CurveRL: Principled Distribution-Aware Context Reweighting for LLM Reasoning

Ke Sun[*1] **Yizhou Zhao**[*1] **Jiayi Xin**[1] **Qi Long**[†1] **Weijie Su**[†1]
[1]University of Pennsylvania
{kesun6, qlong}@upenn.edu, yzzhao@sas.upenn.edu
jiayixin@seas.upenn.edu, suw@wharton.upenn.edu

Abstract

Context or prompt-level reweighting has emerged as a central algorithmic lever in Reinforcement Learning with Verified Rewards (RLVR) for improving the reasoning capability of large language models, yet the principle determining what constitutes an optimal weighting remains poorly understood. We address this gap by formulating prompt reweighting as a functional derivative of a utility functional defined in the pass-rate function space, yielding a unified optimality framework that accommodates existing schemes, including REINFORCE and GRPO. Building on this optimality framework, we propose a distribution-aware prompt reweighting approach, called CurveRL, based on a quantile coordinate transform, in which the weight assigned to each prompt depends not on the absolute value of pass rates but on its rank and density to reflect the distributional structure of the pass rates in the learning dynamics. Extensive experiments across multiple benchmarks demonstrate that our proposed CurveRL consistently outperforms GRPO and other RLVR baselines. Our study identifies context-distribution control as a principled axis for analyzing and designing prompt-reweighted RLVR algorithms.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has been the primary driver behind the recent emergence of reasoning models [44, 37]. As outcome-based rewards become more adopted, the token-level MDP [73] effectively collapses to a contextual bandit [50], in which the entire reasoning trace is absorbed into the response as a one-step decision. Among modern RLVR methods, Group Relative Policy Optimization (GRPO) [79, 37] and its variants, such as [104, 58, 21, 110, 100, 88], have emerged as the dominant family, offering competitive performance with low memory overhead. The mechanisms underlying their success, however, remain poorly understood.

Understanding why these algorithms work requires recognizing a distinctive feature of RLVR that has no direct counterpart in standard RL: the ability to directly shape the context or prompt distribution from which training samples are drawn. In standard RL with an external environment, the state-visitation distribution is mainly shaped indirectly through the agent's actions and exploration strategy [90, 49], as the state or context distribution is often treated as exogenous and cannot be directly manipulated. The contextual-bandit structure of outcome-based RLVR removes this constraint, opening up a new and orthogonal axis of information acquisition, which we refer to as *context distribution control*: the prompt distribution is explicitly controllable during training, and the algorithm can decide which prompts to sample and how heavily to weight their gradients. Recently, a growing body of work exploits this freedom through a wide range of mechanisms, such as sample selection [104, 61, 100], curriculum strategies [71, 74, 18], and prompt reweighting [25, 88]. These approaches are generally grounded by separate heuristics about which prompts deserve more

[*]Equal contribution. [†]Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

gradient signal at each training stage. Yet none provides a principled answer to why its particular intervention is the right one. Recent work [88] introduces the maximum likelihood principle into the RLVR objective by maximizing the log-likelihood of pass rates. While maximum likelihood estimation (MLE) enjoys well-established optimality properties in classical statistics [78, 16], we argue that these guarantees do not transfer to RLVR, as policy optimization is structurally different from statistical estimation. As also briefly mentioned in [25], classical MLE optimality rests on a fixed, exogenous probability measure that characterizes the data-generating population, against which the estimator is evaluated. In RLVR, by contrast, the objective is evaluated under a policy-dependent measure or data distribution that co-evolves with the policy throughout training. There is no fixed population to estimate, and the measure itself is a policy-dependent endogenous object of the optimization. Thus, the classical optimality argument in MLE no longer applies. A detailed discussion of related work is given in Appendix A.

