北京大学 PEKING UNIVERSITY

The *Entropy* Mechanism of Reinforcement Learning for Reasoning Language Models

Ganqu Cui[1]*, Yuchen Zhang[1,4]*, Jiacheng Chen[1]*, Lifan Yuan[3], Zhi Wang[5], Yuxin Zuo[2], Haozhan Li[2], Yuchen Fan[1], Huayu Chen[2], Weize Chen[2], Zhiyuan Liu[2], Hao Peng[3], Lei Bai[1], Wanli Ouyang[1], Yu Cheng[1,6]†, Bowen Zhou[1,2]†, Ning Ding[2,1]†

[1] Shanghai AI Laboratory [2] Tsinghua University [3] UIUC [4] Peking University [5] Nanjing University [6] CUHK

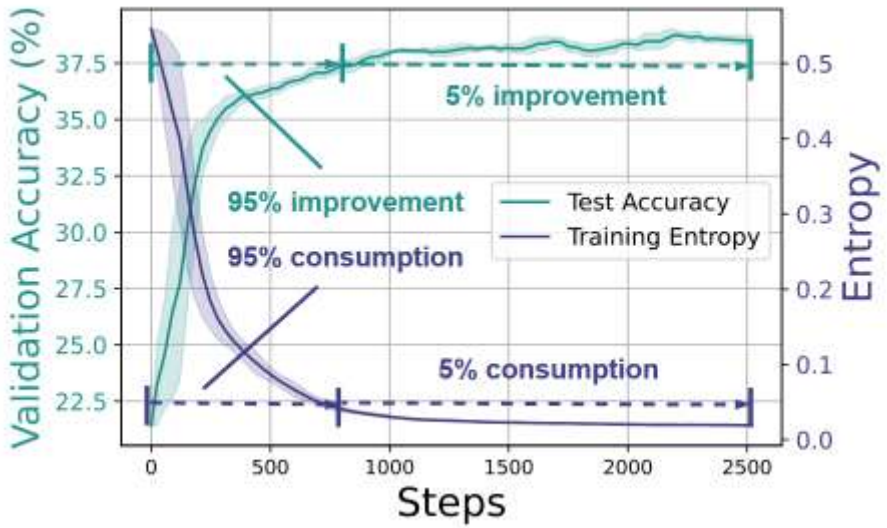
Code: https://github.com/PRIME-RL/Entropy-Mechanism-of-RL

熵崩溃 (Entropy Collapse)

在没有干预的RL训练中，策略熵（Policy Entropy）在早期阶段急剧下降至接近于零 。

探索能力减弱 (Diminished Exploration): 模型变得“过度自信”，停止探索新的、可能有价值的推理路径 。

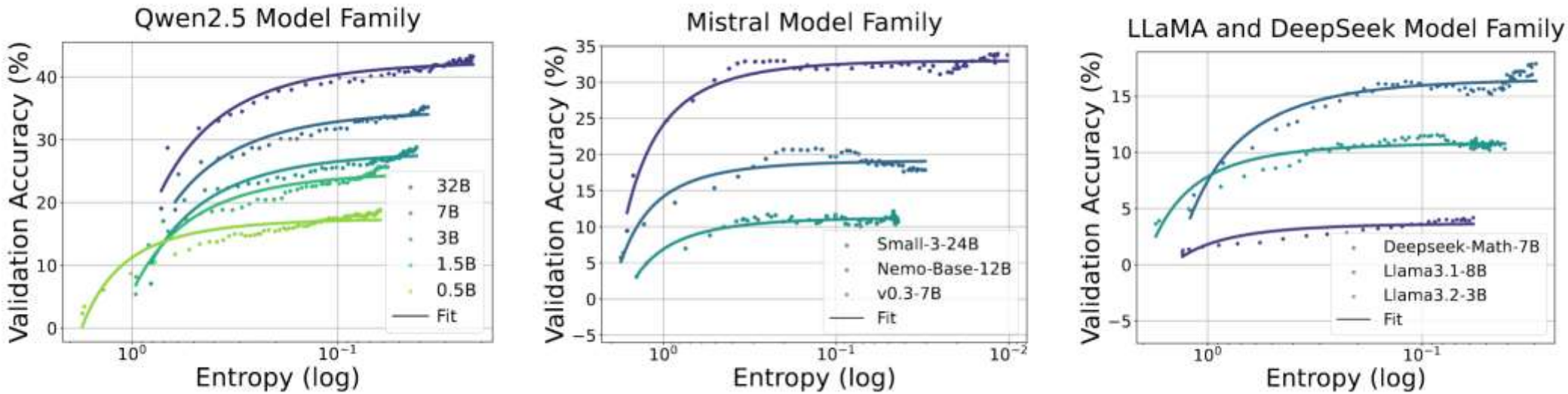
性能饱和 (Performance Saturation): 模型性能（如下游任务准确率）迅速达到一个平台期，无法进一步提升 。



超过95%的熵消耗和性能提升发生在训练的极早期

性能 (R) 和熵 (H) 之间的关系不是随机的，而是高度可预测的

论文发现它们的关系可以通过一个简单的指数函数来拟合： $R = -a\exp(\mathcal{H}) + b,$



当熵耗尽 (H→0) 时，性能的理论上限是可预测的 (R=−a+b) 。

可预测性: 可以在训练早期（高熵时）预测训练后期（低熵时）的最终性能 。

RL 训练过程只是在以一种预定的方式“用熵换取性能” 。

这个定律在Qwen2.5, Mistral, LLaMA等不同模型家族和不同大小上都成立

为什么熵会单调递减?

