NEURAL INFORMATION PROCESSING SYSTEMS

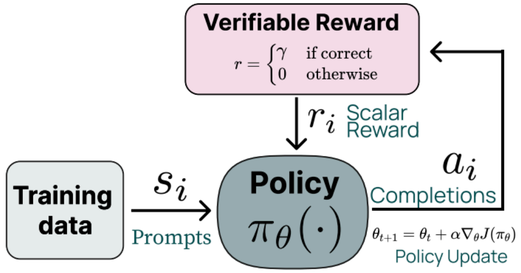
Trust, But Verify: A Self-Verification Approach to Reinforcement Learning with Verifiable Rewards

Xiaoyuan Liu[1,2] **Tian Liang**[2] **Zhiwei He**[2,3] **Jiahao Xu**[2] **Wenxuan Wang**[4]
Pinjia He[1†] **Zhaopeng Tu**[2†] **Haitao Mi**[2] **Dong Yu**[2]

[1]The Chinese University of Hong Kong, Shenzhen [2]Tencent
[3]Shanghai Jiao Tong University [4]Renmin University of China

Speaker: Xiaoyuan Liu

Motivation

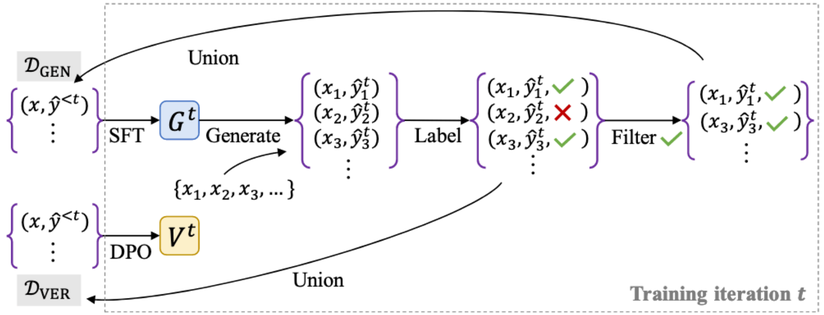


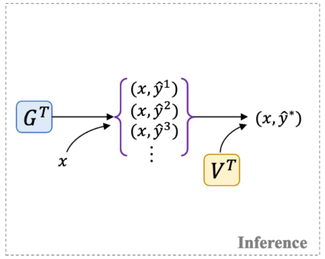
RLVR in Tülu 3 (2025)

$$v(x, y) = \begin{cases} \alpha & \text{if correct,} \\ 0 & \text{otherwise.} \end{cases}$$

Problem	What is the least positive integer multiple of 30 that can be written with only the digits 0 and 2?
Zero-RL	To find the least positive integer multiple of 30 that can be written with only the digits 0 and 2, we need to consider the properties of multiples of 30 . . . Since 222 is divisible by both 2 and 3, it is a multiple of 30. **Let's verify:** 222 is even, so it is divisible by 2. The sum of the digits of 222 is $2+2+2=6$, which is divisible by 3. Thus, 222 is indeed a multiple of 30. Therefore, the least positive integer multiple of 30 that can be written with only the digits 0 and 2 is $\boxed{222}$.

