BEYOND PASS@k: BREADTH-DEPTH METRICS FOR REASONING BOUNDARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm to improve Large Language Models on reasoning tasks such as coding, math or logic. To assess the reasoning boundary (the fraction of problems a model can solve) researchers often report Pass@k at large sampling budgets. Recent results reveal a crossover phenomenon: while RLVR models outperform the base model at small k values, the base model usually outperforms them when sampling a very large number of completions. This has been interpreted as evidence that base models have a larger reasoning boundary. We argue that on tasks with discrete answer spaces, such as math with numeric outputs, Pass@k at large k reflects the increasingly higher chance of success in the limit of the number of trials rather than genuine reasoning, and can therefore be misleading. We propose Cover@τ, which measures the fraction of problems that a model can solve for which at least a τ proportion of completions are correct. Unlike Pass@k, Cover@τ captures reasoning under an explicit reliability threshold: models that rely on random guessing degrade rapidly as τ increases. We evaluate several RLVR models using Cover@τ-based metrics and illustrate how the relative rankings of popular algorithms change compared to Pass@1, offering a different perspective on reasoning boundaries.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (Guo et al., 2025) has become an essential post-training approach for improving the capability of LLMs on math, code and logical reasoning. Recent research (Wang et al., 2025b; Wu et al., 2025a) has called into question the extent to which the RLVR models truly expand the reasoning boundary (i.e., extending the scope of solvable tasks). These works report Pass@k for increasing values of k and the reasoning boundary is defined as Pass@k at large k. While the RLVR model has higher Pass@1, at large k, the base model eventually outperforms it, thus the reasoning boundary of the base model shrinks as a result of applying RLVR. Follow-up works use this crossover plot to illustrate whether RLVR expands the reasoning boundary, showing that it can expand or shrink depending on the domain (Liu et al., 2025a; Cheng et al., 2025).

For tasks with numerical answers, such as math, Pass@k at large k may eventually produce correct answers for all problems, due to random guessing rather than reasoning ability. We thus argue that Pass@k can be problematic as a measure for the reasoning boundary, because it doesn't account for any level of reliability, for any given problem. We thus complement it with a *reliability-controlled reasoning boundary*, that exposes the reliability level explicitly. Concretely, we define Cover@τ, which measures the fraction of problems solved by a model with success rate at least τ. For very small τ, Cover@τ behaves similarly to Pass@k, however, increasing τ tightens the criterion to emphasize consistency over chance. Figure 1 highlights the behavior of Pass@k and Cover@τ for a math dataset from OMEGA (Sun et al., 2025) with numerical answers and small set support.

Our contributions are as follows:

- we introduce Cover@τ, a reliability-thresholded metric that reveals the reasoning abilities under a different reliability level τ. Cover@τ offers a more informative view, highlighting an *explicit breadth-depth trade-off*: low τ captures breadth of problem solving (even if unreliable), while larger τ captures depth (problem solving with high reliability). Measuring

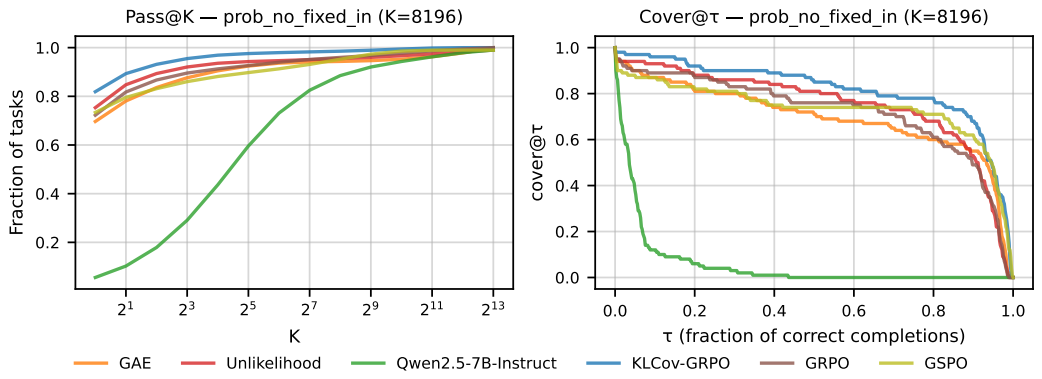

Figure 1: Pass@k and Cover@τ curves for Qwen2.5-7B-Instruct and several RLVR models on the Probability set of OMEGA. **Left:** Pass@k quickly saturates for larger k due to small test support. **Right:** Cover@τ illustrates a more gradual assessment of the models' capabilities, ranging from maximum performance (at low τ values) to very limited capabilities (when requiring models have almost perfect reliability at high τ values).

