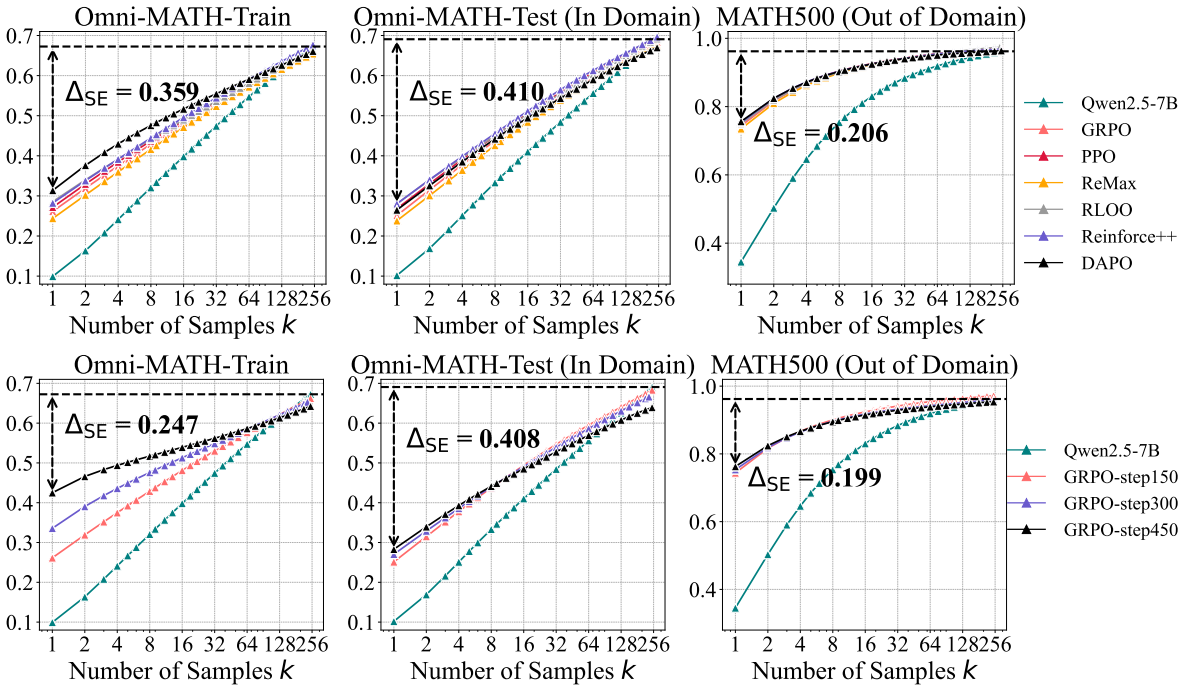

Figure 8: **(Top)** Different RL algorithms. **(Bottom)** Different RL training steps. The detailed values for each point at pass@1 and pass@256 are provided in Table 3 and Table 4.

4.4. Effects of RL Training

Asymptotic Effects. Based on the setup in Section 4.3, we investigate the effect of the training steps on the asymptotic performance of the model. As shown in Figure 1 (right), as RL training progresses, pass@1 on the training set consistently improves from 26.1 to 42.5. However, as RLVR training progresses, pass@256 progressively decreases, indicating a reduced reasoning boundary.

Effect of Number of Rollouts n. The training hyperparameter n, the number of responses per prompt, can affect pass@k by enabling broader exploration during training. We increase n from 8 to 32. As shown in Figure 16, pass@k improves slightly over $n = 8$, but the RL-trained model is still eventually outperformed by the base model. We leave the question of whether scaling RLVR training can eventually surpass the base model to future investigation.

Effect of KL Loss. To control model deviation, some prior work adds a KL penalty. We ablate this by applying a KL term with coefficient 0.001. As shown in Figure 16, the KL-regularized model achieves similar pass@1 to GRPO without KL, but with a much lower pass@128.

4.5. Effects of Entropy

As RL training progresses, the model's output entropy typically decreases (Yu et al., 2025), which may contribute to a reduced reasoning boundary due to less diverse output. To investigate this factor, we increase the generation temperature of the RLVR-trained model to match the output entropy of the base model at $T = 0.6$. As shown in Figure 18, although the RLVR model performs slightly better pass@k at higher temperatures compared to its own performance at $T = 0.6$, it still underperforms the base model across pass@k. This suggests that while reduced entropy contributes to the narrowing of the reasoning boundary, it alone does not fully account for the reduction.

