从 TTS 到 TTRL
无标签数据强化学习探索与展望

报告人：张开颜（博三）

导师：周伯文教授

清华大学电子系·协同交互智能研究中心 (C3I)

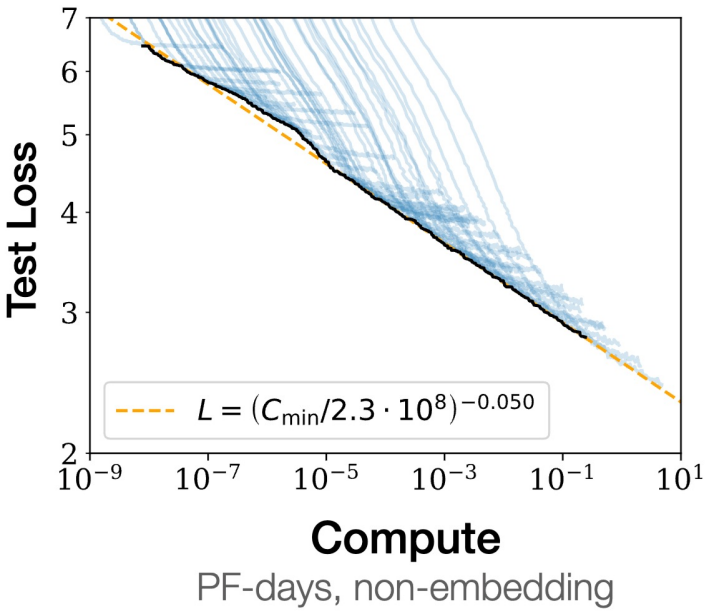
2025.5.14 青稞AI社区

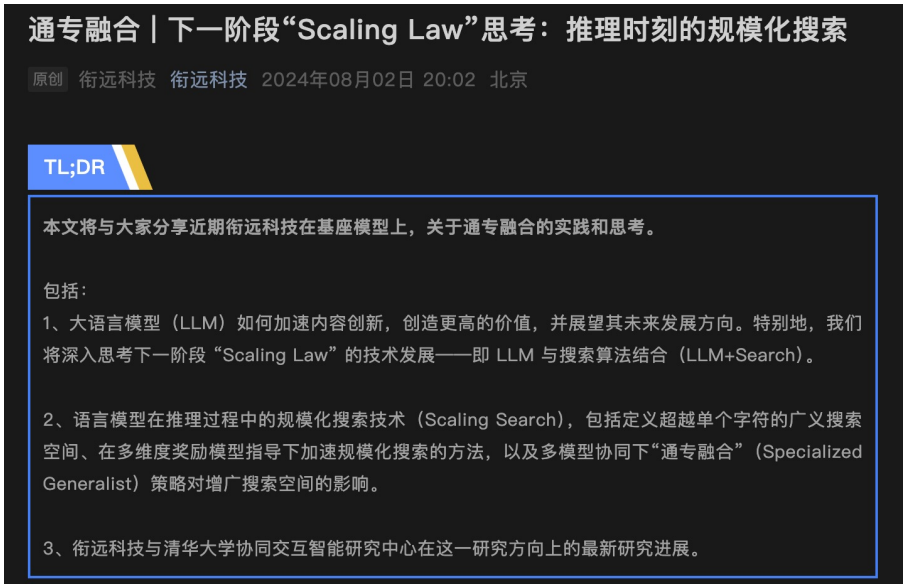
内容大纲

- PART 1: Test-time Scaling (TTS) 与 RL
- PART 2: TTRL: 无标签数据强化学习方法
- PART 3: TTRL 的有效性及局限性讨论
- PART 4: 协同交互视角展望“经验时代”RL

TTS：o1 之前的思考 —— 推理时刻搜索

- “Scaling Law”是驱动大语言模型（LLM）开发和研究的核心经验定律
- 去年7/8月份，我们开始关注从预训练到推理的“Scaling Law”范式转变




通专融合｜下一阶段“Scaling Law”思考：推理时刻的规模化搜索

原创 衍远科技 衍远科技 2024年08月02日 20:02 北京

TL;DR

本文将与大家分享近期衍远科技在基座模型上，关于通专融合的实践和思考。

包括：

1、大语言模型（LLM）如何加速内容创新，创造更高的价值，并展望其未来发展方向。特别地，我们将深入思考下一阶段“Scaling Law”的技术发展——即 LLM 与搜索算法结合（LLM+Search）。

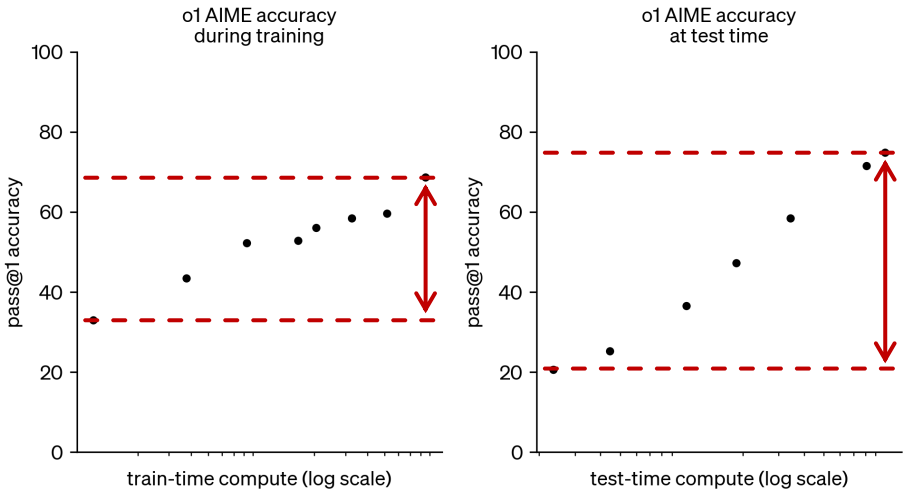
2、语言模型在推理过程中的规模化搜索技术（Scaling Search），包括定义超越单个字符的广义搜索空间、在多维度奖励模型指导下加速规模化搜索的方法，以及多模型协同下“通专融合”（Specialized Generalist）策略对增广搜索空间的影响。

3、衍远科技与清华大学协同交互智能研究中心在这一研究方向上的最新研究进展。

范式转变 (Pre-training → Inference/Test-Time)

TTS：从预训练到测试推理的规模定律

"We have found that the performance of o1 consistently improves with **more reinforcement learning (train-time compute)** and with **more time spent thinking (test-time compute).** The constraints on scaling this approach differ substantially from those of LLM pretraining"

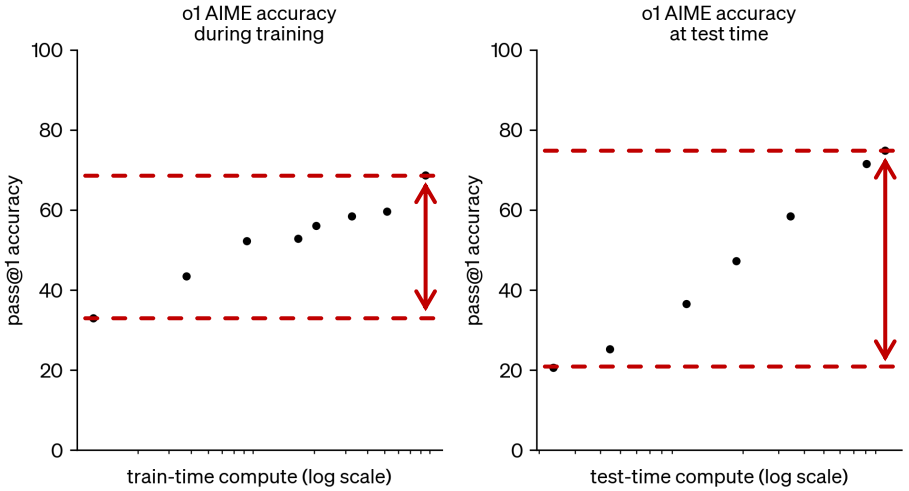


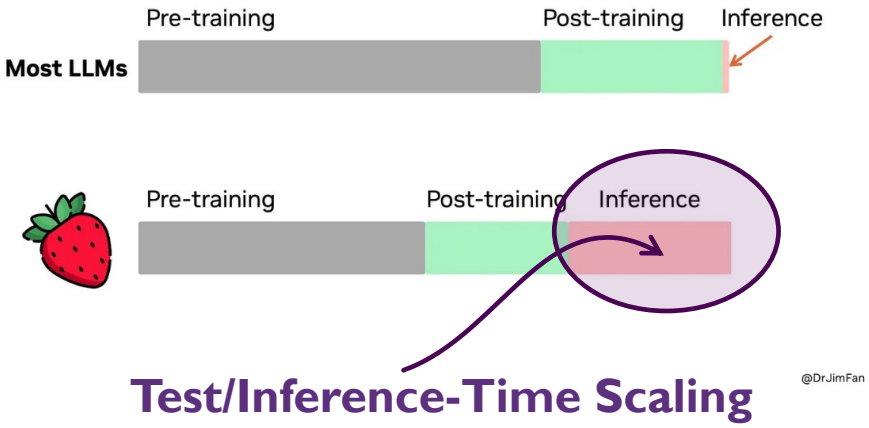
Compute-optimal Scaling
Train-time < Test-time

Learning to reason with llms | openai. Accessed September 12, 2024. https://openai.com/index/learning-to-reason-with-llms/.

TTS：从预训练到测试推理的规模定律

"We have found that the performance of o1 consistently improves with **more reinforcement learning (train-time compute)** and with **more time spent thinking (test-time compute).** The constraints on scaling this approach differ substantially from those of LLM pretraining"





Learning to reason with llms | openai. Accessed September 12, 2024. https://openai.com/index/learning-to-reason-with-llms/.

TTS：不同领域任务场景的探索

Test-Time Scaling (TTS)

- **Open-domain TTS**

  OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025

- **Process Critic TTS**

  GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. UnderReview

- **Model Scale of TTS**

  Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. UnderReview

- **Video Generation TTS**

  Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. UnderReview

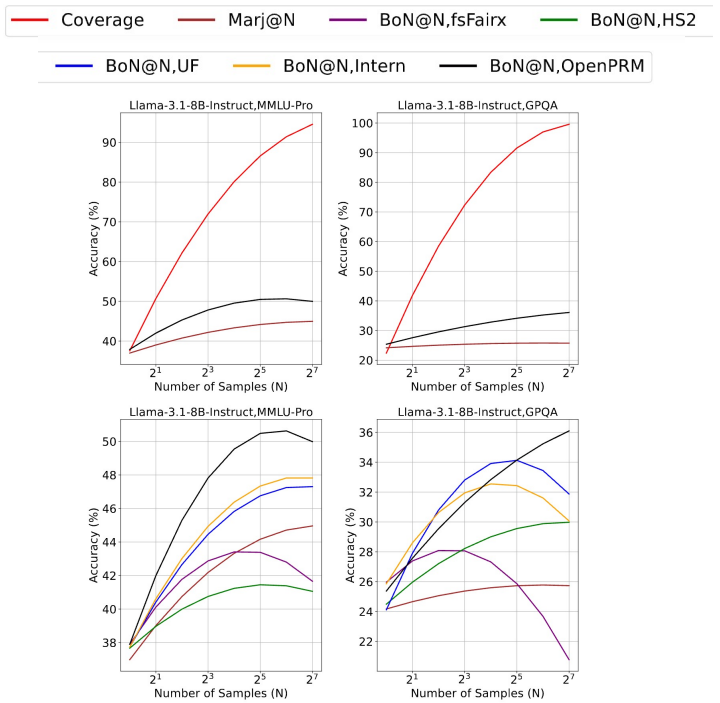
- **More Applications of TTS**

  Coming soon ...

TTS：不同领域任务场景的探索

Test-Time Scaling (TTS)

- **Open-domain TTS**
  OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025
- **Process Critic TTS**
  GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. (UnderReview)
- **Model Scale of TTS**
  Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. (UnderReview)
- **Video Generation TTS**
  Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. (UnderReview)
- **More Applications of TTS**
  Coming soon ...



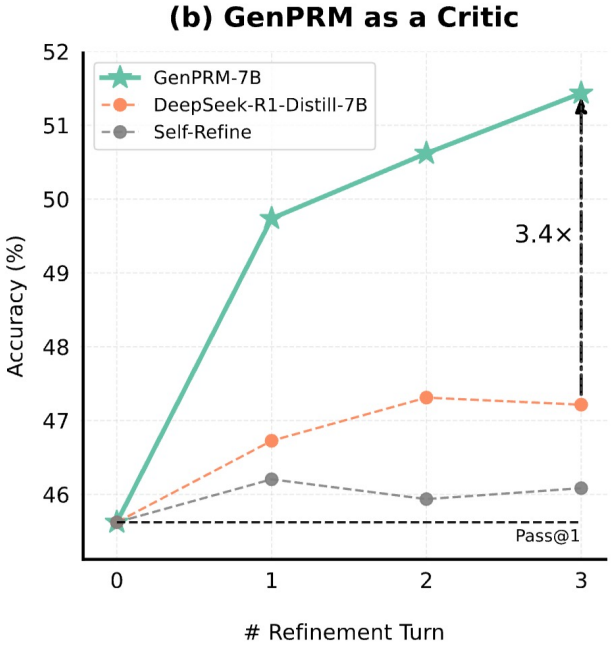
- 针对开放场景任务，构建面向开放领域的过程奖励模型 OpenPRM，在 AlpacaEval / IFEval / MMLU / GPQA 等 Best-of-N 场景验证 TTS 有效性

OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025

TTS：不同领域任务场景的探索

Test-Time Scaling (TTS)

- **Open-domain TTS**
  OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025
- **Process Critic TTS**
  GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. (UnderReview)
- **Model Scale of TTS**
  Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. (UnderReview)
- **Video Generation TTS**
  Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. (UnderReview)
- **More Applications of TTS**
  Coming soon ...



