Reinforcement Learning for Reasoning in Large Language Models with One Training Example

Yiping Wang[1], Qing Yang[2], Zhiyuan Zeng[1], Liliang Ren[3], Liyuan Liu[3], Baolin Peng[3], Hao Cheng[3], Xuehai He[4], Kuan Wang[5], Jianfeng Gao[3], Weizhu Chen[3], Shuohang Wang[3], Simon Shaolei Du[1], Yelong Shen[3]

[1]University of Washington, [2]University of Southern California, [3]Microsoft, [4]University of California, Santa Cruz, [5]Georgia Institute of Technology

Key Takeaway

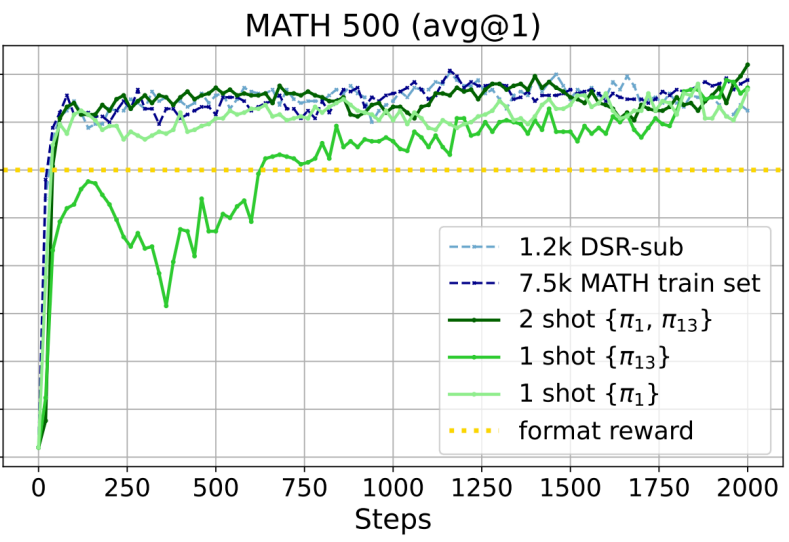
- RLVR with one training examples boosts Qwen2.5-Math-1.5B as good as using 1.2k data. Similarly things happen for 1 (few)-shot RLVR across Qwen2.5-Math-7B/DeepSeek-R1-Distill-1.5B/Llama3.2-3B-Instruct
- Post-saturation generalization, cross-category improvement, more reflection, etc.
- Policy gradient loss primarily drives the improvements observed in 1-shot RLVR, distinguishing it from "grokking", which heavily depends on regularization methods like weight decay.
- Importance of exploration… Future work like **ThetaEvolve**

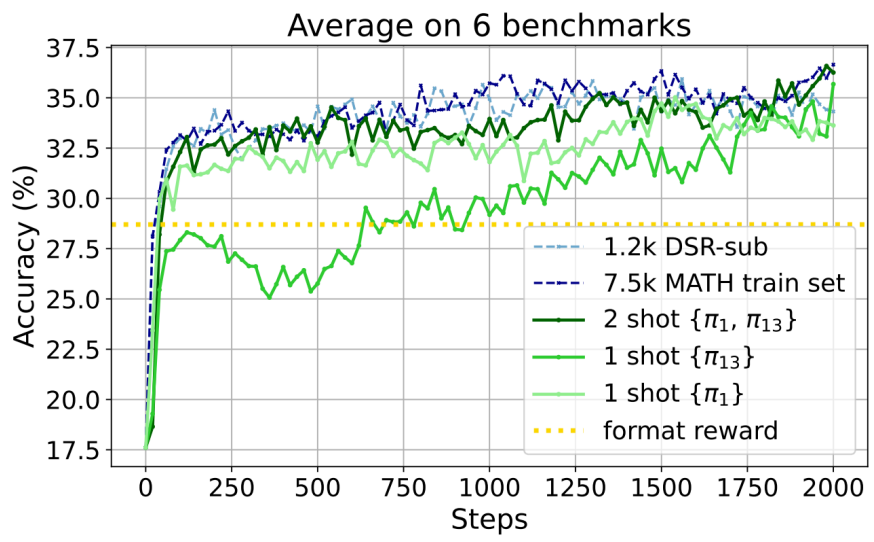
RLVR with 1 example perform as well as using 1.2k dataset