4.6. Effects of Model Size Scaling

Scaling plays a central role in the capabilities of contemporary LLMs. It remains an important question whether the conclusions drawn continue to hold as model size increases. For many large models, isolating the effect of RLVR is not feasible. For example, in the case of GPT-o1, the base model is not publicly

accessible. Qwen3-235B (Yang et al., 2025a) is trained through multiple stages, including RLVR and long-context CoT supervised fine-tuning, which makes it impossible to disentangle the impact of RLVR alone. For Deepseek-R1-Zero, the absence of a publicly hosted API forced us to self-host the model, but throughput was limited to around 50 tokens per second at a maximum sequence length of 32k, rendering pass@k evaluation currently impractical. As a more tractable alternative, we selected the Magistral-Medium-2506 API to conduct a preliminary set of experiments. This model is trained using pure RL, with Mistral-Medium-3-2505 as the starting model (Rastogi et al., 2025). Although the model size is not disclosed, Magistral-Medium performs comparably to Deepseek-R1 and is positioned near the frontier in terms of reasoning capability.

We queried the models using a maximum context length of 40k as the original paper does. Once again, we observed that RLVR provides significant gains at low k, but little or no improvement at higher k. Specifically, at $k = 1$, the RLVR-enhanced model solves approximately 7 more problems on AIME24 and 8 more on AIME25 compared to its base version. However, as k increases, the performance gap steadily narrows. These observations suggest that our conclusion continues to hold even for current, highly capable, near-frontier reasoning models. Whether this trend persists as more compute, such as pre-training scale budgets, is dedicated to RL training remains a critical question for the future of LLM reasoning.

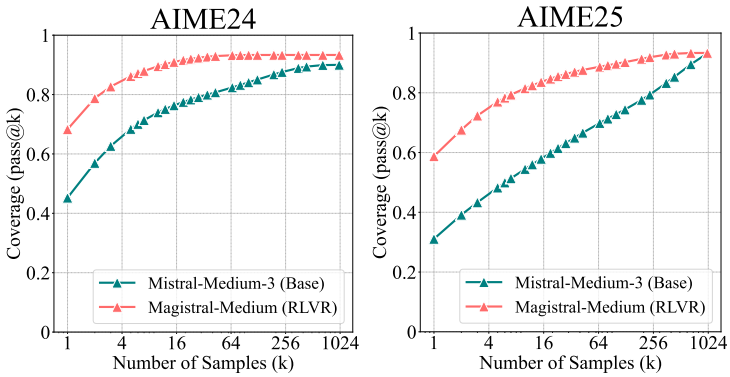

Figure 9: pass@k curves of Magistral-Medium.

5. Discussion

In Section 3 and Section 4, we identified key limitations of RLVR in enhancing LLM reasoning capabilities. In this section, we explore possible underlying factors that may explain why RLVR remains bounded by the reasoning capacity of the base model.

Discussion 1: Key Differences Between Traditional RL and RLVR for LLMs are Vast Action Space and Pretrained Priors. Traditional RL such as AlphaGo Zero and the DQN series (Silver et al., 2017; Mnih et al., 2015; Yue et al., 2023) can continuously improve the performance of a policy in environments like Go and Atari games *without an explicit upper bound*. There are two key differences between traditional RL and RLVR for LLMs. First, the action space in language models is exponentially larger than that of Go or Atari games (Ramamurthy et al., 2023). RL algorithms were not originally designed to handle such a vast action space, which makes it nearly impossible to explore the reward signal effectively if training starts from scratch. Therefore, the second distinction is that RLVR for LLMs starts with a pretrained base model with useful prior, whereas traditional RL in Atari and GO games often begins from scratch. This pretrained prior guides the LLM in generating reasonable responses, making the exploration process significantly easier, and the policy can receive positive reward feedback.

