Start Classifying: Categorical Critics for LLM Reinforcement Learning

Zhijian Zhou[*,1,2,3] **Long Li**[*,4] **Xuan Zhang**[1,2,3] **Zongkai Liu**[2] **Yulei Qin**[3]
Ke Li[3] **Xing Sun**[3] **Xiaoyu Tan**[3,†,‡] **Chao Qu**[1,5,†] **Yuan Qi**[1,5,†]
[1]Fudan University [2]Shanghai Innovation Institute [3]Tencent YoutuLab
[4]Griffith University [5]Shanghai Academy of Artificial Intelligence for Science
arthurtan@tencent.com quchao@fudan.edu.cn qiyuan@fudan.edu.cn
[*]Equal contribution [†]Corresponding authors [‡]Project leader

Abstract

Proximal Policy Optimization (PPO) for large language models typically trains its critic by mean-squared-error (MSE) regression on scalar value targets. Although scalar MSE is statistically valid for estimating the conditional expected return, sparse binary rewards in reinforcement learning with verifiable rewards (RLVR) make critic optimization and calibration especially consequential: small value errors directly distort the scalar advantages used by PPO. We study whether a classification-based training objective can improve this critic signal. **HL-Gauss PPO** replaces the scalar MSE head with a categorical predictor over a discretized value support, trained by cross-entropy against smoothed HL-Gauss targets. Its output is decoded to a scalar expectation for standard GAE and PPO; the actor update is therefore unchanged and is not distributional. Across mathematical reasoning, tool-augmented math, and Search-R1, and on both Qwen2.5 and Qwen3 backbones, HL-Gauss PPO consistently improves over strong PPO and DAPO baselines. Controls with one-hot, two-hot, and Bernoulli two-bin critics show that neither a larger output head nor binary classification alone explains the gains. On a common collection of reasoning prefixes, HL-Gauss improves Brier score and calibration error and yields more symmetric, lower-variance advantages. These results position categorical value learning as an effective optimization surrogate for PPO critics in RLVR.

1 Introduction

Reinforcement learning has emerged as a powerful paradigm for improving the reasoning capabilities of large language models (LLMs). Recent work on reinforcement learning with verifiable rewards (RLVR) — where a deterministic verifier provides binary correctness signals — has demonstrated striking gains on mathematical reasoning (OpenAI, 2024; DeepSeek-AI, 2025; Yu et al., 2025; Shao et al., 2024). Within this paradigm, two broad families of algorithms have gained traction: critic-free methods such as GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025), which estimate advantages from within-group reward comparisons; and actor–critic methods based on PPO (Schulman et al., 2017), which learn an explicit value function to compute advantages via GAE (Schulman et al., 2015). While critic-free methods sidestep the difficulty of value learning, recent studies such as VAPO (Yue et al., 2025) have shown that a well-trained critic can provide stable and sample-efficient training for long-horizon reasoning tasks.

Despite the renewed interest in PPO for LLMs, comparatively little attention has been paid to the critic's *learning objective*. Standard PPO fits scalar value targets with MSE. In the common RLVR setting with binary terminal rewards and $\gamma{=}\lambda{=}1$, the return from an intermediate state is Bernoulli and the population-optimal scalar prediction is its conditional mean, equivalently the probability of eventual success. Scalar MSE is therefore statistically consistent for the value required by PPO; binary or bimodal returns do not make scalar

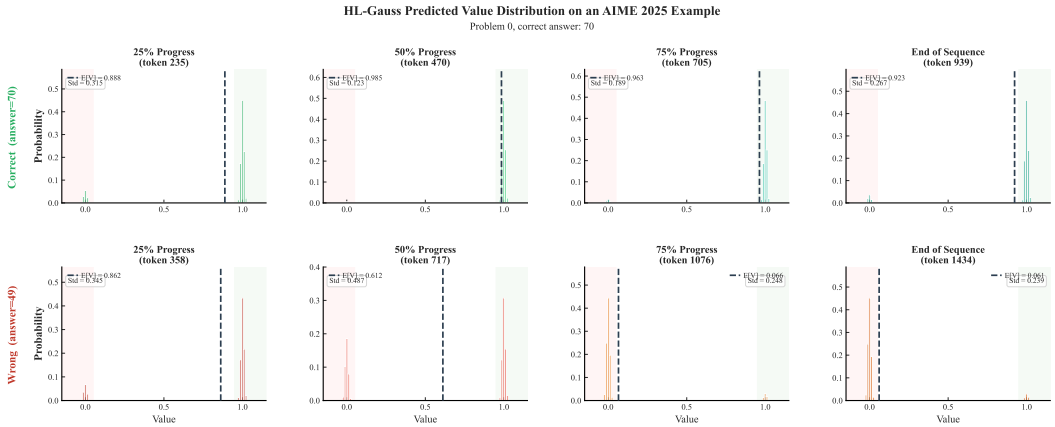

Figure 1: Categorical critic outputs across problem-solving progress on a paired AIME 2025 example. Early prefixes have diffuse mass near both endpoints; the output subsequently concentrates near 1 for the correct rollout and near 0 for the incorrect rollout. PPO uses only the decoded expectation of each distribution, so this visualization describes the critic parameterization rather than a distributional actor update.

value learning invalid. The practical issue is optimization and calibration. Because the critic enters the actor update through $\hat{A}_t = G_t - V_\phi(s_t)$, modest value miscalibration can create substantially asymmetric positive and negative advantages, as we show in Section 5.4.

Recasting regression as classification is well established: histogram losses, categorical value heads, and broader empirical studies show optimization benefits in conventional RL (Imani & White, 2018; Schrittwieser et al., 2020; Farebrother et al., 2024). DisPPO instead models the full return distribution in LLM PPO via quantile regression (Zhou et al., 2026). We use classification only as a critic-training surrogate and decode a scalar expectation before an unchanged GAE/PPO update; our method is therefore not a distributional policy-improvement algorithm. This isolates whether a smoothed categorical critic objective can improve calibration and advantage estimation without changing policy optimization.

We propose **HL-Gauss PPO**, which replaces the scalar MSE critic head with a categorical value head trained via HL-Gauss cross-entropy (Imani & White, 2018), while leaving the actor-side PPO update entirely unchanged. This is a minimal, drop-in modification: only the critic's output layer ($\mathbb{R}^d \to \mathbb{R}^1$ becomes $\mathbb{R}^d \to \mathbb{R}^m$) and critic loss change. Empirically, HL-Gauss PPO consistently outperforms strong PPO and DAPO baselines across mathematical reasoning, tool-augmented math, and Search-R1. The trend also holds on Qwen3-4B-Base, extending the main Qwen2.5-7B evaluation.

We study two possible optimization differences—the per-sample gradient channel induced by the output head and the local geometry induced by target smoothing—without treating either as a complete causal explanation. More directly, calibration measurements and advantage dynamics show that the MSE critic becomes overconfident on likely-failure prefixes and produces strongly asymmetric raw advantages, whereas HL-Gauss yields better calibrated scalar predictions and more balanced signals. A Bernoulli two-bin baseline and matched categorical-head ablations further separate binary classification, head capacity, and target smoothing.

Our contributions are as follows:

- **Controlled evaluation in LLM RLVR.** We apply the established HL-Gauss value-learning objective to PPO critics for LLM reasoning and show consistent gains across mathematical reasoning, tool use, search, and Qwen2.5/Qwen3 backbones while leaving the actor update unchanged.

- **Controls and diagnostics.** One-hot, two-hot, Bernoulli two-bin, support, and smoothing controls isolate the role of target construction. Direct calibration metrics and advantage analyses connect the critic objective to the scalar learning signal received by PPO.

2 Preliminaries

This section reviews the two foundations underlying our method: PPO-style reinforcement learning for large language models, together with several practical refinements used in recent reasoning-RL systems, and the regression-as-classification perspective for value learning.

2.1 PPO-style RL for LLMs and practical refinements

We model autoregressive generation as a sequential decision process. Let x denote the input prompt and $y_{1:T}$ the generated token sequence. At decoding step t, the state is the prefix $s_t = (x, y_{1:t-1})$ and the action is the next token $a_t = y_t$. The policy $\pi_\theta(a_t \mid s_t)$ is parameterized by the LLM with parameters θ. The learning objective is to maximize expected return $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right]$, where rewards may be provided at token or trajectory level.

