Step-wise Rubric Rewards for LLM Reasoning

SRaR assigns rubric feedback to the reasoning steps that earn it.

Weichu Xie[1,2,*] · Haozhe Zhao[1,*] · Wenpu Liu[1,2,*] · Yongfu Zhu[2] · Liang Chen[1] · Minghao Ye[3] · Zirong Chen[4] · Yuqi Xu[1,2] · Shuai Dong[5,2] · Ziyue Wang[1,2] · Xinbo Xu[1] · Kean Shi[1] · Ruoyu Wu[1] · Xiaoying Zhang[2] · Wenqi Shao[2] · Baobao Chang[1,†] · Nan Duan[2] · Jiaqi Wang[2]

[1] Peking University [2] JD Explore Academy [3] Shanghai Jiao Tong University [4] Tsinghua University [5] Shanghai Innovation Institute

* Equal contribution † Corresponding author: Baobao Chang

NeurIPS 2026

REASONING · RLVR

1 One scalar rewards the wrong steps

Outcome rewards supervise the final answer. Response-level rubric rewards also give **every step the same signal**, even when its reasoning is wrong.

18.2%	49.9%	1,000
Wrong steps in correct answers	Correct steps in wrong answers	DAPO solutions diagnosed

2 A rubric keeps three kinds of evidence

- **SUGGEST:** reward standard reasoning and explicit derivations.
- **PITFALL:** penalize observed arithmetic or logical errors.
- **BONUS:** credit useful alternative paths and cross-validation.

Budgets: **+0.8 / −1.0 / +1.0**. Divide each budget by its category's item count. Unmet criteria receive zero reward.

A triggered PITFALL is an error to penalize.

3 Build rubrics from contrasting paths

GPT-5 paths on DAPO yield a **16K-problem corpus**. Rubrics contrast correct and flawed solutions after verifying their final answers against ground truth.

Each rubric is problem-specific. Criteria grounded in observed failure modes target errors the policy is likely to make during training, rather than relying on a single reference answer.

SAMPLE Reasoning paths → VERIFY Answer matches → DISTILL Typed criteria

4 Less repetitive self-correction

AIME 2025	RaR	SRaR
Loop rate (%) ↓	**48.1**	**26.5**
Mean self-correction phrases ↓	25.4	14.8

Average count of self-correction phrases per generated response, including expressions such as "Wait, actually..." or "Wait, no...".

Table 3. 960 rollouts per model. Loop detection uses correction phrases, repeated headings and duplicate paragraphs.

5 SRaR: rubric item → step → tokens

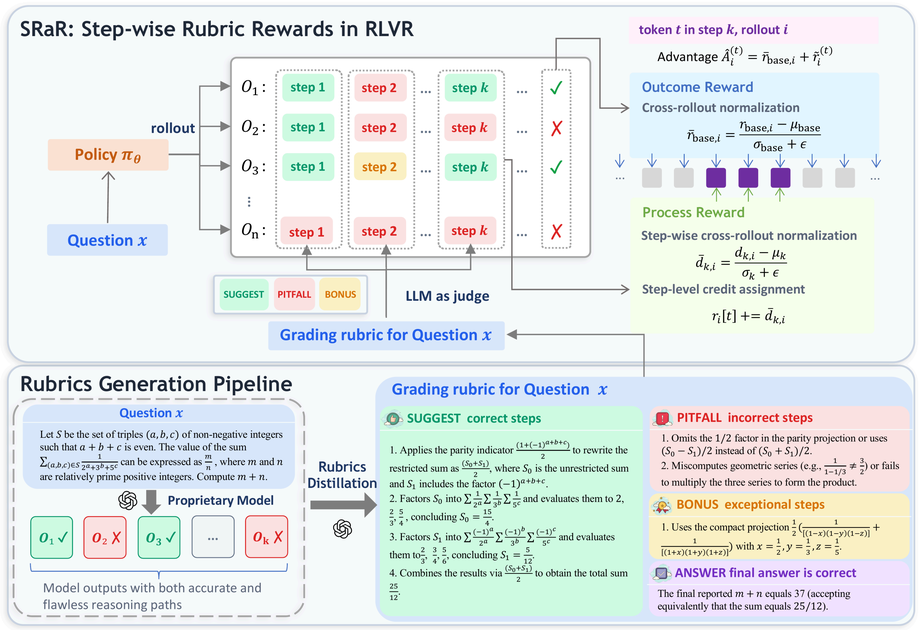


Figure 1. Step-attributed judging and contrastive rubric distillation, reproduced from the paper.

GPT-OSS-20B judges each criterion and identifies its target step. Tokenizer offsets map that step to its tokens.

6 Normalize steps; keep the base separate

Sum rubric deltas by step. Compare each step only across rollouts of its prompt that have criteria attributed to it.

GATHER THE CRITERION DELTAS FOR ONE STEP

$$d_{k,i} = \sum_{j:k_j=k} \delta_j$$

STEP-SPECIFIC RELATIVE CREDIT

$$\bar{d}_{k,i} = \frac{d_{k,i} - \mu_k}{\sigma_k + \epsilon}$$

Set $\bar{d}_{k,i} = 0$ if at most one rollout has attributed items. Scores that do not vary across the group produce zero signal.

FOR EACH TOKEN T IN STEP K

$$\widehat{A}_i^{(t)} = \underbrace{\frac{r_{\text{base},i} - \mu_{\text{base}}}{\sigma_{\text{base}} + \epsilon}}_{\text{outcome / format}} + \underbrace{\bar{d}_{k,i}}_{\text{step rubric}}$$

$r_{\text{base}} = 0.9\, r_{\text{acc}} + 0.1\, r_{\text{fmt}}$. Rubric scores adjust their own step's tokens and stay out of the base reward's group statistics.

Train Qwen3-8B / 32B Non-Thinking on **8 × NVIDIA H200**, using step headers, a boxed answer and clipped PPO updates.

Use the resulting token advantages in the clipped PPO update.

7 Higher mean accuracy at both scales

Mean accuracy (%)	8B	32B
GRPO	58.54	61.65
GRPO-VPS	61.68	61.40
RGR-GRPO	63.24	67.73
DAPO	67.26	72.31
RaR	68.96	73.06
SRaR	**72.53**	**75.81**
Gain over RaR (pp)	**+3.57**	**+2.75**

Table 1. Qwen3 Non-Thinking. AIME24/25 and AMC23: Avg@32. MATH500, Minerva Math and Olympiad Bench: single generation.

8 Correct answers with correct reasoning

AIME 2025 (%)	RaR	SRaR
Faithful reasoning ↑	**34.5**	**46.7**
Step accuracy ↑	80.8	85.2
Misaligned trajectory ↓	20.7	10.0

Faithful reasoning requires a correct answer **and all-correct steps**. SRaR: 14 faithful traces out of 30.

Table 2. Single-pass diagnostic: 29 parseable RaR traces and 30 SRaR traces. This is a sample observation.

9 Normalization prevents collapse

8B ablation (Table 11)	Mean (%)
Without step normalization	56.76
Merged advantage	72.06
Full SRaR	**72.53**

Code transfer: SRaR is best on **3 of 4** benchmarks; DAPO wins MBPP+ (Table 12). Judge quality and explicit step boundaries remain practical dependencies.