Discussion 2: Priors as a Double-Edged Sword in This Vast Action Space. Since the sampling of responses is guided by the pretrained prior, the policy may struggle to explore new reasoning patterns beyond what the prior already provides. Specifically, in such a complex and highly combinatorial space, most responses generated by *naive token-level sampling exploration* are constrained by the base model's prior. Any sample deviating from the prior is highly likely to produce invalid or non-sensical outputs, leading to negative *outcome reward*. As discussed in Section 2.1, policy gradient algorithms aim to maximize the log-likelihood of responses within the prior that receive positive rewards, while minimizing the likelihood of responses outside the prior that receive negative rewards. As a result, the trained policy tends to produce responses already present in the prior, constraining its reasoning ability within the boundaries of the base model. From this perspective, training RL models from a distilled model may temporarily provide a beneficial solution, as distillation helps inject a better prior.

Possible Future Work. As discussed above, inefficient exploration mechanisms in a vast action space and the reliance on binary outcome rewards may be the root causes of the limitations observed in current RLVR settings. To fundamentally address these challenges, several directions may be worth exploring:

- **Efficient exploration strategies in high-level abstraction.** High-level exploration mechanisms such as AlphaEvolve (Novikov et al., 2025), which perform self-evolution in a program-level abstraction space, may be crucial for navigating the vast action space. Such strategies could facilitate the discovery of out-of-prior reasoning patterns and previously unseen knowledge structures.
- **Data scale via curriculum.** A curriculum can begin by training on easier subproblems, allowing the model to improve sampling efficiency and acquire essential meta-skills. By increasing success rates on simpler tasks before tackling harder ones, such a curriculum may hierarchically reduce the exploration space and lift performance from nearly zero to non-zero on challenging parent tasks, thereby enabling RLVR to obtain meaningful rewards (Zhang et al., 2025; Li et al., 2025). Although traces of such hierarchical relationships occasionally appear in current RLVR training data, and their effects have been observed in recent work (Chen et al., 2025b), realizing their full potential will require a more deliberate, large-scale data-RL iteration pipeline that ensures sufficient coverage of meta-skills as well as appropriate relationships between easy and hard problems.
- **Process reward and fine-grained credit assignment.** Compared to purely binary outcome rewards, incorporating intermediate signals to guide the reasoning trajectory may significantly improve exploration efficiency and steer exploration toward more promising solution paths.
- **Agentic RL.** Current RLVR reasoning are limited to single-turn response, whereas iterative refinement based on feedback is crucial for IMO-level reasoning (Huang & Yang, 2025). It also lacks the ability to actively collect new information by using search tools or conducting experiments. A multi-turn agentic RL paradigm, featuring richer interactions with environment feedback, could allow models to generate novel experiences and learn from them. This emerging agent framework has been described as the beginning of an "era of experience" (Silver & Sutton, 2025).

6. Related Work

We summarize key related works on the analysis of RLVR here and provide a more comprehensive discussion in Appendix B. While recent RLVR methods have achieved impressive empirical results (Guo et al., 2025; Lambert et al., 2024), their fundamental impact on reasoning remains underexplored. Several studies (Liu et al., 2025a; Zhao et al., 2025b; Shah et al., 2025) suggest that reflective behaviors in RLVR models originate from the base models rather than being learned through reinforcement learning. Dang *et al.* (Dang et al., 2025) observed a decline in pass@k performance post-RLVR training, but their analysis was limited in scope. More importantly, they did not explore the relationship between the base model and the RL model. Deepseek-Math (Shao et al., 2024) also observed similar trends, but their study was limited to a single instruction-tuned model and two math benchmarks. In contrast, our work systematically investigates a wide range of models, tasks, and RL algorithms to accurately assess the effects of current RLVR methods and models. We further provide in-depth analyses, including accuracy distributions, reasoning coverage, perplexity trends, and comparison against distilled models, offering a comprehensive understanding of RLVR's capabilities and limitations.

7. Conclusion and Limitations

RLVR is widely regarded as a promising approach to enable LLMs to continuously self-improve and acquire novel reasoning capabilities. In this paper, we systematically investigate the effect of current RLVR methods on the reasoning capacity boundaries of LLMs. Surprisingly, our findings reveal that current RLVR rarely elicits fundamentally new reasoning patterns; instead, the reasoning capabilities of RLVR-trained models remain bounded by those of their base models. These results indicate that current RLVR methods have not fully realized the potential of reinforcement learning to elicit novel reasoning abilities in LLMs through exploration and exploitation. This limitation may stem from the lack of effective exploration strategies in the vast language space as we discussed in Section 5. Exploration in high-level abstraction, fine-grained credit assignment, and multi-turn agent-environment interactions may