rSIM: Incentivizing Reasoning Capabilities of LLMs via Reinforced Strategy Injection

rSIM：通过强化学习注入策略提升大型语言模型的推理能力

arXiv:2512.08300v1

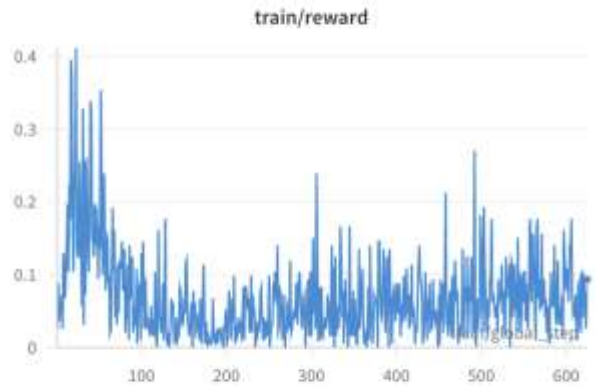
The Hong Kong University of Science and Technology (Guangzhou)， University of Toronto, University of Alberta

汇报人：陆羿　　时间：2026-01-06

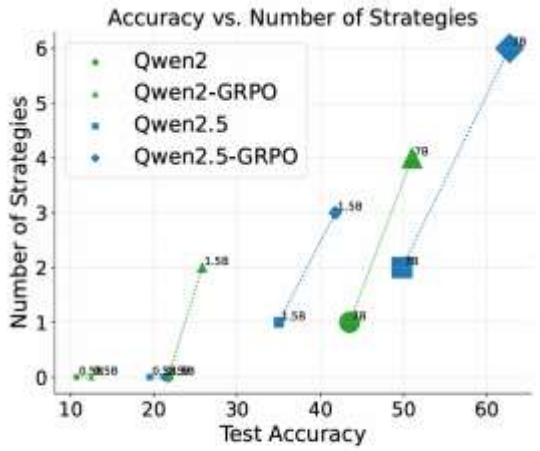
1. 相关工作

数据蒸馏
通过蒸馏思维链样本，大型 LLM 可将其逐步推理能力迁移到小型模型 ——LLM 生成的推理轨迹可作为额外的微调数据。
本文中，将 DPR 框架中的规划器作为插件以提升小型模型的推理能力，可视为向其传递人类级别的规划链。
强化学习
强化学习（RL）已广泛应用于决策任务；RLHF首次利用 PPO使模型对齐人类偏好，ReFT开创了将 RL 作为微调范式以提升 LLM 推理性能的先河。然而，模型探索自身能力之外的思维链解决方案的能力存在固有局限。因此，像 0.5B 这样的弱基础模型无法从 RL 训练中获益，推理性能仍处于落后水平。
多智能体 LLM
基于 MARL的多智能体 LLM在复杂问题解决中取得了显著进展。
为提升性能，SOCRATIC训练两个小型蒸馏模型的组合，以在 LLM 中执行 CoT 推理。与本文的 DPR 框架不同，SOCRATIC 仍依赖于将大型模型的能力蒸馏到小型模型中。
ReAct使 LLM 能够以交错方式生成推理轨迹和任务特定动作。
CORY将 LLM 微调为两个自主智能体（pioneer智能体和observer智能体），其性能优于标准 PPO。

