CPMöbius: Iterative Coach–Player Reasoning for Data-Free Reinforcement Learning

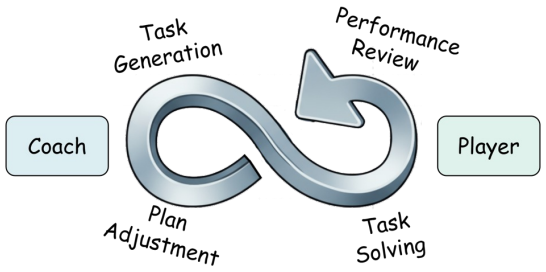
Ran Li*, **Zeyuan Liu***, Yinghao Chen, Bingxiang He, Jiarui Yuan, Zixuan Fu, Weize Chen, Jinyi Hu, Chen Qian, Zhiyuan Liu, Maosong Sun

Tsinghua University, University of Cambridge, Shanghai Jiao Tong University

ICML International Conference On Machine Learning

Motivation

- LLM reasoning performance heavily depends on:
  - supervised fine-tuning (SFT)
  - reinforcement learning with external rewards
- But these rely on:
  - human-curated datasets
  - expensive verifiers ⇒ leads to **supervision bottleneck**
  - limited scalability



Can LLM continuously improve their reasoning ability through self-interaction without external training data?

Full Paper

Wechat

Let's connet:

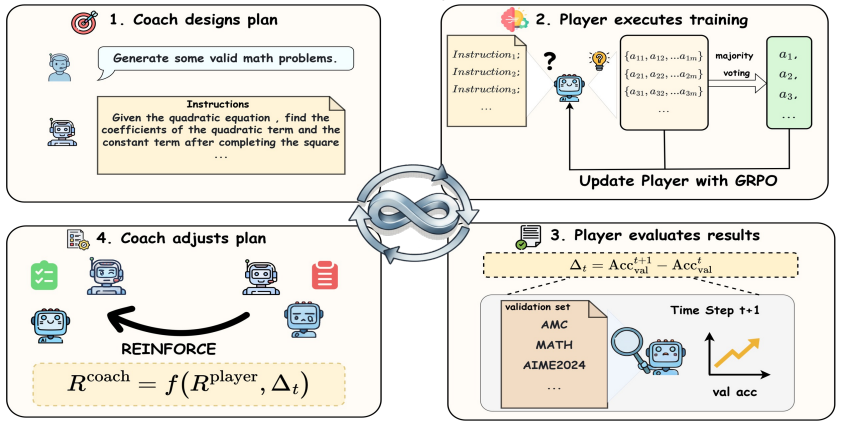
rl810@cam.ac.uk

in/li-ran-9b5801411/

CPMöbius

- **Coach designs plan**. The Coach generates a batch of m task instructions $\{x_i\}_{i=1}^{m} \sim \pi_{\theta_t}^{\mathrm{C}}(\cdot)$, where $\pi_{\theta_t}^{\mathrm{C}}$ is the current Coach policy.
- **Player executes training.** For every x_i the current Player produces n independent answers $\{y_{i,j}\}_{j=1}^{n} \sim \pi_{\phi_t}^{\mathrm{P}}(\cdot\,|\mathrm{x_i})$. Majority voting over the n answers yields a **pseudo-label** y_i^*. Then each answer receives a verifiable reward $r_{i,j} = \mathbb{I}[y_{i,j} = y_i^*]$ as well as a GRPO advantage $A_{i,j}$ computed w.r.t. the n samples for question i. The **instruction-level training reward** is obtained by averaging:

$$R_i^{\text{Player}} = \frac{1}{n}\sum_{j=1}^{n} r_{i,j}$$



- **Player evaluates results.** The updated Player receives **environment feedback**

$$\Delta_t = \text{Acc}\left(\pi_{\phi_{t+1}}^{\mathrm{P}}; \mathcal{D}_{val}\right) - \text{Acc}\left(\pi_{\phi_t}^{\mathrm{P}}; \mathcal{D}_{val}\right)$$

which measures the Player's accuracy difference after receiving environment feedback.

- **Coach adjusts plan.** Each instruction x_i is assigned an **instruction reward**

$$R_i^{\text{Coach}} = R_i^{\text{Player}} \cdot \Delta_t,$$

i.e., instructions that produced high Player rewards and coincided with a global accuracy improvement are reinforced. A group of m instruction-level **REINFORCE** steps update Coach parameters θ_t using each instance in the batch $\{(x_i, R_i^{\text{Coach}})\}_{i=1}^{m}$.

Results

- We evaluate the proposed method on standard mathematical reasoning benchmarks, reporting both overall performance and out-of-distribution (OOD) performance, where OOD excludes AMC use as validation for relevant baselines.
- We compare entropy minimization method **RENT-RL** and self-play method **R-zero** as baselines, under the same evaluation protocol, demonstrating consistent improvements in both overall and OOD settings.

Models	Average	OOD Average	AMC	AIME 2024	AIME 2025	Minerva	MATH	Olympiad
Qwen2.5-Math-1.5B								
Base Model	23.3	19.8	34.6	6.2	2.8	16.3	56.2	23.4
R-Zero (Iter 3)	27.1	24.7	39.2	9.8	5.0	19.3	62.4	26.8
RENT	27.1	24.7	39.3	**10.0**	5.0	19.0	62.2	**27.1**
CPMöbius	**28.8**	**26.8**	**39.4**	9.8	**5.4**	**28.0**	**63.1**	26.9
OpenMath-Nemotron-1.5B								
Base Model	59.5	54.9	82.3	**55.6**	43.3	**25.1**	89.4	61.0
R-Zero (Iter 3)	-	-	-	-	-	-	-	-
RENT	61.7	56.5	**87.7**	55.0	46.0	24.2	90.7	66.7
CPMöbius	**62.1**	**57.0**	87.5	54.9	**46.9**	24.3	**91.2**	**67.9**
OctoThinker-3B-Hybrid-Zero								
Base Model	21.3	20.6	24.6	3.9	1.7	16.3	57.9	23.4
R-Zero (Iter 3)	20.5	19.5	25.9	2.0	0.3	14.6	58.1	22.3
RENT	23.0	21.7	**29.2**	**7.3**	**2.1**	15.0	60.2	24.1
CPMöbius	**23.6**	**22.0**	28.0	4.8	1.7	**22.1**	**60.4**	**24.7**
Qwen2.5-Math-7B-Instruct								
Base Model	35.8	33.0	49.2	9.0	6.3	34.6	78.0	37.4
R-Zero (Iter 3)	36.9	34.2	50.5	9.5	7.4	32.7	83.3	38.1
RENT	39.2	37.6	53.1	10.8	**9.9**	38.8	83.8	**38.8**
CPMöbius	**40.7**	**38.4**	**55.6**	**11.8**	9.6	**44.9**	**84.2**	38.3

Insights

- The results suggest that meaningful reasoning improvements can emerge from interaction learning signals without relying on external supervision.
- Both optimization-based and self-play baselines are insufficient to fully capture the benefits of structured cooperative learning.
- This highlights the potential of co-evolutionary frameworks as a scalable alternative for improving LLM reasoning.