Standard PPO uses a clipped surrogate objective optimized via policy gradients, with advantages estimated using generalized advantage estimation (GAE) (Schulman et al., 2015; 2017): $\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$ where $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$. The critic $V_\phi(s)$ predicts a scalar value estimate and is typically trained by MSE regression to a bootstrapped return target.

In modern reasoning-RL systems for LLMs, this basic PPO recipe is often augmented with practical refinements (e.g., asymmetric clipping, token-level policy gradients, group sampling, value pretraining) (Yue et al., 2025) motivated by long-horizon reasoning and sparse rewards. Our focus is on changing how the critic represents and learns value targets, not on the actor objective or training recipe.

2.2 Regression as classification

Farebrother et al. (2024) propose to train value functions as a classification problem. The key idea is to project a scalar target y onto a fixed discrete support $\{z_i\}_{i=1}^{m}$, predict a categorical distribution $\mathbf{p}_\phi(s)$ via an m-way softmax head, and optimize a cross-entropy loss instead of an MSE regression loss. Given a value range $[v_{\min}, v_{\max}]$, the interval is discretized into m bins with centers $\{z_i\}$. The value prediction is the expected value of the categorical distribution:

$$\hat{V}_\phi(s) = \sum_{i=1}^{m} p_\phi(z_i \mid s)\, z_i.$$

The target distribution $\mathbf{q}(y)$ can be constructed via sharp (one-hot) or smooth (Gaussian) encodings. **HL-Gauss** (Imani & White, 2018) applies a Gaussian kernel:

$$q_i(y) \propto \exp\left(-\frac{(z_i - y)^2}{2\sigma^2}\right), \quad \sum_i q_i(y) = 1.$$

The critic is trained with cross-entropy loss: $L_V(\phi) = -\sum_i q_i(y) \log p_\phi(z_i \mid s)$.

3 Method: HL-Gauss PPO

We propose HL-Gauss PPO, which replaces the scalar regression critic in PPO with a categorical value head trained via the HL-Gauss cross-entropy objective. The actor-side PPO update, advantage estimation, and rollout pipeline remain unchanged; only the critic architecture and its learning target differ. The method is therefore a classification-based surrogate

for learning the scalar value consumed by PPO, not a distributional policy-improvement algorithm. Architectural and hyperparameter details are deferred to Appendix D.

3.1 Categorical value head

In PPO, the critic appends a linear projection $\mathbf{w}^\top \mathbf{h} + b$ to the backbone's last hidden state $\mathbf{h} \in \mathbb{R}^d$, producing a scalar value estimate $V_\phi(s) \in \mathbb{R}$. We replace this with an m-dimensional linear layer followed by a softmax, yielding a categorical distribution over a fixed discrete support.

Concretely, we partition the value range $[v_{\min}, v_{\max}]$ into m equally spaced bins with width $\Delta = (v_{\max} - v_{\min})/m$ and centers $z_i = v_{\min} + (i - \frac{1}{2})\,\Delta$ for $i = 1, \ldots, m$. The critic outputs logits $\boldsymbol{\ell}(s) = W\mathbf{h} + \mathbf{b} \in \mathbb{R}^m$, and the predicted distribution is $p_i(s) = \mathrm{softmax}(\boldsymbol{\ell}(s))_i$. The scalar value estimate is recovered as the expectation: $\hat{V}_\phi(s) = \sum_{i=1}^{m} p_i(s)\, z_i$. This scalar value is used downstream for GAE advantage computation (Section 2.1); no higher-order distributional statistic is passed to the actor.

3.2 Target distribution construction

At each training step, the critic receives a scalar bootstrapped return $y_t = \hat{V}_\phi(s_t) + \hat{A}_t$ computed via GAE. We project y_t onto a target distribution $\mathbf{q}(y_t)$ over the m bins. For the HL-Gauss projection, each bin's target probability is computed as the Gaussian CDF mass falling within that bin:

$$q_i(y_t) \propto \Phi\left(\frac{z_i + \Delta/2 - y_t}{\sigma}\right) - \Phi\left(\frac{z_i - \Delta/2 - y_t}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal CDF and σ controls the smoothness of the target (Imani & White, 2018; Farebrother et al., 2024). The resulting distribution is normalized to sum to one.

3.3 Training objective

The critic is trained by minimizing the cross-entropy between the target distribution and the predicted distribution:

$$L_{\text{critic}}(\phi) = -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \sum_{i=1}^{m} q_i(y_t)\, \log p_i(s_t),$$

where $\mathcal{M}$ is the set of valid (non-padded) response tokens. The actor objective remains the standard PPO clipped surrogate $L_{\text{clip}}(\theta)$. Actor and critic are updated alternately within each PPO epoch, following standard practice.

4 Analysis: Optimization and advantage calibration

In binary-reward RLVR, scalar MSE is population-consistent for the conditional expected return. Our analysis therefore does not argue that scalar critics are representationally invalid. Instead, we describe two local optimization differences between scalar and categorical heads and then connect practical critic calibration to the advantages used by PPO.

4.1 Per-sample gradient channels

Standard PPO trains a scalar critic $V_\phi(s) = \mathbf{w}^\top \mathbf{h} + b$, where $\mathbf{h} \in \mathbb{R}^d$ is the last-layer hidden state of the LLM backbone. Minimizing the MSE loss $\mathcal{L}_{\text{MSE}} = (V_\phi(s) - y)^2$ yields a gradient with respect to the backbone representation:

$$\nabla_{\mathbf{h}} \mathcal{L}_{\text{MSE}} = 2(V_\phi(s) - y) \cdot \mathbf{w} \in \mathrm{span}(\mathbf{w}). \tag{1}$$

For a fixed scalar head and one sample, this gradient is confined to the direction $\mathbf{w}$. This is a statement about the local gradient passed to the final hidden state; it does *not* imply that cumulative backbone updates across samples, steps, or layers are globally rank one.

In contrast, a categorical critic with m bins uses logits $\mathbf{z} = W\mathbf{h} + \mathbf{b}$ with $W \in \mathbb{R}^{m \times d}$ and $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$. Given target $\mathbf{q}$, cross-entropy yields $\nabla_{\mathbf{h}} \mathcal{L}_{\mathrm{CE}} = W^{\top}(\mathbf{p} - \mathbf{q})$, which can combine multiple rows of W. The categorical head therefore provides a broader *per-sample* gradient channel. This observation is generic to multi-output heads and is not, by itself, an explanation specific to RLVR. Indeed, the weaker one-hot results in Table 1 show that a multi-output head alone is insufficient.

4.2 Target smoothing and local geometry

The difference between one-hot and HL-Gauss controls points to target construction rather than output dimensionality alone. Sharp one-hot targets encourage predictions near simplex corners, whereas HL-Gauss assigns probability mass to neighboring bins and encourages an interior solution.

HL-Gauss mitigates this by applying Gaussian smoothing to the value targets. From an information-geometric perspective, HL-Gauss keeps the target distribution away from the corners of the probability simplex. For a categorical head with logits $\boldsymbol{\ell}$ and probabilities $\mathbf{p} = \mathrm{softmax}(\boldsymbol{\ell})$, the Fisher information matrix with respect to the logits is

$$F(\boldsymbol{\ell}) = \mathbb{E}_{y \sim \mathbf{p}} \left[\nabla_{\boldsymbol{\ell}} \log p(y; \boldsymbol{\ell}) \nabla_{\boldsymbol{\ell}} \log p(y; \boldsymbol{\ell})^{\top} \right] = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\top}. \tag{2}$$

When $\mathbf{p}$ approaches a corner of the simplex, this matrix becomes nearly degenerate. Near the population optimum, smoothed targets encourage less extreme predictions and can yield better-conditioned local geometry. We treat this as a plausible optimization mechanism, supported by the one-hot, two-hot, and bandwidth controls, rather than as a complete causal account of the policy gains.

4.3 Calibration and advantage asymmetry

Critic calibration directly shapes the PPO update. In the illustrative case $\gamma{=}\lambda{=}1$, GAE reduces to $\hat{A}_t = G_t - V_\phi(s_t)$ with $G_t \in \{0, 1\}$. If the scalar prediction overestimates the conditional success probability, failed trajectories receive negative advantages of magnitude $V_\phi(s_t)$ while successful trajectories receive positive advantages of magnitude $1 - V_\phi(s_t)$. Thus even a valid scalar estimator can produce asymmetric learning signals when imperfectly optimized or calibrated under a changing on-policy distribution. Our empirical question is whether the HL-Gauss training surrogate reduces this error in practice. Section 5.4 answers affirmatively with direct Brier/ECE measurements and advantage statistics. The categorical output itself is not passed to the actor; the benefit appears through the quality of its decoded scalar value.

