Reasoning with RL vs Sampling

Distribution Sharpening or New Capabilities?

Lezhi (Carrie) Tan

Motivation

- Reinforcement learning has become a dominant approach for improving LLM reasoning.
- Methods like GRPO power models such as DeepSeek-R1.
- But an important question remains:

Does RL actually create new reasoning abilities?

Two Competing Hypotheses

Hypothesis 1: Learning new capabilities

- RL training teaches the model new reasoning patterns.

Hypothesis 2: Distribution sharpening

- Base models already contain reasoning traces.
- RL only reshapes the probability distribution.

Group Relative Policy Optimization (GRPO)

- a given prompt x
- samples a group of solutions y_i, receiving reward r_i
- $$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(y_i|x)}{\pi_0(y_i|x)}A_i, \text{clip}\left(\frac{\pi_\theta(y_i|x)}{\pi_0(y_i|x)}, 1-\epsilon, 1+\epsilon\right)A_i\right) \quad (1)$$

  where

$$A_i = \frac{r_i - \text{mean}(r_1, ..., r_G)}{\text{std}(r_1, ..., r_G)}$$

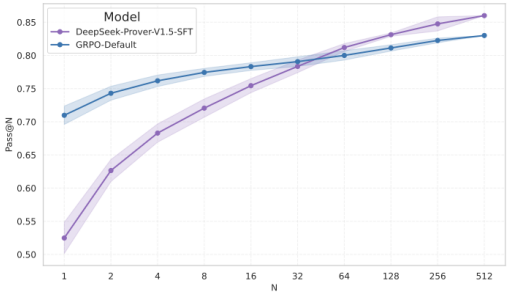
Evaluation Metric: pass@N

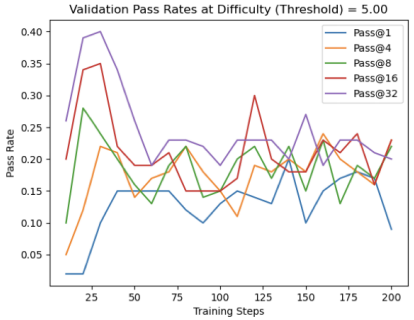
Reasoning tasks use sampling.

$$pass@N = 1 - (1 - p)^N$$

- success if **any** of N samples is correct
- commonly used in
  - code generation
  - theorem proving,
  - math reasoning

Empirical Observation





- GRPO improves pass@1
- But performance degrades for large N

Why RL Mostly Sharpens the Distribution

RL increases probability of correct solutions by a factor:

$$\frac{\pi_{RL}(y|x)}{\pi_0(y|x)} \approx 1 + \epsilon$$

Therefore

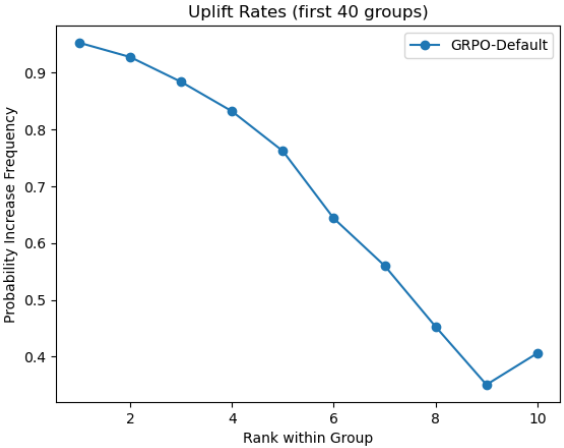
$$p_{RL} \approx (1 + \epsilon)p_0$$

Expected pass rate:

$$E[pass@N] \approx 1 - (1 - (1 + \epsilon)p_0)^N$$

- Increase in passN requires increasing the probability of solutions with $p_0 \approx 1/N$.

Rank Bias



- High-probability solutions are reinforced
- Rare correct solutions are ignored

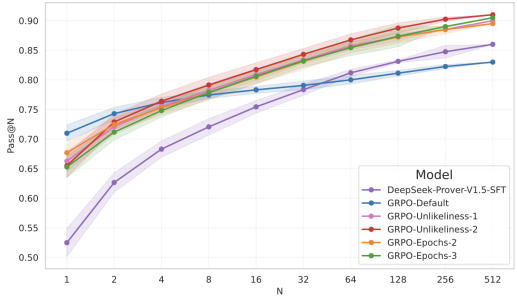
Unlikeliness Reward

Solution proposed in Paper 1:

- Penalize high-probability correct samples

$$r_i = R(x, y_i) \left(1 - \beta_{\text{rank}} \frac{G - \text{rank}(y_i)}{G}\right)$$

- Encourage learning from rare correct solutions



Key Takeaway from Paper 1

RL improves reasoning mainly through

distribution sharpening

Paper 2: Reasoning with Sampling

Key idea:

If RL only sharpens distributions...

can we reproduce this effect without training?

