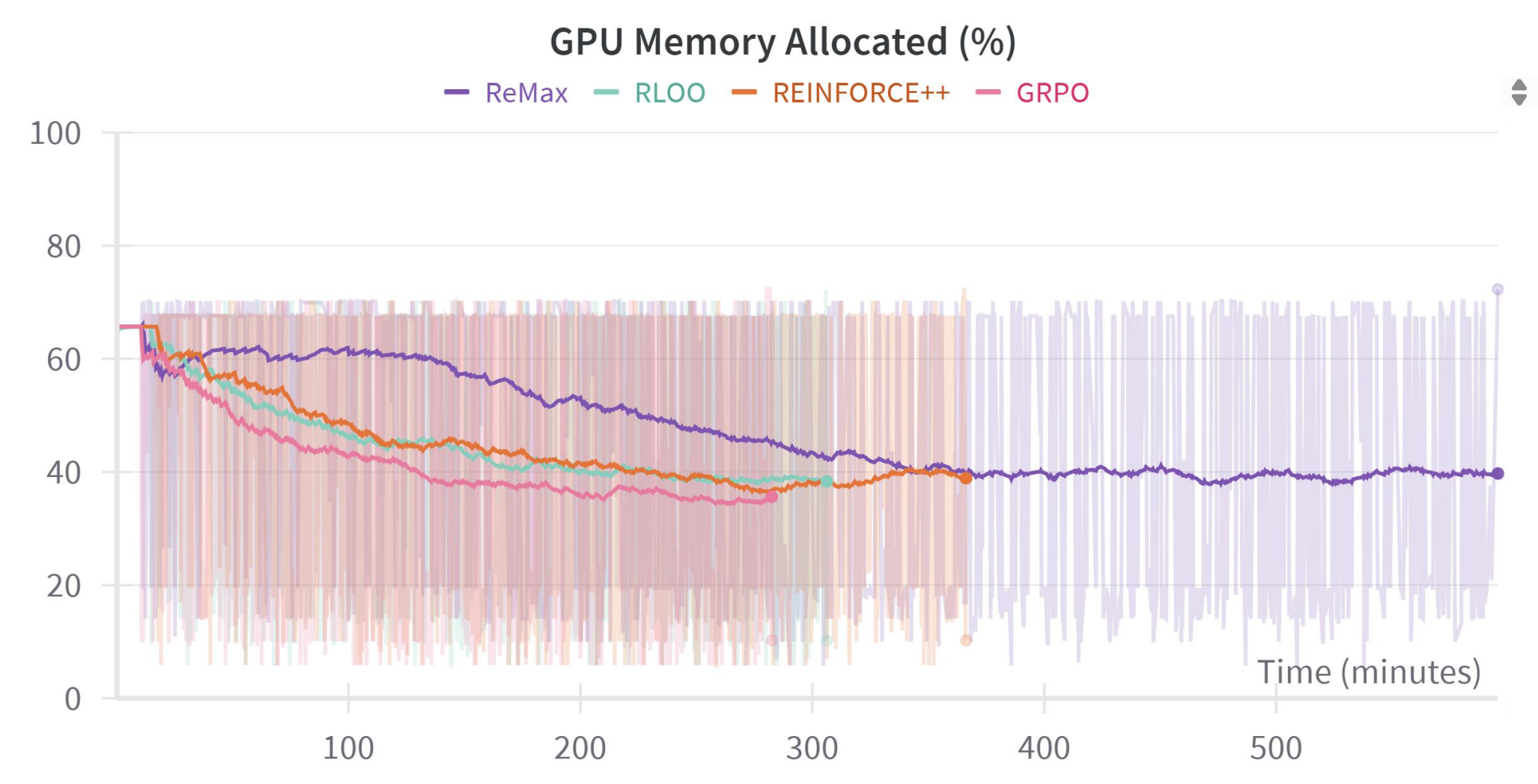

Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k



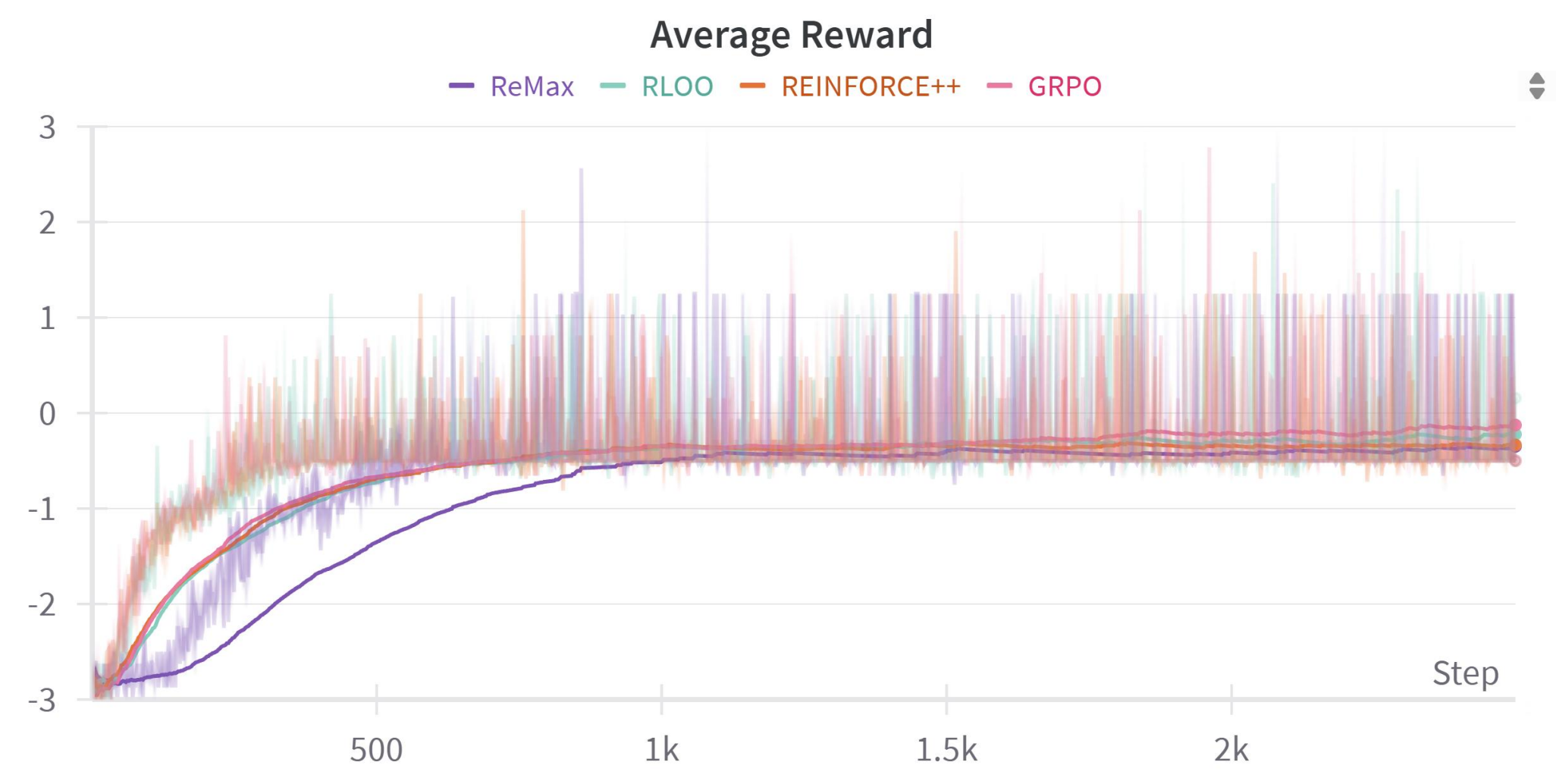
Experiments Results And Discussion

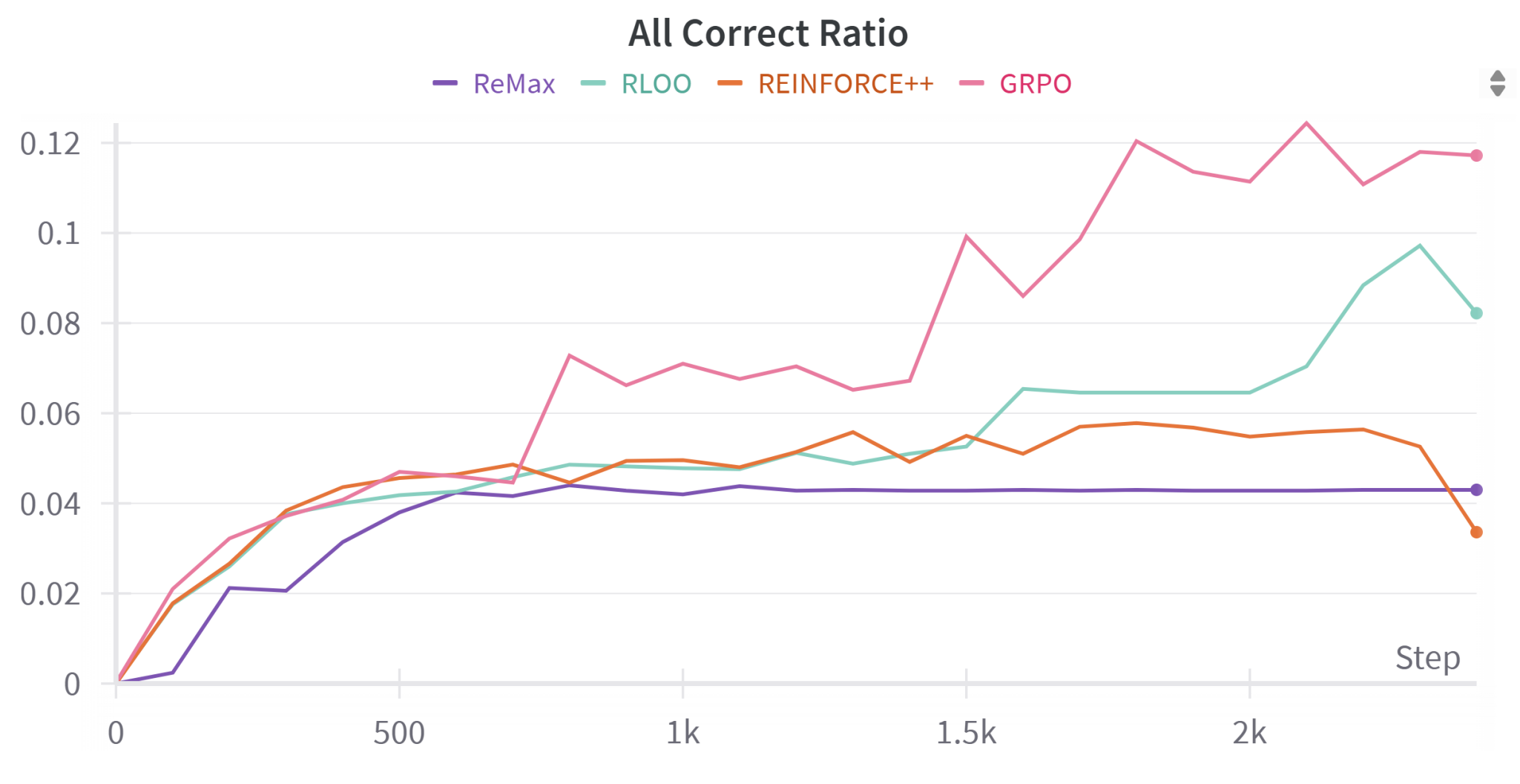
Efficient Consumption Resource



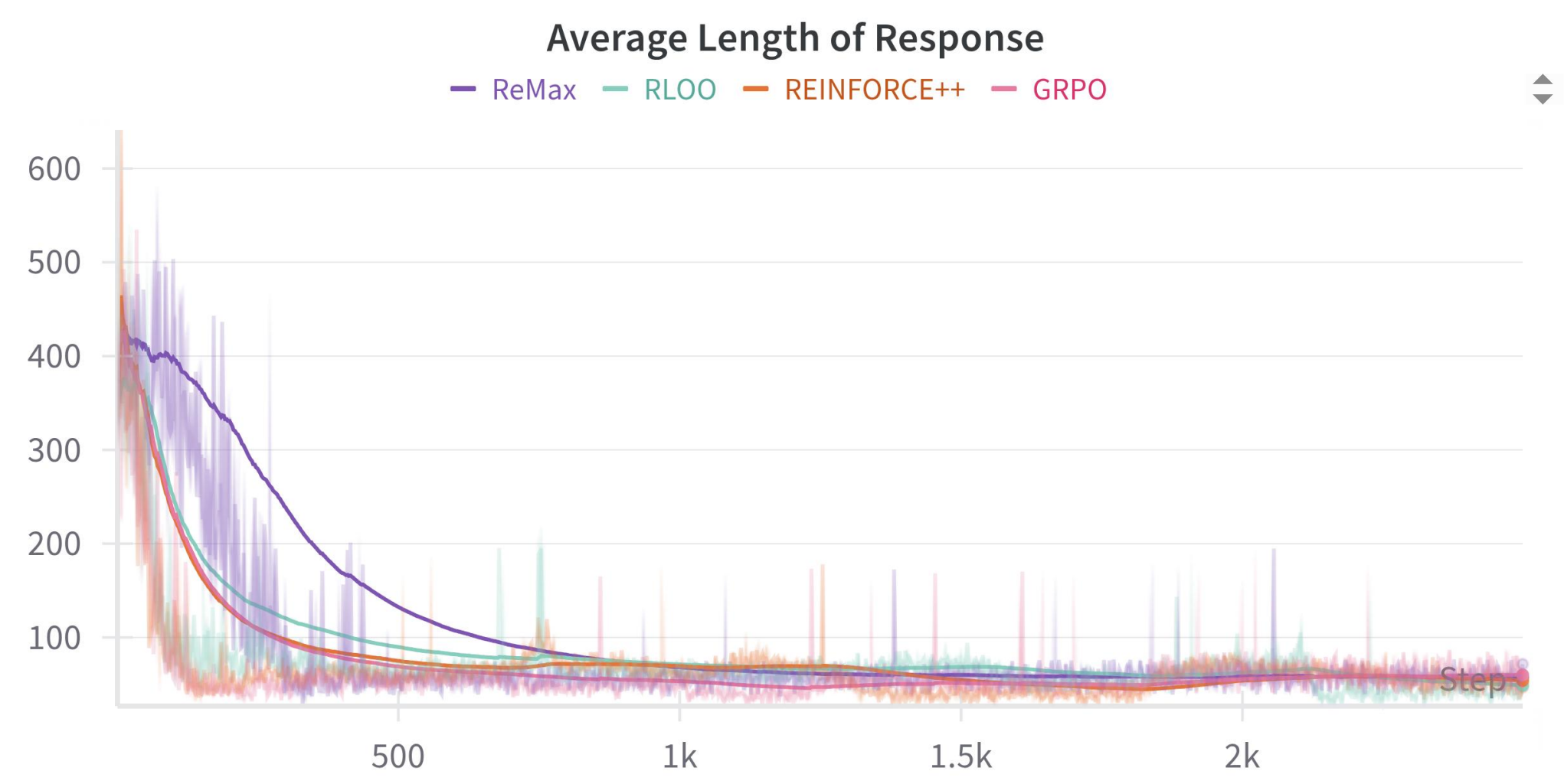


