Logic-Verified GRPO: Graded Z3 Process Rewards for Logical Reasoning in Small LLMs

Anonymous authors
Paper under double-blind review

Abstract

Recent work integrates symbolic solvers into reinforcement learning for LLM reasoning, but existing approaches typically use binary chain-level verification: a full reasoning trace is either correct or not. We introduce Logic-Verified GRPO, which uses the Z3 SMT solver to provide graded, step-level process rewards within GRPO training—each step is independently verified and receives proportional credit based on its formal validity. We evaluate on FOLIO and ProntoQA using Qwen2.5-3B-Instruct. While both GRPO variants improve accuracy over the baseline (+8–10pp), our key finding is that graded step verification produces markedly better epistemic calibration: the Z3-verified model achieves +14pp improvement on Unknown (unprovable) conclusions (55.6% vs. 41.7% baseline), while outcome-only GRPO actually degrades Unknown recognition (38.9%). This suggests that graded symbolic process rewards teach models to distinguish "valid proof found" from "no valid derivation exists"—a distinction invisible to outcome-only or binary verification rewards.

1 Introduction

Recent work on reinforcement learning with verifiable rewards (RLVR) has demonstrated that symbolic solvers can provide reward signals for training LLM reasoning (DeepSeek-AI, 2025; Shao et al., 2024). Several concurrent approaches integrate formal verification into GRPO training for logical reasoning (Chen et al., 2025; Xu et al., 2025), typically using binary chain-level verification where a complete reasoning trace receives a single pass/fail reward.

However, for logical reasoning—where the goal is to determine whether a conclusion is entailed, contradicted, or unprovable given premises—binary verification may be insufficient. A model that produces five valid steps and one invalid step should receive different feedback than one that produces entirely unsound reasoning, yet binary rewards treat both as failures. Moreover, the ability to recognize genuinely Unknown conclusions requires the model to understand when no valid proof can be constructed—a distinction that coarse reward signals may not teach.

We propose Logic-Verified GRPO, which provides graded, step-level Z3 process rewards: each reasoning step is independently parsed into first-order logic and verified against the premises using the Z3 SMT solver (de Moura & Bjørner, 2008), receiving proportional credit ($v \in \{0, 0.3, 0.5, 1.0\}$) based on verification outcome. Unlike binary chain-level approaches, this provides dense per-step supervision that rewards partially valid reasoning chains.

Our contributions are:

1. A graded Z3 step-level reward for GRPO where each step receives independent proportional credit, in contrast to binary chain-level verification (§2).
2. Evidence that graded step verification produces markedly better epistemic calibration: +14pp on Unknown conclusions vs. baseline on FOLIO (Han et al., 2022) and ProntoQA (Saparov & He, 2023), while outcome-only GRPO degrades Unknown recognition.

3. Analysis that the graded reward teaches qualitatively different reasoning: the model learns to distinguish provable from unprovable conclusions, with increasing step validity throughout training.

2 Method

2.1 Problem Setup

Given premises $P = \{p_1, \ldots, p_n\}$ and a conclusion c, the task is to determine whether c is True (entailed by P), False (contradicted by P), or Unknown (neither entailed nor contradicted). The model must output structured reasoning in a <reasoning>...</reasoning><answer>...</answer> format, with numbered steps citing logical rules.

2.2 Reward Functions

We define three reward components:

Correctness reward r_c: Binary signal—$r_c = 2.0$ if the extracted answer matches the ground truth, 0.0 otherwise.

Format reward r_f: Incremental credit (0–1.0) for structural elements: <reasoning>/</reasoning> tags (+0.2 each), <answer>/</answer> tags (+0.2 each), and ≥ 2 numbered reasoning steps (+0.2).

Z3 step reward r_z (our contribution): For each reasoning step s_i in the model's output:

1. Parse the step to identify the logical rule being applied (e.g., modus ponens, universal instantiation).
2. Extract the derived proposition.
3. Translate it to a Z3 assertion and check: is $\neg s_i$ unsatisfiable given the premises? If so, s_i is a valid consequence (full credit). If satisfiable (indeterminate), partial credit is given for correctly citing a valid rule.

The Z3 reward is the mean step validity: $r_z = \frac{1}{|S|} \sum_i v(s_i)$ where $v(s_i) \in \{0, 0.3, 0.5, 1.0\}$ depending on verification outcome. Crucially, this is a graded reward—unlike binary chain-level approaches that assign pass/fail to entire traces, each step independently contributes to the reward, providing dense supervision even for partially valid reasoning chains.

