IOWA STATE UNIVERSITY

Meta

RESTRAIN: From Spurious Votes to Signals — Self-Driven RL with Self-Penalization

Zhaoning Yu
Email: zn.yu1029@gmail.com

IOWA STATE UNIVERSITY

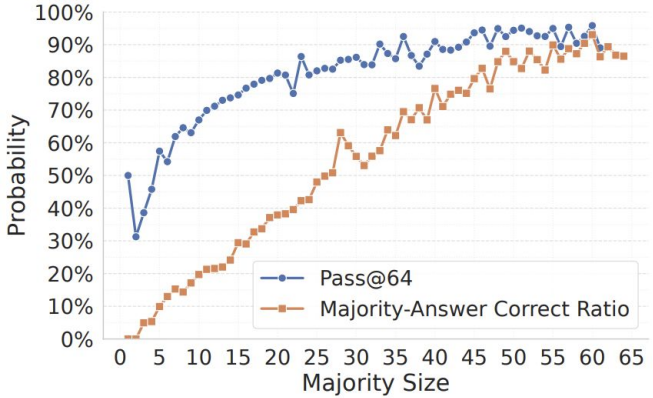
Background

- **The RLVR Breakthrough:** Reinforcement Learning with Verifiable Rewards recently unlocked massive reasoning gains.
- **The Supervised Ceiling:** High-quality human reasoning data is scarce, expensive, and ultimately finite.

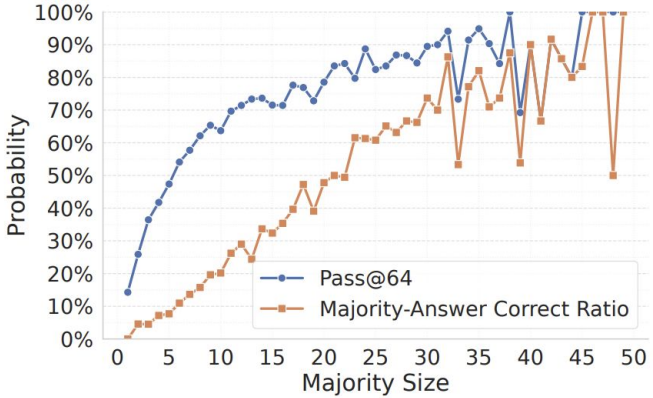
The Core Question: Can an AI autonomously improve its reasoning capabilities using exclusively unlabeled prompts, without falling into a destructive feedback loop of its own hallucinations?

Motivation

The large gap between Pass@64 and majority-vote shows that correct answers often diverge from majority votes. Accuracy also drops sharply when the majority size is small, revealing that majority votes can carry spurious signals. These observations motivate our self-penalizing framework, which seeks robust promising reasoning paths beyond unreliable majority votes.



(a) Qwen3-4B-Base



(b) Octothinker Hybrid 8B base

Introducing RESTRAIN

Meta

1. Pseudo-Labeling

RESTRAIN weights all generated answers proportionally to their frequency, capturing nuanced signals.

2. Rollout Penalization

When the majority consensus is low, the model zeroes out rewards and applies a negative penalty to all rollouts to deter spurious paths.

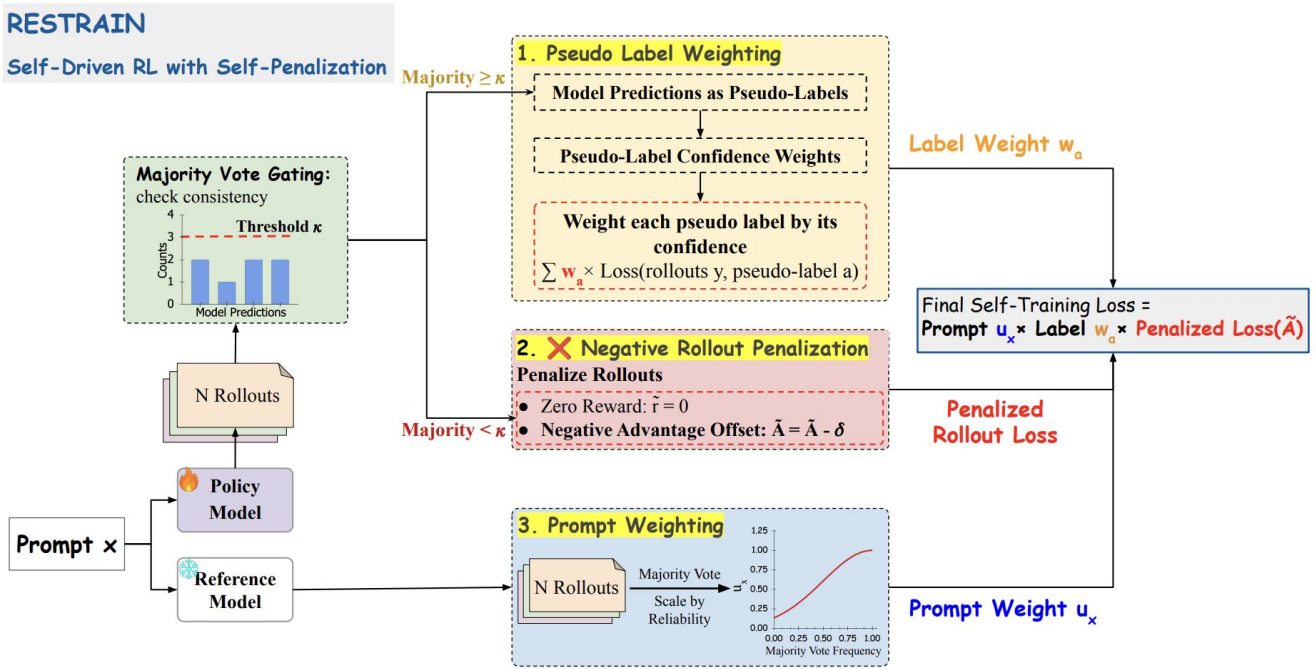
3. Prompt Weighting

Updates are scaled based on the prompt's self-consistency score. High-confidence prompts yield larger updates, mitigating noise.