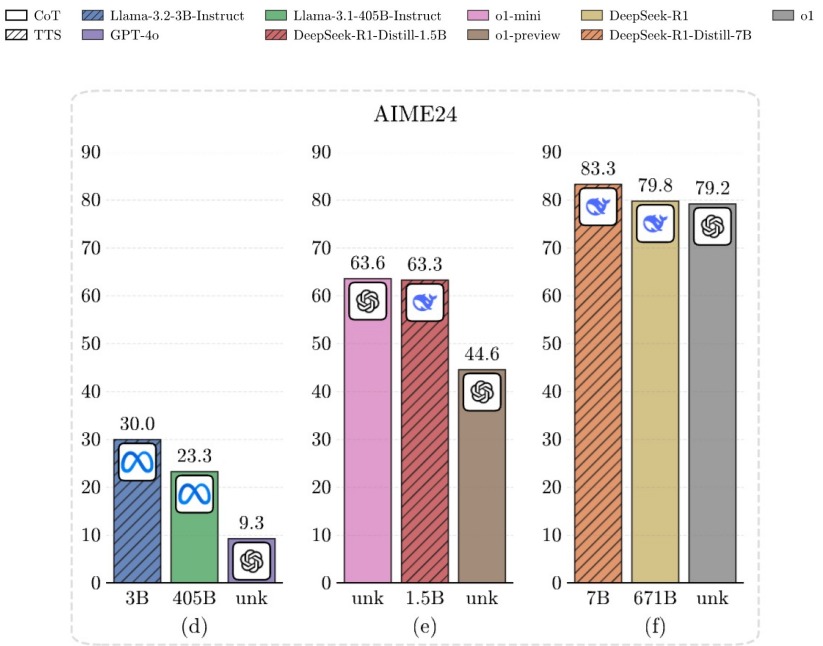
- 进一步将 PRM 扩展到 Generative PRM，基于 R1-7B 构建生成式过程奖励模型 GenPRM，通过在生成过程中进行多轮改错，显著提升模型推理准确性

GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou.

TTS：不同领域任务场景的探索

Test-Time Scaling (TTS)

- **Open-domain TTS**
  OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025
- **Process Critic TTS**
  GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. (UnderReview)
- **Model Scale of TTS**
  Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. (UnderReview)
- **Video Generation TTS**
  Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. (UnderReview)
- **More Applications of TTS**
  Coming soon ...



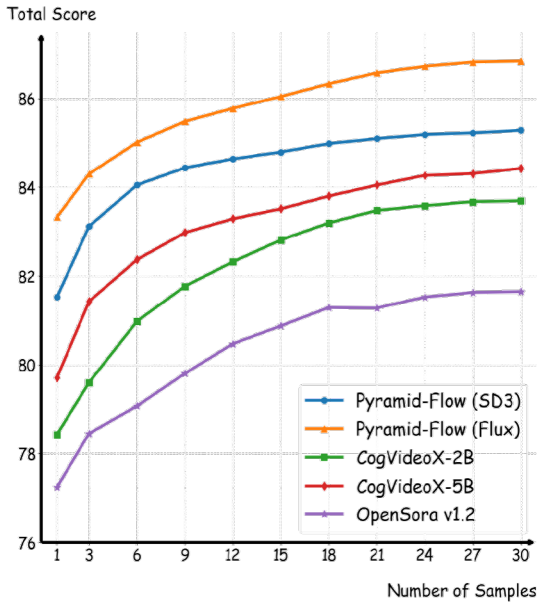
- 针对不同规模尺寸的 Policy Model 和 ORM/ PRM，探索不同 Policy Model 和 Reward Model 组合算法对于 TTS 的影响，最终1B 小模型超过 405B 大模型

Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. (UnderReview)

TTS：不同领域任务场景的探索

Test-Time Scaling (TTS)

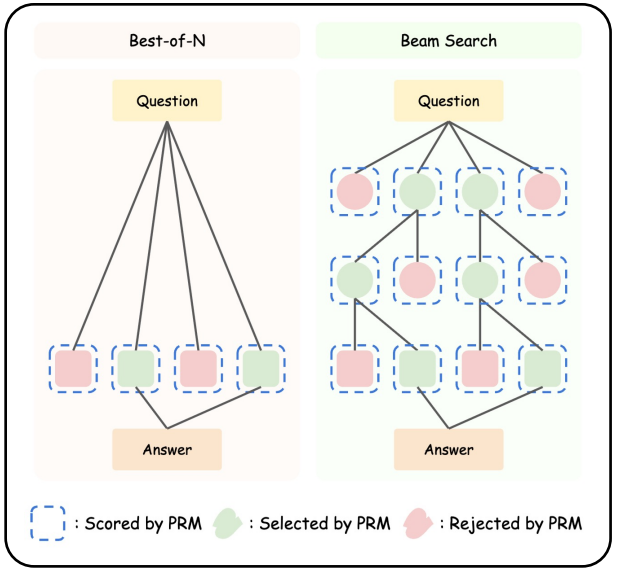
- **Open-domain TTS**
  OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025
- **Process Critic TTS**
  GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. (UnderReview)
- **Model Scale of TTS**
  Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. (UnderReview)
- **Video Generation TTS**
  Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. (UnderReview)
- **More Applications of TTS**
  Coming soon ...



- 针对视频生成任务，探索 Diffusion Model 和 Autoregressive Model 的 TTS 效果，并针对 AR Model 提出 Tree-of-Frames (ToF) 显著提升视频生成效果

Video-T1: Test-Time Scaling for Video Generation. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, Xiaohang Zhan, Yueqi Duan. (UnderReview)

TTS：如何将并行搜索内化到模型?



Parallel Search (w/ PRM/ORM)

Majority Voting, Best-of-N, Step Beam Search, MCTS, …

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a-\sqrt{a+x}} = x$

First, let's square both sides:

$a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

Sequential Search (SFT / RL)

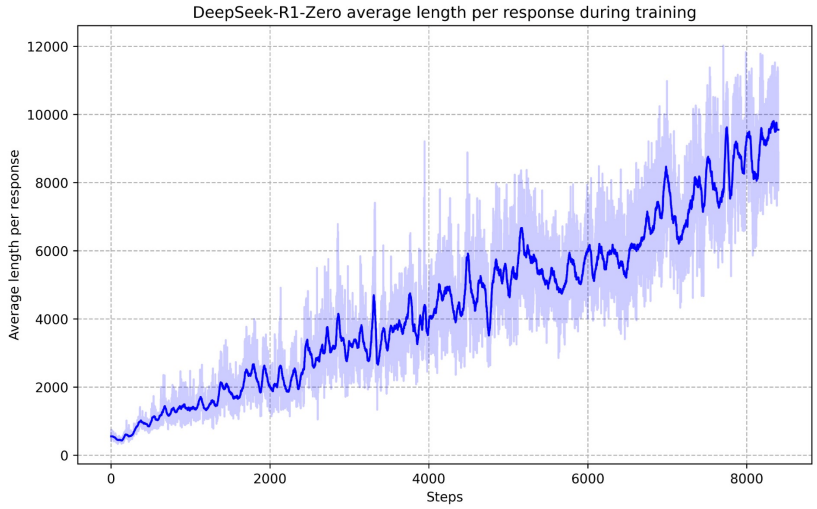
o1-like, Long CoT (Self-Critic/Reflection)

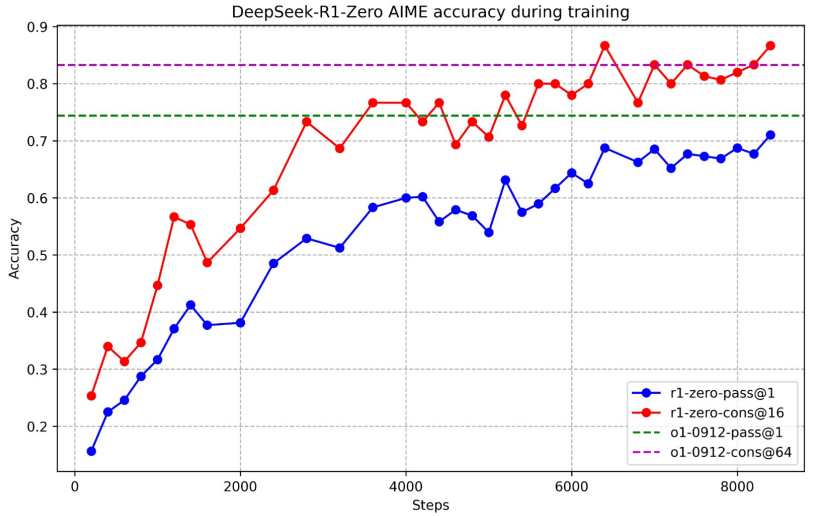
范式转变 (Parallel → Sequential)

TTS：强化学习（RL）

- DeepSeek-R1：基于 Rule-based 奖励的 GRPO 训练， 实现长思维链深度推理

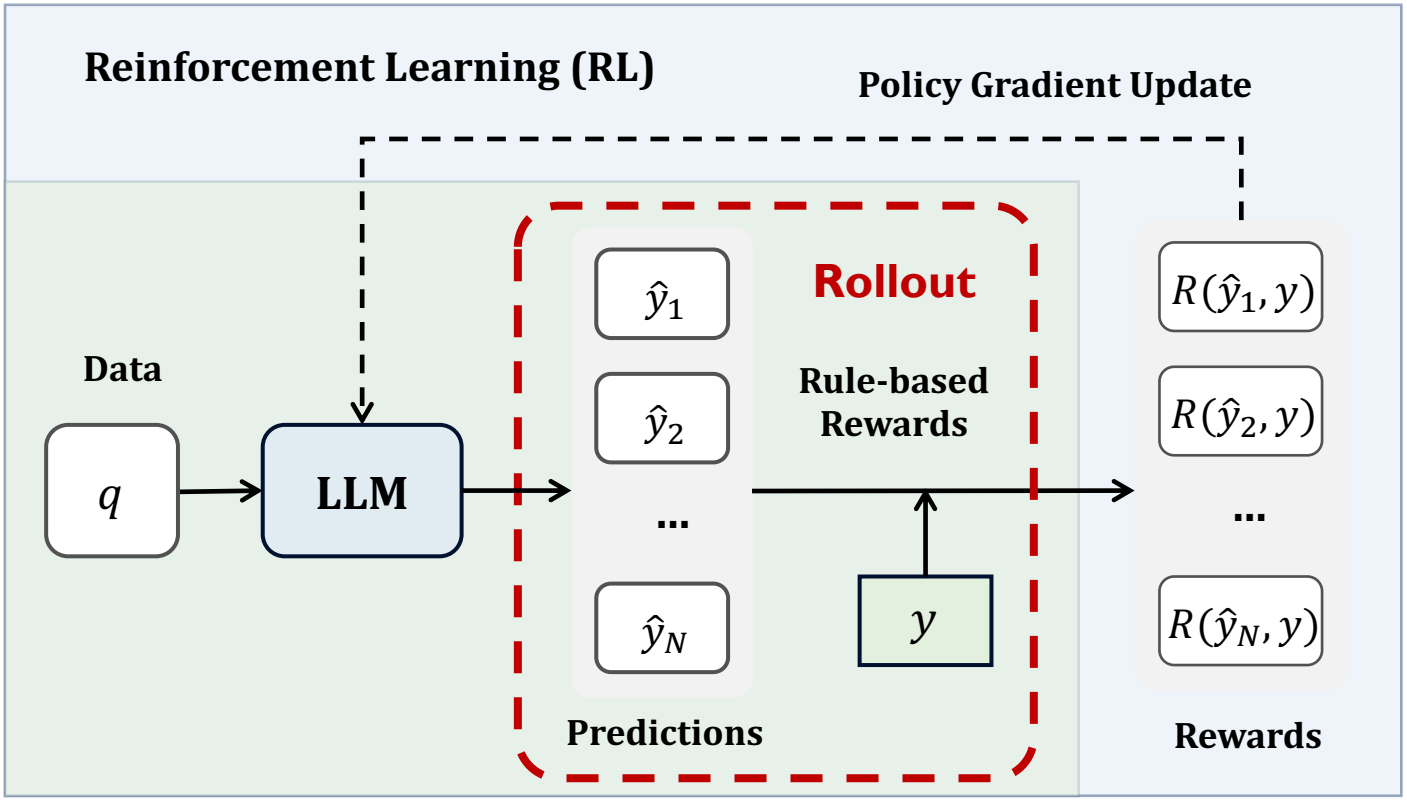
> A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:





TTS：强化学习（RL）

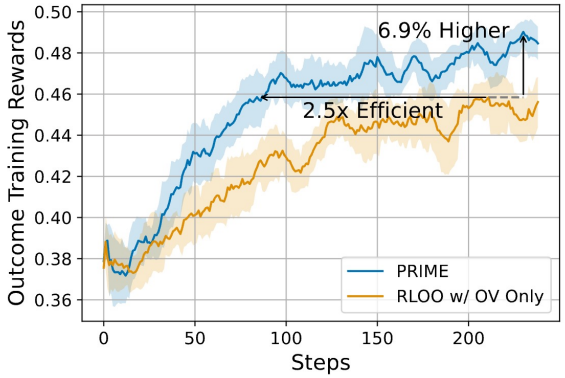
- Rollout 阶段采样大量回复，基于正确答案进行规则打分（0/1）
- 不断强化 LLM 在预训练阶段习得的解题能力，如反思、推理