Cover@τ reveals a different ranking of popular RLVR algorithms when compared to Pass@1 or Pass@k at large k; this shows a complementary perspective of the model capabilities

- we demonstrate that Pass@k is a weighted average of Cover@τ, with weights from a Beta(1,k) distribution; this reveals that Pass@k is biased towards low-τ regions of Cover@τ, emphasizing lucky hits rather than reliability
- we illustrate the usefulness of Cover@τ by characterizing the performance of different RLVR methods trained on math datasets from OMEGA and Reasoning Gym.

2 Related work

DeepSeekR1 (Guo et al., 2025) has paved the way for reasoning-focused LLMs, by finetuning LLMs with reinforcement learning from verifiable rewards. The model employs the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024), which is a critic-free variant of PPO (Schulman et al., 2017), that estimates the baseline from the group average rewards.

2.1 Evaluation metrics

Models are usually evaluated via Pass@k (Chen et al., 2021), using $k = 1$ as the most common value. Higher values of k (e.g. 16 or 64) are also used to test the upper bound capabilities of the models. Recent work argues that the relative decrease in reasoning capacity of RLVR models compared to the base model is an artefact of Pass@k and proposes CoT-Pass@k (Wen et al., 2025) to account for the correctness of both the thinking tokens and the final answer. Finally, some papers report maj@k or cons@k (Guo et al., 2025; Shao et al., 2024), which counts a problem as solved based on aggregation over k samples, either majority (>50% of completions correct) or mode (most frequent). Both metrics target whether a model is consistent and are related to our proposal. Maj@k is equivalent to Cover@τ with $\tau = 0.5$. Cons@k has no fixed reliability threshold: the effective τ varies per problem with the mode's frequency, so it does not correspond to a single Cover@τ. Our Cover@τ is structurally similar to the performance profiles advocated by Agarwal et al. (2021), where the fraction of solved games is reported relative to varying human-normalized score levels. The key distinction is we evaluate coverage under varying reliability levels using a fixed (and sufficiently large) sample budget of k completions, without normalizing to human performance.

2.2 Exploration-promoting methods

Several shortcomings in GRPO have been reported, notably optimization biases (Liu et al., 2025b; Zheng et al., 2025) and entropy collapse (Yu et al., 2025). The latter is particularly problematic,

because it reduces exploration and quickly saturates performance (Cui et al., 2025). Follow-up work addresses the entropy collapse and encourages exploration. Simpler approaches add an explicit entropy loss term (Wang et al., 2025b) or increase the upper clipping threshold to favor low-probability exploration tokens (Yu et al., 2025). More fine-grained techniques such as KL-cov (Cui et al., 2025) suppress the high-covariance tokens that correlate with large decreases in entropy. Other works identify forking tokens (Wang et al., 2025a), which are high entropy tokens serving as logical connectors. Restricting policy gradient updates to these high-entropy tokens maintains entropy and enhances exploration. Finally, GRPO-Unlikeliness (He et al., 2025) promotes low-probability solutions by penalizing the reward of high-probability ones.

2.3 Procedurally generated datasets

Progress of RLVR methods on math reasoning has been questioned, due to potential contamination in popular static benchmarks. Portion of these datasets may appear in the pretraining data of popular LLMs (Wu et al., 2025b), which can conflate reasoning abilities with memorization. To limit contamination concerns, recent math and logic datasets are procedurally generated and designed with increasing difficulty and structural variation (Sun et al., 2025; Stojanovski et al., 2025).

