Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.39405

Reinforcement Learning with Verifiable Rewards (RLVR) for Mathematical Reasoning in Large Language Models

Tao Yan

College of Science and Technology, Ningbo University, Cixi, China
TaoYan-1@outlook.com

Abstract. Reinforcement learning with verifiable rewards (RLVR) is the mainstream technology to improve the mathematical reasoning ability of large language models, but evaluation protocols across papers are not unified, and low-sampling and high-sampling pass@k metrics often yield inconsistent conclusions, and there is a lack of systematic review of the evaluation practice. Focusing on the core issue of evaluation completeness, this paper systematically combs three RLVR technical routes: outcome supervision, process supervision, and verifier-free, analyzes the principles of representative algorithms, summarizes the experimental results of public papers, and designs a structured coding scheme covering 9 fields. The 9 representative papers are coded one by one from the aspects of technical route, base model, sampling scale, and training configuration. Statistics show that the current RLVR field is deeply in the dilemma of fragmentation of evaluation protocols: the caliber of evaluation indicators is incompatible with each other, high-sampling evaluation data are scarce, and there are obvious differences on whether low sampling can expand the reasoning boundary. The disclosure of key hyperparameters such as temperature and KL coefficient is scattered and unsystematic, and the horizontal comparison lacks a common foundation. Based on this, this paper proposes a structured coding analysis framework to provide a unified benchmark for the horizontal comparison of evaluation results for subsequent research.

Keywords: large language model, mathematical reasoning, reinforcement learning with verifiable rewards, RLVR

1. Introduction

With the rapid development of large language model technology, the ability of logical derivation and complex problem solving has become the key index to measure the intelligent level of the model. Mathematical reasoning is widely used to evaluate the deep reasoning ability of large-scale models by virtue of its rigorous derivation and automatic verification of answers [1]. The large model relying only on supervised fine tuning (SFT) performs well on simple problems, but it is easy to fail in the face of difficult problems. In order to improve the mathematical reasoning ability, the academic community has proposed RLVR, which relies on the verifiable reward iterative optimization model output. The representative algorithms are divided into three types of routes: outcome supervision, process supervision, and verifier-free based on the way of reward generation.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.39405

In the outcome-supervised route, the accuracy rate of ReFT on GSM8K was 68.91%, which was about 10% higher than that of SFT [2]. In the direction of process supervision, Lightman et al. achieved 78.2% Best-of-N accuracy on MATH based on PRM800K [3]. Math-Shepherd raised Mistral-7B from 77.9% to 84.1% on GSM8K, VPPO reports complete sampling results from pass@1 to pass@128 [4, 5]. ProRL shows that RLVR can expand the reasoning boundary [6]. Although various algorithms have made progress, there are significant differences in the evaluation practice.

The evaluation defects gradually appear, which are reflected in incomplete sampling disclosure, inconsistent protocols, opaque baseline, etc. Different paper bases, data sets and parameters are inconsistent, and most of them rely on low sampling index demonstration, lacking high sampling support. Yue et al. find that pass@1 improves while pass@128 decreases, low-sampling improvement is not equivalent to reasoning upper bound expansion [7]. Chen et al. further pointed out that the evaluation benchmark is difficult to reveal the failure mode, and it is difficult to compare the algorithms fairly [8]. To sum up, although previous studies have achieved rich results at the algorithm level, the fragmentation of the evaluation protocol has greatly reduced the credibility of the conclusion. Therefore, the systematic review of the evaluation practice is of great significance to the research in the field of RLVR.

This paper aims to systematically sort out the current situation of evaluation practice in the field of RLVR, reveal the inconsistency and incompleteness of evaluation protocols through structured coding analysis, and give the idea of standardization research.

2. RLVR technical pathways and representative methods

2.1. Outcome-supervised ORM-RL

The outcome-supervised route generates rewards based solely on whether the final answer is correct, and does not evaluate the intermediate steps. For a complete output $y=(y_1, y_2,..., y_T)$ containing T reasoning steps, the reward function of ORM-RL only depends on the comparison result of the final answer y_T and the standard answer y*:

$$r_{\mathrm{ORM}}(y) = \mathbb{I}(y_T = y^*) \tag{1}$$

Where $r_{ORM}(y)$ is the reward given by ORM-RL for the complete output y, and $\mathbb{I}$ ( ) is the indicator function. If the answer is correct, take 1, otherwise take 0. y_t is the reasoning output at step t, and y* is the standard answer.