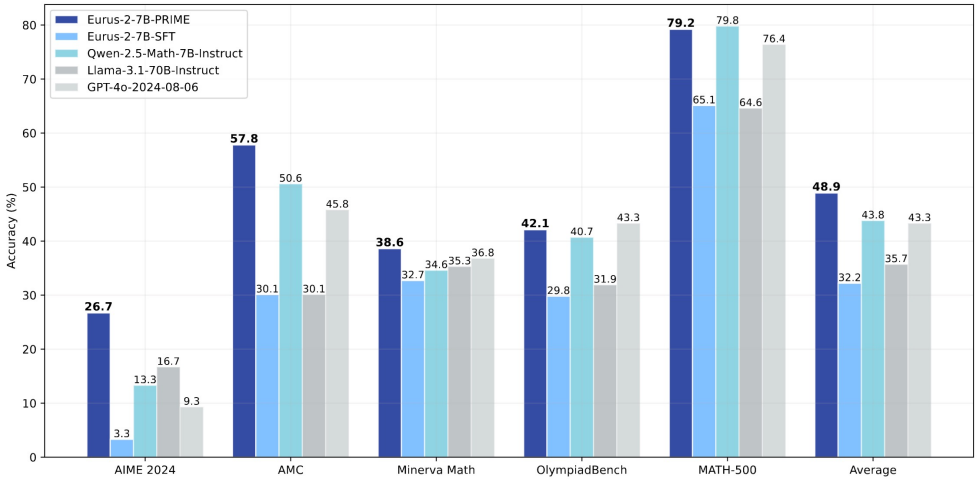


TTS：强化学习（RL）

- PRIME：结合 Rule-based 奖励和隐式过程奖励模型监督的 RL 算法

$$A_t^i = \underbrace{\sum_{s=t}^{|\mathbf{y}^i|} \gamma^{s-t} \cdot \left[r_\phi(y_s^i) - \frac{1}{K-1} \sum_{j \neq i} r_\phi(\mathbf{y}^j) \right]}_{\text{RLOO with implicit process rewards}} + \underbrace{r_o(\mathbf{y}^i) - \frac{1}{K-1} \sum_{j \neq i} r_o(\mathbf{y}^j)}_{\text{RLOO with outcome rewards}}$$



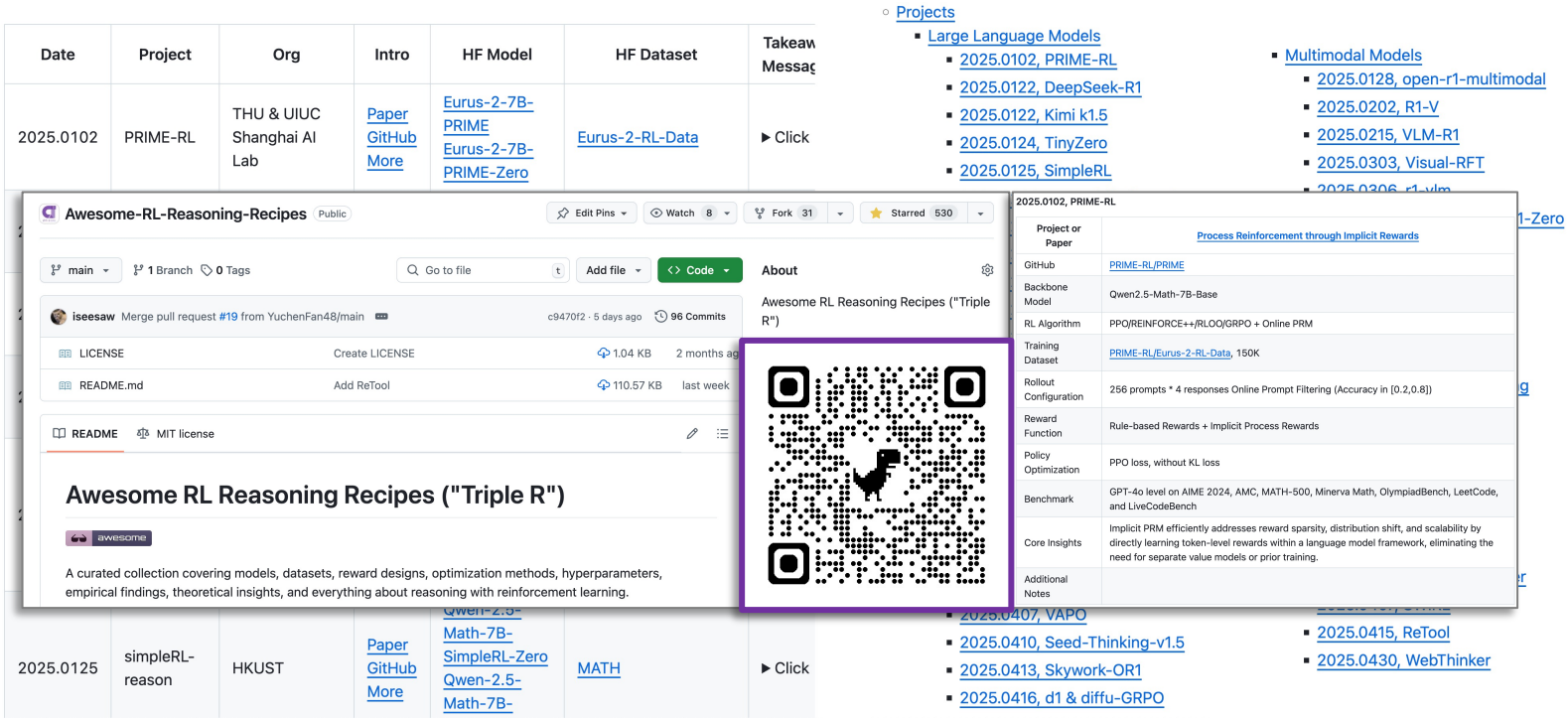


Process Reinforcement through Implicit Rewards. Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, **Kaiyan Zhang**, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, Ning Ding.

Free Process Rewards without Process Labels. Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, **Kaiyan Zhang**, Bowen Zhou, Zhiyuan Liu, Hao Peng. ICML 2025

TTS：强化学习（RL）

- Awesome-RL-Reasoning-Recipes 持续跟踪 2025年 RL 提升 LLM 推理工作
- 全方位对比数据、模型、奖励设计、优化算法、超参数、实验与理论贡献等等



Date	Project	Org	Intro	HF Model	HF Dataset	Takeaw Messag
2025.0102	PRIME-RL	THU & UIUC Shanghai AI Lab	Paper GitHub More	Eurus-2-7B-PRIME Eurus-2-7B-PRIME-Zero	Eurus-2-RL-Data	▶ Click
2025.0125	simpleRL-reason	HKUST	Paper GitHub More	Qwen-2.5-Math-7B-SimpleRL-Zero Qwen-2.5-Math-7B-	MATH	▶ Click

- Projects
  - Large Language Models
    - 2025.0102, PRIME-RL
    - 2025.0122, DeepSeek-R1
    - 2025.0122, Kimi k1.5
    - 2025.0124, TinyZero
    - 2025.0125, SimpleRL
    - 2025.0407, VAPO
    - 2025.0410, Seed-Thinking-v1.5
    - 2025.0413, Skywork-OR1
    - 2025.0416, d1 & diffu-GRPO
  - Multimodal Models
    - 2025.0128, open-r1-multimodal
    - 2025.0202, R1-V
    - 2025.0215, VLM-R1
    - 2025.0303, Visual-RFT
    - 2025.0415, ReTool
    - 2025.0430, WebThinker

Awesome-RL-Reasoning-Recipes Public

Edit Pins · Watch 8 · Fork 31 · Starred 530

main · 1 Branch · 0 Tags · Go to file · Add file · Code

About: Awesome RL Reasoning Recipes ("Triple R")

iseesaw Merge pull request #19 from YuchenFan48/main c9470f2 · 5 days ago · 96 Commits

LICENSE	Create LICENSE	1.04 KB	2 months ag
README.md	Add ReTool	110.57 KB	last week

README · MIT license

Awesome RL Reasoning Recipes ("Triple R")

awesome

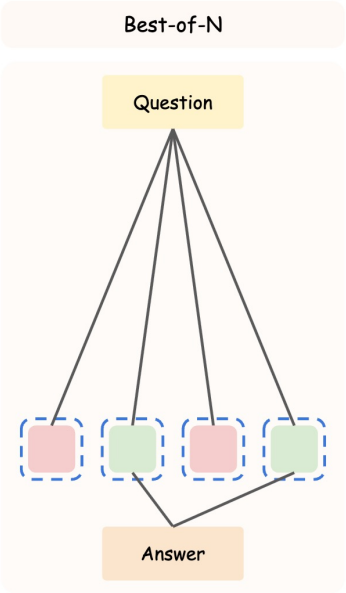
A curated collection covering models, datasets, reward designs, optimization methods, hyperparameters, empirical findings, theoretical insights, and everything about reasoning with reinforcement learning.

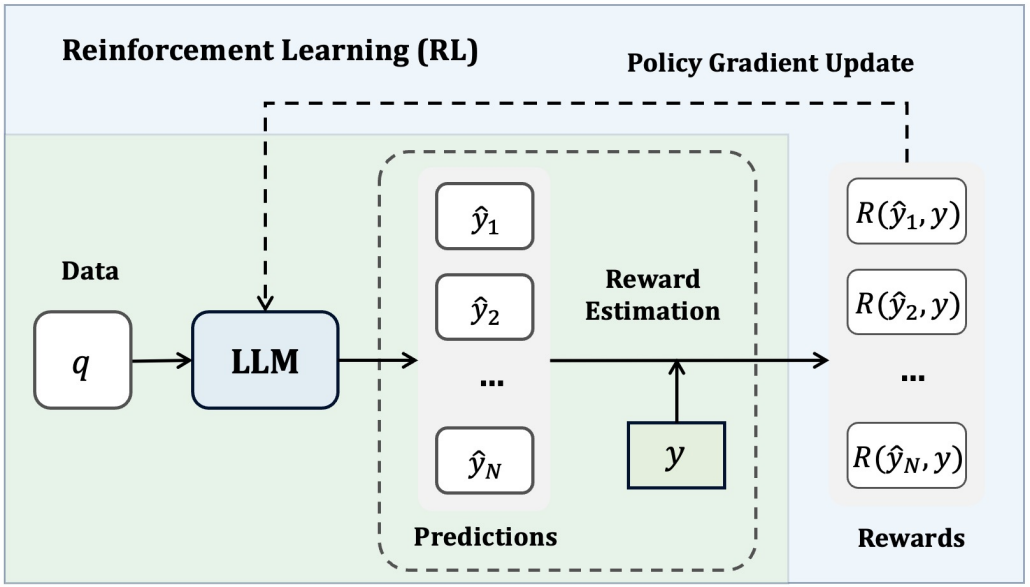
2025.0102, PRIME-RL

Project or Paper	Process Reinforcement through Implicit Rewards
GitHub	PRIME-RL/PRIME
Backbone Model	Qwen2.5-Math-7B-Base
RL Algorithm	PPO/REINFORCE++/RLOO/GRPO + Online PRM
Training Dataset	PRIME-RL/Eurus-2-RL-Data, 150K
Rollout Configuration	256 prompts * 4 responses Online Prompt Filtering (Accuracy in [0.2,0.8])
Reward Function	Rule-based Rewards + Implicit Process Rewards
Policy Optimization	PPO loss, without KL loss
Benchmark	GPT-4o level on AIME 2024, AMC, MATH-500, Minerva Math, OlympiadBench, LeetCode, and LiveCodeBench
Core Insights	Implicit PRM efficiently addresses reward sparsity, distribution shift, and scalability by directly learning token-level rewards within a language model framework, eliminating the need for separate value models or prior training.
Additional Notes	

Awesome RL Recipes for Reasoning. ***Kaiyan Zhang****, Yuchen Fan, Yuxin Zuo, Guoli Jia, Kai Tian, Xingtai Lv, Xuekai Zhu, Ermo Hua, Ning Ding, Biqing Qi, Bowen Zhou. Link:* https://github.com/TsinghuaC3I/Awesome-RRT

TTS：TTS + RL?





Test-Time Scaling ? (Train-Time) Reinforcement Learning

在 Train-Time 进行 RL 训练，训练和测试之间存在数据分布差异！
能否**在 Test-Time 进行 RL 训练**，解决训练与测试数据分布漂移问题？

内容大纲

- PART 1: Test-time Scaling (TTS) 与 RL
- **PART 2: TTRL: 无标签数据强化学习方法**
- PART 3: TTRL 的有效性及局限性讨论
- PART 4: 协同交互视角展望“经验时代”RL

TTRL: TTS = TTI + TTT !!!