3 Pass@k can be misleading

We consider T tasks, where task i has per-trial success probability $p_i \in [0, 1]$ under i.i.d. trials. Pass@k is defined as the average probability that a task is solved within k independent attempts.

Definition 1 (Pass@k)**.**

$$\textit{Pass@k} \;=\; \frac{1}{T}\sum_{i=1}^{T}\Big(1-(1-p_i)^k\Big).$$

Pass@k has been used to assess the reasoning boundaries of RLVR models by comparing the performance of a finetuned model against the base model at varying values of k. At large k values, the performance of the base model typically approaches or even surpasses that of the finetuned model, seemingly closing any gap that may exist at small k values. This pattern, however, does not necessarily reflect genuine reasoning ability. Instead, it primarily highlights the diversity of output trajectories in the base model. In the mathematical reasoning setting, where tasks often involve numerical answers with very limited support, this effect can create a misleading impression about model reasoning.

Figure 1 illustrates this point by plotting Pass@k curves for the base model and several RLVR variants on the Probability (No Fixed) task from the OMEGA dataset. Given enough trials (2^{13}), the base model achieves Pass@$k = 1$, due to the very limited size of the output space (only 30 possible values in the test set). More generally, this phenomenon is inevitable: provided there is nonzero probability of producing the correct answer, Pass@k will always converge to 1 in the limit of infinite trials.

Remark 1 (Degeneracy of Pass@k at Large k)**.** *For any success probability* $0 < p \leq 1$*,*

$$\lim_{k\to\infty}\big(1-(1-p)^k\big) = 1.$$

We thus argue that using Pass@k as a proxy for reasoning boundaries can be misleading, because it confounds true capability with random chance.

4 Cover@τ

Whereas Pass@k captures the binary likelihood of success within k attempts, we propose examining how dataset coverage changes when accounting for the consistency of predictions. We define Cover@τ as fraction of problems for which at least a proportion τ of the generated completions are correct.

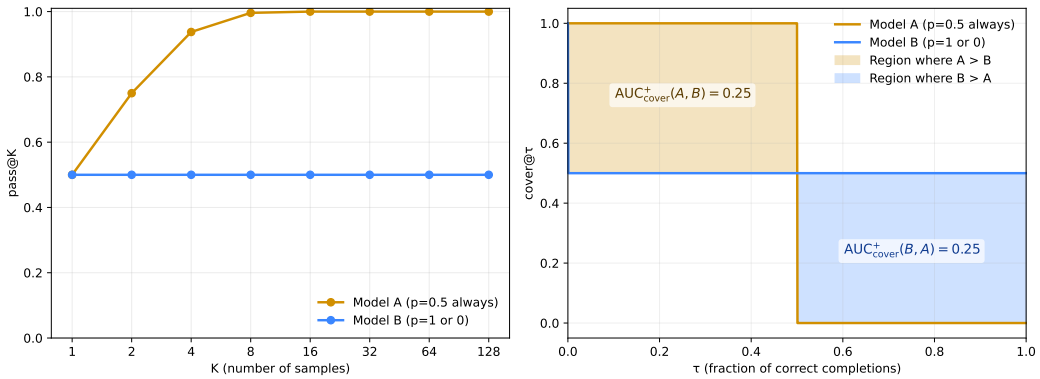

Figure 2: **Left:** pass@K for two models A and B. Both have the same pass@1=0.5, but model A's reasoning boundary increases with more tries, while model B stay flat. **Right:** Cover@τ curves for the same models A and B. Model A solves more problems overall, while Model B solves fewer problems but with higher consistency. When comparing their excess AUC (areas where each curve dominates), their overall advantages balance out.

Formally:

Definition 2 (Cover@τ)**.** *For any threshold $\tau \in [0, 1]$, define*

$$G(\tau) \;=\; \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{p_i \geq \tau\},$$

i.e. the fraction of tasks that can be solved with per-trial success probability at least τ.

