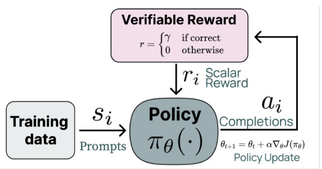
Trust, But Verify: A Self-Verification Approach to Reinforcement Learning with Verifiable Rewards

Xiaoyuan Liu, Tian Liang, Zhiwei He, Jiahao Xu, Wenxuan Wang, Pinjia He, Zhaopeng Tu,Haitao Mi, Dong Yu

Motivation: LLM lacks a notion of correctness



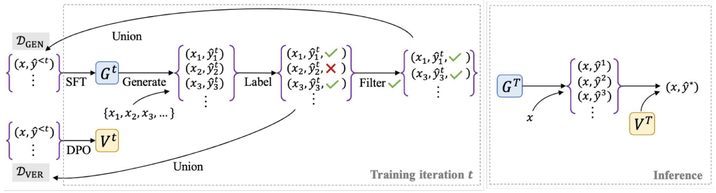
RLVR in Tülu 3 (2025)

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

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

- RLVR improves LLM reasoning, but **not robustly**, which can result in *flawed reasoning* and *suboptimal performance*
  - Lack of self-verification
  - "Superficial Self-Reflection (SSR)" (Liu et al., 2025)



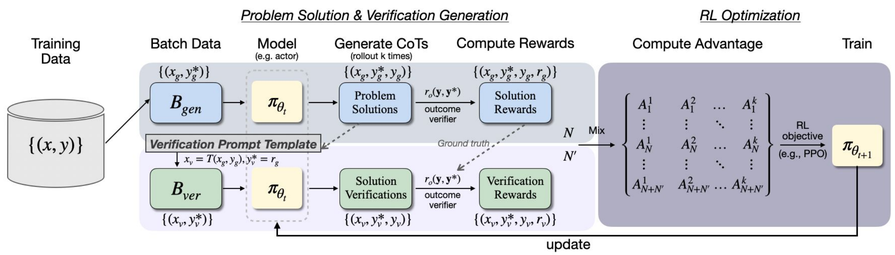
V-STaR (2024)

- This "cognitive gap" is usually covered by a verifier model, which is often trained and used separately
- **Question: Can we teach LLMs to verify their own solutions to improve the correctness of their reasoning?**

Conclusion

To address the above gap, we propose an RLVR training method, *RISE*, that jointly optimizes reasoning and self-verification capability on-the-fly. It internalizes the notion of correctness within the policy, which can be leveraged both internally (self-verified reasoning) and externally (test-time verification).

Method: RISE (Reinforcing Reasoning with Self-Verification)



RISE 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.

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

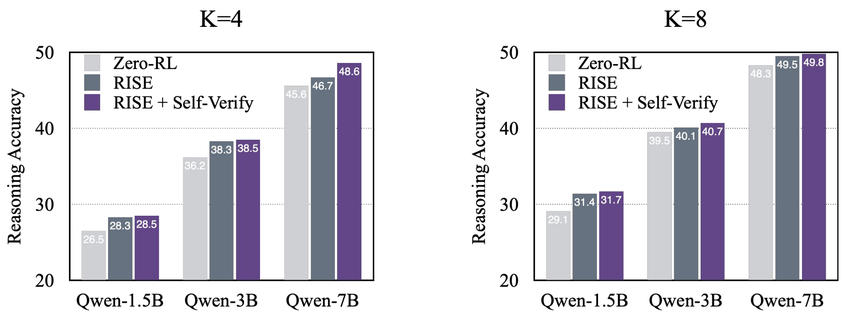
Briefly summarize your analysis, then clearly state your final rating value enclosed in \\boxed{} at the end.

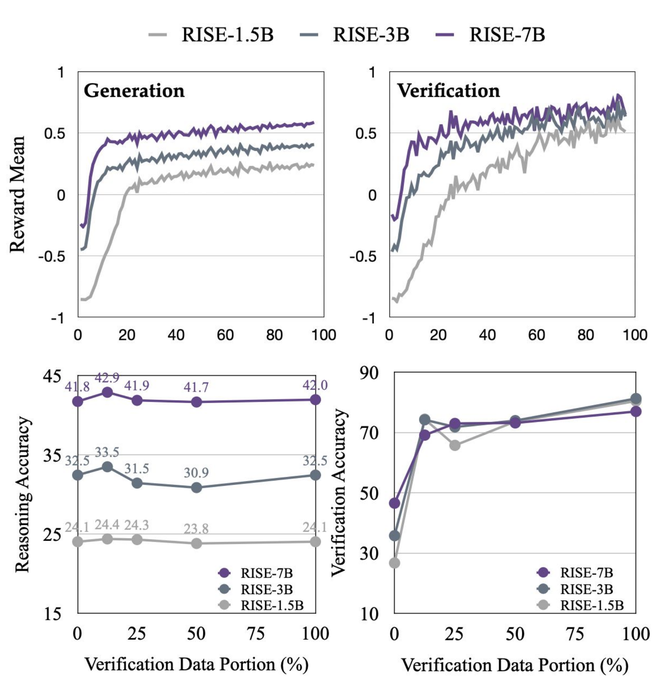
Self-Verification Prompt Template

Results and Analysis: Self-Verification Boosts Reasoning

	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
Qwen2.5-7B												
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**
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**

(RISE-1.5B and RISE-3B results can be found in the paper)





- Learning to self-verify further improves reasoning
- Self-verification improves faster than problem-solving and scales well with more training compute
- RISE enhances test-time scaling performance through self-consistency and weighted majority voting