Prompt of example π_1:

```
The pressure \\( P \\) exerted by wind on a sail varies jointly as the area \\( A \\) of the sail and the
cube of the wind's velocity \\( V \\). When the velocity is \\( 8 \\) miles per hour, the pressure on a
sail of \\( 2 \\) square feet is \\( 4 \\) pounds. Find the wind velocity when the pressure on \\( 4 \\)
square feet of sail is \\( 32 \\) pounds. Let's think step by step and output the final answer within \\boxed
```

Ground truth (label in DSR-sub): 12.8.

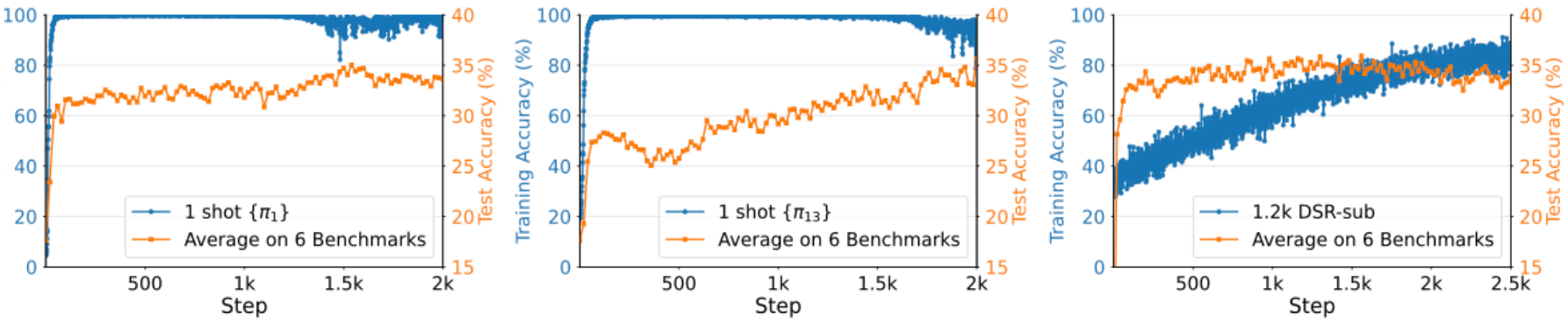


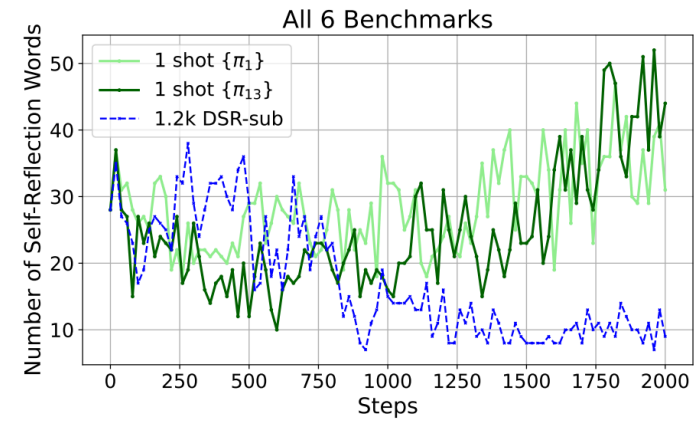


Dataset	Size	ARC-E	ARC-C
Base	NA	48.0	30.2
MATH	7500	51.6	32.8
DSR-sub	1209	42.2	29.9
$\{\pi_1\}$	1	52.0	32.2
$\{\pi_{13}\}$	1	**55.8**	**33.4**
$\{\pi_1, \pi_{13}\}$	2	52.1	32.4

- **Even generalize to other cross-category or non-math tasks**

Post-saturation Generalization: Generalization After Training Accuracy Saturation





- Analysis: Policy Gradient Loss is the Main Contributor
- Entropy Loss Further Improve Post-Saturation Generalization
- Not only format improvements
- Works on AIME25/Minerva, so not only contamination

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	74.8	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	17.1
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6