Motivation: Prompt Reweighting and Its Optimality. Prompt reweighting offers a concrete way to implement context distribution control in RLVR. To learn the LLM policy π_θ, we consider RLVR algorithms whose policy gradient update assigns a policy-dependent prompt weight $w_\theta(x)$ to each prompt x. Let $r(x, y)$ denote the rule-based binary reward function for prompt x and response y, d_0 denote the initial prompt distribution. Define the **pass rate** $p_\theta(x) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r(x, y)]$. A broad class of prompt-reweighted RLVR algorithms employ the following policy gradient update:

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim d_0}\left[w_\theta(x)\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x, y)\nabla_\theta \log \pi_\theta(y|x)\right]\right] = \mathbb{E}_{x\sim d_0}\left[w_\theta(x)\nabla_\theta p_\theta(x)\right]. \quad (1)$$

For instance, the population counterpart of GRPO corresponds to $w_\theta(x) = 1/\sqrt{p_\theta(x)(1 - p_\theta(x))}$. A detailed explanation is given in Section 2 with the proof deferred to Appendix B. Importantly, this generic form in Eq. (1) raises the central question of our study:

What are the principle and the optimality to determine $w_\theta(x)$ in the prompt-reweighted RLVR?

Our Contributions. In this paper, we cast prompt reweighting as context distribution control, where the algorithm directly reshapes the effective prompt distribution. Under this view, we formulate the optimal weight as a functional derivative of a utility function over the pass-rate function space. The resulting optimality framework subsumes existing pointwise weighting rules and reveals their *weight collapse* limitation. We then instantiate the principle with CurveRL, with the optimal weight derived by a distribution-aware utility function in pass-rate quantile space, which uses rank and density information of the evolving pass-rate distribution. Empirically, CurveRL consistently improves the pass@1 and pass@k trade-off across multiple reasoning benchmarks. The contributions of our study can be succinctly summarized as follows:

- We formulate prompt reweighting in RLVR as context distribution control and define optimal weights through utility-dependent functional derivatives.
- We instantiate this principle with a distribution-aware utility in pass-rate quantile space and propose CurveRL that characterizes the distributional structure of the pass-rate distribution.
- We perform extensive experiments, showing that CurveRL improves the pass@1 and pass@k Pareto frontier over standard baselines. The underlying mechanism of CurveRL is also analyzed.

2 Preliminaries and Technical Background

REINFORCE Objective in the Pass-Rate Space. RLVR can be formulated as a contextual bandit problem with a rule-based binary reward $r(x, y) \in \{0, 1\}$ for prompt x and response y. To learn π_θ guided by these rewards, the policy gradient method [86], e.g., REINFORCE [97, 87], is generally utilized to maximize $J_{\text{RL}}(\theta)$. $J_{\text{RL}}(\theta)$ and its gradient $\nabla_\theta J_{\text{RL}}(\theta)$ are defined by the following:

$$J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r(x, y)]\right] := \mathbb{E}_{x\sim d_0}[p_\theta(x)], \ \nabla_\theta J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim d_0}\left[\nabla_\theta p_\theta(x)\right]. \quad (2)$$

Thus, in the pass-rate space, REINFORCE corresponds to a constant prompt weight $w_\theta(x) = 1$.

Group Relative Policy Optimization (GRPO) [79, 37]. We next rewrite the gradient of GRPO in the same prompt-reweighting form. Under a contextual bandit or one-step decision abstraction analysis adopted in [25], GRPO can be written in a sequence-level form, which is closely related to Group Sequence Policy Optimization (GSPO) [112]. By replacing the group-wise empirical baseline and

normalization by their population counterparts (i.e., an infinite group size) and ignoring the clipping and the policy-ratio term, the gradient of GRPO can be approximated by an idealized form [25]:

$$\nabla_\theta J_{\mathrm{GRPO}}(\theta) \approx \mathbb{E}_{x\sim d_0}\left[\frac{1}{\sqrt{p_\theta(x)(1-p_\theta(x))}}\nabla_\theta p_\theta(x)\right]. \tag{3}$$

Therefore, the population-level GRPO's weight is $1/\sqrt{p_\theta(x)(1-p_\theta(x))}$, emphasizes prompts with either very small (i.e., $p_\theta(x)\to 0$) or large (i.e., $p_\theta(x)\to 1$) pass rates in the learning dynamics.