5 Experiments

We evaluate HL-Gauss PPO to study whether a classification-based critic objective improves reinforcement learning for LLMs. Our central comparison is against PPO with a standard scalar MSE critic, which isolates the effect of replacing scalar value regression with categorical value prediction. We additionally compare against DAPO (Yu et al., 2025) and conduct *ablations* over a Bernoulli two-bin critic and one-hot, two-hot, and HL-Gauss targets.

Our experimental design is intentionally controlled: unless otherwise stated, all methods share the same actor objective, rollout and sampling pipeline, optimization schedule, and evaluation protocol, and differ only in how the critic target is parameterized and trained. This setup attributes differences to the critic objective rather than to changes in the policy optimization recipe. We additionally report tool-augmented math (Appendix B.1), Qwen3 generalization (Appendix B.2), smoothing-bandwidth ablations (Appendix B.3), and computational overhead (Appendix C).

5.1 Experimental setup

Unless otherwise stated, all methods are trained under a unified PPO-style framework, where the only change across variants is the critic learning target: standard PPO uses scalar regression with an MSE loss, while our categorical variants use one-hot, two-hot, or HL-Gauss targets. In particular, the actor-side PPO loss, rollout generation procedure, sampling budget, optimizer settings, learning-rate schedule, and evaluation scripts are kept fixed across all methods.

To make the comparison meaningful, we adopt a strong PPO baseline rather than a minimal vanilla implementation. Specifically, our baseline incorporates several optimization techniques that have become common in recent large-scale RL for reasoning models, following *DAPO* (Yu et al., 2025) and parts of the VAPO training recipe (Yue et al., 2025), including Clip-Higher, Token-level Loss, Critic Pretraining, and Group Sampling. These components are applied uniformly across all critic variants.

We nevertheless refer to this baseline as *PPO* rather than *VAPO*, since we do not adopt the full VAPO formulation. In particular, we exclude Length-Adaptive GAE and Positive Example Loss. Empirically, we find that using $\lambda = 1$ in actor performs best in our setting, which leaves little room for additional gains from Length-Adaptive GAE. We also find that Positive Example Loss has a disproportionate effect on pass@k, making it less suitable for our main controlled comparison, whose goal is to isolate the effect of the critic parameterization.

Model setup. We use Qwen2.5-Math-7B (Yang et al., 2024b) as the backbone for mathematical reasoning and Qwen2.5-7B-Instruct (Yang et al., 2024a) for tool-augmented math and Search-R1 experiments. We additionally test Qwen3-4B-Base (Yang et al., 2025) under the same PPO protocol (Appendix B.2). For each backbone, PPO and HL-Gauss PPO start from the same initialization and use the same tokenizer, context setting, and generation/evaluation pipeline. All experiments are implemented in veRL, an open-source RLHF framework built on HybridFlow (Sheng et al., 2024).

Dataset setup. For training on mathematical reasoning, we use DAPO-Math-17K (Yu et al., 2025). We evaluate on AIME24, AIME25 (MAA, 2025), and further report HMMT (HMMT, 2025), BeyondAIME, and Brumo (Yu et al., 2025) for broader coverage. For an agentic task, Search-R1, we report mean@1 accuracy on NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). Across all tables, we report avg@k and pass@k where applicable, and use Δ to denote absolute gains over PPO under matched settings.

5.2 Mathematical reasoning

Table 1 shows that HL-Gauss PPO consistently outperforms PPO and DAPO under the same experimental setup. On Qwen2.5-Math-7B, HL-Gauss PPO improves the overall avg@256 score from 15.88 (PPO) to 18.74 (+2.86), with the largest single-dataset gain on AIME25 (+5.20). The same trend appears in pass@256, where the average score increases from 38.48 to 48.06 (+9.58). On Qwen3-4B-Base, the corresponding averages improve from 15.33 to 17.18 for avg@256 and from 53.40 to 57.13 for pass@256 (Appendix B.2), showing that the trend is not confined to Qwen2.5-7B.

What the Bernoulli control isolates. The Bernoulli two-bin critic predicts $P(G{=}1 \mid s)$ with a binary classification loss and is distinct from two-hot interpolation over the 101-bin support. It improves pass@256 over MSE (45.67 vs. 38.48) but does not improve avg@256 (15.50 vs. 15.88), and remains below HL-Gauss on both aggregates. Binary classification therefore helps recover rare successful trajectories, but does not fully account for the benefit of a smoothed, metric-aware value support.

Why pass@k improves more. If a sample succeeds with probability p, then pass@$k = 1 - (1-p)^k$ and its sensitivity is $k(1-p)^{k-1}$. Improvements on hard problems, where p is small, can therefore be amplified in pass@k even when their effect on avg@k is modest.

Table 1: Performance on mathematical reasoning with avg@k. Best results in **bold**, second best underlined, within each backbone group. Δ rows denote the performance gain of HL-Gauss PPO over PPO.

(a) avg@k results

	AIME24 *avg@256*	**AIME25** *avg@256*	**HMMT** *avg@256*	**BeyondAIME** *avg@256*	**Brumo** *avg@256*	**Avg.**
Qwen2.5-Math-7B	10.70	4.40	0.30	2.00	5.72	4.62
+ DAPO	31.90	17.30	4.28	7.50	19.00	16.00
+ PPO	34.30	18.60	1.00	7.30	18.20	15.88
+ Bernoulli 2-bin PPO	34.10	17.40	3.00	7.50	15.50	15.50
+ One-hot PPO	27.80	17.40	1.50	7.80	19.80	14.86
+ Two-hot PPO	**37.40**	19.20	3.23	7.60	20.30	17.55
+ HL-Gauss PPO (ours)	35.20	**23.80**	**4.78**	**8.50**	**21.40**	**18.74**
Δ	+0.90	+5.20	+3.78	+1.20	+3.20	+2.86

(b) pass@k results

	AIME24 *pass@256*	**AIME25** *pass@256*	**HMMT** *pass@256*	**BeyondAIME** *pass@256*	**Brumo** *pass@256*	**Avg.**
Qwen2.5-Math-7B	73.33	46.67	16.70	33.00	34.00	40.74
+ DAPO	63.33	46.67	**20.00**	35.00	46.70	42.34
+ PPO	63.33	46.67	16.70	29.00	36.70	38.48
+ Bernoulli 2-bin PPO	73.33	53.33	16.67	35.00	50.00	45.67
+ One-hot PPO	66.67	**56.67**	16.70	31.00	46.70	43.55
+ Two-hot PPO	70.00	**56.67**	13.30	35.00	50.00	45.00
+ HL-Gauss PPO (ours)	**76.67**	53.33	**20.00**	**37.00**	**53.30**	**48.06**
Δ	+13.34	+6.66	+3.30	+8.00	+16.60	+9.58

This is consistent with our calibration analysis: the largest critic gap occurs on likely-failure prefixes, where rare successful continuations are most vulnerable to under-reinforcement.

5.3 Search-R1 benchmark

To further evaluate reasoning with external search, we add experiments on Search-R1 (Jin et al., 2025) using the Qwen2.5-7B-Instruct (Yang et al., 2024a) backbone. We compare three training strategies under the same setup: DAPO, PPO, and HL-Gauss PPO. We report per-test-set metrics on seven datasets (NQ, TriviaQA, PopQA, HotpotQA, 2Wiki, MuSiQue, and Bamboogle) for transparent comparison.

5.4 Empirical analysis of advantage dynamics

We now examine the *advantage signals* produced by each critic, since these directly determine the PPO policy update. In our RLVR setting (γ=1, λ=1), with sparse binary rewards $r \in \{0, 1\}$ assigned only at the terminal token, the GAE estimator reduces to a Monte Carlo residual: $\hat{A}_t = r - V_\phi(s_t)$. This makes advantage analysis a direct probe of critic calibration. We collect 64 rollouts per problem on AIME24 and AIME25 and evaluate critics from matched training protocols.

