Agent RL Scaling Law: Spontaneous Code Execution for Mathematical Problem Solving

Xinji Mai[1,2†] **Haotian Xu**[2†] **Xing W**[2] **Weinong Wang**[2]
Yingying Zhang[3*] **Wenqiang Zhang**[1*]
[1]Fudan University
[2]Xiaohongshu
[3]East China Normal University
xjmai23@m.fudan.edu.cn,{xuhaotian,wuxing,wangweinong}@xiaohongshu.com

简介

开源代码：HTTPS://GITHUB.COM/YYHT/OPENRLHF_ASYNC_PIPLINE

从qwen 2.5 base model出发，使用RLVR的方式，用PPO/Reinforce++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：严格意义上，本文并不是SCALING LAWS，因为并没有给出具体的幂律关系公式。

背景

LLM虽然擅长语言推理，但在涉及精确、可验证的数学计算时，表现依然不理想。而经过TIR sft后的模型一般有更强的数学推理能力，能否让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？

实验设置

- 框架：OpenRLHF和Open-Reasoner-Zero
- 实验对象：Qwen2.5 1.5B/7B/32B
- 强化学习算法：PPO和Reinforce++
- RLVR
- 训练集：ORZ-57k 与 DeepMath，均为可验证答案的数学题

能力涌现

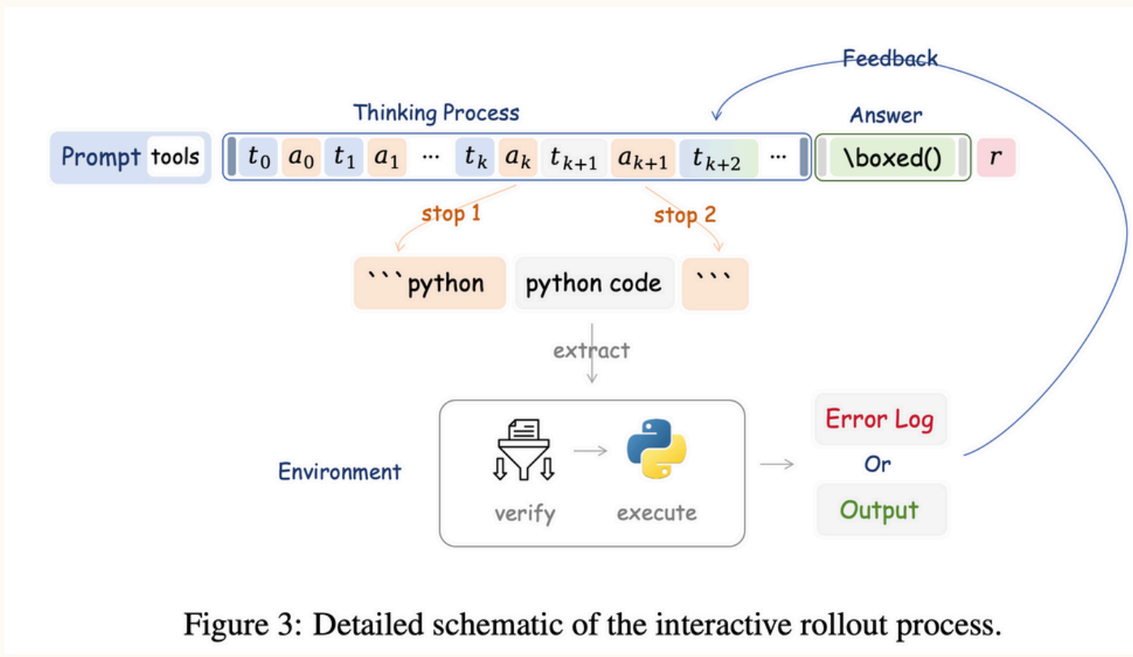
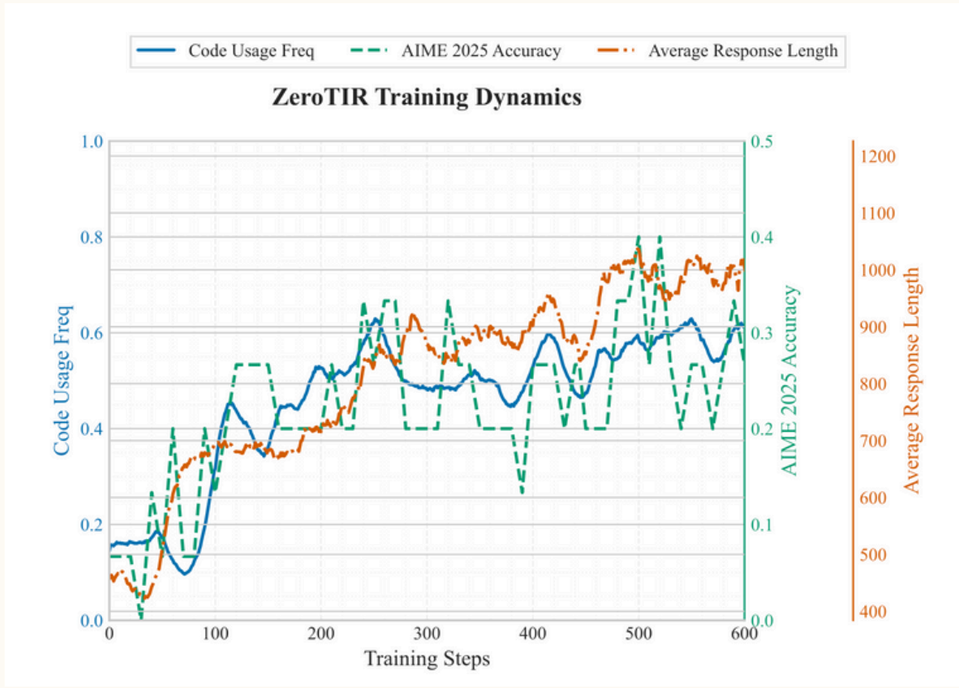


Figure 3: Detailed schematic of the interactive rollout process.