Consider two LLMs, tested on the same set of problems: A has probability of success of 0.5 on all problems, while B has probability of success 0 on half the problems and 1 on the other half. Both models have $Pass@1 = 0.5$, but the **Cover@τ** plot in the right side of Figure 2 also shows model performances at explicit reliability levels: model A solves more problems, while model B solves fewer problems, but more consistently. Thus, Cover@τ captures a fine-grained view of the performance regime of the models:

- low τ values: Cover@τ highlights the **breadth** of the capabilities (how many problems are at least sometimes solvable)
- higher τ values: Cover@τ indicates **depth** (how reliably a problem is solved)

Two models A and B can cross over in their Cover@τ curves: model A performs better at lower reliability thresholds, but B dominates at higher thresholds. To capture such cases where models trade dominance at different τ levels, we define $\text{AUC}^{+}_{cover}(A, B)$. This is the excess AUC between model's A coverage over model's B coverage across reliability thresholds:

$$\text{AUC}^{+}_{cover}(A, B) = \int_0^1 \max\big(G_A(\tau) - G_B(\tau),\, 0\big)\, d\tau,$$

where $G_M(\tau)$ denotes the Cover@τ curve of model M. This pairwise metric captures the total coverage advantage of model A relative to model B, ignoring regions where coverage of A is worse. In the right side of Figure 2, both models have equal AUC^{+}_{cover}, indicating a similar level of performance.

5 Relating Pass@k and Cover@τ

We now demonstrate the connection between Pass@k and Cover@τ. Specifically, Pass@k is a Beta-weighted average of the Cover@τ metric, and thus represents only one particular projection of the richer information contained in Cover@τ.

Proposition 1. *Pass@k as Weighted Average of Cover@τ.*

For any $k \geq 1$,

$$\textit{Pass@k} = \int_0^1 k(1-\tau)^{k-1}\, G(\tau)\, d\tau,$$

where $G(\tau)$ denotes the Cover@τ curve.

Proof. For a single task with success probability p, define $f_k(p) = 1 - (1-p)^k$. Since $f_k'(p) = k(1-p)^{k-1}$ and $f_k(0) = 0$, the fundamental theorem of calculus gives

$$f_k(p) = \int_0^p f_k'(\tau)\, d\tau = \int_0^1 k(1-\tau)^{k-1}\, \mathbf{1}\{p \geq \tau\}\, d\tau.$$

Averaging over tasks,

$$\text{Pass@}k = \frac{1}{T}\sum_{i=1}^{T} f_k(p_i) = \int_0^1 k(1-\tau)^{k-1} \left(\tfrac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{p_i \geq \tau\}\right) d\tau.$$

The term in parentheses is exactly $G(\tau)$, which yields the claim. □

Corollary 1. *Pass@k as Expectation under Beta Weights.*

Let $\tau \sim \mathrm{Beta}(1,k)$ with density $p_k(\tau) = k(1-\tau)^{k-1}$ on $[0,1]$. Then

$$\textit{Pass@k} = \int_0^1 G(\tau)\, p_k(\tau)\, d\tau = \mathbb{E}_{\tau\sim\mathrm{Beta}(1,k)}[\, G(\tau)\,].$$

Corollary 2. *Uniform AUC Equals Pass@1.*

The unweighted area under the Cover@τ curve satisfies

$$\int_0^1 G(\tau)\, d\tau = \frac{1}{T}\sum_{i=1}^{T} p_i = \text{Pass@1}.$$

Remark 2. *As $k \to \infty$, the weighting distribution $\mathrm{Beta}(1,k)$ concentrates at $\tau = 0$. Therefore*

$$\lim_{k\to\infty} \textit{Pass@k} = G(0^+),$$

the fraction of tasks with nonzero success probability. If every $p_i > 0$, then Pass@$k \to 1$.

Proposition 2. *Cover@τ dominance implies Pass@k dominance.*

Given two models A and B, if Cover@$\tau(A) \geq$ Cover@$\tau(B)$ for all $\tau \in [0,1]$, then Pass@$k(A) \geq$ Pass@$k(B)$ for every $K \geq 1$.