分析了Softmax策略（LLM）的熵变化。发现熵的一步变化（ΔH）与动作对数概率和logits变化（Δz）的协方差成负相关 。

$$\mathcal{H}(\pi_\theta^{k+1}) - \mathcal{H}(\pi_\theta^k) \approx \mathbb{E}_{s\sim d_{\pi_\theta}}\left[\mathcal{H}(\pi_\theta^{k+1}|s) - \mathcal{H}(\pi_\theta^k|s)\right] \approx \mathbb{E}_{s\sim d_{\pi_\theta}}\left[-Cov_{a\sim\pi_\theta^k(\cdot|s)}\left(\log \pi_\theta^k(a|s),\ z_{s,a}^{k+1} - z_{s,a}^k\right)\right]$$

在策略梯度和自然策略梯度算法中，logits的变化（Δz）又与其优势值 (Advantage A(s,a)) 成正比 。

最终机制:

$$\mathcal{H}(\pi_\theta^{k+1}|s) - \mathcal{H}(\pi_\theta^k|s) \approx -\eta \cdot Cov_{a\sim\pi_\theta^k(\cdot|s)}\left(\log \pi_\theta^k(a|s)\ , A(s,a)\right)$$

直观解释:

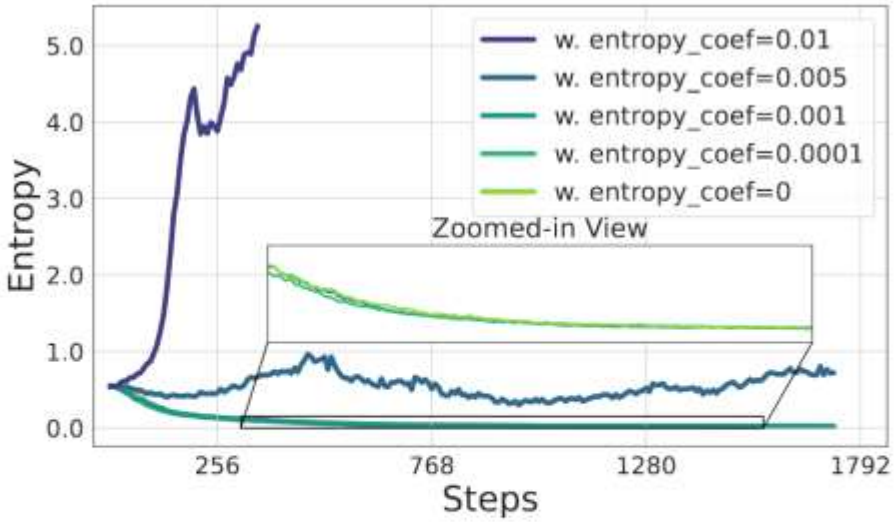
- 高概率 + 高优势 动作 (Easy Win): Cov(·) 为正 ⟹ΔH 为负 ⟹ 熵减少 。
- 低概率 + 高优势 动作 (Rare Discovery): Cov(·) 为负 ⟹ΔH 为正 ⟹ 熵增加 。

现实: 训练中，模型不断利用 “Easy Wins”，导致协方差持续为正，熵因此单调下降 。

正则化防止熵崩溃

1. 最大熵RL (MaxEnt RL): 直接添加熵正则化项（即熵损失）

   对超参极其敏感（容易导致熵爆炸或无效），且不提升性能 。



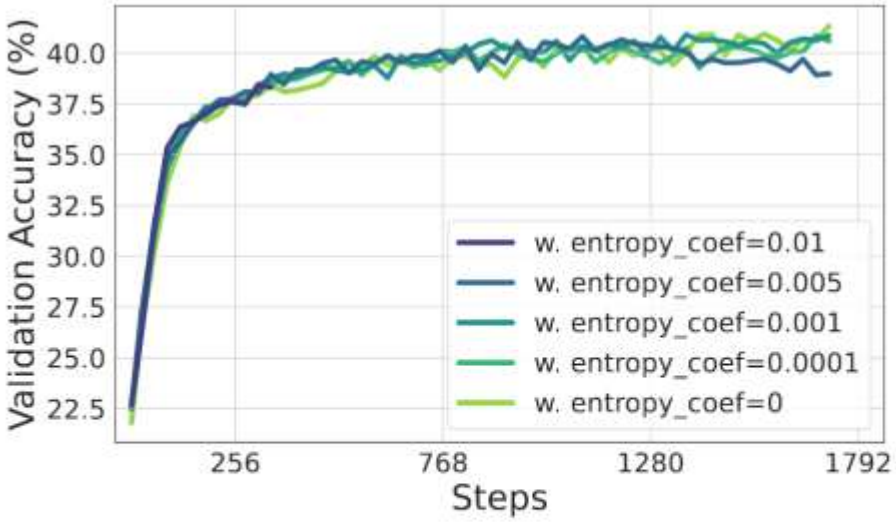


Figure 9: The policy entropy and validation accuracy of adding entropy loss where $L_{\text{ent}} = L - \alpha\mathcal{H}(\pi_\theta)$. L is the original loss and α is the coefficient of entropy loss.