Table 3 and Figure 2 reveal a pronounced skew in the advantages produced by the MSE critic: penalties for incorrect rollouts (r=0) are 2–3$\times$ larger in magnitude than rewards for correct ones (r=1). On AIME24, for example, the mean raw advantage for failures is -0.597, whereas successful rollouts receive only $+0.227$ (ratio: 2.63$\times$). The resulting policy update places more magnitude on failures than successes. In contrast, the HL-Gauss critic produces nearly symmetric advantages (ratio $\approx$ 1.0$\times$). Batch-wise advantage normalization applies

Table 2: Search-R1 results with mean@1 accuracy (%). The upper block reports Search-R1-instruct numbers quoted from Search-R1 (Jin et al., 2025) for reference only. The lower block reports our runs on Qwen2.5-7B-Instruct, where **bold**/underlined indicate best/second-best within this block. Δ denotes the performance gain of HL-Gauss PPO over PPO in percentage points.

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Reported in Search-R1 (Jin et al., 2025) (reference only)								
Search-R1-instruct (GRPO)	42.90	62.30	42.70	38.60	34.60	16.20	40.00	39.60
Search-R1-instruct (PPO)	39.30	61.00	39.70	37.00	41.40	14.60	36.80	38.50
Qwen2.5-7b-instruct	32.13	53.32	25.37	27.63	7.50	28.00	40.00	30.56
DAPO	48.95	65.54	42.23	47.74	35.44	19.69	45.60	43.60
PPO	49.30	65.30	44.40	48.12	41.00	21.43	44.80	44.91
HL-Gauss PPO (ours)	**50.86**	**67.29**	**46.26**	**49.64**	**43.38**	**22.06**	**50.40**	**47.13**
Δ	+1.56	+1.99	+1.86	+1.52	+2.38	+0.63	+5.60	+2.22

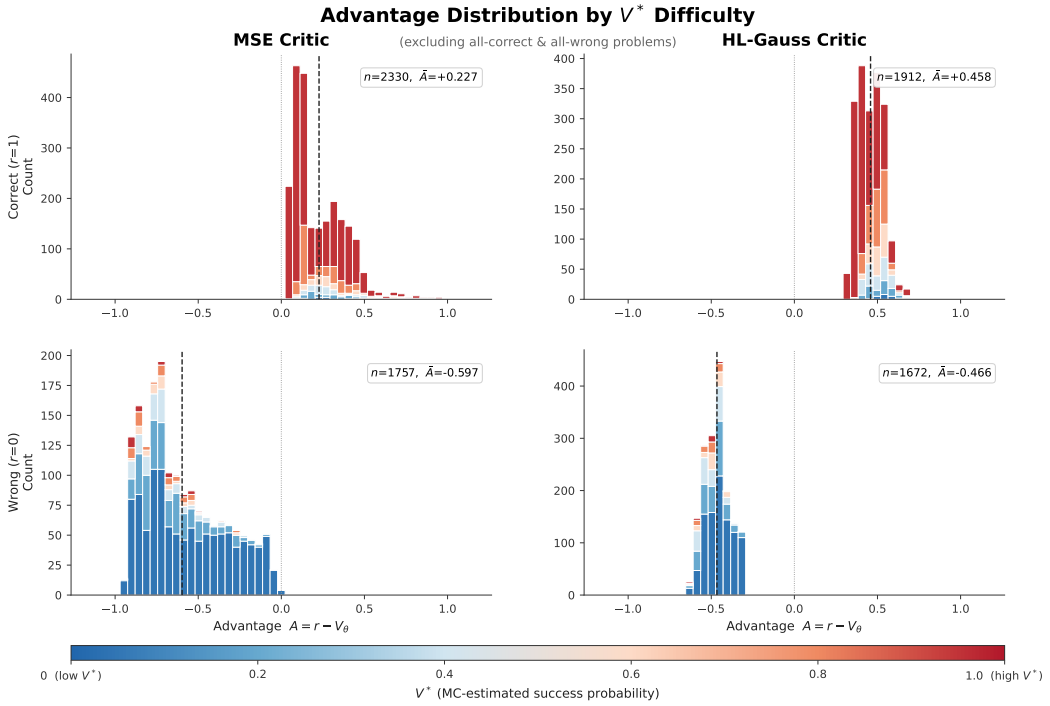

Figure 2: Advantage distribution on AIME24 conditioned on episode outcome and oracle value V^* (estimated via 256 Monte Carlo continuations). Each bar is colored by V^*, indicating the reasoning state's success probability. **Left:** MSE critic produces asymmetric advantages — mean positive advantage (+0.23) is much smaller than mean negative advantage (−0.60). **Right:** HL-Gauss critic yields near-symmetric advantages (≈±0.46). All-correct and all-wrong problems are excluded.

only an affine rescaling, $\tilde{A}_i = (A_i - \mu_B)/\sigma_B$; it standardizes scale but does not remove this relative asymmetry. The ratio is not expected to remain constant as the policy and critic evolve. Measurements at steps 30, 90, 150, and 300 show that MSE becomes increasingly negative-skewed while HL-Gauss remains closer to one (Appendix B.5).

Direct calibration. We additionally evaluate AIME24 prefixes at four positions per rollout, with oracle success probability $V^*(s)$ estimated from 256 Monte Carlo continuations. Table 4 reports the retained sample count and calibration metrics on this common prefix collection. Expectation decoding is the operational comparison because it is the scalar passed to

Table 3: Raw advantage statistics for MSE vs. HL-Gauss critics. Ratio near 1.0 indicates symmetric learning; higher values indicate the critic penalizes errors more than rewards successes.

Dataset	Critic	$\bar{A}_{r=0}$	$\bar{A}_{r=1}$	Ratio
AIME24	MSE	-0.597	$+0.227$	$2.63\times$
	HL-Gauss	-0.466	$+0.458$	$1.02\times$
AIME25	MSE	-0.486	$+0.175$	$2.78\times$
	HL-Gauss	-0.450	$+0.456$	$0.99\times$

Table 4: Critic calibration on AIME24 reasoning prefixes; lower is better. Oracle V^* is estimated with 256 Monte Carlo continuations. "Exp." is the scalar expectation used by PPO, while "mode" is auxiliary. Clamp is the fraction of predictions requiring projection to the valid binary range.

Critic / decoder	N	Brier	ECE	MCE	Clamp
MSE	7,671	0.2031	0.2612	0.4191	0.72%
HL-Gauss exp.	7,680	0.1637	0.2131	0.3021	0.00%
HL-Gauss mode	7,680	**0.1470**	**0.1160**	**0.1624**	0.00%

GAE/PPO; mode decoding is included only as an auxiliary diagnostic. Relative to MSE, HL-Gauss expectation improves Brier score by 19.4%, ECE by 18.4%, and MCE by 27.9%. The largest Brier gap occurs for $0 < V^* < 0.5$ (0.296 for MSE vs. 0.121 for HL-Gauss expectation), the region in which value overestimation most strongly inflates negative advantages. These post-hoc diagnostics support a calibration-based account, while not establishing advantage symmetry as the sole causal mediator of policy performance.

6 Related work

Reinforcement learning for LLMs. RLHF commonly uses PPO with learned reward models (Ouyang et al., 2022; Stiennon et al., 2020; Schulman et al., 2017). Critic-free alternatives include DPO (Rafailov et al., 2023), adaptive DPO sample scheduling (Huang et al., 2025), REINFORCE-style optimization (Ahmadian et al., 2024), and GRPO (Shao et al., 2024). DPO optimizes preference pairs directly, REINFORCE uses Monte Carlo returns, and GRPO uses within-group reward comparisons. The success of DeepSeek-R1 has renewed interest in RLVR (DeepSeek-AI, 2025): DAPO adds dynamic sampling and clip-higher (Yu et al., 2025), whereas VAPO improves actor–critic training through value pretraining and length-aware stabilization (Yue et al., 2025). Other work targets exploration and diversity using policy divergences (Li et al., 2026), count-based rewards (Zhang et al., 2026), or asymmetric group-relative weighting (Yu et al., 2026). The last notion concerns symmetry of GRPO group weights, unlike the positive/negative scalar GAE balance analyzed here. These methods change data selection or policy optimization; we instead change only the critic learning objective, not rewards, policy regularization, or group weighting.

Regression as classification for value functions. Discretizing continuous targets and training by classification has appeared throughout supervised learning and RL. Histogram losses provide smooth target distributions (Imani & White, 2018), MuZero uses two-hot categorical value and reward heads (Schrittwieser et al., 2020), and broader studies show favorable scaling under non-stationary targets (Farebrother et al., 2024). These works establish the loss family; our contribution is its controlled study in PPO-based LLM RLVR. One-hot, two-hot, Bernoulli, smoothing, calibration, and advantage controls separate target construction from output dimensionality.