Advanced Prompt Reweighting RL Objectives. Several recent RLVR variants can also be interpreted as pointwise transformations of the pass rate [93, 88, 100]. A representative example is Maximum Likelihood RL (MaxRL) [88], where the Maximum Likelihood (ML) principle [11, 16] is heuristically introduced into the RL objective:

$$J_{\mathrm{ML}}(\theta) = \mathbb{E}_{x\sim d_0}\left[\log p_\theta(x)\right], \quad \nabla_\theta J_{\mathrm{ML}} = \mathbb{E}_{x\sim d_0}\left[\frac{1}{p_\theta(x)}\nabla_\theta p_\theta(x)\right]. \tag{4}$$

In summary, these algorithms can be viewed as assigning different pointwise weights to prompts in the pass-rate function space, providing the technical foundation for the utility-functional framework developed in the remainder of the paper. We defer the detailed proof and discussions to Appendix B.

3 Prompt Reweighting as Utility-Dependent Context Distribution Control

3.1 Prompt Reweighting as Context Distribution Control

Problem Setting: Policy-Reweighted Contextual Bandit (PRCB). Introducing prompt reweighting into the RLVR policy-gradient update makes the *effective context distribution* policy-dependent in the learning dynamics. Given a generic non-negative weight $w_\theta(x)$, define $d_\theta(x) = d_0(x)w_\theta(x)/Z_\theta$, where $Z_\theta = \int d_0(x)w_\theta(x)dx$ is the normalization factor. Under this framework, the general gradient update can be written as

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim d_0}\left[w_\theta(x)\nabla_\theta p_\theta(x)\right] = Z_\theta \mathbb{E}_{x\sim d_\theta}\left[\nabla_\theta p_\theta(x)\right], \tag{5}$$

where the positive scalar Z_θ only changes the update scale rather than the gradient update direction at a fixed policy iteration. Thus, prompt reweighting can be interpreted as optimizing under an effective prompt measure d_θ rather than the fixed base measure d_0. A special case is the pointwise objective $J_g(\theta)$ with its gradient form:

$$J_g(\theta) = \mathbb{E}_{x\sim d_0}\left[g\left(p_\theta(x)\right)\right], \quad \nabla_\theta J_g(\theta) = \mathbb{E}_{x\sim d_0}\left[g'\left(p_\theta(x)\right)\nabla_\theta p_\theta(x)\right], \tag{6}$$

which induces the pointwise weight $w_\theta(x) = g'(p_\theta(x))$ for a deterministic function g, and includes REINFORCE, GRPO, and MaxRL as examples from Section 2. We call this setting a *Policy-Reweighted Contextual Bandit (PRCB)*: the contextual-bandit interaction remains one-step and outcome-based, but the optimization measure over contexts becomes policy-dependent through reweighting. Unlike MDP and non-Markovian environments that are treated as the temporal structure extension of classical contextual bandit, PRCB highlights a distinctive axis of information acquisition in RLVR in the direction of measure adaptivity via direct context distribution control. Notably, a direct context distribution control is typically unavailable in standard RL with an external environment. We provide a detailed discussion and list new challenges in the PRCB setting in Appendix F.1.

3.2 Optimality of Prompt Reweighting is Utility-Dependent

Policy Optimization in RLVR as Pass-rate Distribution Transport. Let $\mathcal{X}$ be the prompt space and π^* be an oracle policy. Define an oracle pass-rate function $p^*: \mathcal{X}\to[0,1]$ by $\mathbb{E}_{y\sim\pi^*(\cdot|x)}\left[r(x,y)\right] := p^*(x)$. Let f_θ and f^* be the corresponding density functions of the random variables $p_\theta(x)$ and $p^*(x)$ under $x\sim d_0$. The learning dynamics when performing policy optimization in RLVR can be interpreted as a distribution transport problem from f_θ to f^* in the pass-rate space:

$$f_\theta(t) \to f^*(t), \quad t\in[0,1]. \tag{7}$$

Importantly, perturbations to $p_\theta(x)$ at individual prompts induce heterogeneous effects on the resulting distribution transport, which are governed not only by local sensitivity about pass rates, but also by cross-prompt interactions, coupling, and interference in the induced distributional dynamics [40, 7].