正则化防止熵崩溃

2. 直接添加策略模型和参考模型的KL散度正则项
对超参极其敏感（容易导致熵爆炸或无效），且不提升性能 。

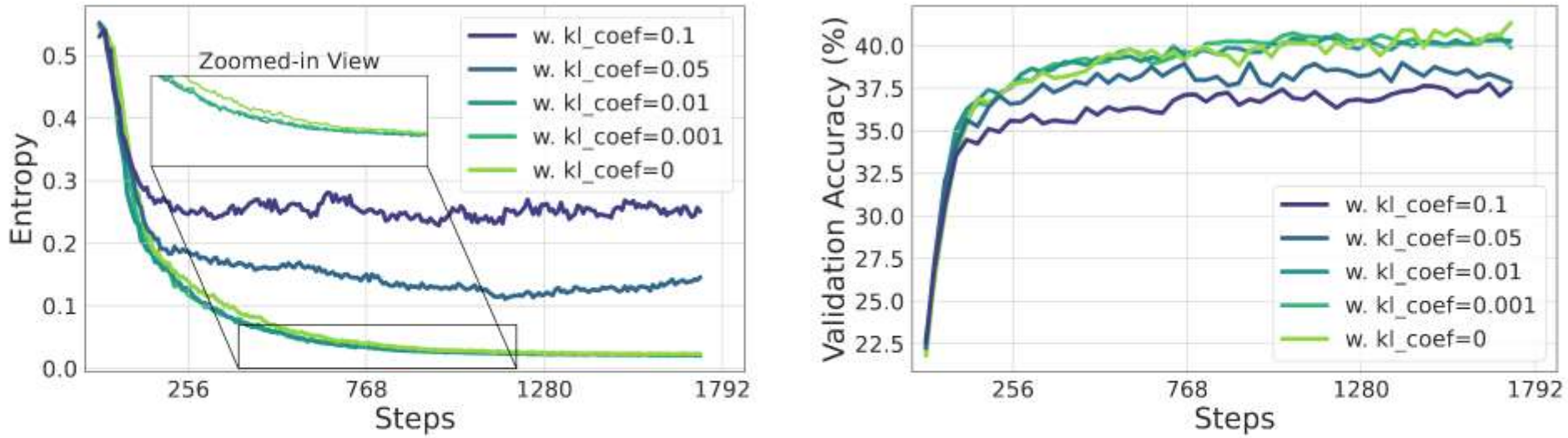


Figure 10: The policy entropy and validation accuracy of adding KL penalty between policy and reference model where $L_{\mathrm{KL}} = L + \beta \mathbb{D}_{\mathrm{KL}}(\pi_\theta || \pi_{\mathrm{ref}})$. L is the original loss and β is the coefficient of KL loss.

抑制高协方差

论文的新思路: 既然问题出在“高协方差”上，那就直接控制它 。

分析发现 (Table 1)，极少数 (Top 0.02%) 的 "离群" Token 贡献了极高的协方差值，它们是导致崩溃的主因 。

1. Clip-Cov (裁剪协方差):

机制: 识别高协方差的 Token。

操作: 随机选择其中一小部分 (r)，并分离它们的梯度 (detach gradients) 。

2. KL-Cov (KL惩罚协方差):

机制: 识别协方差最高的 Top-k% 的 Token 。

操作: 仅对这些 Token 施加 KL 散度惩罚（使其与旧策略保持一致）。

Table 1: Covariance distribution of Qwen2.5-7B in training step 1.

Group	Mean Value
Top 0.02%	5.654
Top 0.2%	3.112
Top 2%	1.385
Top 20%	0.351
Top 50%	0.152
All	0.003