Distributional reinforcement learning. Full-return methods include C51 (Bellemare et al., 2017), QR-DQN (Dabney et al., 2018), D4PG (Barth-Maron et al., 2018), nonlinear distributional gradient TD (Qu et al., 2019), and LLM-oriented DisPPO (Zhou et al., 2026). They explicitly model $Z(s, a)$ and may exploit information beyond its mean for risk or exploration. Our categorical head instead serves only as a training objective and is decoded to the scalar $\hat{V} = \sum_i p_i z_i$ used by an unchanged GAE/PPO update; the actor never consumes quantiles,

variance, or another higher-order statistic. It is therefore a drop-in replacement for the MSE critic rather than a full-return distributional method.

7 Conclusion

We study HL-Gauss as a drop-in critic-training surrogate for PPO in RLVR. The method replaces scalar MSE regression with smoothed categorical cross-entropy, then decodes a scalar expectation for an otherwise unchanged GAE/PPO actor update. Across mathematical reasoning, tool-augmented math, Search-R1, and Qwen2.5/Qwen3 backbones, this modification yields consistent gains over strong PPO and DAPO baselines. Bernoulli two-bin and matched categorical-head controls show that classification alone and output dimensionality do not fully explain the improvement. Direct calibration measurements further show that HL-Gauss produces better scalar value estimates on reasoning prefixes and more symmetric, lower-variance advantages. The evidence therefore supports a practical optimization-and-calibration benefit, rather than a claim that scalar expected-return critics are theoretically invalid.

Limitations. Our backbone coverage remains within the Qwen family and 4B–7B scale; broader architectures and larger models remain to be tested. The calibration and advantage analyses are diagnostic and do not prove that advantage symmetry is the only causal mediator of the policy gains. Finally, the best support margins and smoothing bandwidth can depend on the task's reward range, although the corresponding sensitivity and task-dependent settings are reported in the appendix.

8 Ethics Statement

This work proposes a methodological improvement to the critic training objective within PPO for LLMs. Our method does not introduce new data, models, or applications; it modifies an internal training component of existing RLVR pipelines. The benchmarks used are publicly available mathematical reasoning and open-domain QA datasets that do not contain sensitive or personally identifiable information. We do not foresee specific ethical risks arising from replacing MSE with classification-based value learning. Broader ethical considerations surrounding RLVR-trained LLMs — such as the potential for generating plausible but incorrect reasoning — apply equally to all methods in this space and are not unique to our contribution.

9 Reproducibility statement

We have made every effort to ensure that the results presented in this paper are reproducible. Base models and evaluation datasets are publicly available, and we will also release our checkpoints and datasets. Our code is openly available at: https://github.com/ZhijianZhou/HL-guass-ppo. The experimental setup, including training steps, training configurations, and hardware details, is described in detail in the appendix. We have also provided a full description to assist others in reproducing our experiments.

LLM usage disclosure. Large language models were used to assist with language polishing. All technical content, claims, and references were reviewed and verified by the authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82394432 and 92249302) and the Shanghai Municipal Science and Technology Major Project (Grant No. 2023SHZDZX02).

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.

Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taiga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop regressing: Training value functions via classification for scalable deep RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13049–13071. PMLR, 2024. URL https://proceedings.mlr.press/v235/farebrother24a.html.

HMMT. Harvard-MIT Mathematics Tournament (HMMT), 2025. https://www.hmmt.org/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. Adaptive batch-wise sample scheduling for direct preference optimization. *arXiv preprint arXiv:2506.17252*, 2025. doi: 10.48550/arXiv.2506.17252. URL https://arxiv.org/abs/2506.17252.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2157–2166. PMLR, 2018.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. doi: 10.48550/arXiv.2503.09516. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Klein Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6a1b224b153e55c40a6359f9c9fb9d8c-Paper-Conference.pdf.

MAA. American invitational mathematics examination (aime), 2025. https://maa.org/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2023.

OpenAI. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. doi: 10.48550/arXiv.2412.16720. URL https://arxiv.org/abs/2412.16720.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Chao Qu, Shie Mannor, and Huan Xu. Nonlinear distributional gradient temporal-difference learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5251–5260. PMLR, 2019. URL https://proceedings.mlr.press/v97/qu19b.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. doi: 10.48550/arXiv.1506.02438. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024. doi: 10.48550/arXiv.2409.19256. URL https://arxiv.org/abs/2409.19256.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024a. URL https://arxiv.org/abs/2412.15115.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math Technical Report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b. doi: 10.48550/arXiv.2409.12122. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Zhiqi Yu, Zhangquan Chen, Mengting Liu, Heye Zhang, and Liangqiong Qu. Unveiling implicit advantage symmetry: Why GRPO struggles with exploration and difficulty adaptation. *arXiv preprint arXiv:2602.05548*, 2026. doi: 10.48550/arXiv.2602.05548. URL https://arxiv.org/abs/2602.05548.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, et al. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025. doi: 10.48550/arXiv.2504.05118. URL https://arxiv.org/abs/2504.05118.

Xuan Zhang, Ruixiao Li, Zhijian Zhou, Long Li, Yulei Qin, Ke Li, Xing Sun, Xiaoyu Tan, Chao Qu, and Yuan Qi. Count counts: Motivating exploration in LLM reasoning with count-based intrinsic rewards. In *International Conference on Learning Representations*, volume 2026, pp. 107571–107604, 2026. URL https://openreview.net/forum?id=9xIBbfItGP.

Zhijian Zhou, Long Li, Xuan Zhang, Zongkai Liu, Yanting Miao, Yuchen Liu, Deshu Chen, Ke Li, Xing Sun, Ruoxi Jiang, Xiaoyu Tan, Chao Qu, and Yuan Qi. DisPPO: Quantile-based distributional reinforcement learning for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tkdgg5uqNK.

A Detailed Training Configurations

A.1 Training Data

Mathematical Reasoning For both training and evaluation datasets, we use the following system prompt:

System Prompt

Please reason step by step, and put your final answer within \boxed{}.

Math-with-tool For tool-augmented mathematical reasoning, we follow the standard chat template and use the built-in tool instruction prompt. Specifically, we prepend the following system prompt to each conversation in both training and evaluation:

System Prompt

You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"type": "function", "function": {"name": "code_interpreter", "description": "A tool for executing code.", "parameters": {"type": "object", "properties": {"code": {"type": "string", "description": "The code to execute."}}, "required": ["code"]}}
</tools>
For each function call, return a JSON object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

Search-R1 For search-augmented reasoning, we follow the standard web information seeking paradigm. We prepend the following system prompt to each query in both training and evaluation:

System Prompt

You are a Web Information Seeking Master. Your task is to thoroughly seek the internet for information and provide accurate answers to questions. No matter how complex the query, you will not give up until you find the corresponding information.
As you proceed, adhere to the following principles:
1. Persistent Actions for Answers: You will engage in many interactions, delving deeply into the topic to explore all possible aspects until a satisfactory answer is found.
2. Repeated Verification: Before presenting a Final Answer, you will **cross-check** and **validate the information** you've gathered to confirm its accuracy and reliability.
3. Attention to Detail: You will carefully analyze each information source to ensure that all data is current, relevant, and from credible origins.

A.2 RL Training Configuration

Mathematical Reasoning We use the hyperparameters in Table 5 for RL training on mathematical reasoning.

We use an outcome-based reward function that assigns $+1$ for correct final answers and 0 otherwise.

Math-with-tool We use the hyperparameters in Table 6 for RL training on tool-augmented mathematical reasoning.

For tool-augmented reasoning, we define an outcome-based reward with an auxiliary tool-usage term. Let s denote the generated solution and y the ground-truth answer. We first compute a base correctness score using strict boxed-answer verification:

$$r_{\text{ans}}(s, y) \in \{0, 1\}, \tag{3}$$

where $r_{\text{ans}}(s, y) = 1$ if the final boxed answer is correct and 0 otherwise.

To account for tool usage, we define a lightweight turn-based bonus. Let T denote the total number of interaction turns in the trajectory. The tool-related adjustment is given by

$$r_{\text{tool}}(T) = \frac{T-2}{2} \cdot \alpha, \tag{4}$$

Table 5: RL training configurations for mathematical reasoning (Qwen2.5-Math-7B).

