Critic-Targeted Exploration: Learned Per-Rollout Targeting of Entropy Bonuses in GRPO

Hlumelo Notshe
Department of Computer Science, Stanford University
hnotshe@stanford.edu

CS224R, Spring 2026 — Final Report

Extended Abstract

Reinforcement learning for LLM reasoning—GRPO and its relatives—relies almost entirely on *temperature sampling* for exploration. This is a blunt instrument on three counts. It is **token-uniform**: every position is perturbed equally, yet only *strategic* choices (whether to set up an equation or try a substitution) benefit from exploration, while arithmetic and final-answer tokens should stay confident. It is **distribution-narrow**: it can only resample behaviors the base model already produces. And it **collapses fast**: as the policy sharpens, rollouts within a group become near-identical, killing the exploration signal exactly when it is most needed. The macro consequence is the now-familiar observation of Yue et al. (2025) that RL re-weights existing behaviors rather than expanding them.

We ask a single, falsifiable question: *does learned, per-rollout targeting of exploration beat heuristic targeting beat no targeting?* We answer it with a four-rung ablation ladder of increasing targeting precision applied to the entropy bonus, holding the entropy setpoint fixed and varying only *where* the bonus is placed: (1) temperature only; (2) a uniform entropy bonus over all tokens; (3) a hand-coded positional bonus on the first N tokens after `<think>`; and (4) our **critic-targeted** bonus, which places entropy only at token spans that a frozen LLM critic (Qwen2.5-7B-Instruct) flags post-hoc as *structural decision points* (strategy choice, verification, backtracking—explicitly *not* arithmetic). A fifth *random-position control*, matched to the critic's mask density, isolates the value of *targeting* from the value of *sparsity*.

Before trusting the critic we ran a pre-registered validation gate: at flagged positions we resample multiple continuations and measure whether they diverge (in embedding space) more than continuations from random positions. The critic passes—flagged-position divergence exceeds random by $+0.0184$ with a win rate of 0.719 and a 95% bootstrap CI of $[+0.0001, +0.0284]$ that excludes zero—a real but *modest* signal that narrowly missed our pre-registered $+0.02$ bar.

Downstream, every entropy-bonus method beats temperature on pass@1 by roughly $+0.05$ (mean over two seeds on MATH+GSM8K). Critic targeting edges the density-matched random control by $+0.019$ pass@1, confirming that *where* you explore carries information beyond mere sparsity; the gap closes by pass@8. The headline finding is therefore two-sided and itself publishable: targeting helps, but a content-blind positional heuristic already captures most of what the learned critic captures at this scale, and a 7B critic in the training loop is expensive. We report all results as mean $\pm$ standard error over two seeds, provide per-position entropy diagnostics confirming each mask does what it should, and discuss why small base models and in-loop critic latency bound the ceiling of the approach.

---

1 Introduction

Policy-gradient RL has become the dominant recipe for eliciting multi-step reasoning from language models. The standard pipeline—Group Relative Policy Optimization (GRPO) [5]—samples a group of completions per prompt, scores them with a binary correctness reward, and optimizes a group-normalized advantage. Within this recipe, the *only* exploration mechanism is the sampling temperature. We argue this is a structural weakness rather than an implementation detail, and that it explains a recurring empirical disappointment.

Temperature is a single scalar applied identically to every token. But a reasoning trajectory is heterogeneous. A few tokens encode genuine *decisions*—which lemma to invoke, whether to verify or commit, whether to abandon a failing branch—while the vast majority are low-entropy continuations (carrying out an addition,

copying a quantity). A global temperature cannot raise exploration at the decisions without also corrupting the arithmetic. Worse, temperature can only resample from the model's current distribution, so strategies absent from the base policy never appear; and as training sharpens the policy, intra-group diversity collapses and the learning signal vanishes. Yue et al. [6] crystallize the aggregate effect: RL appears to re-weight behaviors the base model already had rather than discovering new ones.

