ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models

Mingjie Liu **Shizhe Diao** **Ximing Lu** **Jian Hu** **Xin Dong**
Yejin Choi **Jan Kautz** **Yi Dong**
NVIDIA
{mingjiel, sdiao, ximingl, jianh, xind, yejinc, jkautz, yidong}@nvidia.com

Abstract

Recent advances in reasoning-centric language models have highlighted reinforcement learning (RL) as a promising method for aligning models with verifiable rewards. However, it remains contentious whether RL truly expands a model's reasoning capabilities or merely amplifies high-reward outputs already latent in the base model's distribution, and whether continually scaling up RL compute reliably leads to improved reasoning performance. In this work, we challenge prevailing assumptions by demonstrating that prolonged RL (**ProRL**) training can uncover novel reasoning strategies that are inaccessible to base models, even under extensive sampling. We introduce ProRL, a novel training methodology that incorporates KL divergence control, reference policy resetting, and a diverse suite of tasks. Our empirical analysis reveals that RL-trained models consistently outperform base models across a wide range of pass@k evaluations, including scenarios where base models fail entirely regardless of the number of attempts. We further show that reasoning boundary improvements correlates strongly with task competence of base model and training duration, suggesting that RL can explore and populate new regions of solution space over time. These findings offer new insights into the conditions under which RL meaningfully expands reasoning boundaries in language models and establish a foundation for future work on long-horizon RL for reasoning. We release model weights to support further research:
https://huggingface.co/nvidia/Nemotron-Research-Reasoning-Qwen-1.5B

1 Introduction

Recent advances in reasoning-focused language models, exemplified by OpenAI-O1 [1] and DeepSeek-R1 [2], have marked a paradigm shift in artificial intelligence by scaling test-time computation. Specifically, test-time scaling enables long-form Chain-of-Thought (CoT) thinking and induces sophisticated reasoning behaviors, leading to remarkable improvements on complex tasks such as mathematical problem solving [3–6] and code generation [7, 8]. By continuously expending compute throughout the reasoning process—via exploration, verification, and backtracking—models boost their performance at the cost of generating longer reasoning traces.

At the heart of these advances lies reinforcement learning (RL), which has become instrumental in developing sophisticated reasoning capabilities. By optimizing against verifiable objective rewards rather than learned reward models, RL-based systems can mitigate the pitfalls of reward hacking [9–11] and align more closely with correct reasoning processes. However, a fundamental question remains under active debate within the research community: *Does reinforcement learning truly*

Preprint. Under review.

《PRORL：長期強化學習
拓展大型語言模型的推理邊界》
NVIDIA 2025.5.30

「菜，就多練」

SA

You

What is ‘lollipop’ backwards

ChatGPT

The word “lollipop” spelled backward is “poolpillol.”


思考中...
我操，用户彻底怒了。
You're absolutely right!