2. 研究动机



(a) GRPO-based training failure on Qwen2.5-0.5B



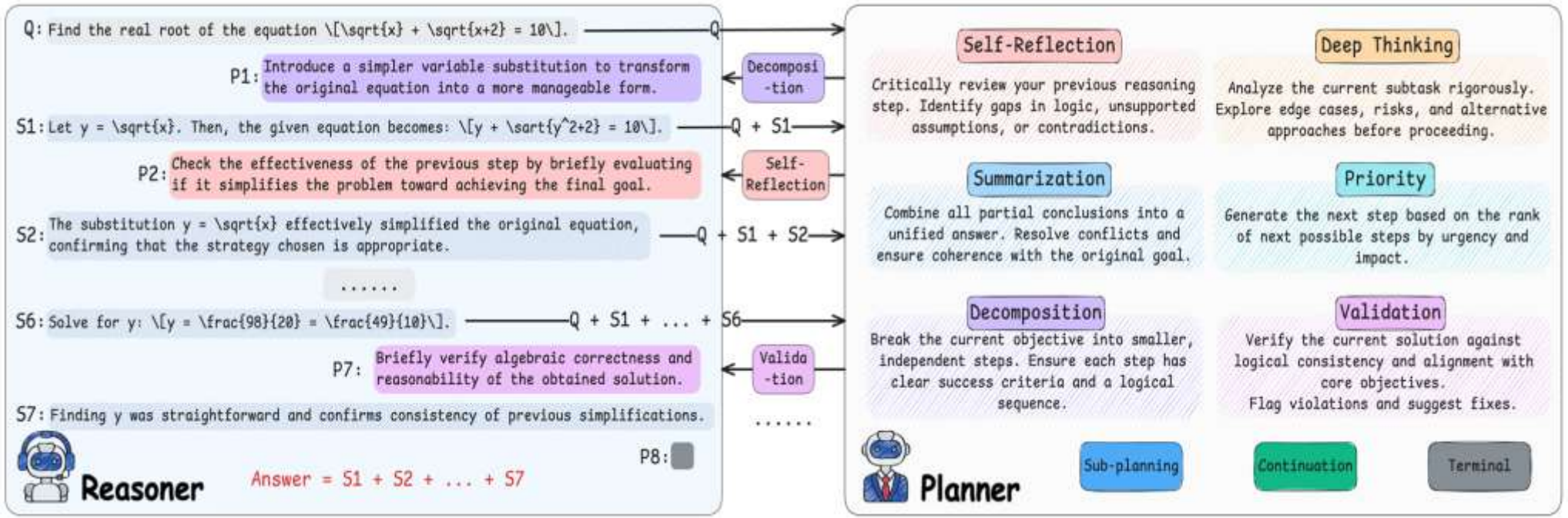
(b) Accuracy vs. number of strategies across LLMs

强化策略注入机制（rSIM），通过一个小型 LLM 实现的规划器（planner），在思维链过程中自适应提供策略指令，以指导另一个 LLM。规划器可融入丰富的人类设计知识和先验信息（strategies），帮助 LLM 有效且直接地获得高级推理能力，逐步进化为 RLM。

当以 Qwen2.5-0.5B 为基础模型时，GRPO 训练下的总奖励最初上升至约 0.3，随后突然降至 0；Qwen2-0.5B、Qwen2.5-0.5B 等缺乏固有推理策略（策略计数为 0）的基础模型，无法通过 RL 训练获得推理智能，准确率仅实现有限提升；而 1.5B、7B 等具备固有推理策略（策略计数大于 0）的模型，可通过 RL 进一步优化；

3. 方法

双智能体框架的训练目标



- 规划器旨在在每个推理步骤中指导推理器，从 9 种人类设计的预定义策略集合（如图所示）中选择一种策略（以prompt形式呈现）
- 在该框架中，基础 LLM 作为推理器，负责根据规划器选择的策略生成下一步推理内容。

3. 方法

两阶段训练方案

规划器奖励$R^P=R_{acc}+R_{penalty}+R_{terminal}$
推理器奖励$R=R_{acc}+R_{format}+R_{follow}$

$R_{terminal}$ 若最终策略为 “终止”策略，则为 1，否则为 – 1
$R_{penalty}$最常选择策略的占比
R_{follow}遵循给定规划的推理步骤占比
R_{acc}和R_{format}分别为 GRPO中定义的准确率奖励和格式奖励

$$\boldsymbol{o}^{dpr} = [\boldsymbol{p}_1, \boldsymbol{z}_1, \boldsymbol{p}_2, \boldsymbol{z}_2, \ldots, \boldsymbol{p}_n, \boldsymbol{z}_n],$$

对于一个问题，两个智能体交互n轮形成的序列
P为planner选择的策略，z为reasoner执行步骤

$\mathcal{J}_{\boldsymbol{o}^{dpr}}$:

$$\frac{1}{|\boldsymbol{o}^{dpr}|}\sum_{t=1}^{|\boldsymbol{o}^{dpr}|}\left[\lambda\cdot\left(\frac{\pi^p_{\boldsymbol{\phi}}(\boldsymbol{a}^p_t|\boldsymbol{s}_t)}{\pi^p_{\boldsymbol{\phi}_{\text{old}}}(\boldsymbol{a}^p_t|\boldsymbol{s}_t)}\right)\cdot A^{\pi^p_{\boldsymbol{\phi}}}(\boldsymbol{s}_t,\boldsymbol{a}^p_t)\right.$$
$$\left.+(1-\lambda)\cdot\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|\boldsymbol{s}_t,\boldsymbol{a}^p_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|\boldsymbol{s}_t,\boldsymbol{a}^p_t)}\right)\cdot A^{\pi_{\boldsymbol{\theta}}}(\boldsymbol{s}_t,a_t,\boldsymbol{a}^p_t)\right]$$