实验结果

我们的方法有效保持了探索性。

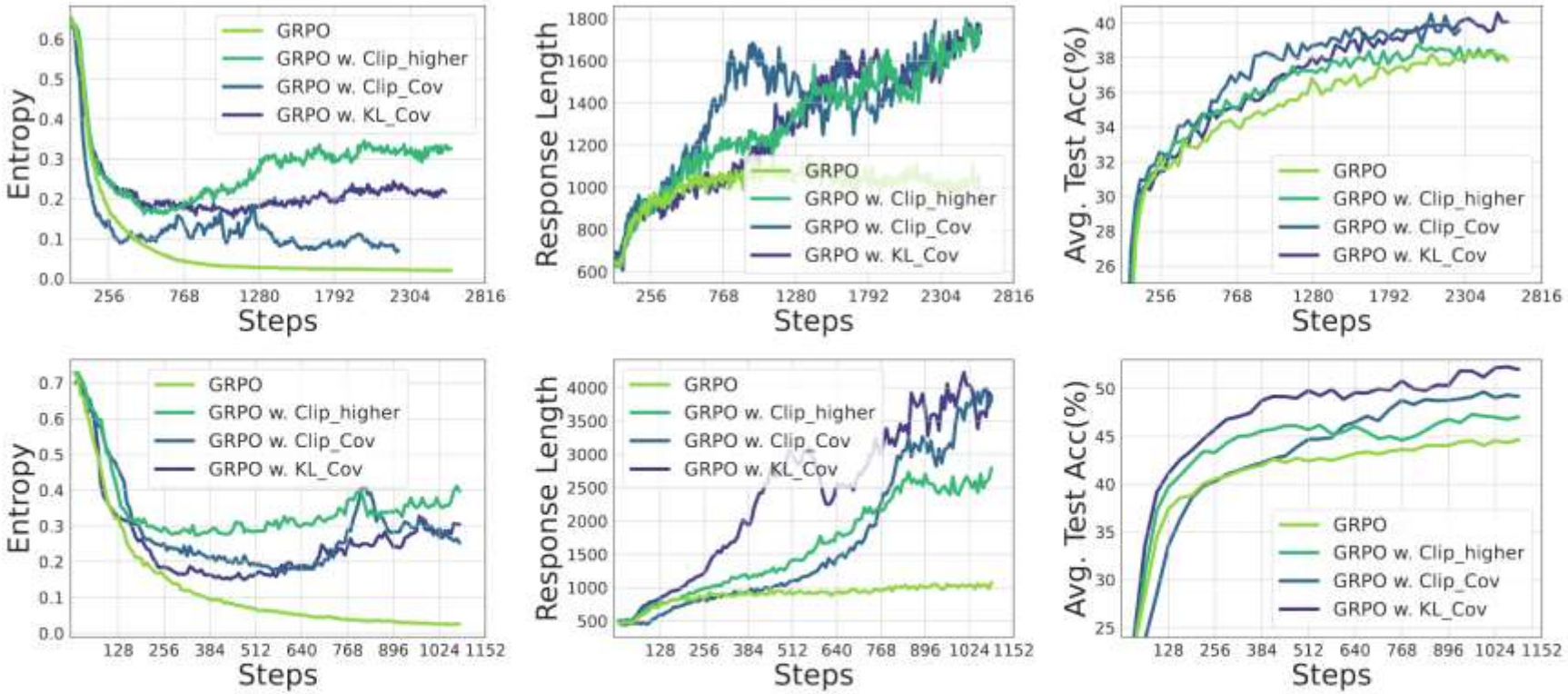


Figure 11: Training Qwen2.5-7B (*Top*) / Qwen2.5-32B (*bottom*) with GRPO with/without our methods. *Left:* Entropy dynamics. Our methods uplift policy entropy from collapse, enabling sustained exploration. *Middle:* Our method also incentivizes longer responses compared with vanilla GRPO. *Right:* The policy model consistently outperforms the baseline on testsets, avoiding performance plateaus.

总结

1: 熵崩溃是真实存在的。 在LLM的RL训练中，策略熵会迅速崩溃，导致性能饱和，这是扩展RL的主要障碍 。

2: 性能与熵的交换是可预测的。 性能与熵遵循$\mathrm{R} = -aexp^{H} + b$的规律，这揭示了一个可预测的“性能天花板” 。

3: 协方差是entropy collapse问题的根源。 熵的下降是由 "高概率、高优势" 动作带来的正协方差驱动的 。

4: 精准控制是关键。 通过 Clip-Cov 和 KL-Cov 直接抑制高协方差 Token，可以有效防止熵崩溃，实现持续的性能提升，尤其对大模型效果显著 。

北京大学 PEKING UNIVERSITY

Reasoning with Sampling: Your Base Model is Smarter Than You Think

Aayush Karan[1], Yilun Du[1]

