TTRL: Test-Time Reinforcement Learning

Yuxin Zuo[1,2*], Kaiyan Zhang[1*], Li Sheng[1,2], Shang Qu[1,2], Ganqu Cui[2], Xuekai Zhu[1], Haozhan Li[1,2], Yuchen Zhang[2], Xinwei Long[1], Ermo Hua[1], Biqing Qi[2], Youbang Sun[1], Zhiyuan Ma[1], Lifan Yuan[1], Ning Ding[1,2†], Bowen Zhou[1,2†]

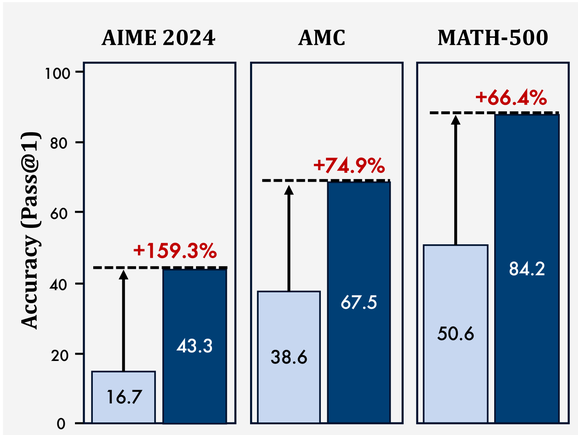
[1]Tsinghua University [2]Shanghai AI Lab *Equal Contribution. †Corresponding Authors.

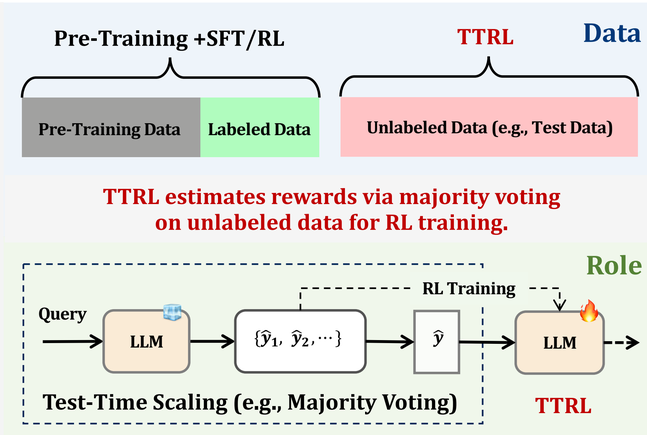
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

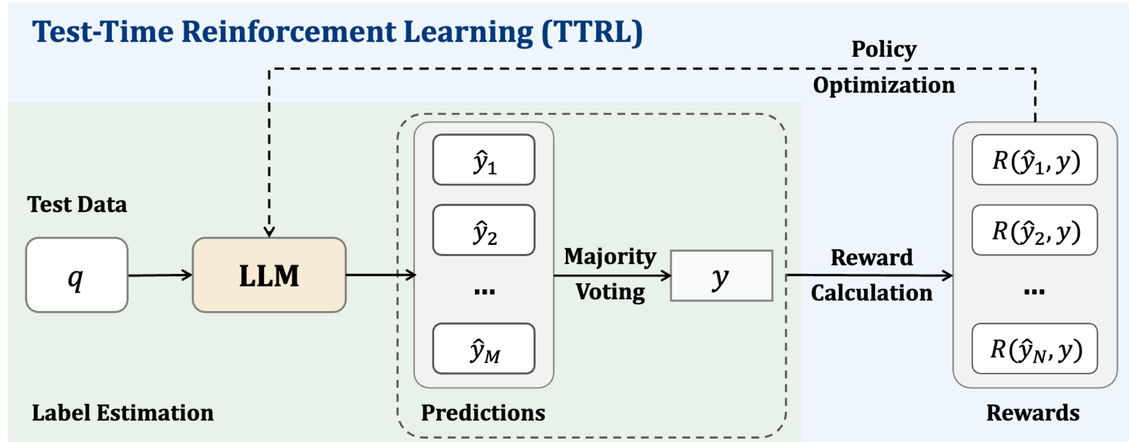
Scale up RL with unlabeled data to turn more computation into intelligence, and move toward the RL-driven self-evolution method.





Test-Time Reinforcement Learning

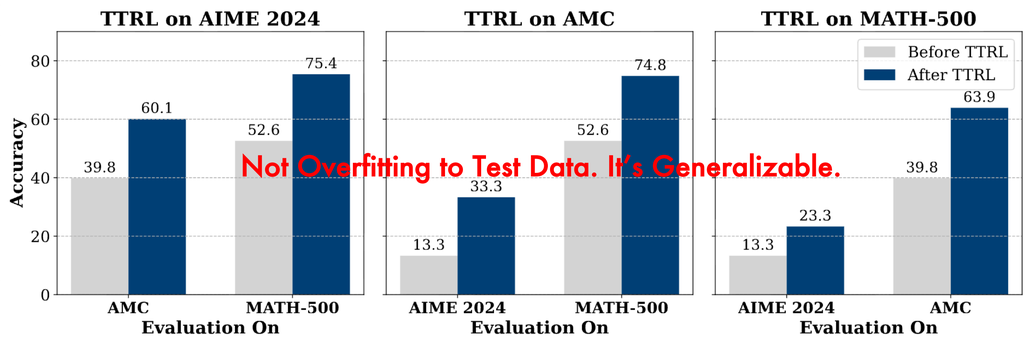
We study the problem of training a pre-trained model during test time using RL without ground-truth labels. We call this setting Test-Time Reinforcement Learning.