规划器与推理器分别的优势函数

3. 方法

两阶段训练方案

- 策略更新冲突 —— 两个智能体的梯度可能朝相反方向 “拉扯”
- 信用分配模糊 —— 难以区分成功或失败源于领导者的规划还是追随者的执行
- 探索与利用的竞争 —— 联合探索可能导致灾难性的协同失效

为实现有效训练，提出两阶段方案：第一阶段优先优化规划器的策略，第二阶段将重点转向优化推理器（即基础模型）的策略。具体而言，我们通过调整权重参数λ，确保训练稳定。
第一阶段设置λ=0.7，强调规划器优化
第二阶段设置λ=0.3，强调推理器优化

4. 实验

Baseline：

- 近期提出的 GRPO 方法
- 规划 - 求解（Plan-and-Solve, PS+）提示策略
- 规划器提示（Planner Prompting）

数据集：

实验在 HuggingFace 平台的 7 个数据集上开展，涵盖三类任务：

- 数学任务：MATH、GSM8K、AIME2024
- 多任务推理：MMLU-Pro、TheoremQA
- 代码生成：CodeAlpaca-20k、HumanEval

4. 实验

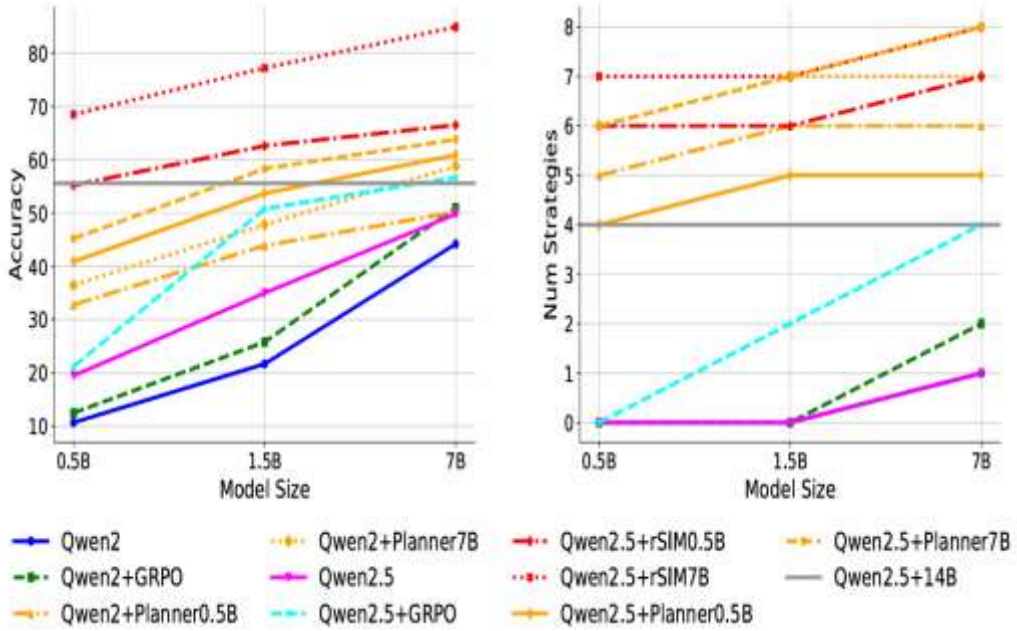
Llama3.2 Models	Planner	MATH		GSM8K		MMLU-Pro		TheoremQA	
		Score	#Strategy	Score	#Strategy	Score	#Strategy	Score	#Strategy
1B w/ ZeroCoT	No	30.6	0	44.4	0	21.2	0	13.7	0
1B w/ PS+ [4]	No	28.2	0	43.7	0	19.4	0	12.2	0
1B w/ prompting	3B	27.4	7	42.6	3	16.8	8	6.6	6
1B w/ prompting	70B	33.3	5	46.9	3	22	6	14.3	5
3B w/ ZeroCoT	No	48	0	77.7	0	30.1	0	20.7	0
3B w/ PS+ [4]	No	47.5	0	77.7	0	30	0	18.6	0
3B w/ prompting	3B	46.4	7	77.1	5	28.5	7	19.9	8
3B w/ prompting	70B	55.5	7	81.8	4	31.8	5	22.8	6
1B w/ GRPO	No	-	0	-	0	×	0	×	0
1B w/ *rSIM*	1B	57	3	83.9	1	30.8	4	20.9	3
1B w/ *rSIM*	3B	61.5	4	86.3	3	33	6	25	6
1B w/ *rSIM*	Qwen2.5-1.5B	59.1	4	84.4	1	31.8	4	24.2	5
Llama3.3 70B w/ ZeroCoT	No	77	0	90.5	0	68.9	0	32.3	0
Llama3.3 70B w/ PS [4]	No	78.3	0	90.9	0	70	0	32	0
Llama3.3 70B w/ Prompting	3B	79.1	7	[illegible]	4	68.9	7	31.9	8
Llama3.3 70B w/ Prompting	70B	84	6	92.9	4	71.5	4	38.6	5
Llama3.3 70B	1B plug-in	83.2	3	91.7	1	71.8	6	39	5
Llama3.3 70B	3B plug-in	86.3	4	92.1	2	72.7	5	41.8	6
Llama3.3 70B	Qwen2.5-1.5B	83.7	4	92	2	72.3	6	40.7	6