Optimal Weight Function as Marginal Contribution of a Utility Functional. Assume $p_\theta \in L^2(d_0)$, the space of square-integrable functions under the prompt distribution d_0. We define a policy-dependent functional $\mathcal{U}_\theta : p_\theta(\cdot) \to \mathbb{R}$ on the pass-rate function space. Note that $\mathcal{U}_\theta$ integrates a broad transformation class of $p_\theta(\cdot)$ over $x \sim d_0$. Define a pushforward measure $\mu_\theta = (p_\theta)_{\#} d_0$ (equivalently, $\mu_\theta = \mathrm{Law}_{x\sim d_0}(p_\theta(x))$), i.e., the distribution of $p_\theta(x)$ by mapping prompts x to the pass rate function $p_\theta(\cdot)$, with the property that $\mu_\theta([0,1]) = \mathbb{P}_{x\sim d_0}(p_\theta(x) \in [0,1]) = 1$. In measure theory, μ_θ induces the Cumulative Distribution Function (CDF) F_θ defined by $F_\theta(t) = \mu_\theta([0,t]) = \mathbb{P}_{x\sim d_0}(p_\theta(x) \le t)$. In Definition 1, we define the optimal weight function $w_\theta^\star(x)$ as the functional derivative [55, 28] of the utility functional $\mathcal{U}_\theta$ over the pass rate function $p_\theta(\cdot)$ evaluated at the prompt x, also interpreted as the marginal contribution of $\mathcal{U}_\theta$ by lifting $p_\theta(x)$.

Definition 1 (Utility-dependent Optimal Prompt Weight $w_\theta^\star$). *For any perturbation $h \in L^2(d_0)$, according to the Riesz representation, the first variation of $\mathcal{U}_\theta$ is given by*

$$\lim_{\epsilon\to 0} \frac{\mathcal{U}_\theta(p_\theta + \epsilon h) - \mathcal{U}_\theta(p_\theta)}{\epsilon} = \int w_\theta^\star(x)\, h(x)\, d_0(x) dx, \tag{8}$$

where $w_\theta^\star(x)$ is the functional derivative denoted by $w_\theta^\star(x) := \frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x)$, indicating the optimal weight.

As the Gateaux derivative defines a linear functional in the perturbation direction h (cf. Eq. (8)), the Riesz representation theorem [14] ensures that, under mild regularity conditions, this functional admits a representation in the Hilbert space $L^2(d_0)$. Particularly, there exists $w_\theta^\star \in L^2(d_0)$ such that the first variation of the objective can be expressed as an inner product with respect to d_0, which naturally defines the optimal weight function $w_\theta^\star(x)$. Our definition is reminiscent of the notion of influence functions widely studied in statistics [39, 75] and recently revisited in modern machine learning [47, 36, 65], where one quantifies the impact of infinitesimal perturbations on a global objective. More broadly, the functional derivative generalizes the ordinary partial derivative from finite-dimensional variables to infinite-dimensional function spaces [55, 28]. Partial derivatives with respect to $p_\theta(x)$ only capture the local sensitivity of each x, but *functional derivatives can characterize how the objective responds to perturbations on the entire pass rate function $p_\theta(\cdot)$*. The functional derivatives-based definition generalizes the partial derivatives, enabling distribution-aware weighting that reflects the data geometry, akin to the geometry-aware formulations in optimal transport [92, 72]. This perspective is also the theoretical foundation of our approach in Section 4.