[1]Harvard University

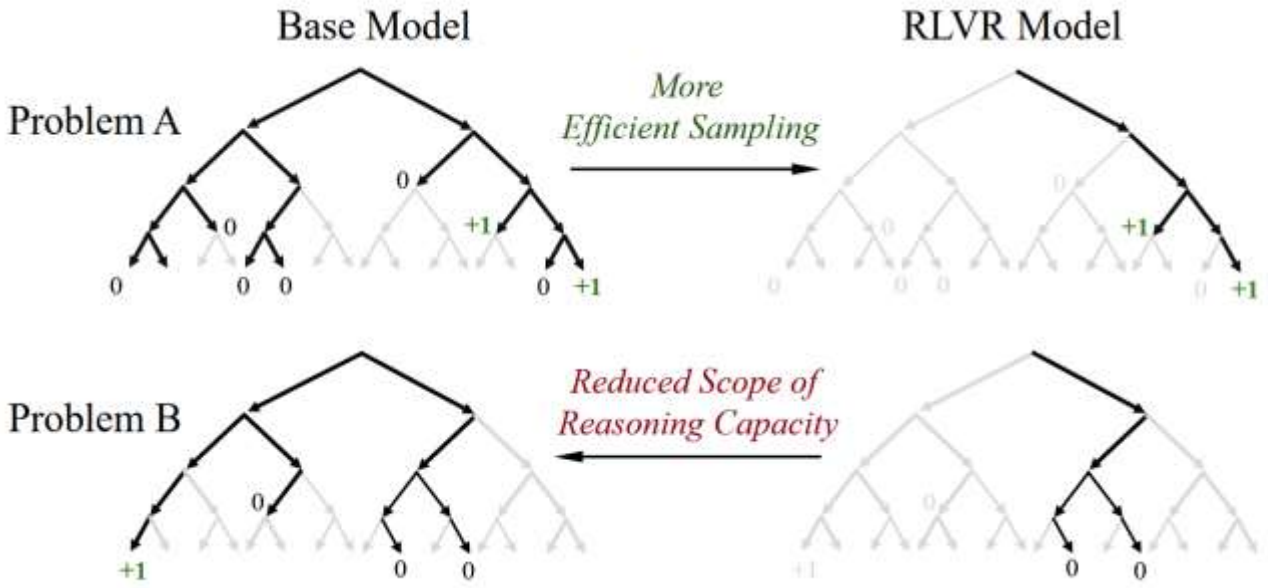
Website Code

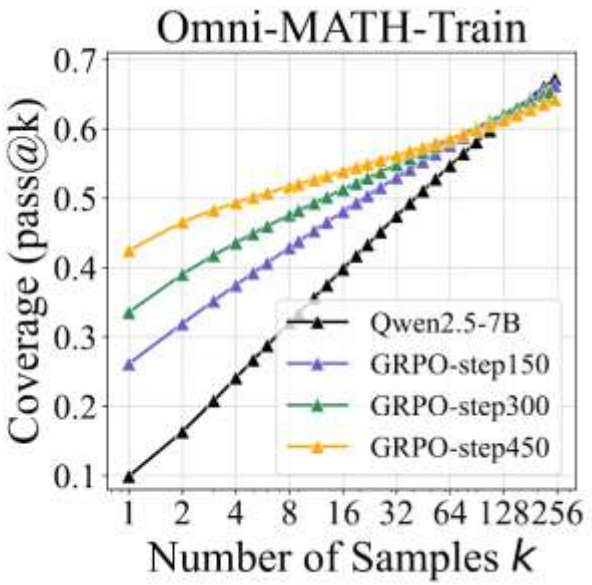
2.1 LLM 推理能力的提升

•RL-post training (如 RLHF, RLVR) 是增强 LLM 推理能力的主流方法 (如 GRPO)。在数学 (MATH)、编程 (HumanEval) 等领域取得显著性能提升。

•疑问：RL 增强的能力是**全新**的，还是仅仅对基模型**已有能力**的“分布锐化” (Distribution Sharpening)?

•RL后训练在单次（single-shot）任务上表现好，但可能牺牲多样本（multi-shot / pass@k）多样性。





2.1幂采样 (Power Sampling)

北京大学 PEKING UNIVERSITY

•**分布锐化:** 提升高似然度序列的相对权重，抑制低似然度序列。

•如果RL模型只是基模型的“锐化”版本，我们可以通过显式指定一个锐化后的目标采样分布来实现相同的效果。

低温度采样：

$$p_{\text{temp}}(x_t|x_0 \dots x_{t-1}) = \frac{p(x_t|x_{t-1} \dots x_0)^\alpha}{\sum_{x'_t \in \mathcal{X}} p(x'_t|x_{t-1} \dots x_0)^\alpha},$$

幂采样：

$$p_{\text{pow}}(x_t|x_0 \dots x_{t-1}) = \frac{\sum_{x_{>t}} p(x_0, \dots, x_t, \dots, x_T)^\alpha}{\sum_{x_{\geq t}} p(x_0, \dots, x_t, \dots, x_T)^\alpha}.$$

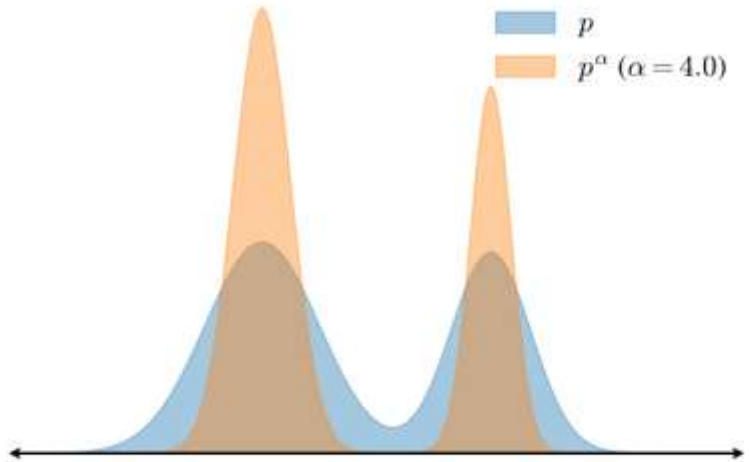


Figure 2: **A toy example of distribution sharpening.** Here p is a mixture of Gaussians, which we plot against p^α ($\alpha = 4.0$).

幂分布采样不等于低温采样

北京大学 PEKING UNIVERSITY

- **低温度采样 (Low-Temperature Sampling)**: 在每一步条件分布上指数化 。
  - 幂采样考虑了未来路径的整体锐化 (Sum of Exponents)。
  - 低温度采样是“贪婪”的，忽略了未来路径的指数锐化 (Exponent of Sums)。

Observation 1. *The power distribution upweights tokens with few but high likelihood future paths, while low-temperature sampling upweights tokens with several but low likelihood completions.*

Example 1. We can observe this phenomenon with a simple example. Let us consider the token vocabulary $\mathcal{X} = \{a, b\}$ and restrict our attention to two-token sequences (x_0, x_1): aa, ab, ba, bb. Let

$$p(aa) = 0.00, \qquad p(ab) = 0.40, \qquad p(ba) = 0.25, \qquad p(bb) = 0.25,$$

so that

$$p(x_0 = a) = 0.40, \qquad p(x_0 = b) = 0.50.$$

Let $\alpha = 2.0$. Under p^α, we have

$$p_{\text{pow}}(x_0 = a) \propto 0.00^2 + 0.40^2 = 0.160, \qquad p_{\text{pow}}(x_0 = b) \propto 0.25^2 + 0.25^2 = 0.125,$$

so p^α prefers sampling a over b. Under low-temperature sampling,

$$p_{\text{temp}}(x_0 = a) \propto (0.00 + 0.40)^2 = 0.160, \qquad p_{\text{temp}}(x_0 = b) \propto (0.25 + 0.25)^2 = 0.250,$$