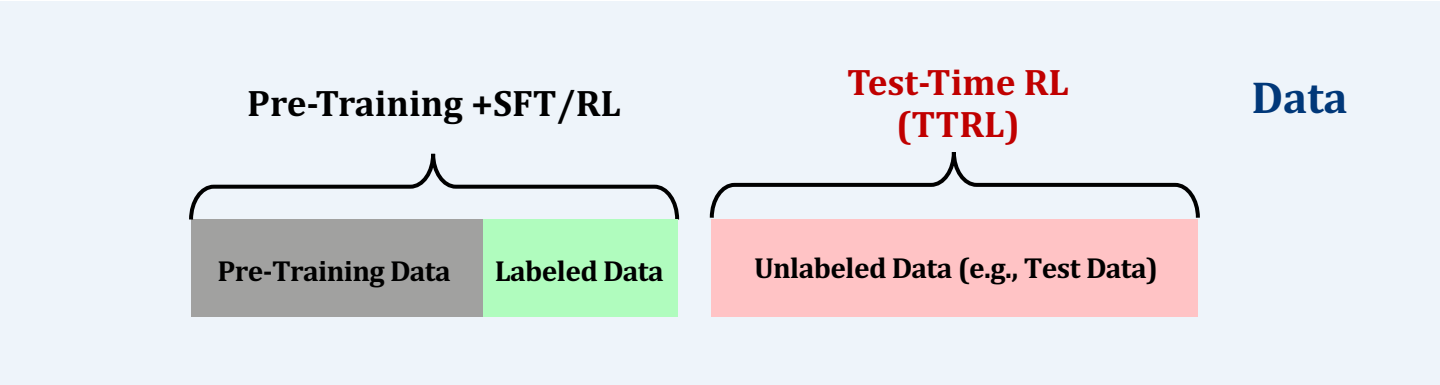
- 目前几乎所有 TTS 都局限为 Test-Time Inference (TTI)，推理时刻不会改变模型参数，完全依靠模型预训练以及后训练数据来注入知识
- 能否在 Test-Time 阶段，通过训练来 Scaling 计算量并进一步提升效果?
  - Test-Time Training (TTT)

Name	Category	Methods
Test-Time-Scaling (TTS)	Test-Time Inference (TTI) Test-Time Training (TTT)	Majority Voting, Best-of-N Test-Time Reinforcement Learning (TTRL)

TTRL:Test-Time Reinforcement Learning.Yuxin Zuo, **Kaiyan Zhang***, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi,Youbang Sun, Ganqu Cui, Ning Ding, Bowen Zhou.*

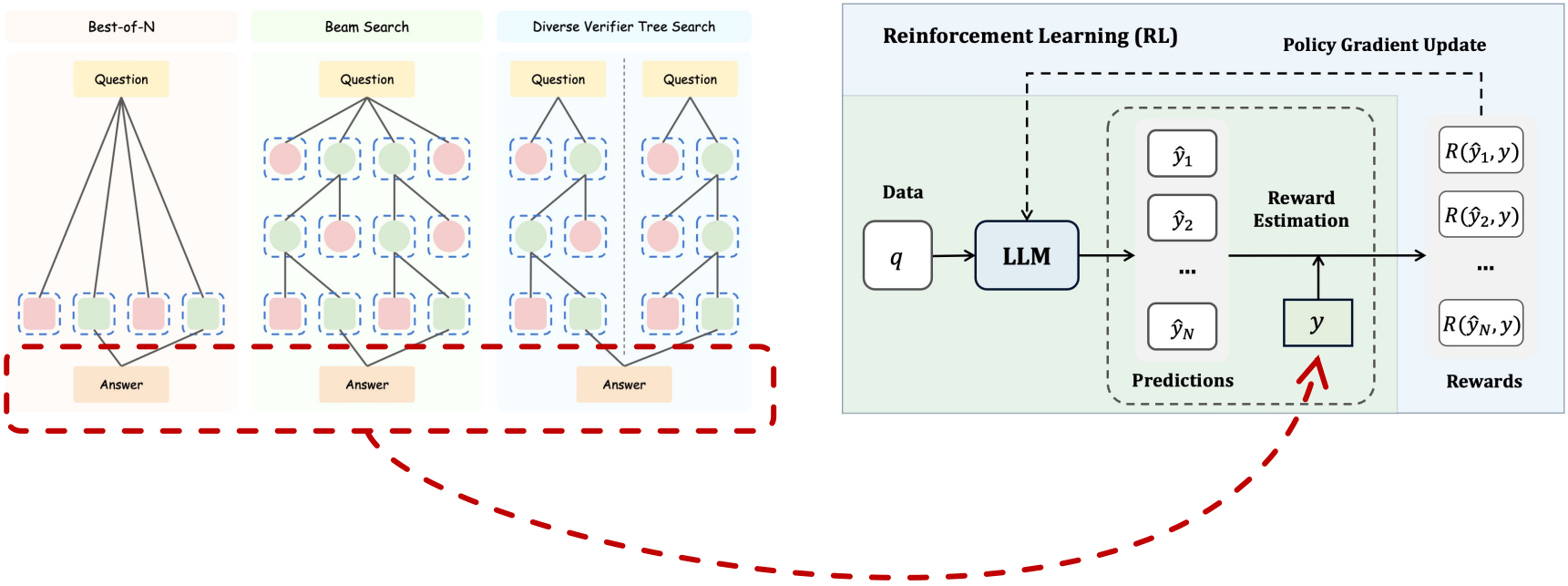
TTRL: RL on Unlabeled Data

- Train-Time RL (Labeled Data) → Test-Time RL (Unlabeled Data)
- **The core challenge of the problem is reward estimation during inference while not having access to ground-truth information.**



TTRL:Test-Time Reinforcement Learning.Yuxin Zuo, **Kaiyan Zhang***, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi,Youbang Sun, Ganqu Cui, Ning Ding, Bowen Zhou.*

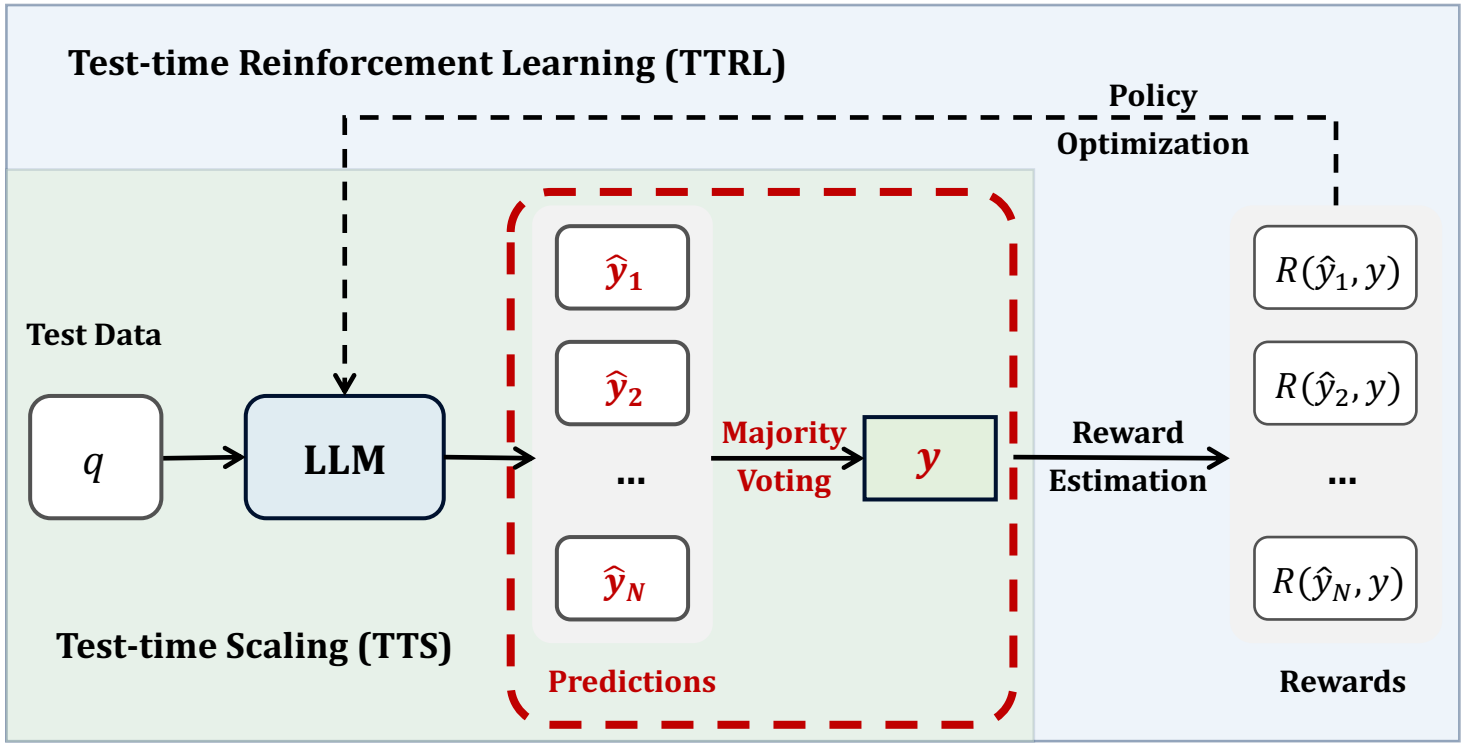
TTRL: Reward Estimation



TTS 方法（如 Best-of-N，Beam-Search）能够获得很好的预测效果，能否将 TTS 获得的预测结果作为伪标签来估计奖励，直接驱动 RL 训练?

TTRL: Reward Estimation

- TTRL 采用 Majority Voting，而非 ~~Best-of-N 或者 MCTS~~ 等答案
- Majority Voting 完全利用 LLM 自身能力，**避免训练 Reward Hacking**

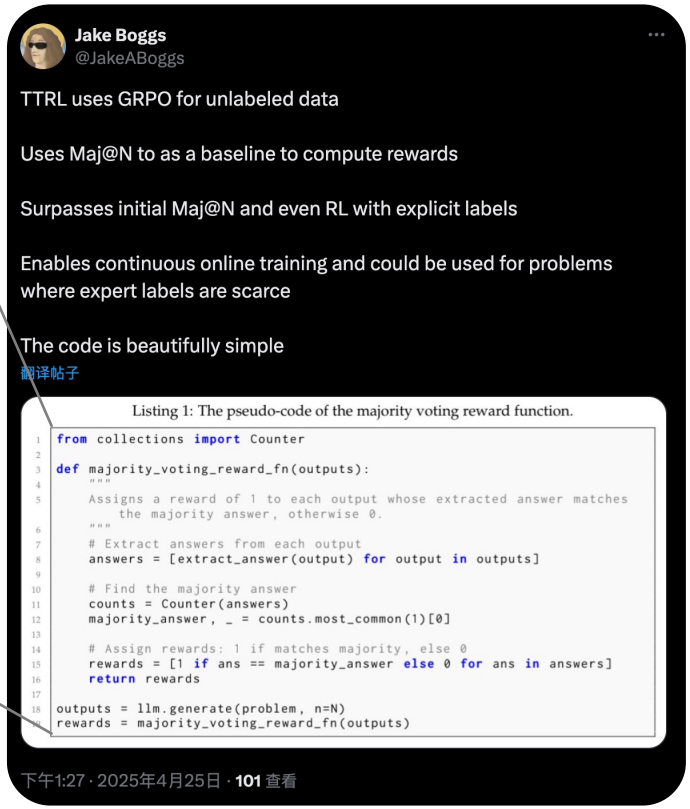


TTRL: "Talk is cheap, Show me the code"

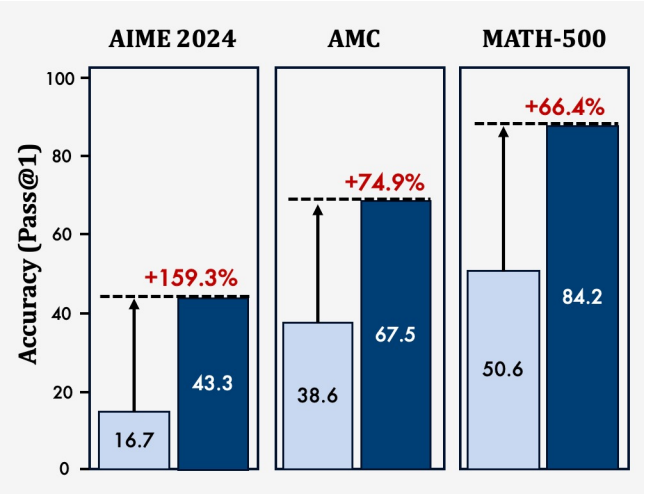
Listing 1: The pseudo-code of the majority voting reward function.

```python
from collections import Counter

def majority_voting_reward_fn(outputs):
    """
    Assigns a reward of 1 to each output whose extracted answer matches
        the majority answer, otherwise 0.
    """
    # Extract answers from each output
    answers = [extract_answer(output) for output in outputs]

    # Find the majority answer
    counts = Counter(answers)
    majority_answer, _ = counts.most_common(1)[0]

    # Assign rewards: 1 if matches majority, else 0
    rewards = [1 if ans == majority_answer else 0 for ans in answers]
    return rewards

outputs = llm.generate(problem, n=N)
rewards = majority_voting_reward_fn(outputs)
```

Twitter:"The code is beautifully simple"