Hyperparameter	Value
Optimizer	AdamW
Actor learning rate	1e-6
Critic learning rate	2e-6
Training batch size	512
Samples per prompt	16
Mini-batch size	32
Max prompt length	1024
Max response length	3072
Rollout temperature	1.0

where $\alpha = 0.1$ is a fixed scaling constant and the first two turns receive no bonus.

The final reward is computed as

$$r(s, y, T) = \begin{cases} r_{\text{ans}}(s, y) + \mathbb{I}[r_{\text{tool}}(T) > 0], & \text{if } r_{\text{ans}}(s, y) = 1, \\ \max(0,\ r_{\text{ans}}(s, y) + r_{\text{tool}}(T)), & \text{otherwise,} \end{cases} \tag{5}$$

where the auxiliary term is bounded so that incorrect final answers do not receive positive reward.

Table 6: RL training configurations for tool-augmented mathematical reasoning (Qwen2.5-7B-Instruct).

Hyperparameter	Value
Optimizer	AdamW
Actor learning rate	1e-6
Critic learning rate	2e-6
Training batch size	128
Samples per prompt	16
Mini-batch size	32
Max prompt length	2048
Max response length	16384
Max conversation turns	8
Rollout temperature	1.0

Search-R1 We use the hyperparameters in Table 7 for RL training on Search-R1.

The reward function for Search-R1 serves as the primary training signal, guiding the optimization process in RL. We adopt a rule-based reward system that consists solely of final outcome rewards, which assess the correctness of the model's response. For factual reasoning tasks, correctness is evaluated using rule-based criteria such as exact string matching:

$$r_\phi(x, y) = \text{EM}(a_{\text{pred}}, a_{\text{gold}}), \tag{6}$$

where a_{pred} is the extracted final answer from response y and a_{gold} is the ground truth answer. The EM score returns 1 for exact matches and 0 otherwise.

A.3 Inference Configurations

Mathematical Reasoning We use a rollout temperature of 0.6, top-p sampling with $p = 0.95$, and a maximum response length of 4096 tokens. We adopt $k = 256$ for AIME24/2025 datasets, $k = 16$ for Minerva, MATH500, and OlympiadBench, and $k = 8$ for additional benchmarks, balancing computational cost and task difficulty.

Table 7: RL training configurations for Search-R1 (Qwen2.5-7B-Instruct).

Hyperparameter	Value
Optimizer	AdamW
Actor learning rate	1e-6
Critic learning rate	2e-6
Training batch size	512
Samples per prompt	5
Mini-batch size	256
Max prompt length	2048
Max response length	16384
Max conversation turns	6
Rollout temperature	1.0

Math-with-tool We use a rollout temperature of 0.6, top-p sampling with $p = 0.95$, and a maximum response length of 16384 tokens.

Search-R1 We use a rollout temperature of 1.0 during training and 0.0 (greedy sampling) during evaluation, with top-p sampling at $p = 1.0$, and a maximum response length of 16384 tokens. We evaluate with mean@1 accuracy on seven datasets (NQ, TriviaQA, PopQA, HotpotQA, 2Wiki, MuSiQue, and Bamboogle).

A.4 HL-Gauss Critic Configuration

The HL-Gauss categorical critic extends the PPO critic with a categorical value head used as a training parameterization for a decoded scalar value. The following hyperparameters are specific to the HL-Gauss formulation:

Table 8: HL-Gauss critic-specific configurations (task-dependent support ranges).

Hyperparameter	Value	
	Math Reasoning	Math-with-tool / Search-R1
Number of bins (m)	101	
Support range $[v_{\min}, v_{\max}]$	$[-0.1, 1.1]$	$[-1.1, 2.2]$ / $[-1.1, 1.1]$
Smoothing bandwidth (σ)	0.009	
Critic warmup steps	30	
Discount factor (γ)	1.0	
GAE parameter (λ)	1.0	
Gradient clipping	1.0	
PPO clip range	[0.2, 0.28]	

The HL-Gauss binning scheme uses a uniform discretization of the value support into m=101 bins. The support range follows each task's attainable reward/value range: for binary mathematical reasoning we use $[-0.1, 1.1]$; for tool-augmented reasoning with bonuses, $[-1.1, 2.2]$; and for Search-R1 with possible negative rewards, $[-1.1, 1.1]$. For binary math, using exactly $[0, 1]$ would place both labels on support boundaries, truncating roughly half of the Gaussian mass before renormalization and making endpoint targets nearly one-hot. The 0.1 margin preserves smoothing at both endpoints. With σ=0.009 $= 0.75\Delta$, the target mass is concentrated over approximately 2–3 nearby bins, balancing sharpness and local smoothness.

B Additional experimental results

B.1 Math-with-tool performance

Beyond pure mathematical reasoning on Qwen2.5-Math-7B, we also evaluate HL-Gauss PPO on tool-augmented mathematical reasoning using Qwen2.5-7B-Instruct. Under the same setup alignment, Table 9 reports tool-augmented math results. Compared with PPO, HL-Gauss PPO improves the average from 37.45 to 39.15 (+1.70), driven primarily by the AIME25 gain (+3.2), while remaining nearly unchanged on AIME24 (+0.2). This shows that the classification-based critic transfers effectively to tool-use reasoning scenarios under the same training and evaluation conditions.

Table 9: Math-with-tool performance (Qwen2.5-7B-Instruct). Results reported as avg@64.

Methods	**AIME24** *avg@64*	**AIME25** *avg@64*	**Avg.**
Qwen2.5-7b-instruct	12.7	5.5	9.10
+ DAPO	37.4	29.3	33.35
+ PPO	<u>43.9</u>	<u>31.0</u>	<u>37.45</u>
+ HL-Gauss PPO (ours)	**44.1**	**34.2**	**39.15**
Δ (vs PPO)	+0.2	+3.2	+1.7

B.2 Generalization to Qwen3-4B-Base

To test whether the main result is confined to the Qwen2.5-7B backbone, we repeat the matched PPO comparison with Qwen3-4B-Base. Table 10 shows improvements on all five benchmarks. The aggregate gains are +1.85 avg@256 and +3.73 pass@256.

Table 10: Matched PPO results on Qwen3-4B-Base. HL-Gauss changes only the critic head and loss.

Metric	Critic	AIME24	AIME25	HMMT	BeyondAIME	Brumo	Avg.
avg@256	MSE	21.91	21.31	7.98	9.50	15.96	15.33
	HL-Gauss	**25.78**	**22.97**	**8.76**	**10.93**	**17.48**	**17.18**
pass@256	MSE	67.70	63.30	33.33	46.00	56.67	53.40
	HL-Gauss	**70.00**	**70.00**	**36.67**	**49.00**	**60.00**	**57.13**

B.3 Ablation: Smoothing bandwidth (σ)

We evaluate the sensitivity of HL-Gauss PPO to the smoothing bandwidth parameter σ, which controls the width of the Gaussian kernel applied to target value distributions. We test three values on mathematical reasoning benchmarks (Qwen2.5-Math-7B): $\sigma = 0.009$ (our default, 0.75Δ), $\sigma = 0.012$ (Δ), and $\sigma = 0.024$ (2Δ), where $\Delta \approx 0.012$ is the bin width.

Analysis. The results clearly demonstrate that smaller smoothing bandwidth values yield better performance across all benchmarks. With $\sigma = 0.009$ (0.75Δ), HL-Gauss PPO achieves avg@256 of 18.74 and pass@256 of 48.06, compared to 17.00 and 45.99 for $\sigma = 0.012$ (Δ), and 16.48 and 44.49 for $\sigma = 0.024$ (2Δ).

This trend is consistent across individual datasets (AIME24, AIME25, HMMT, BeyondAIME, Brumo), suggesting that tighter smoothing is preferable for the RLVR regime. Intuitively, smaller σ provides more discriminative target distributions that better distinguish between value regions, while larger σ over-smooths and blurs fine-grained value distinctions. Our choice of $\sigma = 0.75\Delta$ thus represents an effective balance, placing sufficient probability mass at the target value while maintaining sharpness in the surrounding bins. These results

Table 11: Performance of HL-Gauss PPO under different smoothing bandwidth (σ) values on mathematical reasoning. Best results are shown in **bold** and second-best results are underlined.