幂采样算法：MCMC (Metropolis-Hastings)

由于p^{α}是无法归一化的，无法直接采样，因此采用 **MCMC (Metropolis-Hastings)** 算法近似采样。
分成B块逐段完成

Algorithm 1: Power Sampling for Autoregressive Models

Input : base p; proposal p_{prop}; power α; length T

Hyperparams: block size B; MCMC steps N_{MCMC}

Output : $(x_0, \ldots, x_T) \sim p^{\alpha}$

Notation: Define the unnormalized intermediate target

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

for $k \leftarrow 0$ **to** $\lceil \frac{T}{B} \rceil - 1$ **do**

  Given prefix $x_{0:kB}$, we wish to sample from π_{k+1}. Construct initialization $\mathbf{x}^0$ by extending autoregressively with p_{prop}:

$$x_t^{(0)} \sim p_{\text{prop}}(x_t \mid x_{<t}), \qquad \text{for } kB+1 \leq t \leq (k+1)B.$$

  Set the current state $\mathbf{x} \leftarrow \mathbf{x}^0$.

  for $n \leftarrow 1$ **to** N_{MCMC} **do**

    Sample an index $m \in \{1, \ldots, (k+1)B\}$ uniformly.

    Construct proposal sequence $\mathbf{x}'$ with prefix $x_{0:m-1}$ and resampled completion:

$$x'_t \sim p_{\text{prop}}(x_t \mid x_{<t}), \qquad \text{for } m \leq t \leq (k+1)B.$$

    Compute acceptance ratio (9)

$$A(\mathbf{x}', \mathbf{x}) \leftarrow \min\left\{1, \frac{\pi_k(\mathbf{x}')}{\pi_k(\mathbf{x})} \cdot \frac{p_{\text{prop}}(\mathbf{x} \mid \mathbf{x}')}{p_{\text{prop}}(\mathbf{x}' \mid \mathbf{x})}\right\}.$$

    Draw $u \sim \text{Uniform}(0, 1)$;

    if $u \leq A(\mathbf{x}', \mathbf{x})$ **then accept** and set $\mathbf{x} \leftarrow \mathbf{x}'$

  end

  Set $x_{0:(k+1)B} \leftarrow \mathbf{x}$ to fix the new prefix sequence for the next stage.

end

return $x_{0:T}$

单样本准确率 (Single-Shot Accuracy)

北京大学 PEKING UNIVERSITY

- Baselines: Base Model (基模型), GRPO (RL 基线), Low-temperature, Ours (Power Sampling)。
- 结论: 纯采样方法在不训练的情况下，成功激发了与 SOTA RLVR 相当的单样本推理能力。

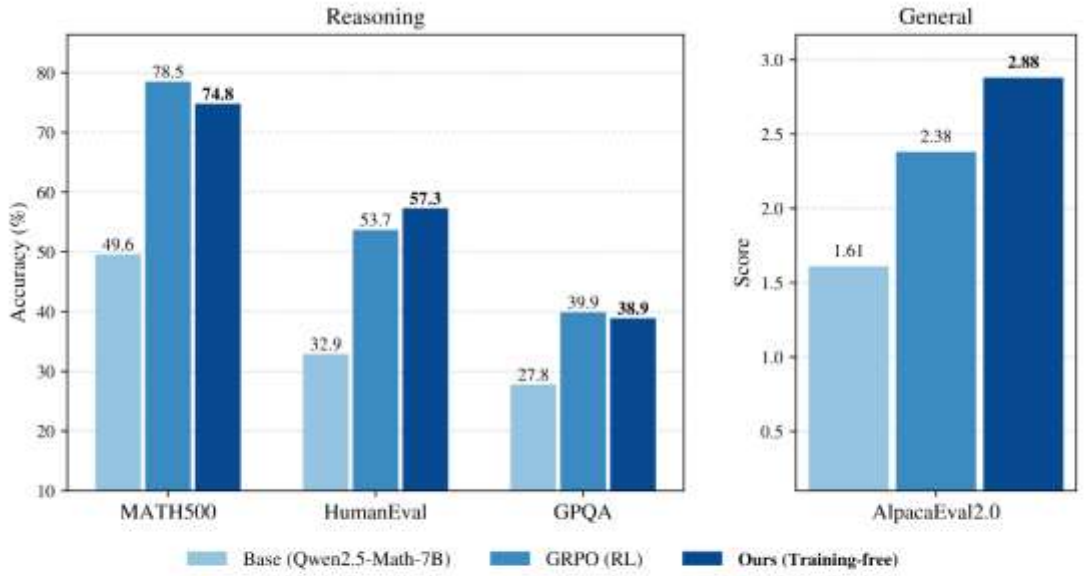


Figure 1: **Our sampling algorithm can match and outperform RL-posttraining.** Left: we compare our sampling algorithm (ours) against the base model (base) and RL-posttraining (GRPO) on three *verifiable reasoning* tasks (MATH500, HumanEval, GPQA). Right: we compare them on an *unverifiable general task* (AlpacaEval2.0). Our algorithm achieves comparable performance to GRPO within the posttraining domain (MATH500) but can *outperform* on out-of-domain tasks such as HumanEval and AlpacaEval.