The advantage of this route is that reward acquisition does not require manual annotation and can directly reuse the existing RL training framework. The representative algorithm is ReFT [2]. Luong et al. used the two-stage process of SFT warm-up and PPO to generate rewards by comparing the answers according to formula (1), without step-by-step manual annotation [2]. ORM-RL is simple to implement and does not need step level annotation, but does not evaluate the reasoning logic. The model may rely on wrong derivation to get the correct answer, resulting in the reward hacking problem.

2.2. Process-supervised PRM-RL

The process-supervised route scores each step of the reasoning chain. For step-t reasoning, y_t, PRM-RL generates fine-grained rewards based on this step and previous history:

DOI: 10.54254/2755-2721/2027.39405

$$r_t = PRM(x, y_{1:t}), R_{PRM} = \sum_{t=1}^{T} r_t \quad (2)$$

Where r_t is the real-time reward given by the process reward model for the t-step reasoning, x is the problem, and y_1: t denotes the prefix of the first t steps. R_{PRM} is the total process reward score of the whole output defined on the right side of the equal sign, which is accumulated by the instant reward r_t of each step as the final reward signal of reinforcement learning update.

Lightman et al. established the process reward paradigm and constructed large-scale manually annotated datasets [3]. Representative algorithms include Math-Shepherd and VPPO [4, 5]. Wang et al. automatically constructed step-level rewards to reduce reliance on manual annotation [4]. Liu et al. proposed a mechanism to locate the first erroneous step and penalize only the subsequent erroneous suffix, preserving the correct prefix [5]. PRM-RL constrains the reasoning logic at each step and reduces guessing of answers, but high-quality step reward construction is costly, and automatically generated signals contain noise.

2.3. Verifier-free RL

The route without a verifier does not train an independent reward model, and rewards are generated based on the answer verification rules. GRPO removes the critic (value) network of the PPO algorithm, and uses the scores of samples within the group for baseline estimation. For the intra-group output y=(y_1, y_2, …, y_G) containing G samples, the advantage of GRPO is estimated as:

$$\widehat{A}_i = \frac{r_i - mean(\{r_j\}_{j=1}^{G})}{std(\{r_j\}_{j=1}^{G})} \quad (3)$$

Where $\widehat{A}_i$ is the advantage estimate of the i-th sample in the group, r_i is the reward of the i-th sample after rule verification, G is the total number of samples in the group, and mean and std are the mean and standard deviation of the reward in the group, respectively.

Representative algorithms include GRPO, DAPO, and ProRL [9, 10, 6]. Shao et al. used equation (3) to replace the critic network and simplify the training process [9]. Yu et al. introduced decoupled clipping and dynamic sample filtering to alleviate entropy collapse [10]. ProRL extends the number of training steps, introduces KL regularization and periodic resetting of the reference policy, and expands the reasoning boundary [6]. Reinforcement learning without verifier eliminates the training of a reward model, and the engineering threshold is low, but the scenarios that rely on the answers to be verifiable and cannot be automatically judged are limited.

3. Quantitative analysis of RLVR evaluation practice based on structured coding

3.1. Coding scheme

Based on the representativeness of the technical route and the completeness of the evaluation data, this paper selects 9 RLVR empirical papers ([2-10]) as the coding objects. Lightman et al. established the PRM paradigm as the foundational work of the process supervision route, although they mainly focused on the training of a reward model [3]. Chen et al. proposed the OPG metric to diagnose benchmark failure modes and reveal the structural defects of the evaluation system with the objective of evaluation audit [8]. Regarding reproducibility and comparability, nine fields are extracted, including six basic fields (literature information, technical route, base model, test data set,

DOI: 10.54254/2755-2721/2027.39405

main index type, training configuration), and three special fields (intermediate sampling index, maximum sampling scale, phenomenon record).