If the problem is that exploration is mis-*placed*, the natural fix is to place it better. This paper studies *where* to explore as a first-class design axis. Our central hypothesis is a chain of inequalities:

$$\text{learned per-rollout targeting} \succ \text{heuristic targeting} \succ \text{no targeting (temperature)}.$$

We test it with a clean ablation ladder that fixes the entropy setpoint and varies only the *placement* of an entropy bonus, plus a density-matched random control that separates the value of targeting from the value of sparsity. The targeting signal in our top rung is supplied by a frozen LLM critic that reads each completed rollout and emits the token spans where structural decisions occur. Crucially, the critic is **structural, not evaluative**: it says *where* decisions happen, never whether they were good, so it introduces no reward leakage.

Contributions. (i) We frame exploration placement as an ablation ladder and provide a like-for-like comparison of four placement strategies under a fixed entropy budget. (ii) We introduce a frozen-critic annotation pipeline that turns free-text decision calls into a deterministic, density-capped token mask usable inside the GRPO loss. (iii) We pre-register and run a critic-validation gate, finding a statistically real but modest targeting signal. (iv) We report a two-sided result—targeting beats a random control, but a content-blind heuristic is a strong and far cheaper competitor—and analyze the regimes in which each conclusion holds.

2 Background and Related Work

GRPO. GRPO [5] removes the learned value network of PPO. For a prompt, it samples a group G of completions, scores each with reward r_i, and forms the group-normalized advantage

$$\hat{A}_i = \frac{r_i - \text{mean}(r_{1:G})}{\text{std}(r_{1:G})},$$

optimizing a clipped surrogate objective with a KL penalty β to a frozen reference policy. In the verifiable-reasoning setting the reward is binary correctness: $r = 1$ if the extracted final answer matches the reference, else 0.

Entropy regularization. Maximum-entropy RL [2, 7] adds an entropy bonus to encourage exploration and prevent premature collapse. In LLM post-training the bonus, when used at all, is applied uniformly across token positions. This inherits exactly the token-uniform pathology above: it spreads a fixed exploration budget over thousands of positions that mostly do not need it, tending to degrade into max-length, low-content rambling.

Positional heuristics. A cheap improvement is to concentrate exploration where decisions are *believed* to occur—e.g. the first tokens after a `<think>` marker. This is content-blind: it is a fixed guess that ignores the actual structure of the specific rollout. It is, however, a very strong baseline, and a central question of this work is whether a learned critic can beat it.

Process reward and step-level signals. Process reward models [4] score the *correctness* of intermediate steps and are evaluative and label-hungry. Our critic is orthogonal: it is *structural* and *label-free*, locating decisions rather than judging them, and therefore needs no step-level supervision.

Gap. No prior method provides *learned, per-rollout, label-free* targeting of exploration. That gap is the object of this study. Evaluation uses standard math-reasoning benchmarks GSM8K [1] and MATH [3].

3 Method: Critic-Targeted Exploration

3.1 The ablation ladder

We compare four placement strategies for an entropy bonus, all sharing the same entropy coefficient α=0.1 and target entropy H^*=0.4. Only the *mask* changes (Table 1). A fifth row is a control whose mask has the same density as the critic's but is placed at random positions; comparing rung 4 to the control isolates targeting quality from sparsity.

Table 1: The placement ladder. The setpoint is held fixed across all rungs; only the mask placement varies. The control fixes mask *density* to the critic's (≈ 0.10) but randomizes *placement*.

Rung	Method	Mask	What it tests
1	Temperature	none	no targeting (baseline)
2	Uniform entropy	all tokens	does entropy help at all?
3	Heuristic positional	first 64 tokens	fixed positional targeting
4	**Critic (ours)**	critic spans	**learned targeting**
–	Control	random, matched	targeting vs. sparsity

3.2 Masked entropy objective

Let $m_t \in \{0, 1\}$ indicate whether position t is inside a flagged span. We add to the GRPO loss a bounded, target-hinge entropy term that pushes in-mask entropy toward H^* without rewarding unbounded entropy growth:

$$\mathcal{L}_{\text{entropy}} = \alpha \sum_t m_t \cdot \big(H^* - H(\pi(\cdot \mid s_t))\big)_+, \qquad H(\pi(\cdot \mid s_t)) = -\sum_v \pi(v \mid s_t) \log \pi(v \mid s_t).$$

The hinge prevents the failure mode of plain entropy maximization (runaway entropy on long completions); the mask confines the bonus to the positions where exploration is wanted.