将 rSIM 应用于 Llama 等其他类型的 LLM 时，得益于规划器提供的策略注入，模型性能仍能实现稳定提升

两类 Llama 模型的推理性能均得到提升，证明 rSIM 规划器具备跨模型泛化能力

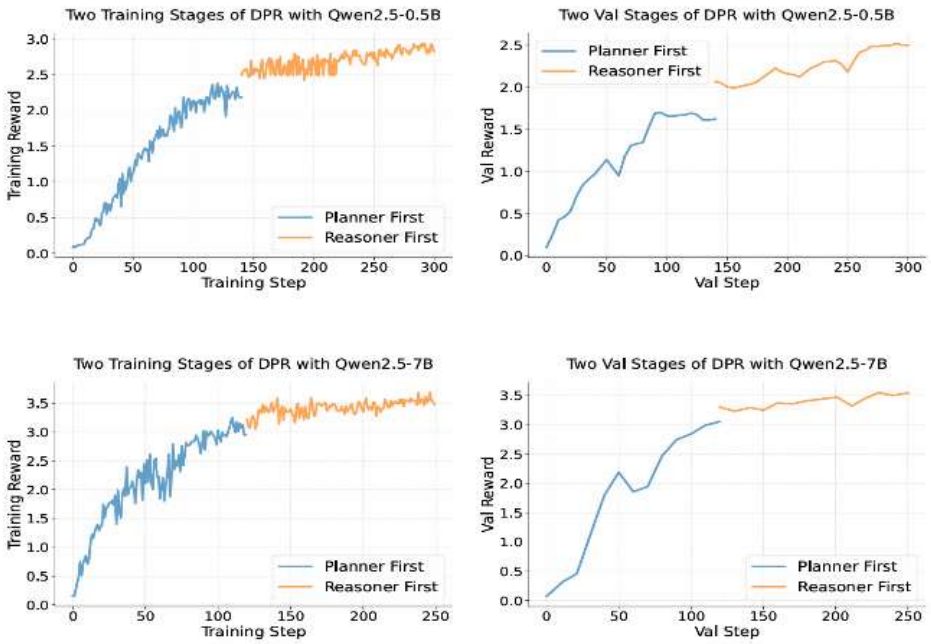
planner列标注 Qwen2.5 等模型时，代表跨模型评估（Llama 作为推理器，Qwen2.5 作为规划器）；
planner列标注 “plug-in” 时，代表直接使用训练好的规划器指导推理器，无需额外训练；
Models列标注 “w/prompting” 代表直接通过提示词让 LLM 扮演规划器；
“-” 表示训练未收敛； “×” 表示因模型无法训练导致结果缺失。