(a) avg@256 results

Smoothing σ	**AIME24** avg@256	**AIME25** avg@256	**HMMT** avg@256	**BeyondAIME** avg@256	**Brumo** avg@256	**Avg.**
$\sigma = 0.024$ (2Δ)	34.90	18.06	3.56	7.40	18.50	16.48
$\sigma = 0.012$ (Δ)	35.05	18.90	4.53	7.50	19.00	17.00
$\sigma = 0.009$ (0.75Δ)	**35.20**	**23.80**	**4.78**	**8.50**	**21.40**	**18.74**

(b) pass@256 results

Smoothing σ	**AIME24** pass@256	**AIME25** pass@256	**HMMT** pass@256	**BeyondAIME** pass@256	**Brumo** pass@256	**Avg.**
$\sigma = 0.024$ (2Δ)	73.33	53.33	16.70	34.10	45.00	44.49
$\sigma = 0.012$ (Δ)	66.67	**60.00**	**20.00**	35.00	48.30	45.99
$\sigma = 0.009$ (0.75Δ)	**76.67**	53.33	**20.00**	**37.00**	**53.30**	**48.06**

validate the default configuration used throughout the main experiments and suggest that further tuning toward even smaller σ may yield diminishing returns due to numerical stability considerations.

B.4 Advantage error decomposition by V* region

To further isolate the bias in critic predictions across different oracle value regions, we stratify advantages by the oracle value $V^*(s_t)$ (estimated via 256 Monte Carlo continuations). Figure 3 shows the signed bias ($V^* - V_\phi$) across ten V* bins. The MSE critic exhibits a systematic shift: it overestimates values in the likely-wrong region ($V^* \approx 0$) and underestimates them in the likely-correct region ($V^* \approx 1$).

This stratification localizes the empirical calibration gap: the MSE critic overestimates likely-failure prefixes and therefore suppresses the positive advantage assigned to rare successful continuations. HL-Gauss predictions stay closer to the oracle estimates across these regions, consistent with the aggregate Brier/ECE results in Table 4.

B.5 Advantage asymmetry over training

The negative-to-positive advantage ratio evolves as the policy distribution and critic calibration change, so it need not be constant. Table 12 reports AIME24 measurements across training. The MSE ratio grows from 0.134 to 2.63, whereas HL-Gauss remains closer to one at the final checkpoint. This temporal pattern shows that the main-text result is not selected from an isolated fluctuation.

Table 12: AIME24 raw advantage means over training. Ratio is $|\bar{A}_{\text{wrong}}|/\bar{A}_{\text{correct}}$; values near one indicate balanced magnitudes.

	MSE			HL-Gauss		
Step	$\bar{A}_c$	$\bar{A}_w$	Ratio	$\bar{A}_c$	$\bar{A}_w$	Ratio
30	+0.777	-0.104	0.134	+0.670	-0.226	0.337
90	+0.512	-0.404	0.790	+0.527	-0.360	0.684
150	+0.312	-0.538	1.724	+0.472	-0.403	0.854
300	+0.227	-0.597	2.630	+0.458	-0.466	1.020

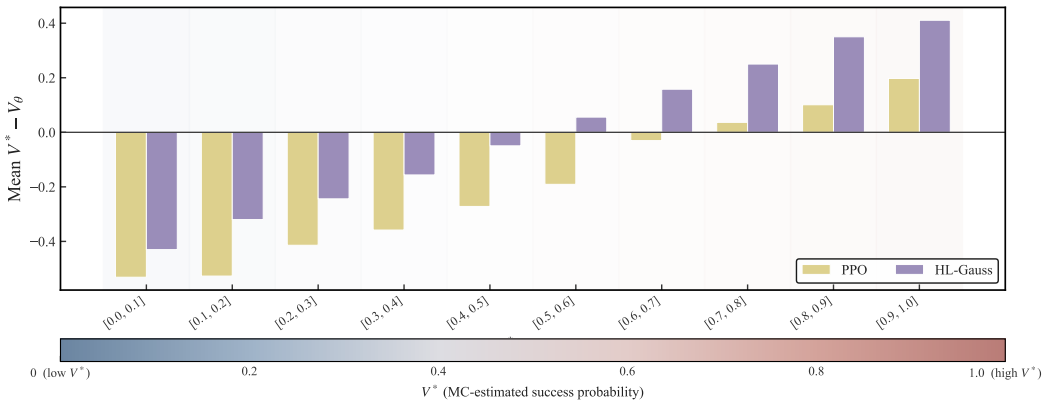

Figure 3: Signed advantage bias ($V^* - V_\phi$) decomposed by oracle value V* region (10 bins, excluding all-correct and all-wrong problems). PPO (MSE critic) shows strong negative bias in low-V* regions (systematic value overestimation) and positive bias in high-V* regions (underestimation), while HL-Gauss critic maintains near-zero bias across all V* regions, indicating superior value calibration.

C Computational Overhead

The HL-Gauss categorical critic introduces minimal parameter overhead compared to the standard MSE critic. The only architectural difference is the value head output dimension: $\mathbb{R}^d \to \mathbb{R}^{101}$ (categorical) versus $\mathbb{R}^d \to \mathbb{R}^1$ (scalar), where d is the hidden dimension. This increases the value head parameters by a factor of 101, but since the value head constitutes a negligible fraction of the LLM backbone ($\sim$ 0.006% for Qwen2.5-7B), the overall model size increase is imperceptible.

In terms of computational cost per training step, the categorical critic adds:

- Softmax computation over 101 bins (negligible)
- Cross-entropy loss instead of MSE (negligible difference)

These overheads are dwarfed by the LLM forward/backward pass, resulting in no meaningful increase in wall-clock training time. Empirically, HL-Gauss PPO and standard PPO exhibit near-identical per-step training speed on the same hardware, confirming that the architectural change is computationally efficient.

D Method Implementation Details

Support range and resolution. In RLVR with binary rewards and γ=1, the theoretical value range is $[0, 1]$. We set $v_{\min}$=−0.1 and $v_{\max}$=1.1 to provide a small margin beyond the reward boundaries, and use m=101 bins (bin width $\Delta \approx 0.0119$). This resolution is sufficient to represent fine-grained value distinctions while adding negligible computational overhead (Appendix C).

Smoothing bandwidth. We set $\sigma = 0.009 = 0.75\Delta$, which provides moderate smoothing across approximately 2–3 bins around each target. This balances between the extremes of one-hot encoding ($\sigma \to 0$, no smoothing) and overly diffuse targets ($\sigma \gg \Delta$, which blur value distinctions).

Critic warmup. Following prior work on PPO training for LLMs (Yu et al., 2025), we warm up the critic for 30 steps before beginning actor updates. During warmup, only the critic is trained on rollout data from the frozen initial policy, allowing the value estimates to stabilize before they influence advantage computation.

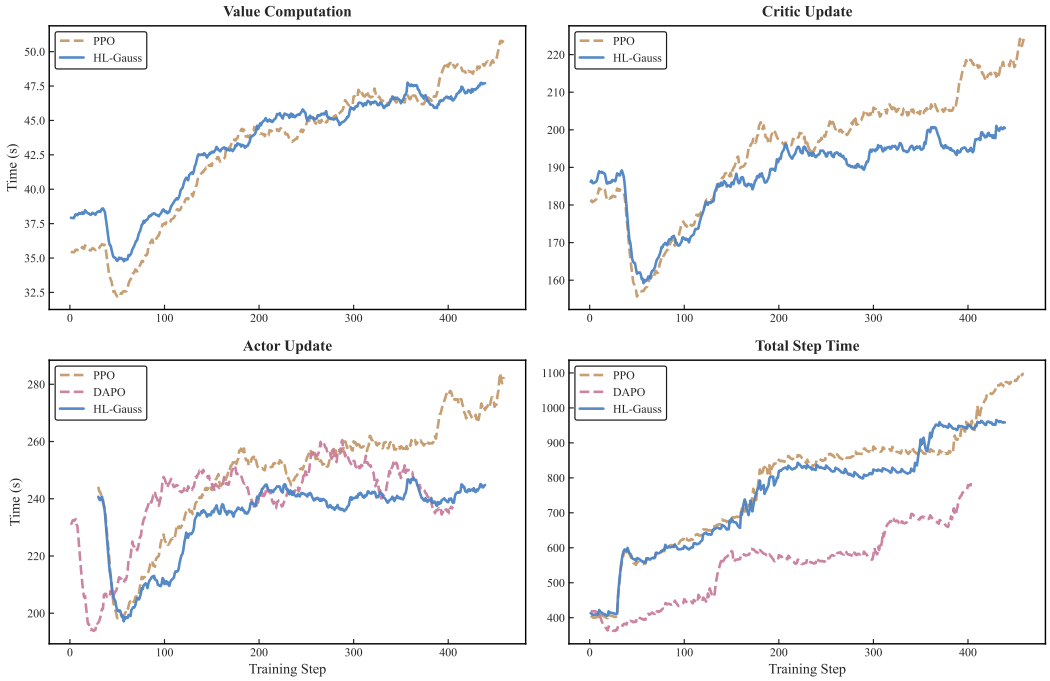

Figure 4: Training step timing breakdown across 400 steps. Value Computation and Critic Update show negligible differences between PPO and HL-Gauss. Actor Update and Total Step Time are nearly identical, confirming minimal computational overhead from the categorical head.

