

Efficient RL for Large Language Models

with Intrinsic Exploration (PREPO)

Yan Sun^{*^}, Jia Guo[^], Stanley Kok^{*}, Zihao Wang[^], Zujie Wen[^], Zhiqiang Zhang[^]

^{*}National University of Singapore, [^]Ant Group

NeurIPS 2025 Efficient Reasoning Workshop



GitHub

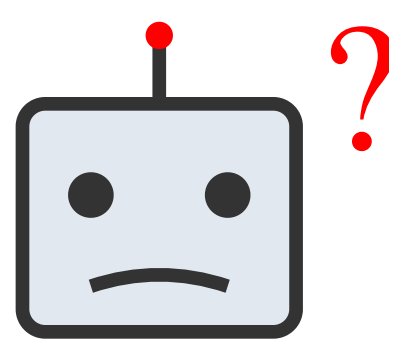


Paper

TL;DR

PREPO reduces Reinforcement Learning (RLVR) training costs by **3x**. It uses "intrinsic" metrics—**Prompt Perplexity** & **Rollout Entropy**—to filter data, guiding exploration.

1. The Problem



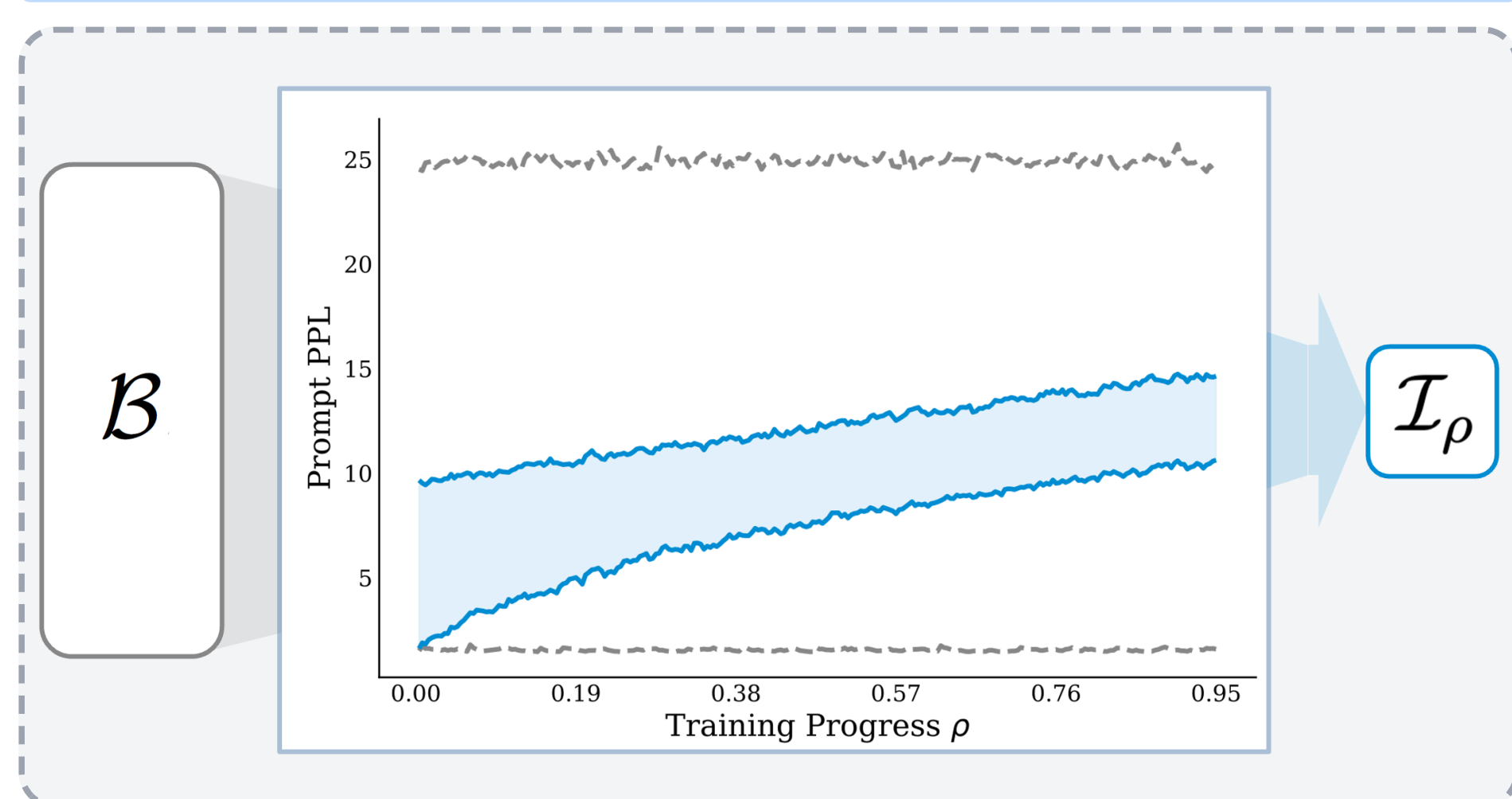
- **Costly**: Standard RLVR generates thousands of rollouts.
- **Inefficient**: Many samples are too easy or too hard (zero advantage).
- **Goal**: Data-efficient RLVR training using data **intrinsic properties**.