This paper carries out coding analysis around three research questions (RQ). RQ1 examines the status quo of high sampling evaluation completeness and the resulting divergence of conclusions. RQ2 examines the disclosure consistency of key hyperparameters (such as temperature and KL coefficient) between training and evaluation. RQ3 compares the evaluation completeness differences of three types of technical routes.

For the determination of phenomenon records, a complete report of pass@k exact numerical tables is the general prerequisite. Only when the paper provides both pass@1 and high-sampling pass@k exact values can it be included in the contraction judgment. Documents that only provide curve displays but are supplemented by clear text descriptions and are sufficient to support the three-level judgment can be included as exceptions. The judgment is divided into three grades; the "yes" category refers to cases where Pass@1 improves while high-sampling pass@k declines or plateaus, typified by Yue et al. [7]. The "no" category refers to the synchronous improvement of high and low sampling, typified by VPPO [5]. The "mixed" category refers to the differential response of different tasks, typified by DeepSeekMath [9]. The literature that only provides curves and lacks clear text description will not be included in the judgment, but its chart trend will be retained as supplementary evidence.

3.2. Statistical results

Based on the coding scheme in Section 3.1, this paper codes and counts 9 empirical papers related to RLVR ([2-10]) one by one, and the statistical results are shown in Table 1 and Table 2.

Statistics found that 9 papers used up to 4 kinds of evaluation indicators and were incompatible with each other. Only two papers fully report exact high-sampling pass@k values [5, 7], and the contraction phenomenon is determined as one showing contraction and one showing no contraction [7, 5]. The disclosure of key hyperparameters such as training and evaluation temperature and KL coefficient is scattered, and there is a blind spot in many papers. In general, there is a lack of unified indicators and sampling protocols in the field, and there is a lack of a common foundation for horizontal comparison.

Table 1. Coding results of basic fields for 9 RLVR papers

Literature	Technology roadmap	Model	Test data set	Main indicator type	Train Config
[2]	ORM-RL	Galactica-6.7B, CodeLLaMA-7B	GSM8K, MathQA, SVAMP	accuracy	SFT warm up+PPO
[3]	PRM-RL	GPT-4(large scale), GPT-4-like(small scale)	MATH	Best-of-N	No RL training
[4]	PRM-RL	Mistral-7B, LLaMA2-7B/70B, DeepSeek-67B	GSM8K, MATH500	accuracy	PPO+automatic construction step reward
[5]	PRM-RL	Qwen3-4B-Base, Qwen3-8B-Base, etc	AIME-24/25, AMC-23, MATH-500, etc	Pass@k	VPPO(protect correct prefix)
[6]	Verifier-Free RL	DeepSeek-R1-Distill-Qwen-1.5B(main),7B(comparison)	AIME24/25, AMC, MATH, Minerva, etc	Pass@k	GRPO+KL regular+periodic reset reference strategy

DOI: 10.54254/2755-2721/2027.39405

Table 1. (continued)

[7]	Verifier-Free RL	LLaMA-3.1-8B, Qwen2.5 series(7B–32B), etc	GSM8K, MATH500, AIME24, AMC23, etc	Pass@k	GRPO, PPO, ReMax and other algorithms
[8]	Verifier-Free RL	Qwen2.5-3B/7B-Instruct	MATH, GSM8K, DeepScaler, etc	Pass@1+OPG	GRPO
[9]	Verifier-Free RL	DeepSeekMath-Base 7B	GSM8K, MATH, HumanEval, MBPP	accuracy(main) +Pass@k (auxiliary)	GRPO
[10]	Verifier-Free RL	Qwen-32B	AIME 2024	accuracy	GRPO+decoupling clipping+dynamic sample filtering

Note: the literature information field has been corresponding to the references by numbers [2-10], and the base model and data set are only representative items.