Example 1: Pointwise Utility Functional. For the pointwise utility function with $\mathcal{U}_\theta = J_g(\theta)$, the prompt weight only depends on $p_\theta(x)$ itself by implicitly assuming the independence among prompts. Therefore, the functional derivative degenerates to the partial derivative: $\frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) = g'(p_\theta(x)) = w_\theta^\star(x)$. This reduction encompasses the aforementioned RLVR algorithms in a straightforward way, including (1) $w_\theta^\star(x) = 1$ in REINFORCE, (2) $w_\theta^\star(x) = 1/\sqrt{(p_\theta(x)(1-p_\theta(x)))}$ in GRPO, and (3)$w_\theta^\star(x) = 1/p_\theta(x)$ in MaxRL, by taking the functional (partial) derivative (see Appendix C.1).

Example 2: Variance-Based Utility Functional. Functional derivatives over the pass-rate function are strictly more expressive than pointwise derivatives, i.e., in general $\frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) \neq \frac{d\mathcal{U}_\theta(p_\theta(x))}{dp_\theta(x)}$. Unlike pointwise utility functionals, functional derivatives can capture global distributional structure of $p_\theta(\cdot)$. As an illustrative example, consider the variance-based utility functional $\mathcal{U}_\theta = \mathrm{Var}_{x\sim d_0}(p_\theta(x)) = \mathbb{E}_{x\sim d_0}\left[p_\theta(x)^2\right] - (\mathbb{E}_{x\sim d_0}[p_\theta(x)])^2$. We can derive $\frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) = 2p_\theta(x) - 2\mathbb{E}_{x\sim d_0}[p_\theta(x)]$, where the second term introduces *a global coupling through the expectation*, reflecting the dependence on the entire pass-rate distribution rather than the local value $p_\theta(x)$ alone. This serves as a motivating example of the general distribution-aware utility functionals introduced in Section 4.

3.3 The Optimality of Utility Functionals is Risk-Dependent

Optimal Choice of the Utility Functional: A Risk-Sensitive Control Perspective. According to Definition 1, the optimal prompt reweighting is determined by the choice of a utility function $\mathcal{U}_\theta$ that influences the performance of the RLVR algorithm. A natural question is whether there exists a single "optimal" utility function $\mathcal{U}_\theta$ that uniformly dominates the alternatives. However, we argue that the answer is negative, as it depends on the choice of risk preference. The RL literature has long recognized that utility selection reflects the designer's risk attitude rather than a universal principle. Specifically, risk-sensitive RL [64] with an objective of some risk measure on cumulative rewards, has not converged on a single optimal risk measure. A typical strategy is to focus on concrete criteria

such as exponential utility and CVaR, treating selection as a tradeoff among interpretability, safety semantics, and tractability [8, 96, 81, 60]. Distributional RL [23, 9] makes this dependence explicit by allowing arbitrary distortion functions. From this perspective, REINFORCE, GRPO, and MaxRL are different points in a continuum of risk preference. A useful summary of this continuum is the widely adopted entropic risk family [42], where the agent seeks to optimize the expectation of an exponential utility function $\mathcal{U}_\theta^{\mathrm{risk}}(\eta)$ with the lower bound J_{RL} by Jensen's inequality:

$$\mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \mathbb{E}_{x\sim d_0}\left[\frac{1}{\eta}\log \mathbb{E}_{y\sim\pi_\theta(\cdot|x)} e^{\eta r(x,y)}\right] \geq \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right]\right] = J_{\mathrm{RL}}(\theta), \quad (9)$$

where $\eta > 0$ controls the degree of risk-seeking: larger values of η place exponentially more weight on high-reward responses. As shown in Proposition 1, this entropic-risk family provides a utility-functional interpolation between standard RL (i.e., REINFORCE) and MaxRL. Increasing η induces stronger reweighting toward low-pass-rate prompts through the resulting weight function $w_\theta^\star$. The proof of Proposition 1 is provided in Appendix C.2. In summary, we argue that no single utility functional $\mathcal{U}_\theta$ is inherently preferred outside of a specific design objective.