RLVR-trained Qwen2.5-7B



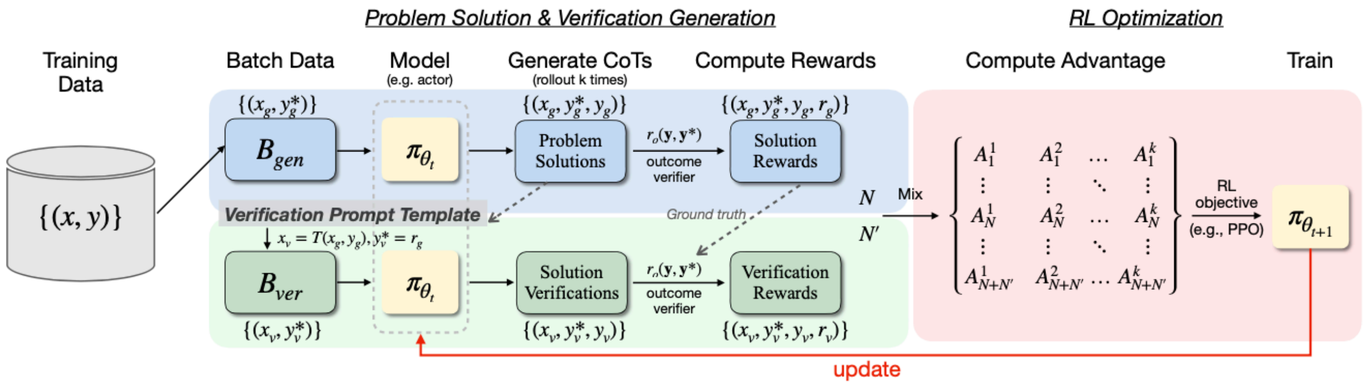


V-STaR (2024)

- RLVR improves LLM reasoning, but not robustly: **LLMs lack a notion of correctness.**
  - Lack of self-verification, or "Superficial Self-Reflection" (Liu et al., 2025)
- This "gap" is usually covered by a **verifier** model, which is often trained and used **separately.**
- Can we teach LLMs to verify their own solutions while reasoning?

Methodology

RISE (**R**einforcing Reason**I**ng with **S**elf-V**E**rification)



Prompt Template

Below you are presented with a question and a tentative response. Your task is to evaluate and assign a rating to the response based on the following clear criteria:

Rating Criteria:

1. Missing final answer enclosed in \\boxed{} at the end: assign \\boxed{-1}.
2. Correct response with the final answer enclosed in \\boxed{} at the end: assign \\boxed{1}.
3. Incorrect response with the final answer enclosed in \\boxed at the end: assign \\boxed{-0.5}.

Question Begin
{**Question**}
Question End

Response Begin
{**Response**}
Response End

Briefly summarize your analysis, then clearly state your final rating value enclosed in \\boxed{} at the end.

PPO-based $$\mathcal{J}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) - \beta KL\left(\pi_\theta || \pi_{ref}\right) \right],$$

GAE $$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t-1}\delta_{T-1},$$
$$\text{where } \delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t).$$

Training Process:

1. Sample a batch of problems and rollout their chain-of-thought solutions. (gt=correct ans)
2. Reconstruct the problems and their solutions as verification problems using a predefined template.
3. Sample a batch of verification problems and rollout their chain-of-thought solutions. (gt=sol reward)
4. Jointly optimize both reasoning and verification via RL.

Experiment

Training dataset: MATH-Hard (Level 3-5), 8.5k problems for Qwen2.5; DeepMath-10K subset for Qwen3
Benchmarks: MATH500, AIME2024, AMC2023, Minerva Math, OlympiadBench

	Reasoning						*Self-Verification*					
Model	**MATH**	**AIME**	**AMC**	**Mine.**	**Olym.**	**Avg.**	**MATH**	**AIME**	**AMC**	**Mine.**	**Olym.**	**Avg.**
GPT-4o	79.0	13.3	55.0	50.0	42.5	48.0	83.4	33.3	67.5	50.4	54.4	57.8
Qwen2.5-1.5B												
Base	2.0	0.0	1.9	0.8	0.6	1.1	19.4	21.9	22.7	15.9	21.1	20.2
Instruct	37.5	0.8	19.4	8.3	11.7	15.5	48.8	22.1	36.5	36.9	29.6	34.8
SFT	10.1	0.0	4.1	1.8	2.0	3.6	19.0	5.8	12.3	10.5	10.9	11.7
Zero-RL	**55.3**	2.1	25.9	**17.4**	19.5	24.0	54.1	5.0	30.7	21.0	23.0	26.8
RISE	54.6	**2.9**	**27.5**	17.2	**19.8**	**24.4**	**75.9**	**85.0**	**70.6**	**66.0**	**74.9**	**74.5**
Qwen2.5-3B												
Base	32.7	1.3	15.3	10.3	10.7	14.1	39.5	13.6	22.5	29.9	21.2	25.3
Instruct	61.0	3.8	34.1	25.6	24.6	29.8	65.6	21.0	45.5	37.6	35.0	40.9
SFT	14.4	0.4	5.3	2.9	2.8	5.2	21.5	2.1	10.9	17.9	13.2	13.1
Zero-RL	64.2	6.7	37.5	**27.4**	**26.6**	32.5	64.9	13.0	39.7	30.3	31.2	35.8
RISE	**64.3**	**7.9**	**42.5**	26.2	**26.6**	**33.5**	**81.0**	**86.3**	**74.4**	**56.1**	**73.6**	**74.3**
Qwen2.5-7B												
Base	38.3	2.1	21.9	11.9	13.2	17.5	58.4	45.9	51.5	48.4	48.4	50.5
Instruct	73.8	10.0	50.6	**35.9**	35.8	41.2	77.2	26.3	57.0	40.2	45.2	49.2
SFT	28.7	0.8	13.8	6.2	7.2	11.3	40.5	36.6	47.4	39.2	36.1	40.0
Zero-RL	74.5	12.1	51.3	34.2	**36.7**	41.7	75.9	21.7	56.5	37.3	41.6	46.6
RISE	**74.8**	**12.5**	**55.9**	34.6	**36.7**	**42.9**	**83.8**	**75.0**	**72.5**	**48.6**	**65.9**	**69.2**