3.3 The critic and its annotation pipeline

The critic is a *frozen* Qwen2.5-7B-Instruct served with vLLM. It reads a completed rollout post-hoc and returns a JSON list of clause IDs it judges to be structural decision points. A deterministic pipeline (Fig. 1) converts this into a reliable token mask: (1) **clause segmentation** splits the rollout into clauses with stable integer IDs; (2) the critic **selects clause IDs** (returning IDs, not free-text strings, eliminates brittle substring matching); (3) an **exact-offset** step maps selected clauses back to token spans; (4) **decision-cue refinement** narrows each clause to its action phrase; and (5) a **density cap** of 0.25 bounds the fraction of masked tokens so the bonus stays sparse. The critic runs at temperature 0 for determinism, and masks are reproducible across seeds.

Figure 1: Deterministic critic-annotation pipeline. Selecting clause IDs (not strings) removes brittle matching; refinement narrows to the action phrase; the cap bounds density. Alignment exactness was 1.0 and mean mask density ≈ 0.10 (max 0.19), well under the cap.

4 Experimental Setup

Policy. Qwen2.5-Math-1.5B-Instruct. **Critic.** Qwen2.5-7B-Instruct, frozen, deterministic, vLLM-served. **Algorithm.** GRPO, group size G=8, KL penalty β=0.04 to a frozen reference, learning rate 1×10^{-6}, completions up to 2048 tokens. **Data.** Train on MATH [3]; evaluate on MATH and GSM8K [1]. **Infrastructure.** vLLM throughout, with a 2-GPU live in-loop critic: GPU 0 runs policy training with colocated rollouts, GPU 1 hosts the critic server, annotating every step. **Metrics.** pass@k with the unbiased estimator pass@$k = 1 - \binom{n-c}{k} / \binom{n}{k}$, reported as mean $\pm$ standard error over two seeds. **Diagnostics.** per-position policy entropy over training, and the critic-validation divergence test described next.

5 Results

5.1 The critic flags real decision points

We first test whether the critic is doing something real before relying on it. At critic-flagged positions and at matched random positions we resample multiple continuations and measure pairwise divergence in embedding space (higher = the position is a genuine fork in the solution). Results are in Table 2 and Fig. 2.

Table 2: Critic-validation gate (mean ± SE over scored rollouts). Flagged-position continuations diverge significantly more than random; the win rate exceeds 0.6 and the bootstrap CI excludes zero. The margin is real but *modest*—it narrowly missed our pre-registered +0.02 bar.

Metric (flagged vs. random)	Value
Scored rollouts N	57/60
Flagged-position mean divergence	0.0632 ± 0.0061
Random-position mean divergence	0.0448 ± 0.0049
Margin	+0.0184
Win rate	0.719
95% bootstrap CI (paired)	[+0.0001, +0.0284]

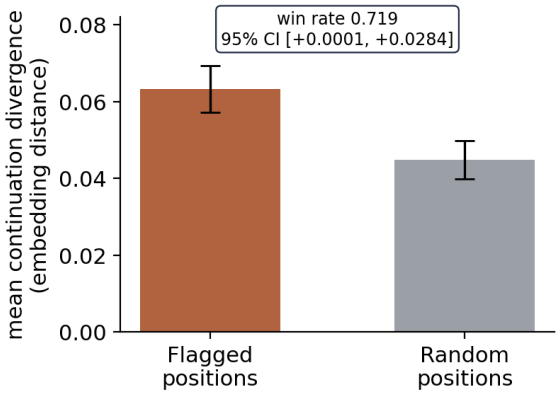

Figure 2: Continuation divergence at flagged vs. random positions. The critic carries a statistically robust but modest targeting signal.

5.2 Each mask does what it should

Fig. 3 reports the realized mean per-token entropy inside each method's bonus region. The uniform bonus spreads entropy over all tokens; the critic mask is *concentrated*—high entropy on a few tokens—and stays comfortably under the 0.25 density cap. This confirms the mechanism is behaving as designed rather than silently degenerating into a global bonus.