Pseudo-Label Weighting

RESTRAIN considers all predicted answers

By weighting each prediction proportionally to its frequency, the model escapes the trap of hard majority collapse. This frequency aware scaling smooths out learning signals and suppresses rare, spurious reasoning paths effectively.



Negative Penalization

Meta

The Trap of Unreliability

When consensus is low, existing unsupervised methods falter. **Majority voting** provides highly unreliable signals when the model lacks certainty. Rewarding these weak majorities severely degrades Pass@n accuracy and induces sudden training collapse.

The RESTRAIN Solution

We assume all responses are incorrect when the majority count falls below a set threshold. We zero out the rewards entirely and apply a uniform negative advantage offset

Prompt-Level Weighting

Meta

Varying Uncertainty: For some prompts, the model exhibits high uncertainty, while for others it produces highly consistent, confident responses.

Scaled Updates: To account for this variation, we scale the policy update for each prompt by a fixed weight that accurately reflects the model's confidence.

Noise Prevention: Low-confidence prompts automatically receive smaller update weights, preventing the model from amplifying internal noise.

Offline Precomputation: To prevent spurious feedback loops, these weights are computed once offline using a frozen base model and kept constant.

Experimental Setup

Datasets

DAPO-14k-MATH: 14k English prompts for core mathematical reasoning.
Synthetic S1k: 5k synthetic reasoning instructions to test generative scaling limits.

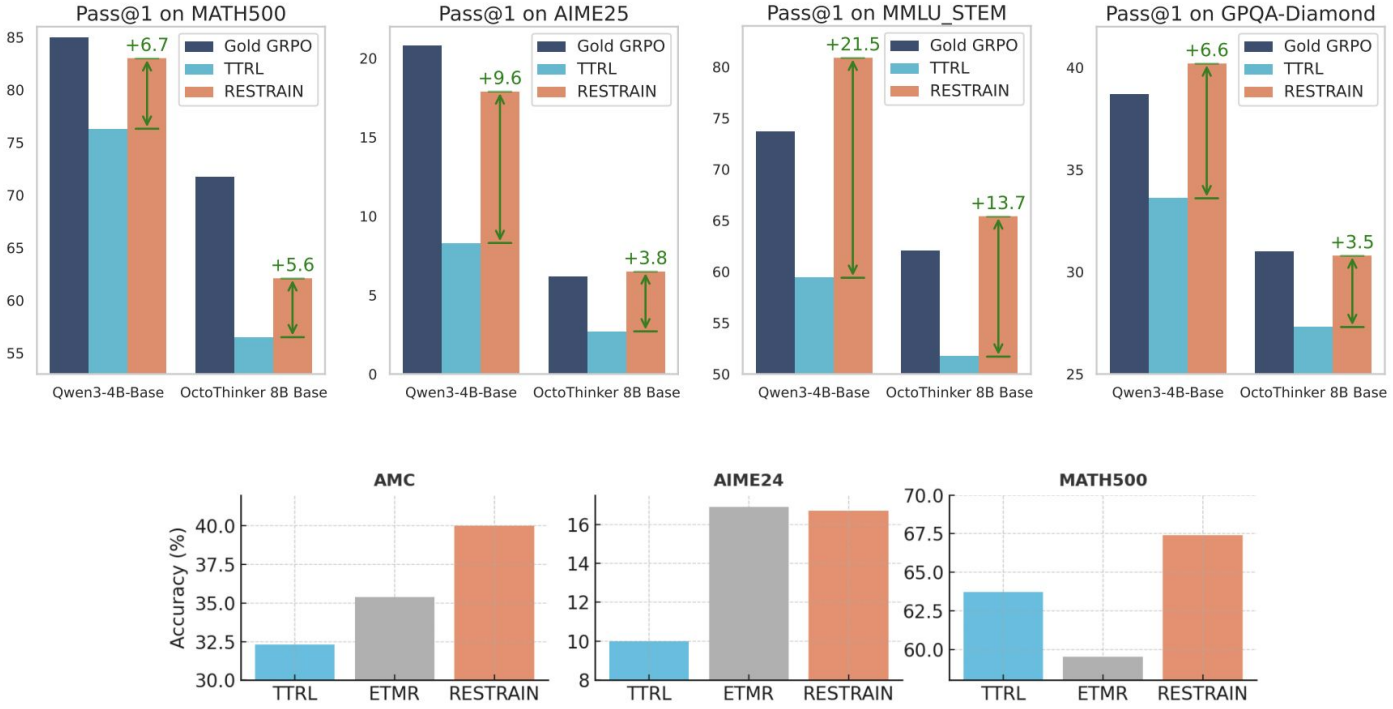
Base Models

Qwen3-4B-Base and **OctoThinker Hybrid-8B-Base**

Benchmarks

MATH-500, AIME25, Olympiad Bench(math), Minerva_math, MMLU STEM and GPQA-Diamond

Experimental Results



Thank you!