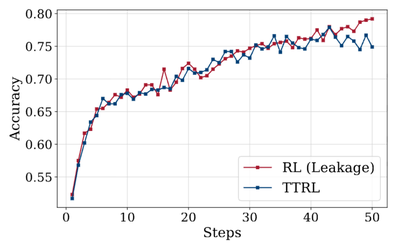
Experiments & Analysis

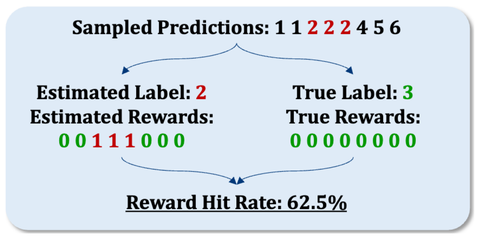
Name	AIME 2024	AMC	MATH-500	GPQA	Avg
Math Base Models					
Qwen2.5-Math-1.5B	7.7	28.6	32.7	24.9	23.5
w/ **TTRL**	15.8	48.9	73.0	26.1	41.0
Δ	+8.1	+20.3	+40.3	+1.2	+17.5
	↑ 105.2%	↑ 71.0%	↑ 123.2%	↑ 4.8%	↑ 74.4%
Qwen2.5-Math-7B	12.9	35.6	46.7	29.1	31.1
w/ **TTRL**	40.2	68.1	83.4	27.7	54.9
Δ	+27.3	+32.5	+36.7	−1.4	+23.8
	↑ 211.6%	↑ 91.3%	↑ 78.6%	↓ 4.8%	↑ 76.5%
Vanilla Base Models					
Qwen2.5-7B	7.9	34.8	60.5	31.8	33.8
w/ **TTRL**	23.3	56.6	80.5	33.6	48.5
Δ	+15.4	+21.8	+20.0	+1.8	+14.7
	↑ 194.9%	↑ 62.6%	↑ 33.1%	↑ 5.7%	↑ 43.7%
Qwen2.5-32B	7.9	32.6	55.8	33.2	32.4
w/ **TTRL**	24.0	59.3	83.2	37.7	51.1
Δ	+16.1	+26.7	+27.4	+4.5	+18.7
	↑ 203.8%	↑ 81.9%	↑ 49.1%	↑ 13.6%	↑ 57.7%
Instruct Models					
LLaMA3.1-8B	4.6	23.3	48.6	30.8	26.8
w/ **TTRL**	10.0	32.3	63.7	34.1	35.0
Δ	+5.4	+9.0	+15.1	+3.3	+8.2
	↑ 117.4%	↑ 38.6%	↑ 31.1%	↑ 10.7%	↑ 30.6%
Qwen3-8B*	26.9	57.8	82.3	48.1	53.8
w/ **TTRL**	46.7	69.1	89.3	53.0	64.5
Δ	+19.8	+11.3	+7.0	+4.9	+10.8
	↑ 73.6%	↑ 19.6%	↑ 8.5%	↑ 10.2%	↑ 20.0%

It performs well on 14 models (1.5B - 32B) spanning the Qwen, LLaMA, Mistral, and DeepSeek series.



Not Overfitting to Test Data. It's Generalizable.





How Well Can TTRL Perform?
Close to training on the benchmark.

Why Does TTRL Work?
"Lucky Hit" plays a crucial role.

Paper Code WeChat

Recent Works

A Survey of Reinforcement Learning for Large Reasoning Models 2025-10-10

A Survey of Reinforcement Learning for Large Reasoning Models

Kaiyan Zhang[1*†], Yuxin Zuo[1*†], Bingxiang He[1*], Youbang Sun[1*], Runze Liu[1*], Che Jiang[1*], Yuchen Fan[2,3*], Kai Tian[1*], Guoli Jia[1*], Pengfei Li[2,6*], Yu Fu[9*], Xingtai Lv[1*], Yuchen Zhang[2,4*], Sihang Zeng[7*], Shang Qu[1,2*], Haozhan Li[1*], Shijie Wang[2*], Yuru Wang[1*], Xinwei Long[1], Fangfu Liu[1], Xiang Xu[5], Jiaze Ma[1], Xuekai Zhu[3], Ermo Hua[1,2], Yihao Liu[1,2], Zonglin Li[2], Huayu Chen[1], Xiaoye Qu[2], Yafu Li[2], Weize Chen[1], Zhenzhao Yuan[1], Junqi Gao[6], Dong Li[6], Zhiyuan Ma[8], Ganqu Cui[2], Zhiyuan Liu[1], Biqing Qi[2‡], Ning Ding[1,2‡], Bowen Zhou[1,2‡]

[1] Tsinghua University [2] Shanghai AI Laboratory [3] Shanghai Jiao Tong University [4] Peking University [5] University of Science and Technology of China [6] Harbin Institute of Technology [7] University of Washington [8] Huazhong University of Science and Technology [9] University College London

† Project Lead. * Core Contributors. ‡ Correspo

zhang-ky22@mails.tsinghua.edu.cn r-LRMs

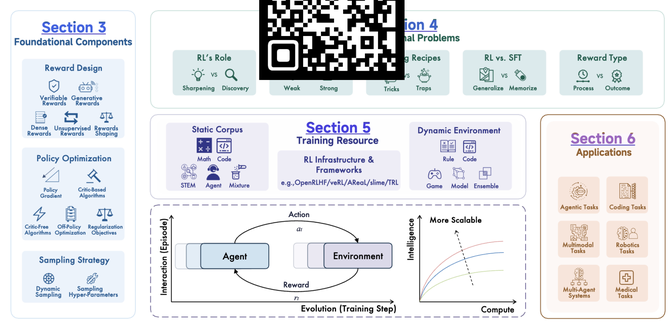


Figure 1 | Overview of the survey. We introduce the foundational components of RL for LRMs, along with open problems, training resources, and applications. Central to this survey is a focus on large-scale interactions between language agents and environments throughout long-term evolution.