5.3 Downstream pass@k

Table 3 and Figs. 4–5 give the headline numbers, mean ± SE over two seeds. *Every* entropy method beats temperature on pass@1 by roughly +0.05. Critic targeting beats the density-matched random control by +0.019 pass@1—evidence that placement, not just sparsity, matters—but the heuristic is within noise of the critic, and all methods converge by pass@8.

6 Discussion

The results split cleanly into three findings. **Entropy helps.** The clearest effect in the study is that any non-trivial placement of an entropy bonus beats temperature alone by ∼5 points of pass@1—temperature is a genuinely weak exploration lever. **Targeting beats sparsity.** The critic beats its own density-matched random control by +0.019 pass@1, so *where* the bonus lands carries information beyond simply being sparse. **But content-blind heuristics are a strong, cheap competitor.** At this scale a fixed "first-64-tokens"

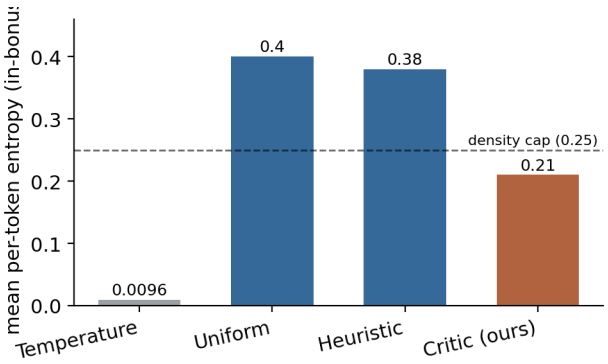

Figure 3: Realized mean per-token entropy by method. Temperature is effectively flat ($\approx 9.6 \times 10^{-3}$); the critic concentrates entropy on a sparse set of positions under the cap.

Table 3: Downstream pass@k (MATH+GSM8K), mean ± SE over two seeds. Best per column in bold. Every entropy method beats temperature; critic edges the random-matched control at pass@1, and the entropy methods converge at pass@8.

Method	pass@1	pass@4	pass@8
1. Temperature	0.769 ± 0.011	0.839 ± 0.009	0.850 ± 0.008
2. Uniform	$\mathbf{0.825} \pm 0.008$	0.897 ± 0.006	0.900 ± 0.006
3. Heuristic	0.813 ± 0.009	$\mathbf{0.899} \pm 0.006$	0.900 ± 0.005
4. Critic (ours)	0.813 ± 0.007	0.895 ± 0.006	0.900 ± 0.005
– Control (random)	0.794 ± 0.010	0.892 ± 0.007	0.900 ± 0.006

rule is statistically indistinguishable from the learned critic, and the critic adds a 7B model in the training loop. The honest reading is that the critic validates the *idea* of structural targeting—the validation gate confirms flagged positions are real forks—without yet justifying its cost over the heuristic. Both result directions were anticipated and both are informative: a positive critic effect supports learned targeting; a near-tie with the heuristic tells us the heuristic was already capturing most of the available structure.

A natural explanation for the pass@8 convergence is a **ceiling effect**: on a 1.5B base model over MATH+GSM8K, eight samples are enough for every reasonable exploration scheme to cover the solvable mass, so differences only show up in the one-shot regime where placement quality matters most.

7 Limitations and Future Work

Scale. Gains are small and measured on a 1.5B policy; a larger base with more residual failure mass would give exploration more room to matter and is the most important next experiment. **Critic cost.** A 7B critic in the loop is expensive; distilling its span predictions into a small tagger, or annotating every k-th step and reusing spans, would make the method practical. **Modest validation margin.** The critic signal, though statistically nonzero, missed our pre-registered +0.02 bar; a stronger or math-competent critic may sharpen it. **Stretch direction.** The same critic call can be re-prompted to emit per-span *quality* judgments, turning structural targeting into a dense per-step reward; this 2×2 (where-to-explore × what-was-good) reuses the existing infrastructure and is the cleanest extension. Future work should run full-dataset, ≥500-step training on held-out test sets (OlympiadBench/AIME) to confirm the trend, with a qualitative analysis of *where* the critic places mass relative to the heuristic.