The outcome-only condition uses $\{r_c, r_f\}$; the Z3-verified condition uses $\{r_c, r_f, r_z\}$. All rewards are passed as separate functions to GRPO, which computes group-relative advantages over $K = 8$ generations per prompt.

2.3 Training Configuration

We train Qwen2.5-3B-Instruct (Qwen Team, 2024) with LoRA (Hu et al., 2022) ($r = 32$, $\alpha = 64$, all linear layers) for 250 GRPO steps. Training uses batch = 8, $K = 8$ generations, lr $= 5 \times 10^{-6}$ with cosine schedule, bf16 precision, and AdamW-8bit on a single H100 GPU ($\sim$90 minutes per run).

3 Experimental Setup

3.1 Datasets

We evaluate on two established logical reasoning benchmarks:

FOLIO (Han et al., 2022): 1,204 examples of natural language reasoning grounded in first-order logic, with human-annotated FOL translations. Problems span complex real-world premises with True/False/Unknown labels.

ProntoQA (Saparov & He, 2023): 500 examples of multi-hop chain reasoning over fictional ontologies (e.g., "Every tumpus is a vumpus. Vumpuses are numpuses. Max is a tumpus. Is Max a numpus?"). Requires 3–15 step deduction chains.

We split into train (2,259), validation (282), and test (283) sets. Only FOLIO and ProntoQA examples are used for evaluation; synthetic examples included in training are excluded from test metrics.

3.2 Conditions

Baseline: Qwen2.5-3B-Instruct zero-shot with a structured reasoning prompt (no training).

Outcome-only GRPO: Trained with $\{r_c, r_f\}$—correctness and format rewards only.

Z3-verified GRPO: Trained with $\{r_c, r_f, r_z\}$—adds the Z3 step verification reward.

4 Results

4.1 Main Results

Table 1 presents results on the 166 FOLIO + ProntoQA test examples.

Table 1: Test accuracy on FOLIO + ProntoQA (166 examples). Format = fraction with correct XML structure. Steps = fraction with ≥ 2 numbered reasoning steps.

Condition	Accuracy	FOLIO	ProntoQA	Format	Steps
Baseline (zero-shot)	40.4%	50.5%	21.1%	11.4%	12.7%
Outcome-only GRPO	50.0%	51.4%	47.4%	65.7%	81.3%
Z3-verified GRPO	48.2%	48.6%	47.4%	69.9%	86.7%

Both GRPO conditions substantially improve over the baseline, with the largest gains on ProntoQA (+26pp), where multi-hop chain reasoning benefits most from RL training. Overall accuracy is similar between the two trained conditions (50.0% vs. 48.2%), but they differ meaningfully in reasoning behavior, as analyzed below.

4.2 Epistemic Calibration: Unknown Handling

Table 2: Accuracy by answer type (FOLIO + ProntoQA, $n = 166$). The Z3-verified model shows substantially improved recognition of unprovable conclusions.

Condition	True ($n = 75$)	False ($n = 55$)	Unknown ($n = 36$)
Baseline	41.3%	38.2%	41.7%
Outcome-only GRPO	60.0%	43.6%	38.9%
Z3-verified GRPO	52.0%	38.2%	55.6%

Table 2 reveals the key qualitative difference between conditions. The outcome-only model improves primarily on True conclusions (+18.7pp over baseline), suggesting it learns to predict "True" more aggressively. In contrast, the Z3-verified model shows its largest gain on Unknown conclusions (+13.9pp over baseline, +16.7pp over outcome-only), indicating it learns to recognize when a conclusion cannot be derived from the premises.

This is the central finding: the Z3 step reward, by verifying each intermediate derivation, teaches the model to distinguish between "I found a valid proof" and "I cannot construct a valid proof"—a distinction invisible to outcome-only rewards.

