ToRL: Scaling Tool-Integrated RL

通过RLVR让QWEN2.5-MATH学会借助PYTHON代码执行器提升数学能力

Xuefeng Li* Haoyang Zou* Pengfei Liu†

SJTU, SII, GAIR

开源代码：HTTPS://GITHUB.COM/GAIR-NLP/TORL

简介

从base model (qwen2.5-math)出发，使用RLVR的方式，用GRPO算法训练模型，让模型自由探索在求解数学问题过程中何时生成Python代码，借助外部代码执行器得到代码执行结果，然后继续求解数学题。

注意：QWEN2.5-MATH-INSTRUCT中的TIR(TOOL-INTEGRATED REASONING)属于SFT阶段做的事情，本文是通过RLVR让BASE MODEL自由探索学会生成PYTHON代码求解数学题

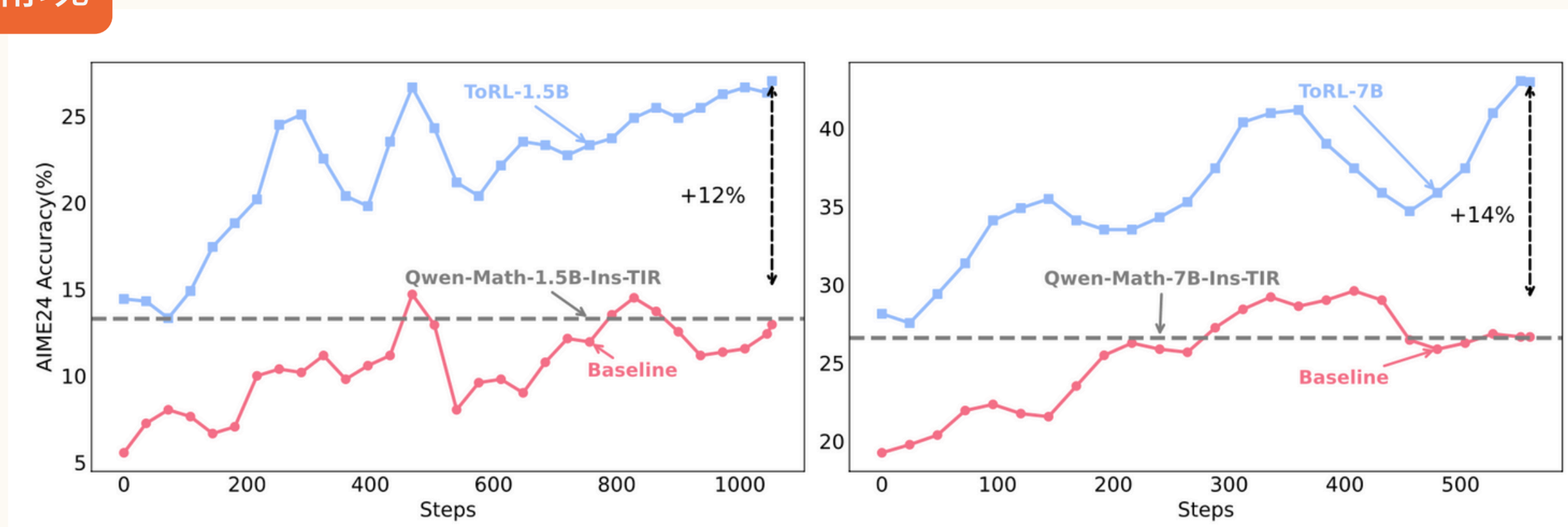
背景

经过TIR sft后的数学模型一般有更强的数学推理能力，能否让math base model学会自己生成Python代码，借助外部代码执行器返回代码结果，然后继续求解数学问题呢？换言之，用RLVR代替sft

实验设置

- 框架：verl
- 实验对象：Qwen2.5-math 1.5B/7B
- 强化学习算法：GRPO
- RLVR
- 为了提升模型探索能力，去除了kl loss
- 代码执行器：字节 Sandbox Fusion

能力涌现



1）随训练进行，使用代码求解问题的比例持续上升
2）生成的代码语法正确性与可执行性也在稳步提升
3）如果运行模型多次(2)生成代码，则效率大打折扣

实验结果

Model	SFT/RL	Tool	AIME24	AIME25	MATH500	Olympiad	AMC23	Avg
Models based on Qwen2.5-Math-1.5B-Base								
Qwen2.5-Math-1.5B-Instruct	RL	✗	10.0	10.0	66.0	31.0	62.5	35.9
Qwen2.5-Math-1.5B-Instruct-TIR	RL	✓	13.3	13.3	73.8	41.3	55.0	41.3
ToRL-1.5B(Ours)	RL	✓	**26.7**$_{+13.3}$	**26.7**$_{+13.3}$	**77.8**$_{+3.0}$	**44.0**$_{+2.7}$	**67.5**$_{+5.0}$	**48.5**$_{+7.2}$
Models based on Qwen2.5-Math-7B-Base								
Qwen2.5-Math-7B-Instruct	RL	✗	10.0	16.7	74.8	32.4	65.0	39.8
Qwen2.5-Math-7B-Instruct-TIR	RL	✓	26.7	16.7	78.8	45.0	70.0	47.4
SimpleRL-Zero	RL	✗	33.3	6.7	77.2	37.6	62.5	43.5
rStar-Math-7B	SFT	✗	26.7	-	78.4	47.1	47.5	-
Eurus-2-7B-PRIME	RL	✗	26.7	13.3	79.2	42.1	57.4	43.1
ToRL-7B(Ours)	RL	✓	**43.3**$_{+10.0}$	**30.0**$_{+13.3}$	**82.2**$_{+3.0}$	**49.9**$_{+2.8}$	**75.0**$_{+5.0}$	**62.1**$_{+14.7}$