TTRL: Main Results

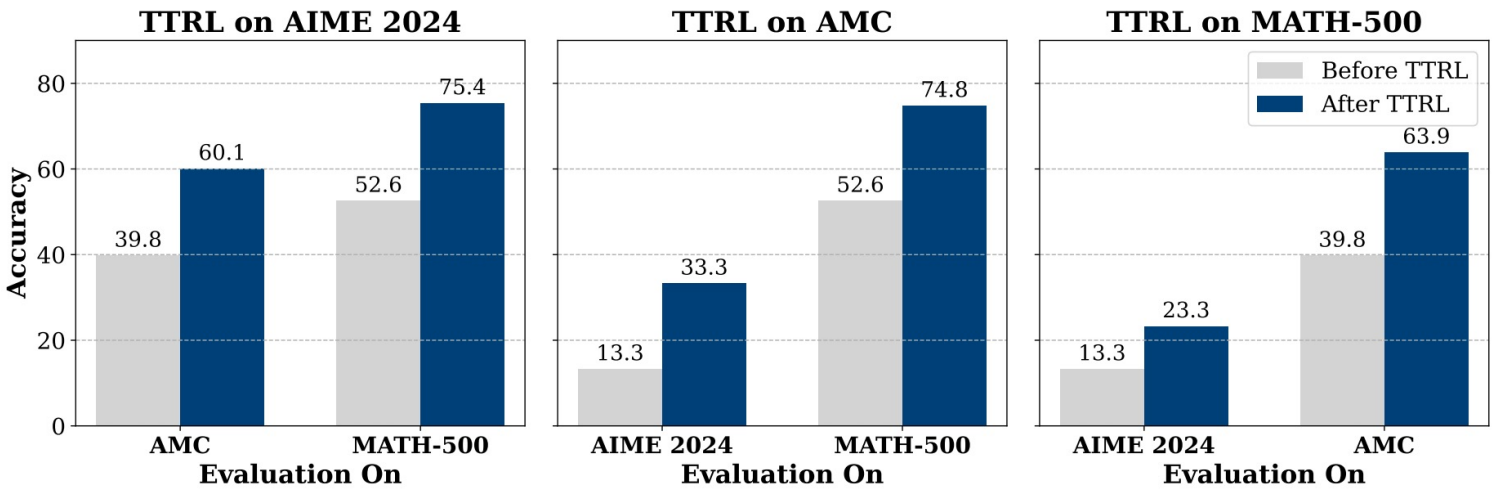


Name	AIME 2024	AMC	MATH-500	GPQA	Avg
Math Base Models					
Qwen2.5-Math-1.5B	7.7	28.6	32.7	24.9	23.5
w/ **TTRL**	15.8	48.9	73.0	26.1	41.0
Δ	+8.1	+20.3	+40.3	+1.2	+17.5
	↑ 105.2%	↑ 71.0%	↑ 123.2%	↑ 4.8%	↑ 74.4%
Qwen2.5-Math-7B	12.9	35.6	46.7	29.1	31.1
w/ **TTRL**	40.2	68.1	83.4	27.7	54.9
Δ	+27.3	+32.5	+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	1.3	20.2	51.6	35.7	27.2
w/ **TTRL**	12.3	54.7	80.3	37.8	46.3
Δ	+11.0	+34.5	+28.7	+2.1	+19.1
	↑ 846.2%	↑ 170.8%	↑ 55.6%	↑ 5.9%	↑ 70.1%

Models
- Qwen, Llama, R1, ...
- 1.5B, 7B, 8B, 32B, ...

Tasks
- AIME, AMC, MATH
- GPQA, ...

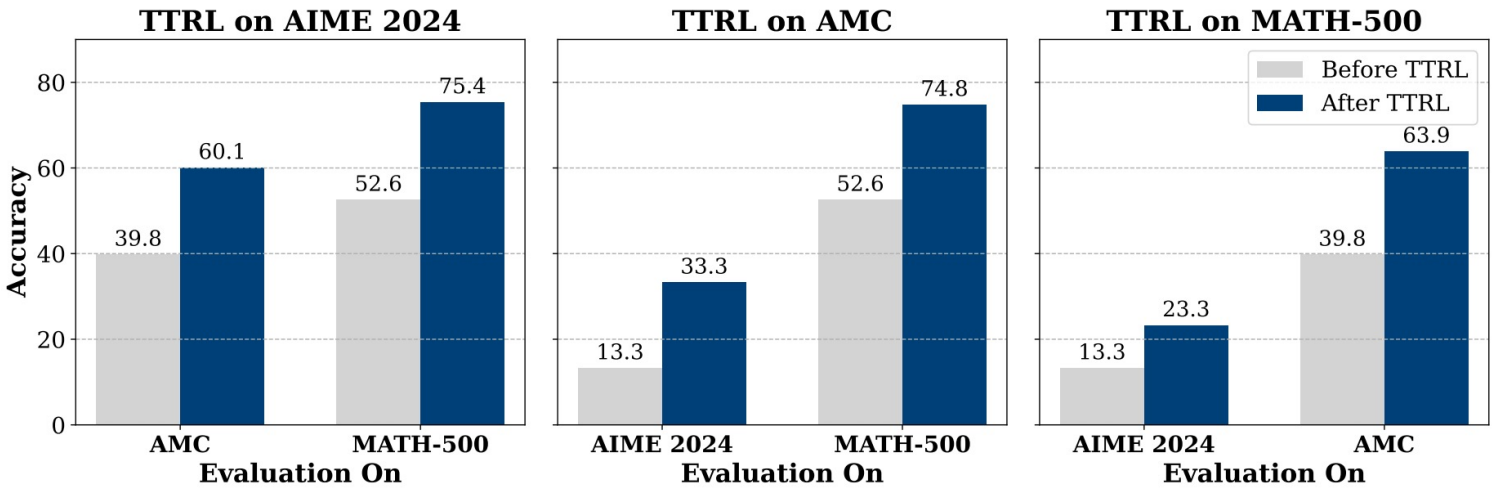
TTRL: Out-of-Distribution (OOD)



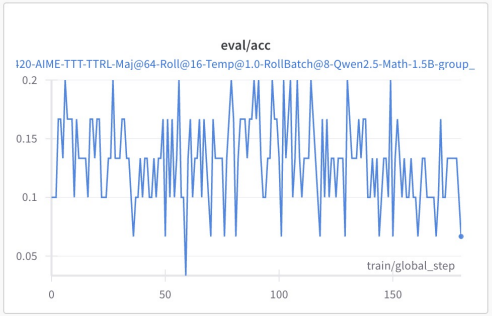
SFT Memorizes, RL Generalizes! [1]

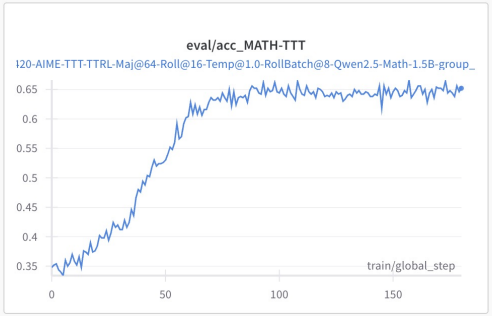
[1] SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, Yi Ma

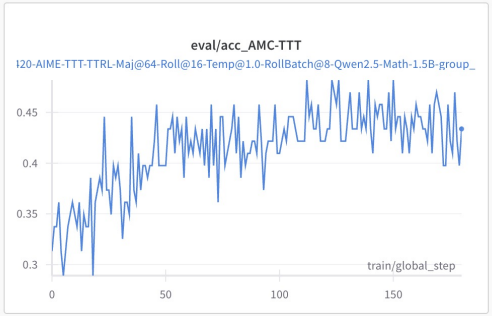
TTRL: Out-of-Distribution (OOD)



Even more surprising

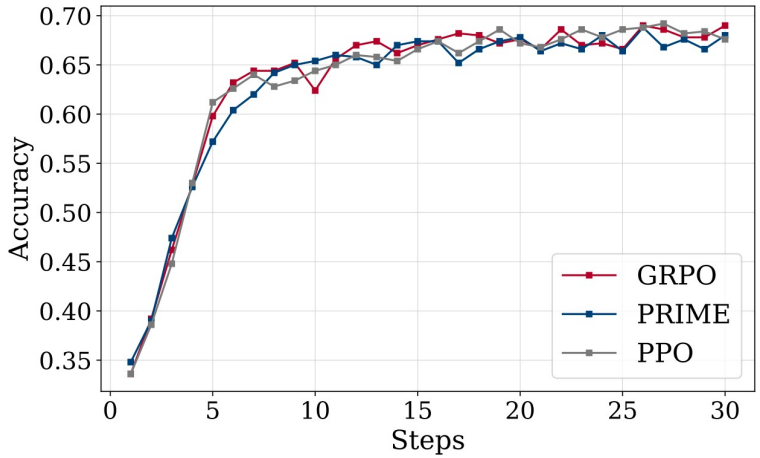




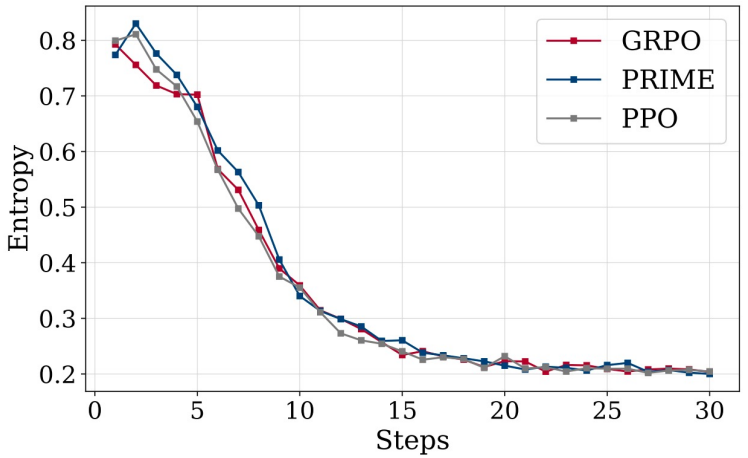


TTRL: Different RL Algorithms

TTRL is compatible with different RL algorithms.



(a) Accuracy Curve.



(b) Entropy Curve.

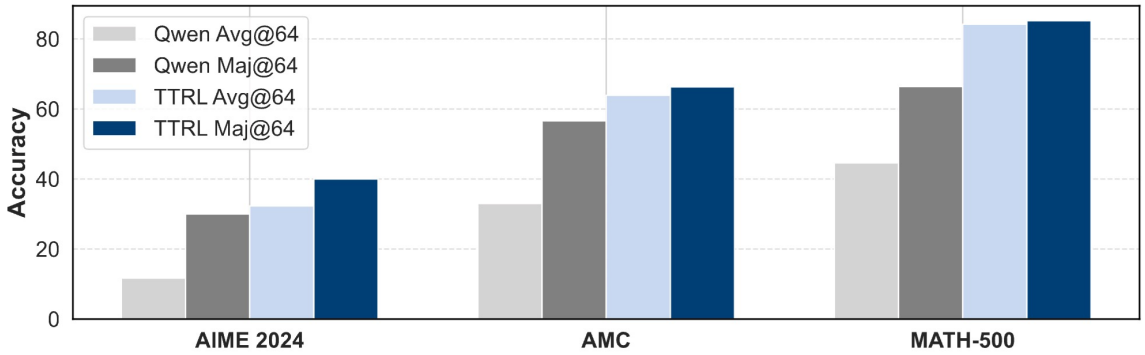
内容大纲

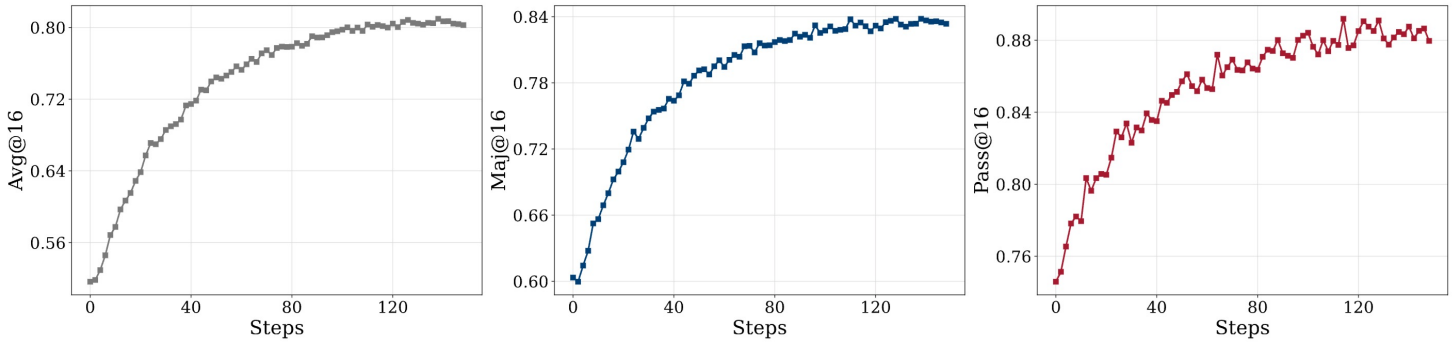
- PART 1: Test-time Scaling (TTS) 与 RL
- PART 2: TTRL: 无标签数据强化学习方法
- **PART 3: TTRL 的有效性及局限性讨论**
- PART 4: 协同交互视角展望“经验时代”RL

TTRL: How, Why, When

Q1: How Well Can TTRL Perform?