Proof. Using proposition 1,

$$\text{Pass@k}(A) - \text{Pass@k}(B) = \int_0^1 k(1-\tau)^{k-1}\, (G_A(\tau) - G_B(\tau))\, d\tau,$$

Since $k(1-\tau)^{k-1} \geq 0$ and $G_A(\tau) - G_B(\tau) \geq 0$, the integral is positive. Therefore Pass@k$(A) \geq$ Pass@k(B). □

OBSERVATIONS

Pass@k as a weighted summary of Cover@τ. Proposition 1 shows that Pass@k is a weighted average of the Cover@τ curve, where the weights are given by a $\mathrm{Beta}(1,k)$ distribution. In other words, Pass@k summarizes only part of the information already contained in Cover@τ.

Bias toward low thresholds. The $\mathrm{Beta}(1,k)$ weighting heavily emphasizes small values of τ as k grows. Thus Pass@k primarily captures whether tasks have *any* nonzero success probability, rather than how reliably they can be solved. In the limit $k \to \infty$, Pass@k collapses to the trivial statistic "fraction of tasks with $p_i > 0$."

Uniform weighting recovers Pass@1. When $G(\tau)$ is weighted uniformly, the resulting area under the Cover@τ curve equals Pass@1, i.e. the average per-trial success probability. This highlights that Cover@τ naturally generalizes both Pass@1 and Pass@k.

Cover@τ ranking is more informative than Pass@k Proposition 2 shows that rankings based on Cover@τ imply the corresponding rankings based on Pass@k. However, the reciprocal is not true, as evidenced by Figure 2. Model A outperforms model B in terms of Pass@k for all K, but the Cover@τ curve reveals ordering differences across reliability levels that the Pass@k curve can hide.

Summary. Cover@τ makes the coverage-reliability trade-offs explicit, avoids the degeneracy of large-k behavior, and reveals finer-grained rankings that Pass@k can obscure.

6 Re-assesing generalization performance on math using Cover@tau

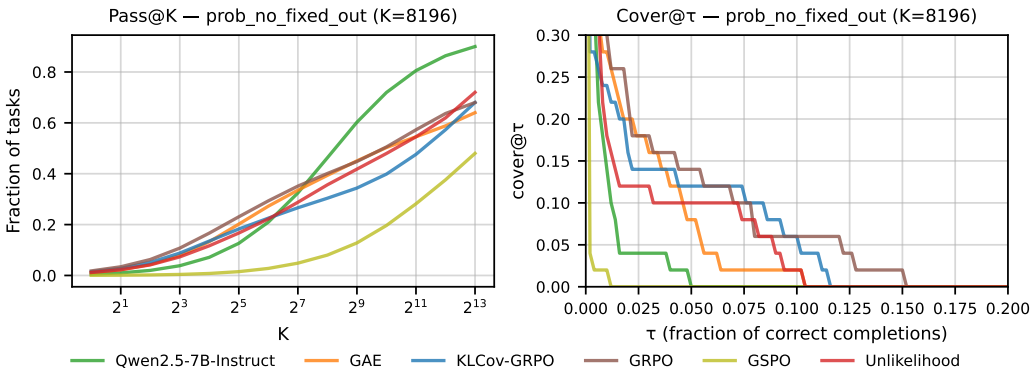

Figure 3: Pass@k and Cover@τ curves for Qwen2.5-7B-Instruct and RLVR models on the Probability (No Fixed) subset of OMEGA, for the OOD test split. **Left:** All models have poor accuracy, and increasing the sampling budget leads to higher Pass@k, especially on the base model. **Right:** GRPO and KL-Cov generalize the best; the base model quickly drops in performance even at low reliability thresholds, suggesting a far more limited reasoning boundary than the $Pass@k$ plot implies.

We evaluate RLVR methods using Cover@τ-based metrics, focusing on: (i) highlighting the trade-offs exposed by different τ values (ii) examining how models perform in mathematical reasoning under OOD settings.

6.1 Experimental setup

Datasets Popular datasets for evaluating mathematical reasoning may suffer from data leakage (Wu et al., 2025b). As a result, training on these datasets may not accurately reflect improvements relative to the base model. We instead focus on two particular datasets: OMEGA (Sun et al., 2025) and Reasoning Gym (Stojanovski et al., 2025). OMEGA provides both IID and OOD test splits across various tasks, from which we select: GCD, Function Intersection, Probability (No Fixed), Digit Sum, and Circle. For all tasks, we train on IID data and test on the OOD splits. For Reasoning Gym, we follow the intra-domain transfer setup: we train on two tasks from the algebra domain (simple equations and polynomial multiplication) and test on a third (intermediate integration).