实验结果

Table 1: Performance comparison on key mathematical reasoning benchmarks.

Model	Params	Tool	AIME24	AIME25	MATH500	Avg.	Code Prop.
Qwen2.5 Ins.	7B	✗	13.3%	20.0%	75.8%	36.4%	0.0
Qwen2.5 Ins.	7B	✓	16.7%	0.0%	76.4%	31.0%	0.0
Qwen2.5 Math Ins.	7B	✗	13.3%	6.7%	83.2%	34.4%	0.0
Qwen2.5 Math Ins.	7B	✓	20.0%	26.7%	78.0%	41.6%	**95%**
SimpleRL-Zero	7B	✗	33.3%	6.7%	77.2%	39.1%	0.0
rStar-Math	7B	✗	26.7%	-	78.4%	52.6%	0.0
Eurus-2-PRIME	7B	✗	26.7%	13.3%	79.2%	39.7%	0.0
TORL	7B	✓	43.3%	**30.0%**	**82.2%**	51.8%	83%
ZTRL	7B	✓	**50.0%**	26.7%	80.2%	**52.3%**	89%

Method	Params	Algorithm	Dataset	Train Env. Iter.	Eval Env. Iter.	aime25	aime24	hmmt feb. 25	hmmt feb. 24	cmimc	olymemath	math500	avg.	code ratio
ZTRL	1.5B	ppo	orz-57k	0	0	10.0%	3.3%	0.0%	0.0%	3.3%	2.0%	55.8%	10.6%	0.000
ZTRL	1.5B	ppo	orz-57k	2	2	3.3%	3.3%	0.0%	0.0%	3.3%	1.25%	60.6%	10.3%	0.073
ZTRL	1.5B	ppo	orz-57k	4	4	10.0%	20.0%	10.0%	0.0%	10.0%	5.0%	59.4%	16.3%	2.161
ZTRL	1.5B	ppo	orz-57k	20	20	13.3%	13.3%	10.0%	0.0%	13.3%	7.75%	62.6%	17.2%	**4.090**
ZTRL	7B	ppo	orz-57k	0	4	26.7%	13.3%	13.3%	6.7%	10.0%	8.2%	80.6%	22.7%	0.143
ZTRL	7B	ppo	orz-57k	20	20	26.7%	50.0%	10.0%	20.0%	16.7%	13.5%	80.2%	31.0%	3.490
ZTRL	7B	Reinforce++	orz-57k	2	2	26.7%	30.0%	16.7%	13.3%	26.7%	12.3%	82.8%	29.7%	3.686
ZTRL	7B	Reinforce++	deepmath	2	2	16.7%	36.7%	20.0%	10.0%	20.0%	13.5%	81.0%	29.6%	1.710
ZTRL	7B	Reinforce++	deepmath	2	4	16.7%	40.0%	16.7%	16.7%	20.0%	13.2%	80.6%	29.1%	2.417
ZTRL	7B	Reinforce++	deepmath	4	2	26.7%	36.7%	16.7%	23.3%	20.0%	12.7%	81.2%	31.3%	2.257
ZTRL	7B	Reinforce++	deepmath	4	4	26.7%	33.3%	20.0%	23.3%	20.0%	12.5%	82.0%	32.1%	2.470
ORZ	7B	PPO	orz-57k	0	0	10.0%	16.7%	0.0%	6.7%	10.0%	7.0%	82.2%	18.9%	0.000
DeepMath-Zero	7B	/	deepmath	0	0	13.3%	23.3%	13.3%	6.7%	10.0%	7.2%	82.4%	22.3%	0.000
ORZ	32B	PPO	orz-57k	0	0	30.0%	40.0%	20.0%	20.0%	30.0%	20.8%	90.6%	35.9%	0.000
ZTRL	32B	Reinforce++	deepmath	2	2	26.7%	53.3%	20.0%	16.7%	20.0%	16.7%	86.2%	34.2%	1.691
ZTRL	32B	Reinforce++	deepmath	2	4	26.7%	50.0%	16.7%	26.7%	23.3%	19.0%	87.8%	35.7%	1.994
ZTRL	32B	Reinforce++	deepmath	4	2	**33.3%**	**56.7%**	**20%**	**26.7%**	33.3%	17.5%	87.8%	**39.3%**	1.558
ZTRL	32B	Reinforce++	deepmath	4	4	30.0%	46.7%	20.0%	23.3%	**36.7%**	**21.8%**	**89.4%**	38.2%	1.863

一点评价

其实本文的工作量真不小，实验做的也很扎实，属于和TORL同期的工作，并且作者直接从BASE MODEL而非MATH BASE 入手，用RLVR的方式来提升模型自主学会使用PYTHON代码求解数学题的能力。

但是，真不应该用SCALING LAW命名，因为并没有具体的幂律关系公式，SCALING LAWS意味着更大量级的实验量，慎重慎重。

@机器爱学习