Table 2. Coding results of special fields and key hyperparameters for 9 RLVR papers

Literature	Intermediate sampling index	Phenomenon record	Maximum sampling scale	Training temperature	Assessment temperature	KL setting
[2]	Not reported	Not included in Statistics	Not reported	1.0 (sampling)	Undisclosed	P-CoT: 0.01, N-CoT: 0.05
[3]	Continuous display of Best-of-N (N=1–1860)	Not included in Statistics (none Pass@k)	Best-of-1860	No RL training	Undisclosed	Not applicable
[4]	Not reported	Not included in Statistics	Not reported	undisclosed	Undisclosed	0.04
[5]	Yes(Pass@2 – Pass@64 Full spectrum value)	No (high and low sampling synchronous lifting)	pass@128	1.0	0.6	0
[6]	Yes (k=2 – 128, curve only)	Hybrid (task-level heterogeneity)	pass@256 (curve only)	1.2	0.6	Yes (coefficient not indicated)
[7]	Yes(pass@1 To pass@256 Value of each sampling point)	Yes (systemic contraction)	pass@256	0.6	0.6	0.001 (ablation)
[8]	Not reported	Not included in Statistics (not involving training)	pass@1	1.0	1.0	1×10^{-2}
[9]	Not reported	Not included in Statistics	pass@64 (curve only)	Not clear	0.7	0.04
[10]	Not reported	Not included in Statistics	Not reported	Undisclosed	1.0	Remove KL

DOI: 10.54254/2755-2721/2027.39405

Note: the coding is based on the contents of the published reports of each literature, "undisclosed" means that the corresponding parameters of the literature are not reported, "not applicable" means that the literature does not involve RL training. Lightman et al. disclosed that the candidate solution sampling temperature is 1.0, which is used for data annotation and solution generation of Best-of-N evaluation at the same time, so the evaluation temperature is not listed separately [3].

3.3. Findings and discussion

Based on the statistical results in Section 3.2 and Tables 1 and 2, this section discusses the above three research questions one by one.

For RQ1, only two papers fully report exact high-sampling pass@k values [7, 5]. Combined with the phenomenon record in Table 2, it is determined that some studies show contraction, some studies show synchronous improvement of high and low sampling, and other studies reveal task-level heterogeneity of contraction, stagnation and expansion [6]. This means the debate over whether pass@1 improvement is equivalent to reasoning boundary expansion is weak, and the samples supporting various conclusions are not enough to be conclusive, and the differences not only stem from the scarcity of highly sampled data, but also from the difference in task response [6].

For RQ2, statistics show that the disclosure of key hyperparameters is generally incomplete. In terms of temperature, there are four incomplete papers, three of which do not specify the evaluation temperature [2-4], among which Wang et al. did not disclose the training temperature, and one of which disclosed the evaluation temperature but did not give the training sampling temperature [10]. In terms of KL, some literatures did not disclose the KL coefficient, some literatures explicitly removed the KL constraint and demonstrated its rationality, and others did not have the concept of KL because they did not involve RL training [3, 4, 10]. This means that the differences in the disclosure of key hyperparameters make the reproduction and comparison lose the common foundation. Under different temperatures, KL constraints and sampling scales, differences in pass@1 or pass@k may stem from training configuration bias rather than algorithmic merit.

For RQ3, statistics show that there are systematic differences in the evaluation completeness of the three types of technical routes. The outcome-supervised ORM-RL route has the lowest completeness. ReFT reports only accuracy-type low-sampling metrics and did not report any intermediate sampling index and the maximum sampling scale [2], and the route only had one sample, which could not be used to judge whether RL training expanded the reasoning boundary. The process-supervised PRM-RL route has significant internal differentiation. Lightman et al. mainly focus on the Best-of-N curve and do not involve RL training [3]. Math-Shepherd likewise reports only accuracy [4], and only VPPO fully reports the complete spectrum from pass@1 to pass@128 [5]; the completeness difference within the route is greater than that between routes. The verifier-free RL route attaches the most importance to pass@k-type high-sampling metrics, because it directly responds to the dispute about "whether RL extends the reasoning boundary" [6, 7], but except for Yue et al. who completely disclosed the high sampling accurate values and key hyperparameters [7], the rest of the literature still lacked high sampling value tables, and the disclosure of hyperparameters was incomplete. Furthermore, the indicators of each route are incompatible with each other. Three papers center on pass@k [5-7], four on accuracy/greedy decode [2, 4, 9, 10], and one each on Best-of-N and pass@1+OPG [3, 8]. The accuracy reported by different papers is not the same concept. Greedy decode and pass@1 use different sampling strategies; Best-of-N relies on reward-model ranking whereas pass@k only checks whether a correct sample exists; OPG further measures the performance gap between the training set and the test set rather than absolute performance. The baseline settings are also different. The base model covers more than 1.5B–32B orders of magnitude, and the data set involves GSM8K, MATH, AIME, Olympiad and other variants. This means that there is neither a unified completeness baseline nor a comparable