	MATH500	HumanEval	GPQA	AlpacaEval2.0
Qwen2.5-Math-7B				
Base	0.496	0.329	0.278	1.61
Low-temperature	0.690	0.512	0.353	2.09
Power Sampling (ours)	**0.748**	**0.573**	**0.389**	**2.88**
GRPO (MATH)	**0.785**	**0.537**	**0.399**	**2.38**
Qwen2.5-7B				
Base	0.498	0.329	0.278	7.05
Low-temperature	0.628	0.524	0.303	5.29
Power Sampling (ours)	**0.706**	**0.622**	**0.318**	**8.59**
GRPO (MATH)	**0.740**	**0.561**	**0.354**	**7.62**
Phi-3.5-mini-instruct				
Base	0.400	0.213	0.273	14.82
Low-temperature	0.478	0.585	0.293	**18.15**
Power Sampling (ours)	**0.508**	**0.732**	**0.364**	**17.65**
GRPO (MATH)	**0.406**	**0.134**	**0.359**	**16.74**

Table 1: **Power sampling (ours) matches and even outperforms GRPO across model families and tasks.** We benchmark the performance of our sampling algorithm on MATH500, HumanEval, GPQA, and AlpacaEval 2.0. We bold the scores of both our method and GRPO, and underline whenever our method outperforms GRPO. Across models, we see that power sampling is comparable to GRPO on in-domain reasoning (MATH500), and can outperform GRPO on out-of-domain tasks.

多样性

- GRPO 样本高度集中于基模型的最高似然度/置信度峰值。
- 幂采样成功地从比基模型高、但比 GRPO 广的区域采样，符合其目标分布。

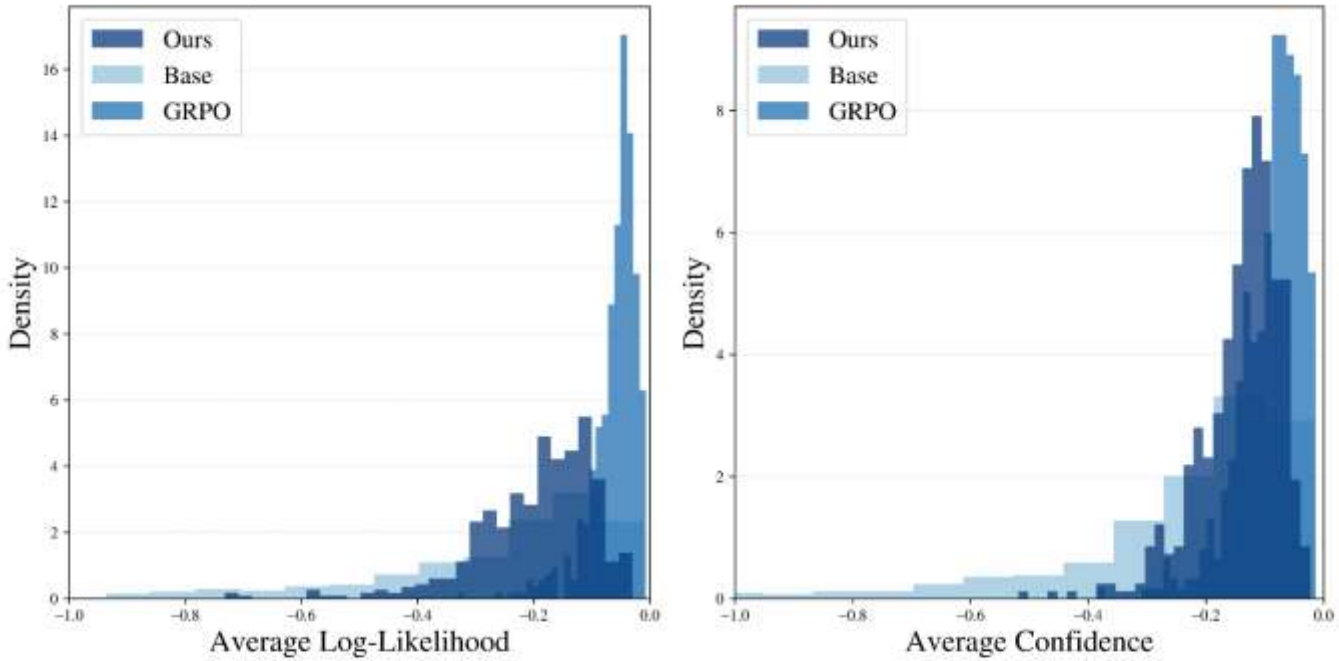


Figure 4: **Base model (Qwen2.5-Math-7B) likelihoods and confidences for MATH500 responses.** Left: We plot the log-likelihoods (relative to the base model) of original, power sampling, and GRPO responses over MATH500. Right: We do the same but for confidences relative to the base model. We observe that GRPO samples from the highest likelihood and confidence regions with power sampling close behind, which correlates with higher empirical accuracy.