Methods We evaluate two commonly used RLVR algorithms: GRPO (Shao et al., 2024) and PPO (Schulman et al., 2017). In addition, we test GSPO (Zheng et al., 2025), which improves upon GRPO by defining importance ratios at the sequence level rather than the token level. Given our out-of-distribution setup, we also consider methods specifically designed to enhance exploration. KL-Cov (Cui et al., 2025) applies a KL penalty to tokens to tokens with high covariance between their log probabilities and advantages, helping to prevent entropy collapse, while GRPO-Unlikeliness (He et al., 2025) introduces an unlikeliness reward, which incentivizes correct but less probable solutions.

Training details We train all models using the VERL framework (Sheng et al., 2024), starting from the Qwen-2.5-7B-Instruct model (Yang et al., 2025). For the Omega tasks, we run training for 30 epochs with a batch size of 500 and a mini-batch size of 96. For Reasoning Gym, we train for 5 epochs, with a batch size of 64 and a mini-batch size of 32. We set PPO epochs to 1 for all experiments. The actor and critic use learning rates of $1e{-6}$ and $1e{-5}$, respectively. For the KL loss between the policy and base model, we use the following: 0.1 on the Probability (No Fixed) task, as well as on Reasoning Gym for the GRPO-Unlikeliness method, 0.05 for the other Omega tasks, and 0.01 for all other Reasoning Gym experiments. Across all methods, we use 32 (OMEGA) and 8 (Reasoning Gym) rollouts per prompt which, for GRPO-style methods, corresponds to the group size. We select the best performing checkpoint on the iid validation sets for evaluation. All experiments were performed on a single node with 8 NVIDIA H200 GPUs.

6.2 Model analysis using the Cover@τ curve

The right side of Figure 1 illustrates the Cover@τ on the IID split of the Probability (no fixed) subset of OMEGA. Our metric highlights the brittleness of the base model: while it solves problems at very small thresholds, its coverage quickly decreases even at modest thresholds, such as $\tau = 0.2$. KL-Cov consistently solves the most problems at almost all the reliability thresholds.

In Figure 3 we plot both the Pass@k and Cover@τ curves for the OOD split of the same dataset. All models have poor Pass@1 performance, but, for large sampling budgets, the base model significantly outperforms the RLVR models. For very small values of $\tau < 0.01$, Cover@τ highlights a similarly strong performance of the base model. This correlates with the theoretical insights from Section 5, showing that as τ approaches 0, it becomes similar to Pass@k at large ks. However, when marginally increasing the threshold to $\tau = 0.025$, the performance of the base model drops significantly, which showcases a less optimistic view of its reasoning capabilities. Additionally, the Cover@τ curve reveals different trade-offs between the RLVR models. GAE solves more problems than GRPO-Unlikeliness for $\tau < 0.050$, while GRPO-Unlikeliness is more reliable for larger τ values.

Cover@τ offers finer granularity. By inspecting the entire curve, we can distinguish between algorithms that (a) succeed rarely across many tasks, and (b) succeed reliably on fewer tasks. Pass@k targets the former, while being uninformative about the latter. Cover@τ exposes important trade-offs between *coverage* and *reliability* of exploration.

6.3 Results

In Table 1 we evaluate several RLVR algorithms on the Reasoning Gym and OMEGA datasets. We report pass@1, as well as Cover@τ for $\tau = 0.2$ and $\tau = 0.8$, to assess both low and high reliability performance. While these point metrics provide insight into the performance at explicit reliability levels, they do not capture how models perform across the entire range of reliability thresholds. To summarize a model's performance across all τ regions, we build on the previously defined pairwise measure $\mathrm{AUC}^{+}_{\mathrm{cover}}(A, B)$. Consider a set of models $\mathcal{M}$, with $|\mathcal{M}| = M$. For a model A, we average

Table 1: Results on the Intermediate Integration task from Reasoning Gym and the OMEGA benchmark (averaged across 5 tasks). Best results per metric are shown in bold, second and third best are highlighted in blue and orange, respectively.