DOI: 10.54254/2755-2721/2027.39405

indicator caliber among the three types of routes, and conclusions about algorithmic superiority are questionable.

The three findings reveal the structural defects of RLVR evaluation system from conclusion divergence, protocol consistency and baseline transparency, which not only makes it difficult to carry out horizontal comparison, but also may lead to the misjudgment of "whether RL extends the reasoning boundary".

4. Analysis of existing difficulties and disputes in the field

4.1. Evaluation index and sampling protocol are not unified

In view of the divergence of conclusions and protocol consistency found in 3.3, a core difficulty in the current RLVR field is the lack of unified standards for evaluation indicators and sampling protocols. The coding statistics in Section 3 show that the representative papers adopt multiple evaluation calibers such as pass@k, accuracy/greedy decode, Best-of-N, and pass@1+OPG, whose definitions and evaluation logics differ, and the results lack direct comparability. More importantly, the current high sampling evaluation paradigm has not yet established a minimum standard at the level of data disclosure. The coexistence of accurate values, curve representation and complete absence makes the core debate of "whether low sampling promotion is equivalent to reasoning boundary expansion" always lack a verifiable common evidence basis. Under this paradigm, different studies have made contradictory observations on the same issue, such as contraction, synchronous improvement and task-level heterogeneity [5-7]. Due to the incompatibility of disclosure patterns, these differences themselves cannot be strictly tested or unified.

To solve this problem, the existing literature has proposed several solutions. Yue et al. suggested reporting multiple k values instead of just reporting Pass@1 [7], the OPG index proposed by Chen et al. provides a quantitative tool for diagnosis and evaluation of benchmark failure [8], while Liu et al. reported the complete set of sampling metrics from pass@1 to pass@128 as an example [5]. However, these schemes still rely on researchers' conscious disclosure and lack of mandatory standards. At the same time, the calculation cost of high sampling evaluation is high. For research teams with limited computing resources, fully reporting pass@64 or even pass@128 poses practical difficulties. To sum up, the fragmentation of the evaluation protocol is rooted in the lack of mandatory disclosure norms in the field, rather than the negligence of individual researchers. The feasible and constructive way out includes three aspects. First, establish a tiered disclosure checklist that lists pass@1, key pass@k sampling points, temperature, and the KL coefficient as mandatory items, and high-sampling full spectra as encouraged items, giving consideration to completeness and affordability of calculation cost. Second, the community level should build low-cost high sampling evaluation protocols and benchmarks, and reduce duplication costs by unifying the sampling scale and criteria. Third, using the structured coding scheme proposed in this paper for reference, the index caliber, sampling scale and hyperparameter disclosure are transformed into horizontally comparable coding data, so that the debate on "whether RL expands the reasoning boundary" can be systematically tested rather than relying on case evidence.

4.2. Opaque base models, datasets, and training configurations

For the baseline transparency problem found in 3.3, another difficulty is the opaque disclosure of the base model, test data set and key training parameters. Statistics show that the base model used in the representative literature covers multiple orders of magnitude from 1.5B to 32B and the test data set

DOI: 10.54254/2755-2721/2027.39405

involves multiple variants such as GSM8K, MATH, AIME, Olympiad, etc. The accuracy reported in different papers is not the same concept. In terms of training configuration disclosure, the disclosure behavior of key hyperparameters such as temperature and KL coefficient is uneven. Some papers do not specify the evaluation temperature, some papers only disclose some parameters, and some papers do not disclose or directly remove KL constraints. This incompleteness and inconsistency in disclosure makes the reproduction and horizontal comparison of different studies lack a common basis.