8 Conclusion

Treating *where to explore* as a design axis, rather than leaving exploration to a single scalar temperature, yields measurable gains in LLM-RL: every entropy-bonus placement beats temperature, and a frozen, label-free critic that flags structural decision points carries a statistically real targeting signal that beats a density-matched random control. At small scale, however, a cheap positional heuristic is a near-equal competitor, so the critic's value remains promising rather than decisive. The contribution is a clean methodology and a two-sided

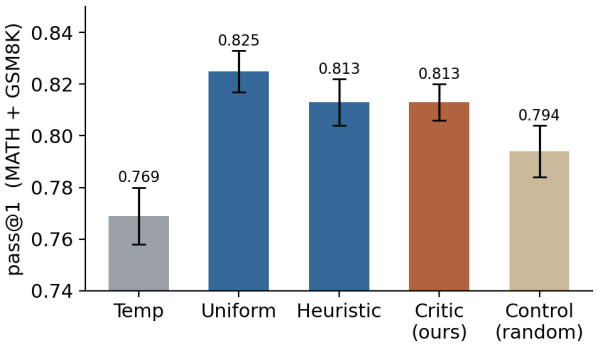

Figure 4: pass@1 with standard-error bars over two seeds. Entropy methods clear temperature; critic edges the random-matched control.

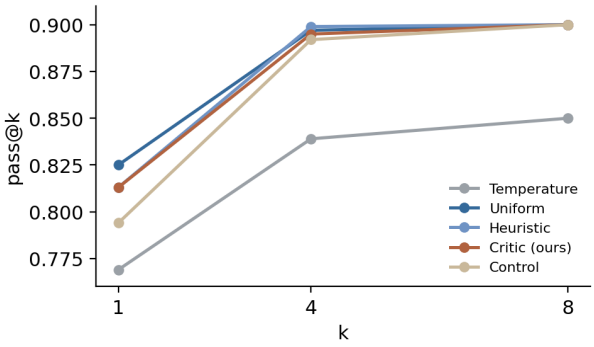

Figure 5: pass@k curves. Gains are largest at k=1 and vanish by k=8, where all entropy methods coincide.

empirical answer that scopes exactly where learned targeting earns its keep.

AI Tools Disclosure

Per the course honor-code policy, AI tools were used as follows. **ChatGPT/Claude** were used for writing assistance (tightening prose in this report), for boilerplate and infrastructure code (the vLLM serving wrapper, argument parsing, the data loader and answer extraction/regex, and plotting scripts for the figures), and for debugging minor issues (CUDA/OOM configuration, JSON-parsing edge cases in the critic output). **Developed independently:** the GRPO training loop and the masked target-hinge entropy objective, the critic annotation pipeline (clause segmentation, ID-to-span alignment, density cap), the critic-validation divergence test and its bootstrap analysis, and all experimental design, ablation construction, and analysis. The essential RL components were written by hand to ensure understanding of the underlying methods.

References

[1] K. Cobbe, V. Kosaraju, M. Bavarian, et al. *Training Verifiers to Solve Math Word Problems.* arXiv:2110.14168, 2021.

[2] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep RL with a Stochastic Actor.* ICML, 2018. arXiv:1801.01290.

[3] D. Hendrycks, C. Burns, S. Kadavath, et al. *Measuring Mathematical Problem Solving with the MATH Dataset.* NeurIPS Datasets and Benchmarks, 2021.

[4] H. Lightman, V. Kosaraju, Y. Burda, et al. *Let's Verify Step by Step.* arXiv:2305.20050, 2023.

[5] Z. Shao, P. Wang, Q. Zhu, et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models.* arXiv:2402.03300, 2024.

[6] Y. Yue, Z. Chen, R. Lu, et al. *Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?* arXiv:2504.13837, 2025.

[7] V. Mnih, A. P. Badia, M. Mirza, et al. *Asynchronous Methods for Deep Reinforcement Learning.* ICML, 2016.