2. Method: Online Prompt Selection

Strategy: Prompt Perplexity

Use perplexity as a proxy for adaptability. Train on **Low PPL** to **High PPL** prompts.

$$P(\rho) = \exp \left(-\frac{1}{N} \sum_{t=1}^N \log p(x_t | x_{prev}) \right)$$



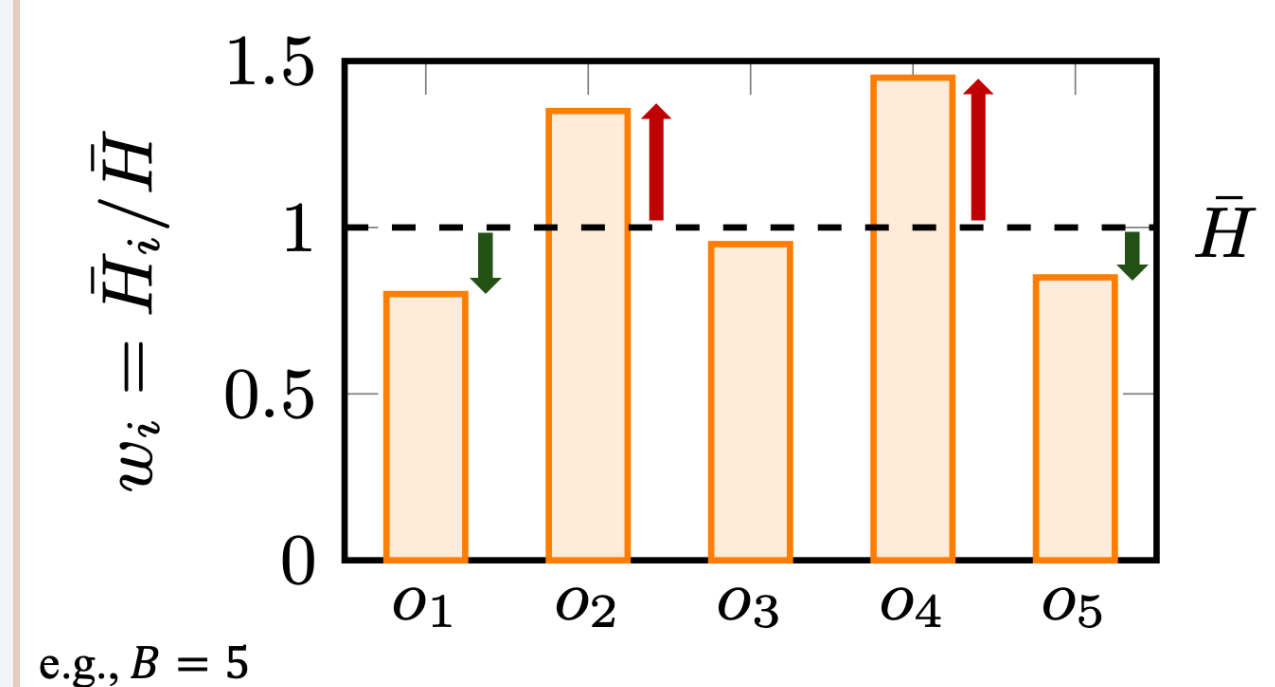
3. Method: Rollout Weighting

Strategy: Relative Entropy

Prioritize diverse reasoning paths. Weight rollouts by their average token-level entropy.

$$w_i = \frac{\bar{H}_i}{\bar{H}}, \quad \bar{H}_i = -\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \sum_{v \in V} p(v|x_t) \log p(v|x_t)$$

mini-batch
(B rollouts)



4. Results

Tested on Qwen & Llama (MATH500, AIME, Olympiad).