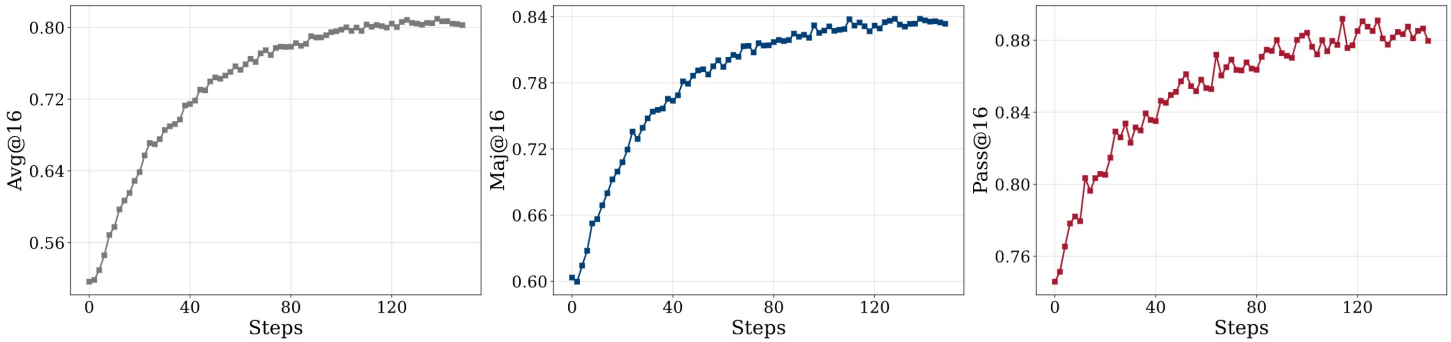
Upper Bound 1: Maj@N of Initial Model



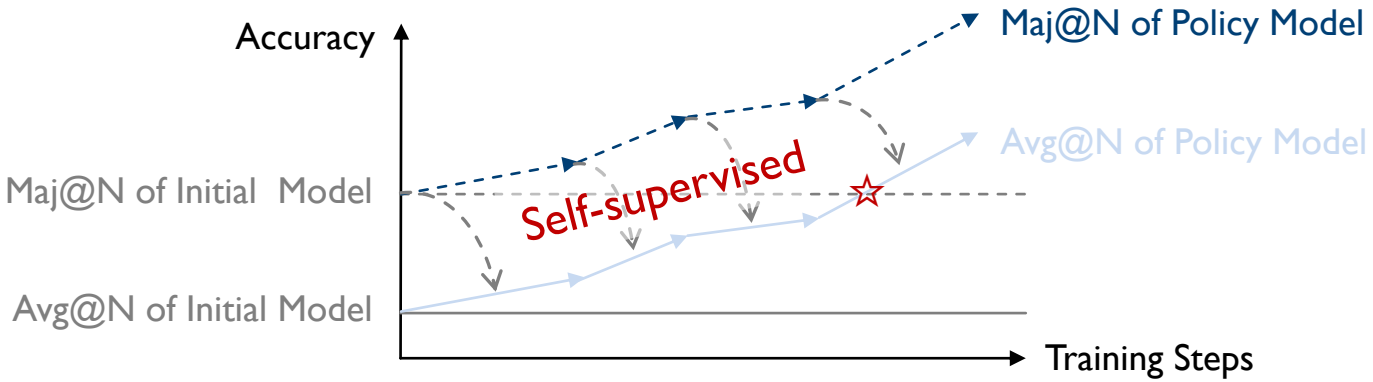


TTRL: How, Why, When

Q1: How Well Can TTRL Perform? Upper Bound 1: Maj@N of Initial Model



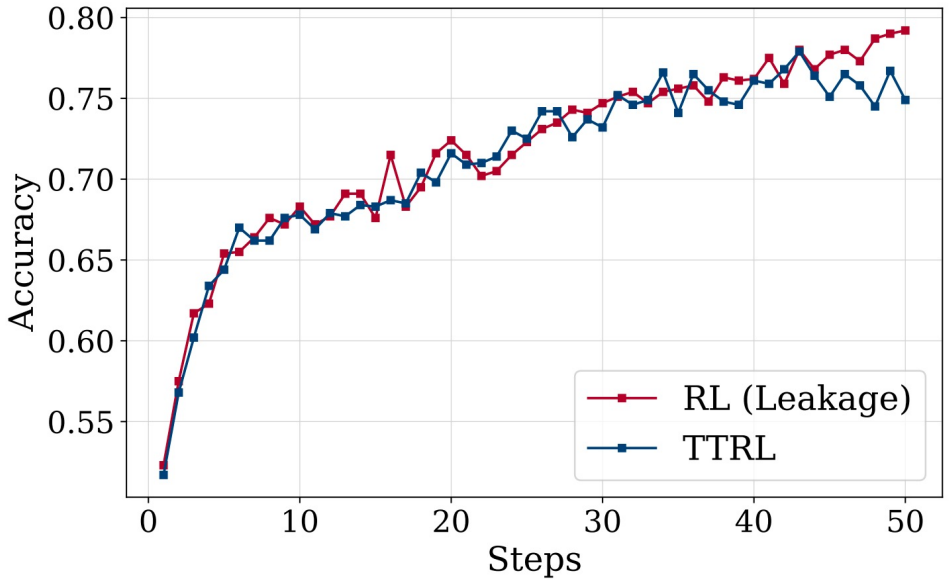
TTRL achieves sustainable self-evolution through "online" and "RL".



TTRL: How, Why, When

Q1: How Well Can TTRL Perform?

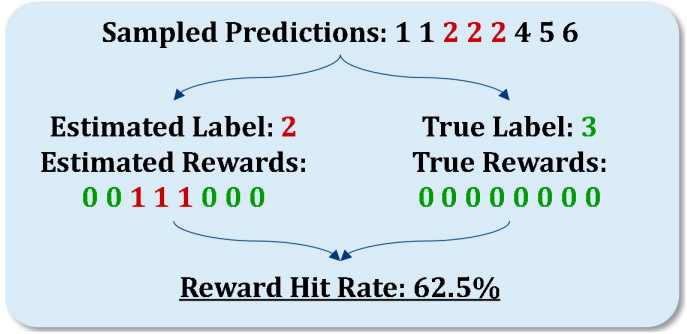
Upper Bound 2: RL with Ground Truth

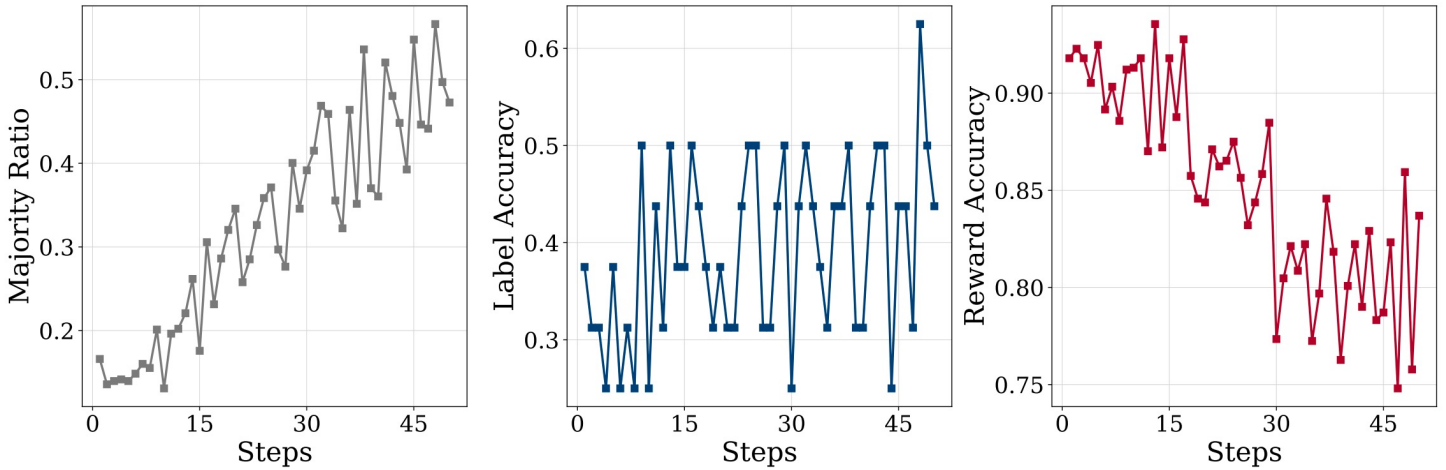


TTRL: How, Why, When

Q2: Why Does TTRL Work?

1. Label Estimations.
2. Reward Calculations.





TTRL: How, Why, When

Q3: When Might TTRL Fail? Lack of Prior Knowledge on Target Task.

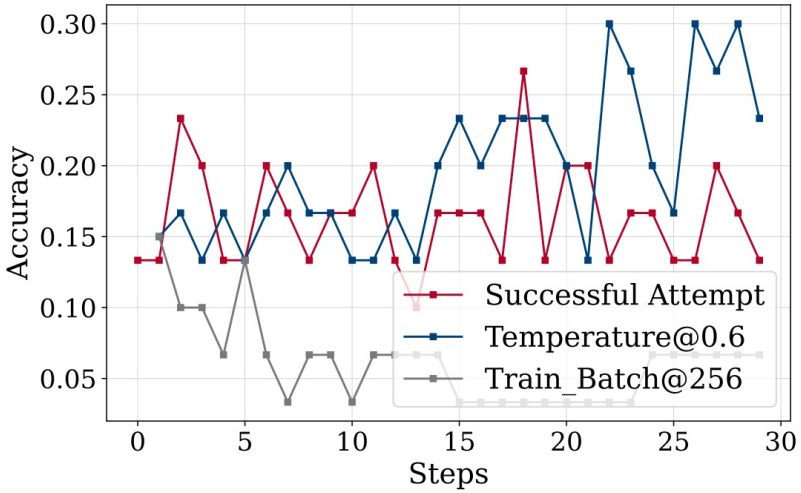
Metric	Name	MATH-500-L1	MATH-500-L2	MATH-500-L3	MATH-500-L4	MATH-500-L5
Accuracy	Backbone	25.9	33.0	36.3	32.5	22.3
	w/ **TTRL**	71.2	76.2	76.3	58.7	39.2
	Δ	+45.4 ↑ 175.3%	+43.2 ↑ 130.8%	+40.0 ↑ 110.2%	+26.2 ↑ 80.4%	+16.8 ↑ 75.3%
Response Len.	Backbone	2,339.2	2,125.1	2,120.6	1,775.1	1,751.3
	w/ **TTRL**	624.3	614.4	672.3	783.5	985.3
	Δ	−1,715.0 ↓ 73.3%	−1,510.6 ↓ 71.1%	−1,448.3 ↓ 68.3%	−991.6 ↓ 55.9%	−766.0 ↓ 43.7%

In RL, there are three key components: **algorithm, environment, and priors**. For a long time, RL researchers focused mostly on the algorithm (e.g. REINFORCE, DQN, TD-learning, actor-critic, PPO, TRPO…) – the intellectual core of how an agent learns – while treating the environment and priors as fixed or minimal. For example, Sutton and Barto's classical textbook is all about algorithms and almost nothing about environments or priors.

Source: https://ysymyth.github.io/The-Second-Half/

TTRL: How, Why, When

Q3: When Might TTRL Fail? Inappropriate RL Hyperparameters.



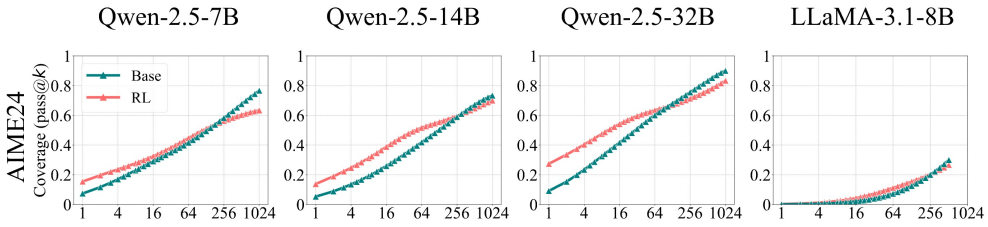
The 37 Implementation Details of Proximal Policy Optimization

25 Mar 2022 | # proximal-policy-optimization # reproducibility # reinforcement-learning # implementation-details # tutorial

Huang, Shengyi; Dossa, Rousslan Fernand Julien; Raffin, Antonin; Kanervisto, Anssi; Wang, Weixun

Source: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/

TTRL: Next Step



Reinforcement Learning Really Can not Incentivize Reasoning Capacity in LLMs Beyond the Base Model? [1]

Future Works Building on our findings, we identify several directions for future research:

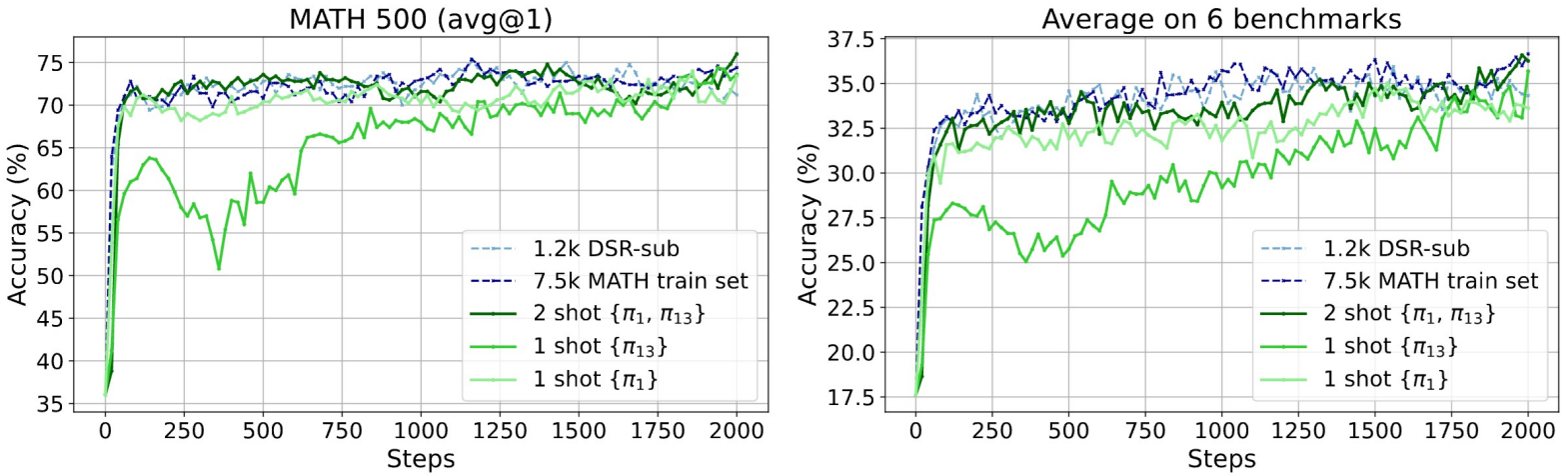
- **Theoretical Analysis:** Developing a formal convergence analysis of **TTRL**, particularly focusing on its ability to optimize toward the two upper bounds in § 4.1.
- **Online Learning with Streaming Data:** Extending **TTRL** to real-time learning scenarios, where models interact with continuously arriving data and adapt dynamically, that is Test-Time Adaptation (Liang et al., 2025).
- **Large-Scale Self-Supervised RL Training:** Scaling up **TTRL** to massive datasets and models to explore its potential in self-supervised regimes without human-labeled data.
- **Agentic Tasks and Scientific Discovery:** Applying **TTRL** to more complex, open-ended domains such as agentic tasks and multi-step scientific reasoning.