Some works in the existing literature have tried to solve this transparency gap. Yu et al. open-sourced a complete training configuration list, Yue et al. disclosed the training and evaluation temperature in detail, and Liu et al. reported the VPPO hyperparameter settings [10, 7, 5].

The solution of the above difficulties depends on a unified disclosure specification and an operable checklist. The structured coding scheme proposed in Section 3 of this paper can be used as a reference framework for writing the experimental sections of future RLVR research, and transform the analysis path of finding problems into the methodological contribution of providing tools.

5. Conclusion

Focusing on the inconsistency and incompleteness of the RLVR evaluation system, this paper carried out structured coding and systematic analysis on 9 papers, and the results revealed the defects of the evaluation system in three aspects. Conclusion divergence, protocol consistency, and baseline transparency. In terms of conclusion divergence, different papers adopt multiple metrics such as pass@k, accuracy, and Best-of-N, which lack a unified caliber. The disclosure of high sampling indicators was seriously insufficient, and only two papers fully report pass@k accurate numerical tables. Most studies rely on low sampling indicators to demonstrate the improvement of reasoning ability, which leads to obvious differences in the conclusion of reasoning boundary expansion. In terms of protocol consistency, the disclosure behavior of key hyperparameters such as temperature parameters and KL settings is highly inconsistent. Some papers do not disclose them at all, and some papers only disclose some parameters, which makes the horizontal comparison lose a common foundation. In terms of baseline transparency, currently there is no mandatory disclosure list and unified evaluation protocol, and most solutions still rely on researchers' voluntary compliance. This paper establishes a structured coding analysis framework, which converts the key information of the literature, such as index type, sampling scale, training configuration, etc., into coding data that can be cross-referenced, so as to provide a reference for subsequent research. The limitation of this paper is that the coding object only covers 9 papers, which can be extended to more base models and evaluation data sets in the future to improve the applicability of the coding scheme.

References

[1] Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. *arXiv* preprint arXiv:2110.14168. DOI: 10.48550/arXiv.2110.14168.

[2] Trung, L., Zhang, X., Jie, Z., et al. (2024). Reft: Reasoning with reinforced fine-tuning. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7601-7614. DOI: 10.18653/v1/2024.acl-long.410.

[3] Lightman, H., Kosaraju, V., Burda, Y., et al. (2024). Let's verify step by step. *International Conference on Learning Representations*, 39578-39601. DOI: 10.48550/arXiv.2305.20050.

[4] Wang, P., Li, L., Shao, Z., et al. (2024). Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9426-9439. DOI: 10.18653/v1/2024.acl-long.510.

DOI: 10.54254/2755-2721/2027.39405

[5] Liu, H., Yu, D., Lu, S., et al. (2026). Save the good prefix: Precise error penalization via process-supervised RL to enhance LLM reasoning. *Findings of the Association for Computational Linguistics: ACL 2026*, 35450-35477. DOI: 10.18653/v1/2026.findings-acl.1767.

[6] Liu, M., Diao, S., Lu, X., et al. (2026). ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Advances in Neural Information Processing Systems, 38*, 17998–18031. https://doi.org/10.52202/085713-0608

[7] Yue, Y., Chen, Z., Lu, R., et al. (2026). Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *Advances in Neural Information Processing Systems, 38*, 57654–57689. https://doi.org/10.52202/085713-1933

[8] Chen, Z., Zhang, Y., Zhou, H., et al. (2026). Rethinking RL evaluation: Can benchmarks truly reveal failures of RL methods? *Findings of the Association for Computational Linguistics: ACL 2026*, 15692–15709. https://doi.org/10.18653/v1/2026.findings-acl.769

[9] Shao, Z., Wang, P., Zhu, Q., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv* preprint arXiv:2402.03300. https://doi.org/10.48550/arXiv.2402.03300

[10] Yu, Q., Zhang, Z., Zhu, R., et al. (2026). DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems, 38*, 113222–113244. https://doi.org/10.52202/085713-3775