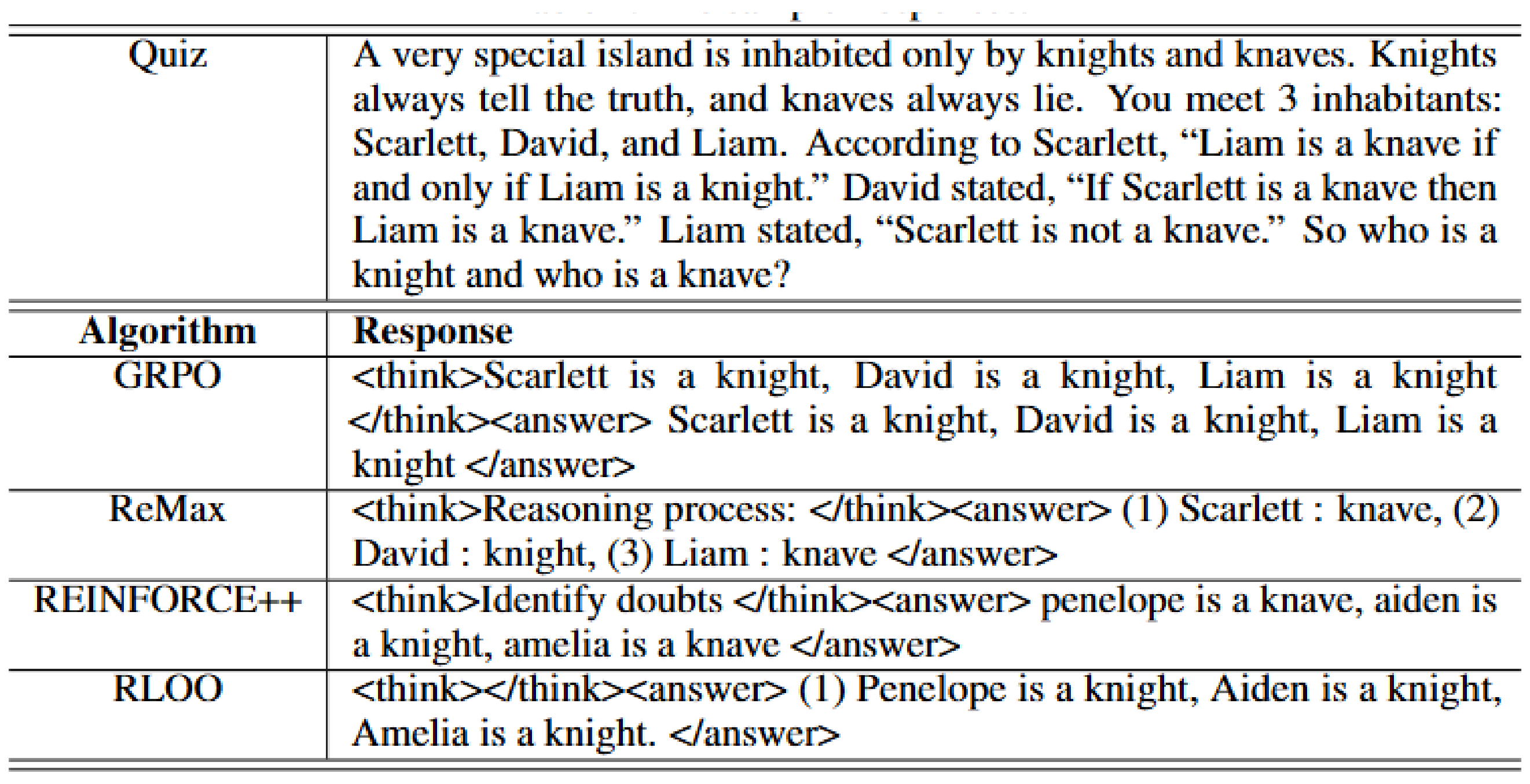
Poor Reasoning Capacity





Is Format Reward a Blessing or a Curse?





Quiz	A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 3 inhabitants: Scarlett, David, and Liam. According to Scarlett, "Liam is a knave if and only if Liam is a knight." David stated, "If Scarlett is a knave then Liam is a knave." Liam stated, "Scarlett is not a knave." So who is a knight and who is a knave?
Algorithm	**Response**
GRPO	<think>Scarlett is a knight, David is a knight, Liam is a knight </think><answer> Scarlett is a knight, David is a knight, Liam is a knight </answer>
ReMax	<think>Reasoning process: </think><answer> (1) Scarlett : knave, (2) David : knight, (3) Liam : knave </answer>
REINFORCE++	<think>Identify doubts </think><answer> penelope is a knave, aiden is a knight, amelia is a knave </answer>
RLOO	<think></think><answer> (1) Penelope is a knight, Aiden is a knight, Amelia is a knight. </answer>

- Whether small-scale models possess reasoning capabilities?
- What kind of rewards should we design to elicit reasoning from small-scale models?

identifying and rewarding certain thinking tokens