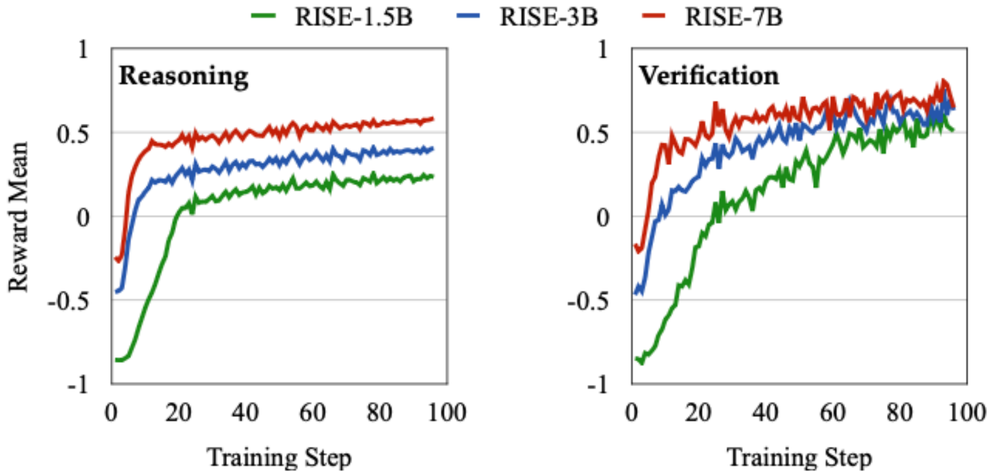
Qwen2.5 Models

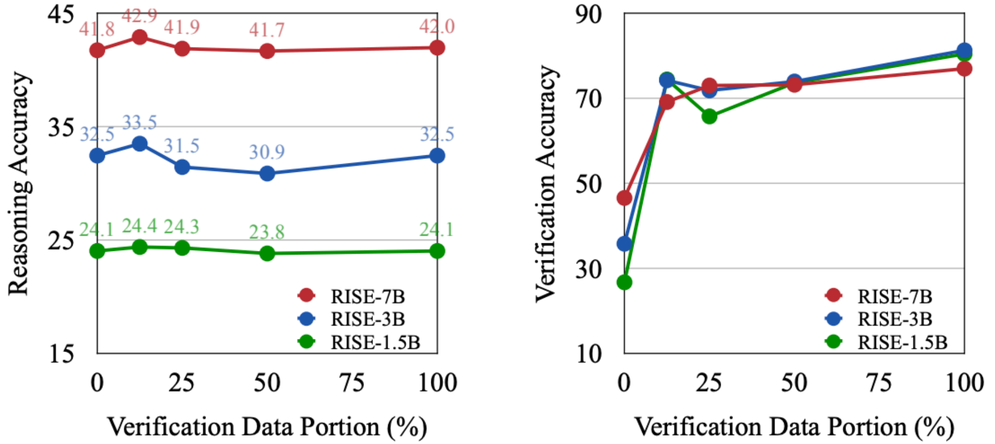
	Reasoning						*Self-Verification*					
Model	**MATH**	**AIME**	**AMC**	**Mine.**	**Olym.**	**Avg.**	**MATH**	**AIME**	**AMC**	**Mine.**	**Olym.**	**Avg.**
Qwen3-4B-Base												
Base	39.4	6.3	24.1	12.6	17.8	20.0	60.9	72.6	61.9	**59.4**	63.8	63.7
Zero-RL	73.7	**13.3**	45.9	29.5	37.2	39.9	73.7	39.9	52.9	37.9	47.8	50.4
RISE	**77.8**	12.9	**52.8**	**43.4**	**40.6**	**45.5**	**87.4**	**79.6**	**70.9**	50.8	**68.0**	**71.3**
Qwen3-8B-Base												
Base	42.5	8.3	28.4	15.4	18.4	22.6	67.0	65.5	64.4	**62.9**	62.0	64.4
Zero-RL	77.6	13.8	58.1	37.7	41.6	45.7	79.7	54.1	68.8	46.9	56.9	61.3
RISE	**83.0**	**21.3**	**59.4**	**48.4**	**44.4**	**51.3**	**91.8**	**85.4**	**87.4**	53.4	**72.2**	**78.1**

