Reinforcement Learning via Self-Distillation

[1] ETH zürich
[2] Max Planck Institute for Intelligent Systems
[3] MIT
[4] Stanford University
Learning & Adaptive Systems

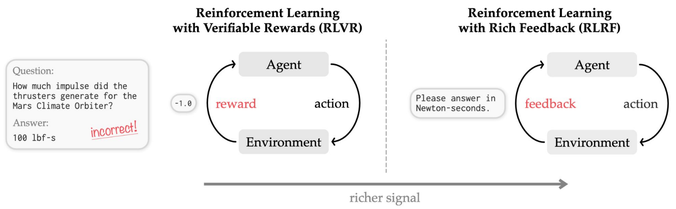
Jonas Hübotter[1], Frederike Lübeck*,[1,2], Lejs Behric*,[1], Anton Baumann*,[1], Marco Bagatella[1,2], Daniel Marta[1], Ido Hakimi[1], Idan Shenfeld[3], Thomas Kleine Buening[1], Carlos Guestrin[4], Andreas Krause[1]

Motivation

- RLVR with binary rewards receives 1 bit signal per rollout → **feedback bottleneck**.
- GRPO assigns the same advantage to each token → **credit assignment bottleneck**.

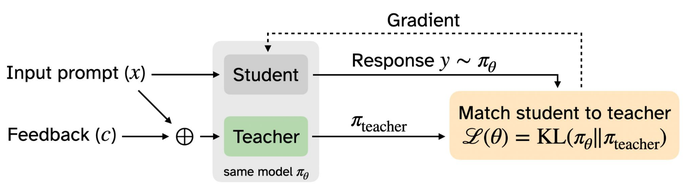
Observation: Models often understand *why* an attempt failed after reflection + feedback:

- Examples of feedback: runtime errors, unit tests, demonstrations, users, …



Self-Distillation (**SDPO**)

"Leverage in-context learning for in-weight learning."



Two equivalent perspectives:

1. On-policy RL with teacher-based advantages

   Maximize $A_i(x, y, c) := \log \frac{\pi_\theta(y_i \mid x, c, y_{<i})}{\pi_\theta(y_i \mid x, y_{<i})}$

2. On-policy distillation with a context-based teacher

   Minimize $\mathscr{L}_{\text{SDPO}}(\theta) := \frac{1}{|y|} \sum_{i=1}^{|y|} \mathrm{KL}(\pi_\theta(\cdot \mid x, y_{<i}) \| \pi_\theta(\cdot \mid x, c, y_{<i}))$

Both yield the same gradient in expectation.

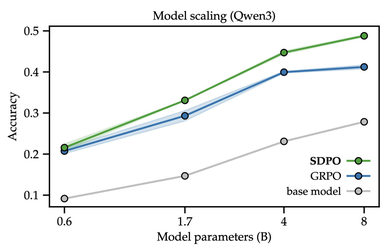
RLVR has a signal and credit assignment bottleneck!

Self-distillation turns **rich feedback** into **dense supervision.**

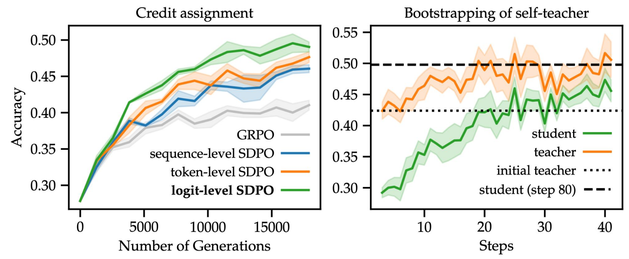
Paper, Code, W&B

Results with Rich Feedback

Qwen3-8B on LiveCodeBench-v6.



Takeaway: SDPO scales with stronger in-context learners.



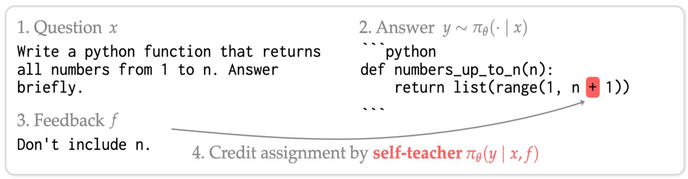
Takeaways: (1) rich feedback helps across varying credit assignment, (2) teacher improves during SDPO.

Results in the RLVR setting

Self-distillation without rich feedback: The teacher reflects on wrong attempts, e.g., by comparing to correct attempts from the same rollout group.

	Chemistry		Physics		Biology		Materials		Tool use	
	1h	5h	1h	5h	1h	5h	1h	5h	1h	5h
Qwen3-8B	41.2		59.2		30.8		58.9		57.5	
+ GRPO	65.9	74.5	63.8	72.7	35.1	**59.9**	**74.3**	77.1	64.9	67.7
+ GRPO (on-policy)	63.3	63.4	63.6	63.6	49.8	49.8	73.9	74.1	60.2	65.7
+ **SDPO** (on-policy)	**73.2**	**80.9**	**66.6**	**75.6**	**50.6**	56.8	72.1	**78.4**	**68.0**	**68.5**
Olmo3-7B-Instruct	22.8		37.7		16.2		36.7		39.3	
+ GRPO	39.7	56.7	55.3	63.3	35.6	**55.8**	70.9	75.0	56.4	**65.0**
+ GRPO (on-policy)	51.4	57.5	**62.7**	62.7	**49.8**	49.8	73.3	73.5	56.8	60.6
+ **SDPO** (on-policy)	**68.0**	**80.0**	59.9	**66.1**	48.0	52.8	**73.7**	**79.1**	**60.8**	62.1

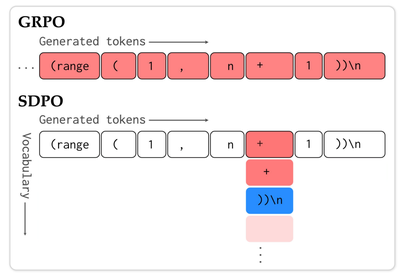
Example of Self-Distillation



with Qwen3-8B

Two differences of SDPO to GRPO:

1. Receiving a rich feedback signal.
2. Turning this into dense supervision.



Test-Time Self-Distillation

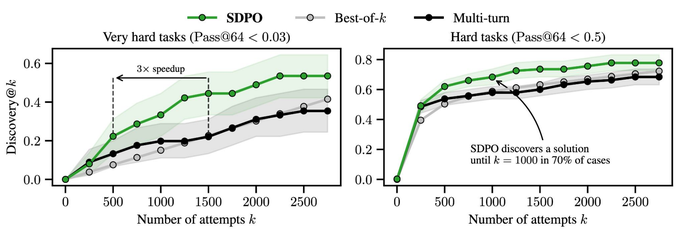
Goal: Accelerate time-to-discovery in hard problems. The *discovery time* is the number of trials to find a solution. We want to increase the probability of discovery:

$$\text{discovery@}k := \mathbb{P}(\text{discovery time} \leq k)$$

We evaluate Qwen3-8B on hard LiveCodeBench-v6 problems (*pass@64 < 0.5*, *pass@64 < 0.03*).

Baselines:

- Best-of-*k* / GRPO: Sampling *k* times from base.
- Multi-turn: Keeping feedback in context.



Takeaways:

- SDPO significantly increases discovery@k: SDPO solves problems not solvable by the base model.
- The initial teacher does not yet solve questions: Teacher feedback is directionally helpful.