[1] Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, Gao Huang.

TTRL: More Recent Works

如何实现在极少样本下进行 TTRL（在线流式学习）？

例如：AIME 30条样本 → 1条样本?
1000条样本训练1次 vs 1条样本训练1000次

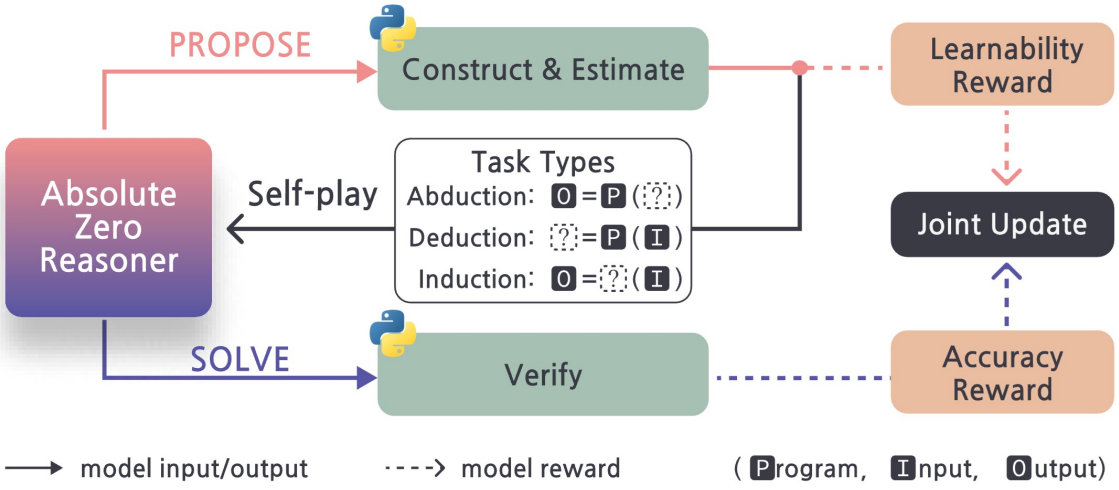


Reinforcement Learning for Reasoning in Large Language Models with One Training Example.Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao,Weizhu Chen, Shuohang Wang, Simon Shaolei Du,Yelong Shen

TTRL: More Recent Works

如何让 LLM 不仅生成答案，更能够自己出题，实现 Self-Play RL?

TTRL: LLM + Prompts → Labels + RL
Absolute Zero: LLM → Prompts + Labels + RL

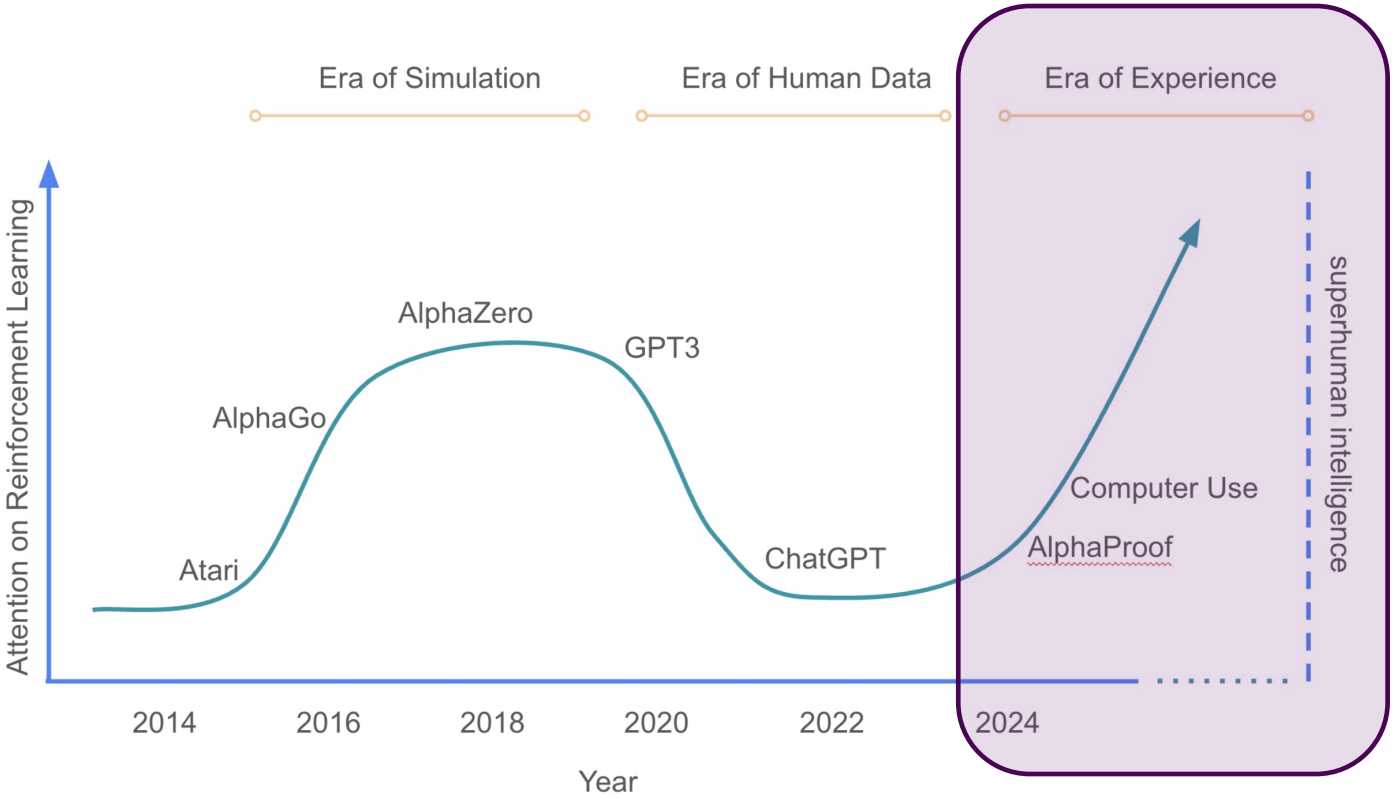


Absolute Zero: Reinforced Self-play Reasoning with Zero Data. Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, Gao Huang.

内容大纲

- PART 1: Test-time Scaling (TTS) 与 RL
- PART 2: TTRL: 无标签数据强化学习方法
- PART 3: TTRL 的有效性及局限性讨论
- **PART 4: 协同交互视角展望“经验时代”RL**

Next: “经验时代” RL



David Silver, Richard S. Sutton* , *Welcome to the Era of Experience*, 2025

Next: “经验时代” RL

Our contention is that incredible new capabilities will arise once the full potential of experiential learning is harnessed. This era of experience will likely be characterised by agents and environments that, in addition to learning from vast quantities of experiential data, will break through the limitations of human-centric AI systems in several further dimensions:

- Agents will inhabit streams of experience, rather than short snippets of interaction.
- Their actions and observations will be richly grounded in the environment, rather than interacting via human dialogue alone.
- Their rewards will be grounded in their experience of the environment, rather than coming from human prejudgement.
- They will plan and/or reason about experience, rather than reasoning solely in human terms

- Agents 应该处于动态变化的环境中，而静态数据或短暂交互
- Agents 的奖励应来源于自身与环境，而不是来源于人类

Next：“经验时代” RL

Our contention is that incredible new capabilities will arise once the full potential of experiential learning is harnessed. This era of experience will likely be characterised by agents and environments that, in addition to learning from vast quantities of experiential data, will break through the limitations of human-centric AI systems in several further dimensions:

- Agents will inhabit streams of experience, rather than short snippets of interaction.
- Their actions and observations will be richly grounded in the environment, rather than interacting via human dialogue alone.
- Their rewards will be grounded in their experience of the environment, rather than coming from human prejudgement.
- They will plan and/or reason about experience, rather than reasoning solely in human terms

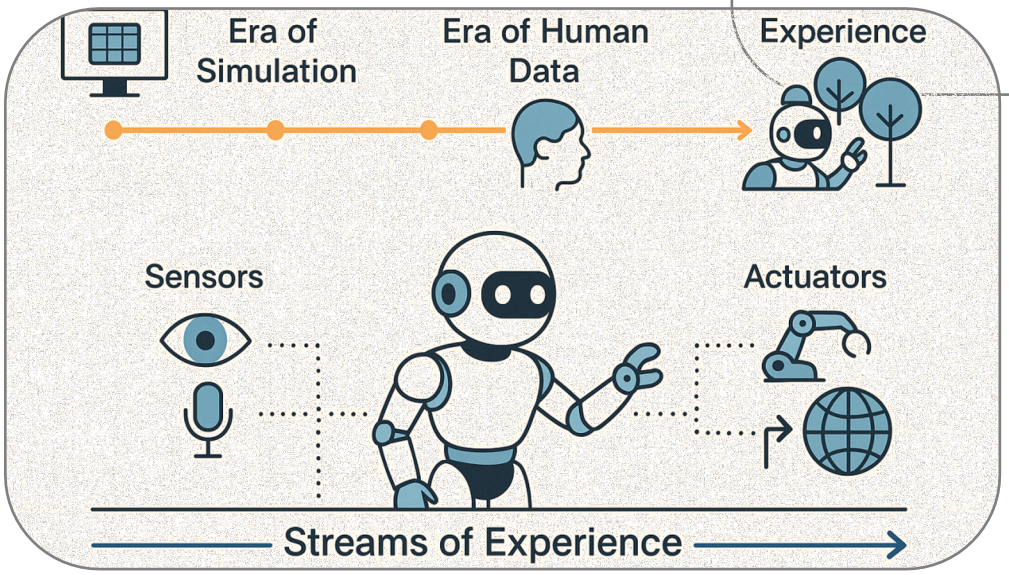
- Agents 应该处于动态变化的环境中，而静态数据或短暂交互
  - TTRL 关注 Test-Time Training，而非静态的训练数据集
- Agents 的奖励应来源于自身与环境，而不是来源于人类
  - TTRL 证明 Majority Voting 是一种自身奖励估计的有效方法

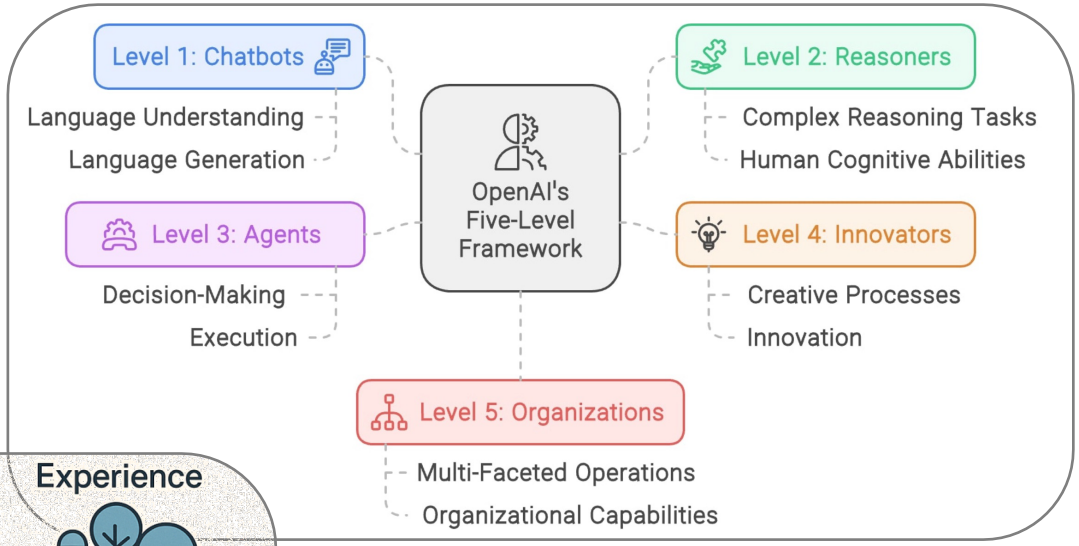
Next：协同与交互

- Agents 与环境如何交互?