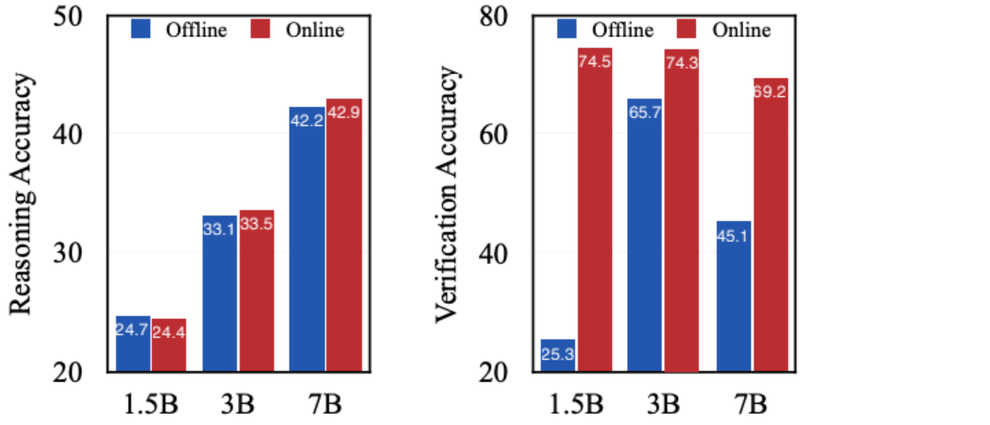
Qwen3 Models

- RISE enhances self-verification capabilities while simultaneously improving reasoning performance.
- Verification improvements outpace problem-solving gains.

Analysis I: Learning Dynamics of RISE







1. Self-verification improves faster than problem solving (First row, Left).
2. Verification scales with training compute (First Row, Right).
3. Online learning is key to self-verification (Second Row).

Analysis II: Advantages of RISE

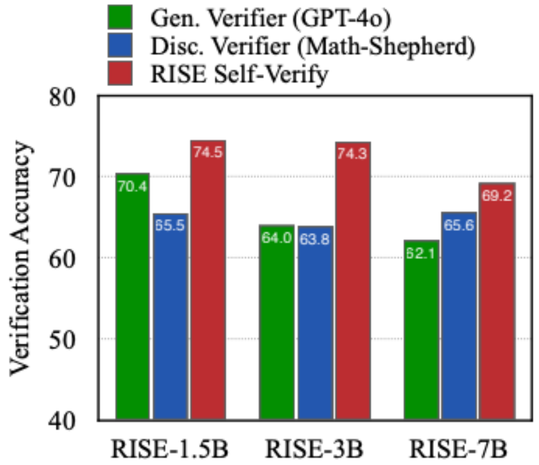
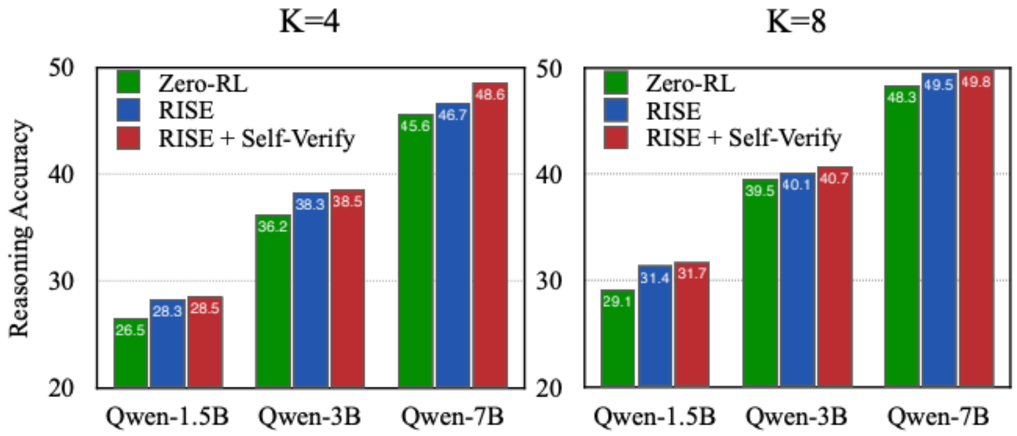