4. 实验

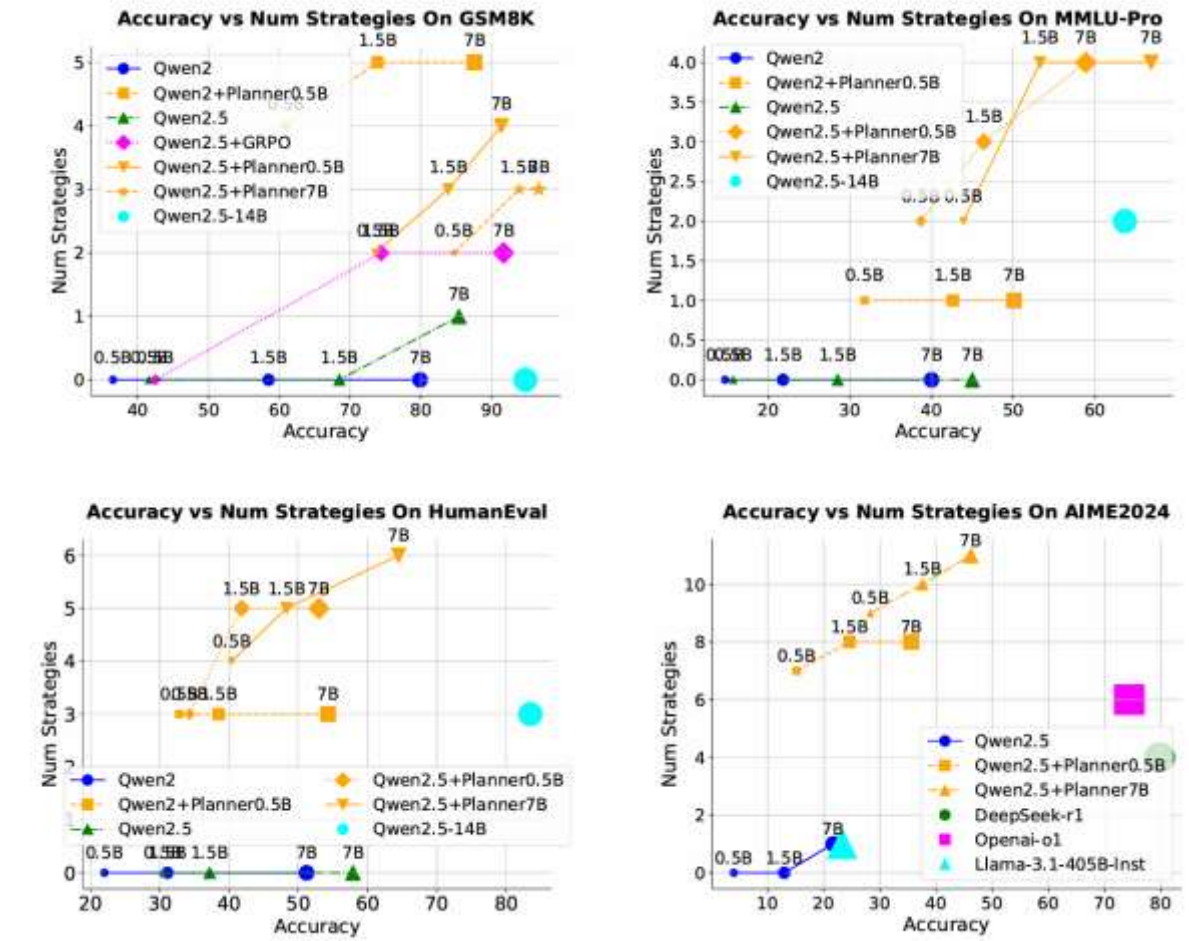


训练曲线（下图）和较高的解决问题的准确率（左图）证明：rSIM 能够使任何 LLM（尤其是小型模型）训练至收敛，推理性能实现显著提升。

经 rSIM 训练后，所有基础模型均能执行 9 种预定义的人类级推理策略（Qwen2.5-0.5B 这类最初无明显策略使用的弱基础模型）



4. 实验



不同规模模型在各数据集上的准确率示意图（采用训练后的 rSIM 规划器作为插件）。标注 “+Planner” 的条目表示基础模型在推理过程中与训练后的规划器协同工作。

- 基础模型集成预训练规划器后，准确率大幅提升，且常优于更大规模的模型
- 训练好的规划器能使任何 LLM 将有效推理策略融入推理过程（对于AIME2024，使 LLM 平均每个问题应用人类级策略超过 8 次）

4. 实验

Methods	Planner	MATH	TheoremQA
0.5B w/ ZeroCoT	No	221.6 ± 172.5	250.3 ± 110.2
14B w/ ZeroCoT	No	261.8 ± 192.2	308.7 ± 137.5
0.5B w/ PS+ [4]	No	327.5 ± 176.7	367.5 ± 153.6
0.5B w/ Prompting	14B	815.2 ± 356.7	993.4 ± 390.5
0.5B w/ *rSIM*	7B	780.3 ± 210.9	800.7 ± 230.2
7B w/ ZeroCoT	No	246.9 ± 189.5	291.2 ± 160.8
7B w/ PS [4]	No	357.2 ± 200.9	390.6 ± 190
7B w/ Prompting	14B	934.5 ± 390.2	1103 ± 487.5
7B w/ *rSIM*	7B	900.6 ± 350.8	970.2 ± 427.1

rSIM 引入的规划器在辅助 LLM 推理时，并未显著增加 token 消耗成本。

Thanks for Your Watching !

rSIM：通过强化学习注入策略提升大型语言模型的推理能力

arXiv:2512.08300v1

The Hong Kong University of Science and Technology (Guangzhou)， University of Toronto, University of Alberta

汇报人：陆羿　　时间：2026-01-06