多样性

- GRPO 表现出多样性崩溃，Pass@k 曲线很快趋平。
  - 幂采样的 Pass@k 曲线**明显优于** GRPO。
  - 幂采样在实现高单样本性能的同时，**避免了多样性崩溃。**
- 幂采样在高似然度区域仍保留了可观的分布广度，而 GRPO 则过度集中于最高峰值。

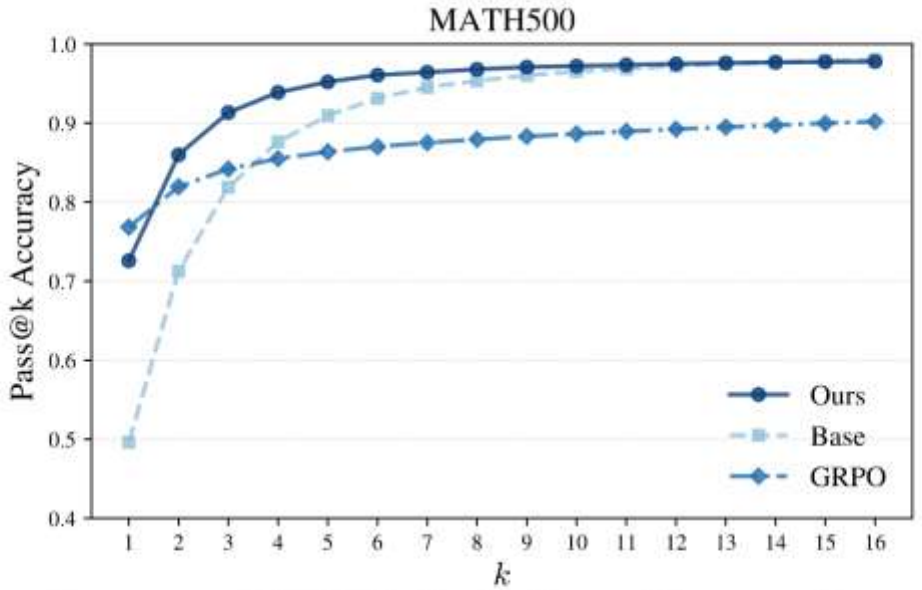


Figure 5: **Pass@k performance on MATH500**. We plot the pass@k accuracy (correct if at least one of k samples is accurate) of power sampling (ours) and RL (GRPO) relative to the base model (Qwen2.5-Math-7B). Our performance curve is strictly better than both GRPO and the base model, and our pass rate at high k matches the base model, demonstrating sustained generation diversity.

超参数的影响

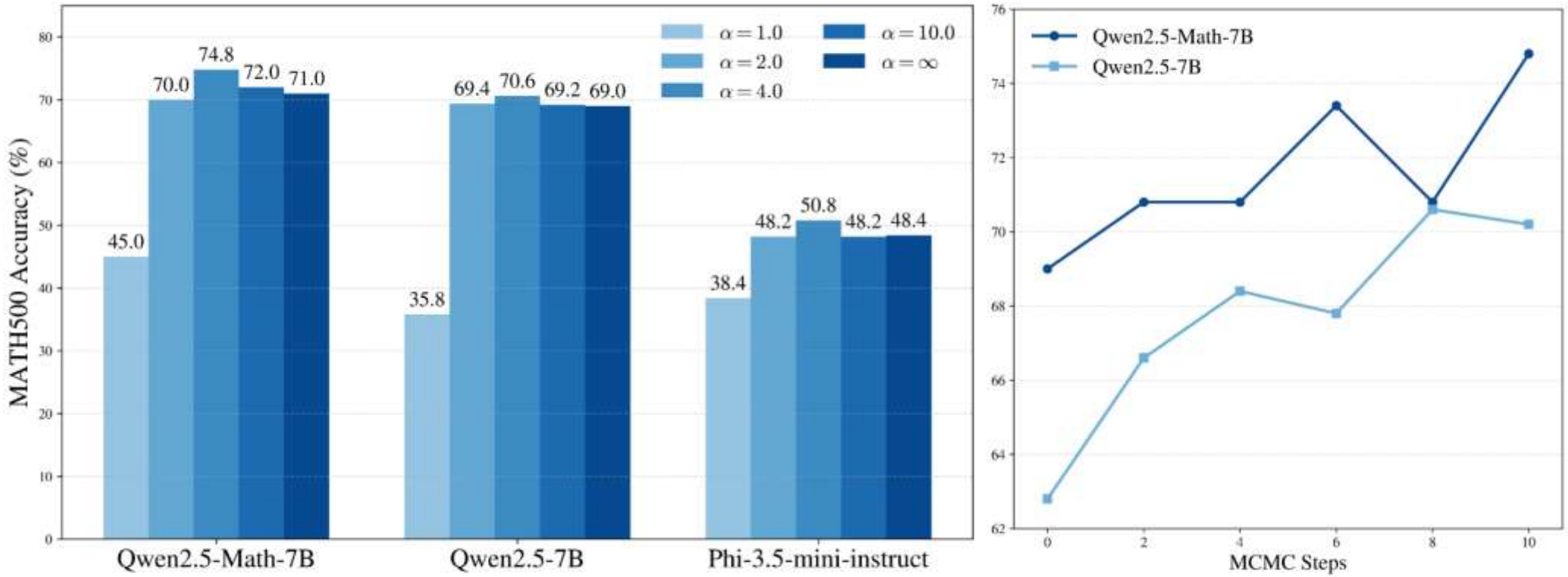


Figure 6: **Effect of hyperparameters on power sampling.** Left: We plot MATH500 accuracy across model families for various values of α. Right: We plot the increase in accuracy of power sampling on Qwen models as the number of MCMC steps increases.

总结

北京大学 PEKING UNIVERSITY

- 基模型在采样阶段的能力被低估，纯采样可以媲美 RLVR。

- 训练无关、数据集无关、无需额外的Verifier。

- 局限性: 1. 以增加推理时计算量为代价来提升性能。
  2. 测试模型不足，且规模较小