Table 10: Verification performance of RISE models and other verifiers on the external solution set.

Verifier Model	Verification Accuracy
RISE-1.5B	67.9
RISE-3B	74.4
RISE-7B	70.7
GPT-4o	57.8
Math-Shepherd	58.9

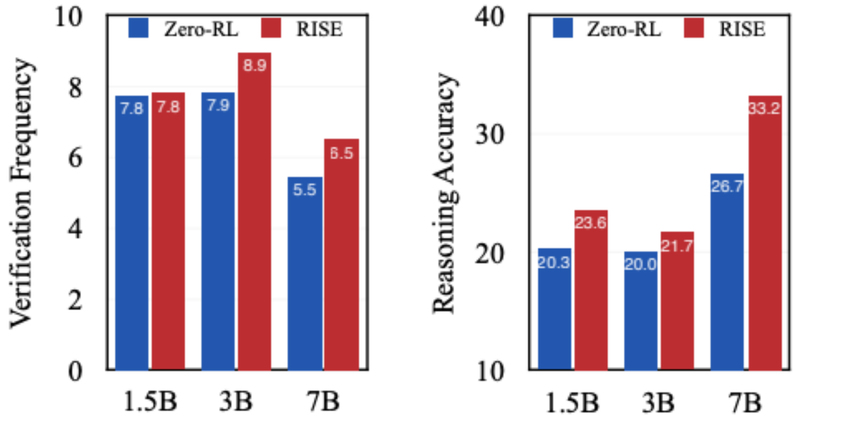


4. **RISE models outperform off-the-shelf verifiers** on both self- and external-solution verification.
5. **RISE enhances test-time scaling performance** through self-consistency and weighted majority voting (RISE + Self-Verify uses verification scores as weights).

Analysis II: Advantages of RISE

Problem	What is the least positive integer multiple of 30 that can be written with only the digits 0 and 2?
Zero-RL	To find the least positive integer multiple of 30 that can be written with only the digits 0 and 2, we need to consider the properties of multiples of 30 ... Since 222 is divisible by both 2 and 3, it is a multiple of 30. **Let's verify:** 222 is even, so it is divisible by 2. The sum of the digits of 222 is $2+2+2=6$, which is divisible by 3. Thus, 222 is indeed a multiple of 30. Therefore, the least positive integer multiple of 30 that can be written with only the digits 0 and 2 is $\boxed{222}$.
RISE	To find the least positive integer multiple of 30 that can be written with only the digits 0 and 2, we need to ... Therefore, the smallest number N is 2220. **To verify:** 2220 ends in 0, so it is a multiple of 10. The sum of the digits of 2220 is $2+2+2+0=6$, which is a multiple of 3. Since 2220 satisfies both conditions, it is the least positive integer multiple of 30 that can be written with only the digits 0 and 2. Thus, the final answer is $\boxed{2220}$.

{"verify", "verifying", "recheck", "validate", "re-evaluate"}



Verification Frequency and Self-verified Reasoning Accuracy

6. **RISE models produce more frequent verifications**, and self-verified responses achieve **higher accuracy.**
7. **RISE enhances verification robustness**, with previously incorrect solutions becoming correct.

Together, RISE helps internalize the notion of correctness within the policy, which can be leveraged both internally and externally.

Thank you for listening! 🙂

Paper Link: https://openreview.net/pdf?id=gA3fFAEXNT

Code Repo: https://github.com/xyliu-cs/RISE