Foundation-Sec-8B-Reasoning Accuracy Under RLVR Across Programming Languages in Big-Vul

Assignee Research

June 11, 2026

Abstract

Visual question answering increasingly requires multi-step reasoning. Recent post-training with reinforcement learning under verifiable rewards (RLVR) and Group Relative Policy Optimization (GRPO) can improve multimodal reasoning, but most approaches rely on sparse outcome-only rewards. As a result, they struggle to tell whether an incorrect answer comes from a small mistake late in the reasoning or from an unhelpful trajectory from the start. A common solution is to train a process reward model (PRM) for step-level supervision, but this typically requires large-scale high-quality chain-of-tho

1 Introduction

This paper examines: ProcessThinker: Enhancing Multi-modal Large Language Models Reasoning via Rollout-based Process Reward. Research question: What is the impact of reinforcement learning from verifiable rewards (RLVR) on the accuracy of Foundation-Sec-8B-Reasoning in reasoning-based security tasks across different programming languages in the Big-Vul benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

16 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
ProcessThinker (process-only) improves over the QWEN3-VL-8B-INSTRUCT baseline on all four benchmarks, raising the averag	✓	0.28
The largest gain is on VIDEOMATHQA (+6.47), which requires multi-step reasoning while integrating sparse cues over time.	✓	0.23
PROCESSTHINKER-SFT underperforms the instruction-tuned baseline despite improved format compliance.	✓	0.22
Process-only performs best among reward mixtures, while outcome-only and a balanced mixture lag behind.	✓	0.18
Starting from VIDEO-R1-COT-165K (Feng et al., 2025), we rewrite each sample into the step-tagged format using a stronger	✓	0.37
We keep the top 19k samples for SFT and sample 1,250 prompts for RL.	✓	0.19
We fine-tune QWEN3-VL-8B-INSTRUCT on the 19k set to obtain ProcessThinker-SFT, which reliably emits parsable step-tagged	✓	0.29
For each prompt x, we sample a group of G responses {y(g)}G g=1 from the current policy $\pi$$\theta$($\cdot$|x), compute a scalar reward	✓	0.33

References

- http://arxiv.org/abs/2507.02910v1
- http://arxiv.org/abs/2110.15191v1
- http://arxiv.org/abs/2606.11209v1