Architecture. The critic is initialized from the same pretrained backbone as the actor. The only architectural difference from the standard PPO critic is the output head: $\mathbb{R}^d \to \mathbb{R}^m$ (categorical) instead of $\mathbb{R}^d \to \mathbb{R}^1$ (scalar). Actor and critic do not share backbone parameters.

E Case Study: Paired Correct and Wrong Rollouts

We conclude with a token-level case study on a single AIME-style prompt for which our sampler produced both a correct rollout and a wrong rollout. The underlying math is straightforward:

$$17_b = b + 7, \qquad 97_b = 9b + 7.$$

Therefore,

$$(b+7) \mid (9b+7) \iff (b+7) \mid \big(9(b+7) - (9b+7)\big) \iff (b+7) \mid 56.$$

Since $b > 9$, we have $b + 7 > 16$, so the only positive divisors of 56 that remain possible are 28 and 56. Hence $b \in \{21, 49\}$, and the required sum is 70. Thus, the rollout ending in 70 is correct, while the rollout ending in 49 is incorrect.

The wrong trace fails only after a locally plausible start. Its base conversion and divisibility setup are correct, but after introducing an auxiliary variable k it effectively narrows $9 - k$ to $\{\pm 1, \pm 7\}$, which is too restrictive and drops the valid case $b = 21$. By contrast, the correct trace keeps the simpler divisor argument $(b + 7) \mid 56$ and directly recovers both valid bases.

The heatmaps below show how the HL-Gauss critic tracks this divergence at the token level. In the correct rollout, predicted values remain high throughout the clean divisor-based derivation and stay high through the final answer 70. In the wrong rollout, the critic assigns moderate value while the setup is still plausible, but its score collapses once the reasoning commits to the incomplete case split and remains low through the final answer 49. This example illustrates that the categorical critic does not merely reward fluent math text; it distinguishes trajectories that stay on the correct proof path from trajectories that only appear locally reasonable.

Figure 5: Correct rollout for the paired case study (page 1/2). The trace quickly reduces the problem to the key divisor condition $(b + 7) \mid 56$, and the critic assigns consistently high value throughout this derivation.

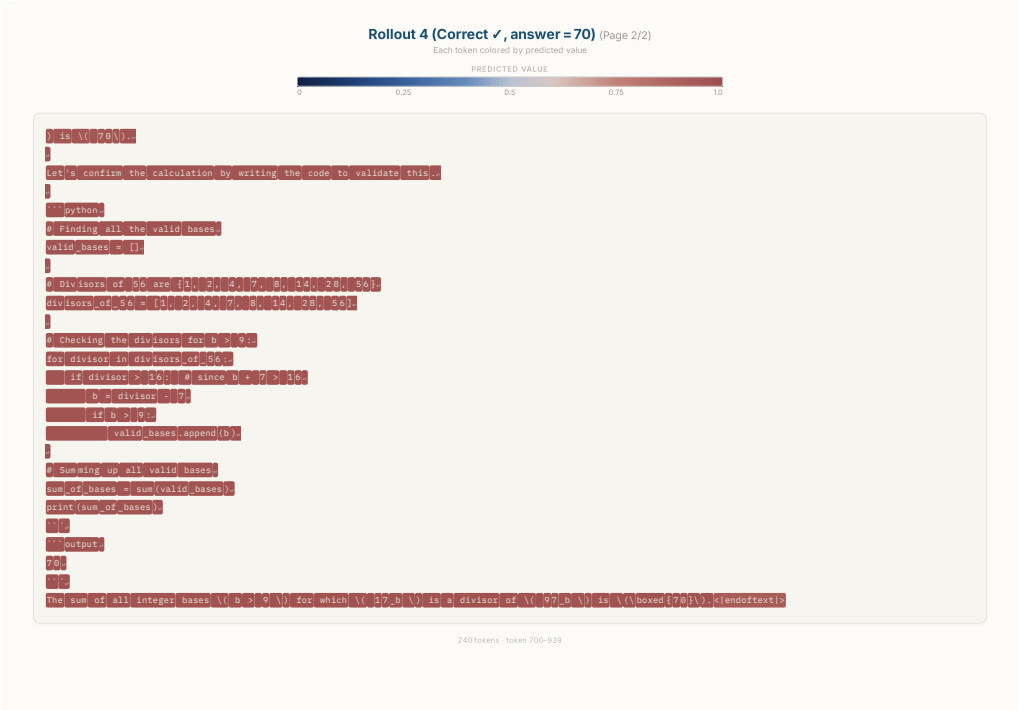

Figure 6: Correct rollout for the paired case study (page 2/2). The critic remains confident through the identification of $b = 21$ and $b = 49$, and stays high through the final correct answer 70.

Figure 7: Wrong rollout for the paired case study (page 1/3). The early setup is partly valid, so the critic does not immediately reject the trace even though the rollout is already drifting toward an overly restrictive analysis.

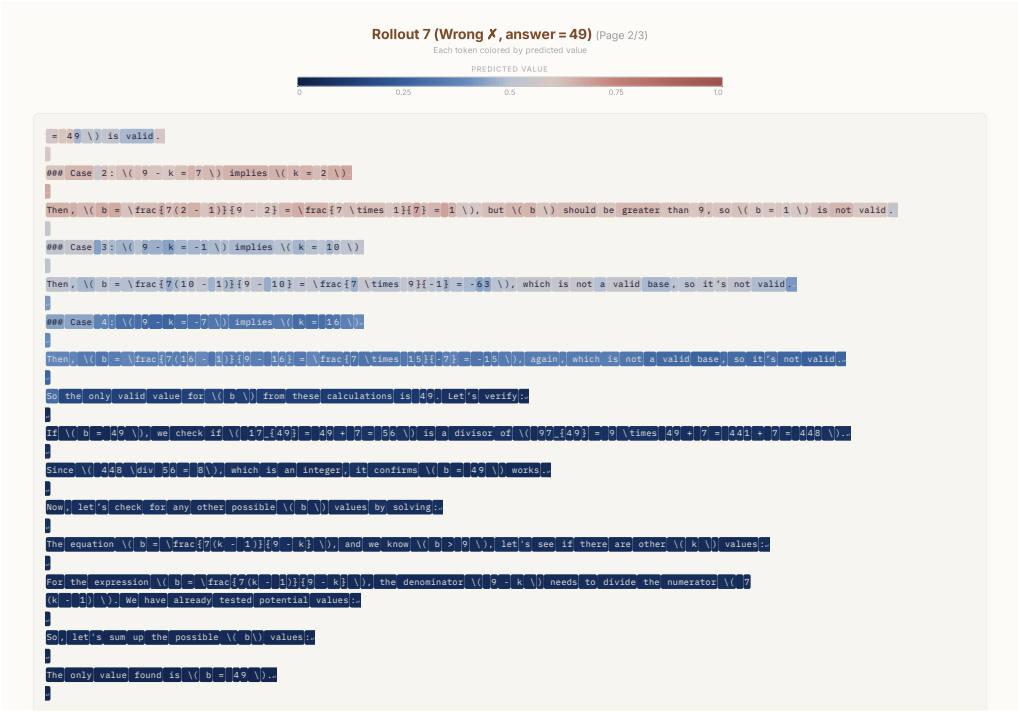

Figure 8: Wrong rollout for the paired case study (page 2/3). Once the derivation commits to the incomplete k-based case split and excludes the valid base $b = 21$, the predicted value drops sharply.

Figure 9: Wrong rollout for the paired case study (page 3/3). The critic remains low through the final incorrect answer 49, showing that it can separate a fluent-looking but invalid continuation from the correct solution path.