Power Distribution

Let the base LLM define a sequence distribution

$$p(x_{0:T}) = \prod_{t=0}^{T} p(x_t \mid x_{<t})$$

We define the **power distribution**

$$p^{\alpha}(x_{0:T}) = \frac{p(x_{0:T})^{\alpha}}{\sum_{x'} p(x'_{0:T})^{\alpha}}$$

where

- $\alpha > 1$ sharpens the distribution
- high-likelihood paths receive more weight
- low-likelihood paths are suppressed

Low-Temperature Sampling

Standard sampling modifies the token distribution:

$$p_{\text{temp}}(x_t \mid x_{<t}) = \frac{p(x_t \mid x_{<t})^{\alpha}}{\sum_{x'_t} p(x'_t \mid x_{<t})^{\alpha}}$$

Important difference:

- Temperature modifies **local token probabilities**
- Power distribution modifies **whole sequence probabilities**

Example: Power Distribution vs Temperature

Vocabulary: $\{a, b\}$ Sequences: aa, ab, ba, bb

$$p(aa) = 0, \quad p(ab) = 0.40, \quad p(ba) = 0.25, \quad p(bb) = 0.25$$

Power distribution ($\alpha = 2$)

$$p^{\alpha}(x_0 = a) \propto 0^2 + 0.40^2 = 0.160, \; p^{\alpha}(x_0 = b) \propto 0.25^2 + 0.25^2 = 0.125$$

Low-temperature sampling

$$p_{temp}(x_0 = a) \propto (0 + 0.40)^2 = 0.160, \; p_{temp}(x_0 = b) \propto (0.25 + 0.25)^2 = 0.25$$

Power sampling prefers tokens with **fewer but higher-likelihood future paths**, while temperature sampling prefers tokens with **many moderate-probability continuations**.

Sampling Challenge

Direct sampling from p^{α} is intractable: need normalization.
Solution:

- Markov Chain Monte Carlo
- Metropolis-Hastings algorithm

Metropolis-Hastings

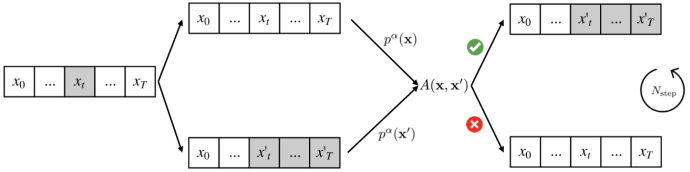


Figure 3: **Illustrating Metropolis-Hastings with random resampling.** A random index t is selected and a new candidate is generated by resampling. Based on the relative likelihoods, the candidate is accepted or rejected, and the process repeats.

Markov chain of sample sequences $(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^n)$ using an arbitrary *proposal distribution* $q(\mathbf{x}|\mathbf{x}^i)$ to select the next candidate $\mathbf{x}^{i+1}$. With probability

$$A(\mathbf{x}, \mathbf{x}^i) = \min\left\{1, \frac{p^\alpha(\mathbf{x}) \cdot q(\mathbf{x}^i|\mathbf{x})}{p^\alpha(\mathbf{x}^i) \cdot q(\mathbf{x}|\mathbf{x}^i)}\right\}, \tag{9}$$

1. propose new sequence
2. compute likelihood ratio
3. accept or reject

Power Sampling Algorithm

Goal: sample sequences from

$$p^{\alpha}(x_{0:T})$$

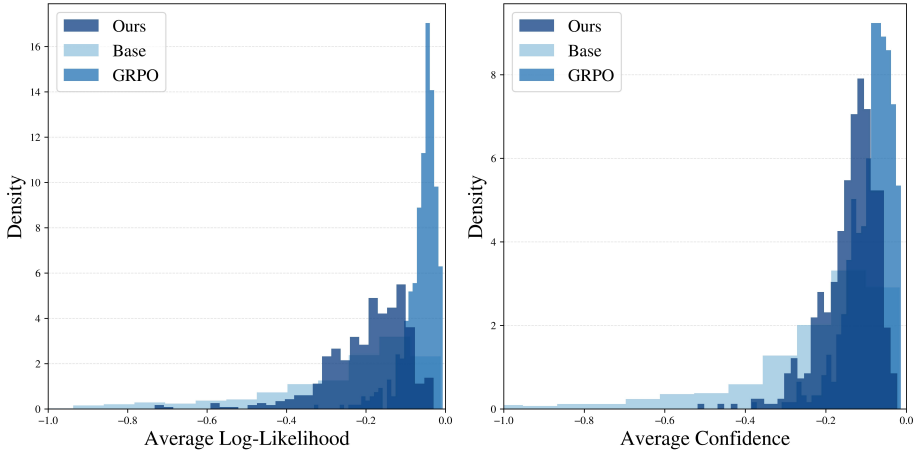
Algorithm structure:

1. Grow the sequence block by block (block length B)
2. Target distribution for prefix:

$$\pi_k(x_{0:kB}) \propto p(x_{0:kB})^{\alpha}$$

3. start from an index m randomly chosen from $[(k+1)B]$, resample completion $x_{m:(k+1)B}$
4. Use Metropolis–Hastings to correct proposals
5. Fix the block and move forward

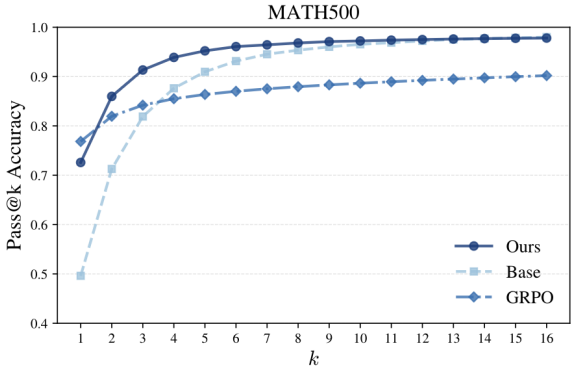
Experimental Results



Power sampling:

- matches RL performance
- sometimes outperforms RL

Diversity Comparison



- RL collapses diversity
- sampling maintains exploration

Takeaways

- RL reasoning may primarily sharpen distributions
- Base models already contain latent reasoning abilities
- Better sampling can unlock these capabilities

Discussion

Open questions:

- When does RL create genuinely new reasoning skills?
- How far can inference-time scaling go?
- How to elicit low-probability correct answers?