构建持续交互的环境反馈
多轮超长记忆的强化学习

……



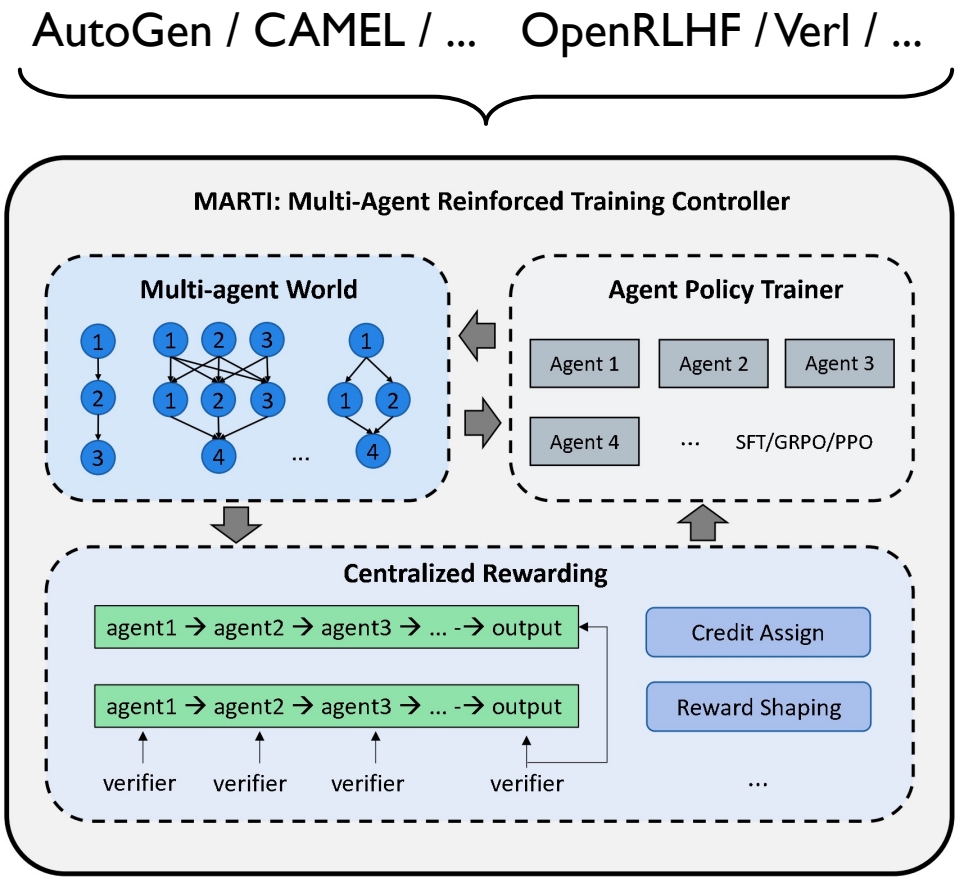


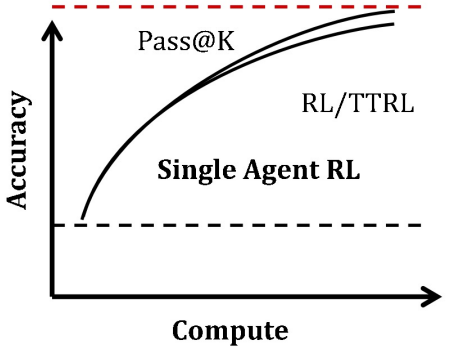
- 多个 Agents 如何协同?

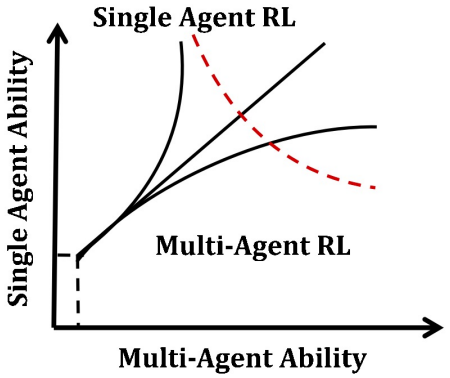
组合不同能力解决复杂问题
协同产生超越单体的经验流

……

Next：多智能体强化训练框架 MARTI







MARTI: A Framework for Multi-Agent LLM Systems Reinforced Training and Inference.

Kaiyan Zhang, Runze Liu, Xuekai Zhu, et al. Work in Progress

Next：多智能体强化训练框架 MARTI

- ✓ 将现有 LLM 直接用于Multi-Agent的协同推理，很多时候会失败 [1]，不如单 Agent TTS

Model	Method	AIME24	AMC	MATH-500	Avg.	Training
Qwen2.5-7B-Instruct	Pass@1	13.3	42.5	75.2	43.7	-
Qwen2.5-14B-Instruct	Pass@1	13.3	60.0	77.2	50.2	-
DeepSeek-R1-Qwen-7B	Pass@1	53.3	87.5	92.0	77.6	-
DeepSeek-R1-Qwen-14B	Pass@1	56.7	92.5	92.0	80.4	-
OpenAI-o1-mini	Pass@1	63.6	92.5	90.0	82.0	-
Qwen2.5-3B	Pass@1	3.3	26.7	53.8	27.9	-
	Maj@4	3.3	37.5	64.0	34.9	-
	Maj@6	6.7	35.0	65.8	35.8	-
	MAD 2x2	3.3	30.0	59.0	30.8	✗
	MoA 3+1	3.3	32.5	62.4	32.7	✗
	CoA 3x1	0.0	35.0	50.6	28.5	✗
Qwen2.5-3B w/ MARTI	MAD 2x2	**16.7**	**49.4**	**70.8**	**45.6**	✓
	MoA 3+1	13.3	47.0	69.0	43.1	✓
	CoA 3x1	13.3	39.7	67.2	40.1	✓
Qwen2.5-3B w/ RL	Pass@1	10.0	36.1	66.7	37.6	✓
	Maj@4	10.0	42.2	69.4	40.5	✓
	Maj@6	13.3	45.4	70.2	42.9	✓
DeepScaleR-1.5B	Pass@1	43.1	73.6	87.8	68.2	-
	Maj@4	46.7	80.0	90.0	72.2	-
	Maj@6	54.0	85.0	90.4	76.5	-
	MAD 3x2	53.3	87.5	90.6	77.1	✗
	MoA 5+1	50.0	82.5	91.2	74.6	✗
	CoA 3x1	53.3	82.5	86.0	73.9	✗
DeepScaleR-1.5B w/ MARTI	MAD 3x2	**66.7**	**87.5**	**92.0**	**82.1**	✓
DeepScaleR-1.5B w/ TTRL	Pass@1	47.3	76.8	89.1	71.1	✓
	Maj@4	54.7	82.2	90.0	75.6	✓
	Maj@6	60.0	84.6	91.2	78.6	✓

[1] Why Do Multi-Agent LLM Systems Fail? Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, et al.

Next：多智能体强化训练框架 MARTI

- ✓ Multi-Agent RL 对比单个 Agent RL，相同推理计算成本，MARTI 训练后能够获得更高的性能上限
- ✓ Multi-Agent 能够更好利用 TTRL

Model	Method	AIME24	AMC	MATH-500	Avg.	Training
Qwen2.5-7B-Instruct	Pass@1	13.3	42.5	75.2	43.7	-
Qwen2.5-14B-Instruct	Pass@1	13.3	60.0	77.2	50.2	-
DeepSeek-R1-Qwen-7B	Pass@1	53.3	87.5	92.0	77.6	-
DeepSeek-R1-Qwen-14B	Pass@1	56.7	92.5	92.0	80.4	-
OpenAI-o1-mini	Pass@1	63.6	92.5	90.0	82.0	-
Qwen2.5-3B	Pass@1	3.3	26.7	53.8	27.9	-
	Maj@4	3.3	37.5	64.0	34.9	-
	Maj@6	6.7	35.0	65.8	35.8	-
	MAD 2x2	3.3	30.0	59.0	30.8	✗
	MoA 3+1	3.3	32.5	62.4	32.7	✗
	CoA 3x1	0.0	35.0	50.6	28.5	✗
Qwen2.5-3B w/ MARTI	MAD 2x2	**16.7**	**49.4**	**70.8**	**45.6**	✓
	MoA 3+1	13.3	47.0	69.0	43.1	✓
	CoA 3x1	13.3	39.7	67.2	40.1	✓
Qwen2.5-3B w/ RL	Pass@1	10.0	36.1	66.7	37.6	✓
	Maj@4	10.0	42.2	69.4	40.5	✓
	Maj@6	13.3	45.4	70.2	42.9	✓
DeepScaleR-1.5B	Pass@1	43.1	73.6	87.8	68.2	-
	Maj@4	46.7	80.0	90.0	72.2	-
	Maj@6	54.0	85.0	90.4	76.5	-
	MAD 3x2	53.3	87.5	90.6	77.1	✗
	MoA 5+1	50.0	82.5	91.2	74.6	✗
	CoA 3x1	53.3	82.5	86.0	73.9	✗
DeepScaleR-1.5B w/ MARTI	MAD 3x2	**66.7**	**87.5**	**92.0**	**82.1**	✓
DeepScaleR-1.5B w/ TTRL	Pass@1	47.3	76.8	89.1	71.1	✓
	Maj@4	54.7	82.2	90.0	75.6	✓
	Maj@6	60.0	84.6	91.2	78.6	✓

Base Model + Zero-Like RL

Large Reasoning Models (LRMs) + TTRL

内容总结

- PART 1: Test-time Scaling (TTS) 与 RL
- PART 2: TTRL: 无标签数据强化学习方法
- PART 3: TTRL 的有效性及局限性讨论
- PART 4: 协同交互视角展望“经验时代”RL

Paper & Project List

- **Reward Model & TTS —— OpenPRM / ImplicitPRM / GenPRM / 1Bvs405B / Video-T1**
  - ✓ *OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees*. **Kaiyan Zhang**, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, Bowen Zhou. ICLR 2025
  - ✓ *Free Process Rewards without Process Labels*. Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, **Kaiyan Zhang**, Bowen Zhou, Zhiyuan Liu, Hao Peng. ICML 2025
  - ✓ *GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning*. Jian Zhao, Runze Liu, **Kaiyan Zhang**, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, Bowen Zhou. UnderReview
  - ✓ *Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling*. Runze Liu, Junqi Gao, Jian Zhao, **Kaiyan Zhang**, Xiu Li, Biqing Qi, Wanli Ouyang, Bowen Zhou. UnderReview
  - ✓ *Video-T1: Test-Time Scaling for Video Generation*. Fangfu Liu, Hanyang Wang, Yimo Cai, **Kaiyan Zhang**, et al. UnderReview
- **RL & Multi-Agent —— TTRL / PRIME /MARTI / AwesomeRL / …**
  - ✓ *TTRL: Test-Time Reinforcement Learning*. Yuxin Zuo*, **Kaiyan Zhang***, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, Bowen Zhou. UnderReview
  - ✓ *Process Reinforcement through Implicit Rewards*. Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, **Kaiyan Zhang**, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, Ning Ding. UnderReview
  - ✓ *Awesome RL Recipes for Reasoning*. **Kaiyan Zhang**, Yuchen Fan, Yuxin Zuo, Guoli Jia, Kai Tian, Xingtai Lv, Xuekai Zhu, Ermo Hua, Ning Ding, Biqing Qi, Bowen Zhou. GitHub
  - ✓ *MARTI: A Framework for Multi-Agent LLM Systems Reinforced Training and Inference*. **Kaiyan Zhang**, Runze Liu, Xuekai Zhu, et al. WIP
  - ✓ More is coming soon …

TTRL Contributors & Our Team

TTRL Contributors:Yuxin Zuo, Kaiyan Zhang(Project Leader), Li Sheng, Shang Qu, Ganqu Cui(Project Leader), Xuekai Zhu, Haozhan Li,Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi,Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding(Corresponding Author), Bowen Zhou(Corresponding Author)

Faculty

Bowen Zhou

Chair Professor

Ning Ding

Assistant Professor

Students

Kaiyan Zhang

Ph.D Student

Xinwei Long

Ph.D Student

Che Jiang

Ph.D Student

Ermo Hua

Ph.D Student

Xuekai Zhu

Ph.D Student

Guoli Jia

Ph.D Student

Xingtai Lv

Ph.D Student

Haozhan Li

Ph.D Student

Shang Qu

Ph.D Student

Siyan Gao

Ph.D Student

Zhekai Chen

Ph.D Student

Kai Tian

M.Sc Student

Zhenzhao Yuan

M.Sc Student

Center for Collaborative & Conversational Intelligence (C3I),Tsinghua University

Website: https://c3i.ee.tsinghua.edu.cn/

课题组长期招收博士后和访问学生，咨询邮箱：wudi1@tsinghua.edu.cn

Q&A
THANKS

zhang-ky22@mails.tsinghua.edu.cn

欢迎合作与讨论交流

研究主题包括不限于
Test-Time Scaling, Reward Models
RL Reasoning, Multi-Agent, ...

个人主页： https://iseesaw.github.io/

课题组公众号