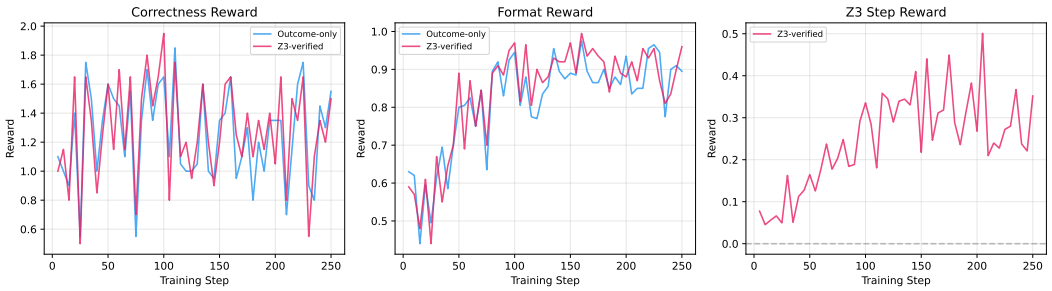

Figure 1: Training reward curves over 250 steps. Left: correctness increases similarly for both conditions. Center: format compliance converges. Right: Z3 step reward increases from 0.08 to 0.35.

4.3 Training Dynamics

Figure 1 shows that both conditions follow similar correctness and format trajectories, but the Z3 step reward (right panel) increases steadily from ∼0.08 to ∼0.35, confirming the model learns to produce formally valid reasoning steps. KL divergence remains below 0.03 throughout.

5 Related Work

RLVR for reasoning. DeepSeek-R1 (DeepSeek-AI, 2025) and DeepSeekMath (Shao et al., 2024) demonstrated GRPO with verifiable rewards for mathematical reasoning. We apply this paradigm to logical reasoning with a fundamentally different reward structure: graded per-step verification rather than binary outcome correctness.

Symbolic verification in RL training. Several concurrent works integrate formal solvers into LLM training. ProSFI (Chen et al., 2025) uses Z3 and other provers within GRPO but employs binary chain-level verification (pass/fail for the entire trace) and requires structured formal intermediates. LogicReward (Xu et al., 2025) uses Isabelle for step-level verification in DPO (not GRPO) training. Our key distinction is graded per-step rewards where each step independently receives proportional credit—enabling denser supervision that, as we show, specifically improves handling of unprovable conclusions.

Post-hoc verification. FoVer (Kamoi et al., 2025) uses Z3 to annotate step-level error labels for training process reward models, not as a live reward during RL. VeriCoT (Feng et al., 2025) applies Z3 for chain-of-thought consistency checking during SFT data curation and inference-time self-correction. Neither integrates Z3 into the RL training loop.

Benchmarks. FOLIO (Han et al., 2022) and ProntoQA (Saparov & He, 2023) both include Unknown labels, which we show are differentially improved by graded step verification—a finding not reported in prior work on formal verification for logical reasoning.

6 Conclusion

We have shown that graded step-level Z3 verification within GRPO produces qualitatively different reasoning behavior compared to outcome-only rewards. While accuracy gains are similar, the Z3-verified model demonstrates markedly better epistemic calibration—particularly +14pp on Unknown conclusions—suggesting that dense per-step symbolic rewards teach models to distinguish provable from unprovable conclusions. This finding complements concurrent work on binary chain-level verification (Chen et al., 2025) by showing that reward granularity matters for epistemic calibration.

Limitations. Our Z3 reward relies on parsing natural language to formal logic, which is imperfect. The accuracy gap is small and within single-seed variance; future work should

include multiple seeds, larger test sets, and direct comparison with binary chain-level verification approaches.

References

Luoxin Chen, Yichi Zhou, and Huishuai Zhang. Learning to generate formally verifiable step-by-step logic reasoning via structured formal intermediaries. In Submitted to International Conference on Learning Representations (ICLR), 2025.

Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In Tools and Algorithms for the Construction and Analysis of Systems (TACAS), volume 4963 of Lecture Notes in Computer Science, pp. 337–340. Springer, 2008.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. VeriCoT: Neuro-symbolic chain-of-thought validation via logical consistency checks. arXiv preprint arXiv:2511.04662, 2025.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Ansong Ni, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. arXiv preprint arXiv:2209.00840, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Generalizable process reward models via formally verified training data. arXiv preprint arXiv:2505.15960, 2025.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In International Conference on Learning Representations (ICLR), 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Jundong Xu, Hao Fei, Huichi Zhou, Xin Quan, Qijun Huang, Shengqiong Wu, William Yang Wang, Mong-Li Lee, and Wynne Hsu. LogicReward: Incentivizing LLM reasoning via step-wise logical supervision. arXiv preprint arXiv:2512.18196, 2025.