	Reasoning Gym				OMEGA OOD			
Method	Pass@1	cov@.2	cov@.8	$\mathrm{AvgAUC}^{+}_{\mathrm{cov}}$	Pass@1	cov@.2	cov@.8	$\mathrm{AvgAUC}^{+}_{\mathrm{cov}}$
base	49.67	55.67	39.67	0.64	8.34	14.94	1.10	0.19
GRPO	**59.66**	**66.00**	48.67	5.52	17.86	22.66	11.98	1.61
GSPO	56.14	57.67	48.00	1.79	18.00	20.26	15.36	2.36
PPO (GAE)	57.94	60.00	55.67	5.13	18.38	22.66	13.68	1.85
KL-Cov	58.55	62.00	**56.33**	**5.70**	**28.34**	**33.58**	**23.34**	**12.78**
Unlikeliness	43.98	51.67	34.00	0.00	17.02	20.94	11.94	0.68

its pairwise excess AUC score against all the other models:

$$\mathrm{AvgAUC}^{+}_{\text{cover}}(A) \;=\; \frac{1}{M-1} \sum_{\substack{B \in \mathcal{M} \\ B \neq A}} \mathrm{AUC}^{+}_{\text{cover}}(A, B) \tag{1}$$

$\mathrm{AvgAUC}^{+}_{\text{cover}}(A)$ quantifies how much A tends to dominate other models on the Cover@τ curve.

In Table 1 we observe that the rankings of the top methods differ between Pass@1, Cover@τ and $\mathrm{AvgAUC}^{+}_{\text{cover}}$. While cover@0.2 yields the same rankings as Pass@1, with GRPO as the top method, cover@0.8 ranks KL-cov first, highlighting its strength in preserving depth at higher thresholds. The $\mathrm{AvgAUC}^{+}_{\text{cover}}$ score (1) averages performance across dominant threshold regions: KL-Cov is ranked first, while GRPO ranks second.

KL-Cov obtains the best results on the OMEGA dataset on all metrics. While GRPO ranks 4th based on Pass@1, it ranks 2nd based on Cover@0.2 (tied with PPO), highlighting better coverage at low reliability. Enforcing higher reliability ($\tau = 0.8$) produces a different ordering compared to $\tau = 0.2$, where breadth is emphasized. GRPO and GRPO-Unlikeliness show good Cover@τ performance at lower thresholds, but their coverage at $\tau = 0.8$ drops out of the top 3. The $\mathrm{AvgAUC}^{+}_{\text{cover}}$ score (1) captures this trade-off: GRPO and GRPO-Unlikeliness lag behind GSPO and PPO, which are ranked 2nd and 3rd, respectively. Overall, the strong performance of KL-Cov suggests that RLVR methods that prevent entropy collapse show stronger generalization abilities.

7 Limitations

Our metric still accounts for the accuracy of the final answer, without evaluating the soundness of the reasoning trace, as was done in Cot-Pass@k(Wen et al., 2025). However, the accuracy of the reasoning trajectory can be combined with the Cover@τ metric.

8 Conclusions

We introduce Cover@τ which emphasizes a *reliability-controlled reasoning boundary*, by taking into account the ratio τ of correct completions for a given task. This fine-grained view of the performance can naturally highlight the trade-off between the coverage of solvable problems and the correct answer consistency.

We connect Pass@k to Cover@τ and demonstrate that Pass@k can be expressed as a weighted average of Cover@τ, that is biased towards low τ regions, making it prone to emphasizing lucky hits rather than reliability. Moreover, Cover@τ reveals ordering differences across reliability levels that Pass@k may hide. Through this new lens, we evaluate several RLVR methods to uncover their reasoning abilities under different reliability thresholds.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *arXiv preprint arXiv:2506.14965*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2505.24760*, 2025.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *arXiv preprint arXiv:2506.18880*, 2025.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025b.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025a.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, et al. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.