Proposition 1 (Gradient of Entropic Risk RL Interpolates Standard RL and MaxRL). *Assume $p_\theta(x) > 0$ for d_0-almost every x and $\frac{\|\nabla_\theta p_\theta(x)\|}{p_\theta(x)} \in L^1(d_0)$, then*

$$\lim_{\eta\to 0^+} \nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \nabla_\theta J_{\mathrm{RL}}(\theta), \ \lim_{\eta\to\infty} \eta \nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \nabla_\theta J_{\mathrm{ML}}(\theta). \quad (10)$$

4 CurveRL: Distribution-Aware Reweighting in Pass-Rate Quantile Space

4.1 From Pointwise Utility to Distribution-Aware Utility in the Quantile Space

Motivation: Weight Collapse of Pointwise Utility Functionals. Previous approaches, such as [25, 88, 100], define utility as a pointwise transformation of the pass rate, i.e., $g(p_\theta(x))$, solely based on the absolute magnitude $p_\theta(x)$. However, such strategies suffer from a systematic limitation that we term *weight collapse*: the induced weights fail to adequately differentiate prompts with distinct learning potential. This issue manifests in both early and late training phases. (1) In the early stage, most prompts have $p_\theta(x) \approx 0$, leading to nearly indistinguishable weights, despite substantial heterogeneity: some prompts lie near the learning frontier, while others are intrinsically difficult and less likely to benefit from further training. (2) In the late stage, many prompts have $p_\theta(x) \approx 1$ and thus yield similar weights, although certain prompts still admit meaningful improvement. The root cause is that pointwise utility functions depend solely on the local information in the pass rate space based on the *absolute* value of $p_\theta(x)$, ignoring the *geometry* in the pass rate space (i.e., the *distributional structure* of $p_\theta(x)$), including the rank, density, and spacing information. To address this limitation, we propose a fundamentally different distribution-aware utility function family. This motivation is analogous to the distinction between geometry-aware distances, such as Wasserstein distance, and pointwise divergences like Kullback–Leibler (KL) divergence (see Appendix F.5).

Distribution-Aware Utility Functional. We aim to incorporate the distributional structure of $p_\theta(x)$ into the prompt weighting $w_\theta(x)$. Consider a reference distribution (i.e., probability measure) μ_{ref}, which induces a CDF F_{ref} with a density f_{ref} defined by $F_{\mathrm{ref}}(t) = \mu_{\mathrm{ref}}([0,t])$. We consider a *quantile coordinate transform* via F_{ref} on $p_\theta(x)$ to develop a new utility function:

$$\mathcal{U}_\theta(F_{\mathrm{ref}}) = \mathbb{E}_{x\sim d_0}\left[\psi\left(F_{\mathrm{ref}}\left(p_\theta(x)\right)\right)\right], \quad (11)$$

where ψ is an increasing function, often called *distortion function* rooted in distortion risk measure in economics and risk theory [101, 1, 94, 27, 5]. In principle, F_{ref} should approximate the distribution of $p_\theta(x)$ to capture meaningful distributional geometry. Nonetheless, it should not coincide with the exact policy-induced CDF F_θ of the same random variable $p_\theta(x)$. This is because, by probability integral transform [16], we have $F_\theta(p_\theta(x)) \sim \mathrm{Uniform}(0,1)$, which removes any dependency on θ in expectation for $\mathcal{U}_\theta(F_\theta)$. Consequently, the utility degenerates to a constant: $\mathbb{E}_{x\sim d_0}\left[\psi\left(F_\theta\left(p_\theta(x)\right)\right)\right] = \int_0^1 \psi(z)dz$, yielding a non-informative $\mathcal{U}_\theta$. Proposition 2 further quantifies this effect by relating the deviation between F_{ref} and F_θ to a distribution-aware 1-Wasserstein distance. The proof is provided in Appendix C.3.

Proposition 2. *Denote W_1 as 1-Wasserstein distance. If ψ is L_ψ-Lipschitz and $\|f_\theta\|_\infty < \infty$, then*

$$|\mathcal{U}_\theta(F_{\mathrm{ref}}) - \mathcal{U}_\theta(F_\theta)| \leq L_\psi \|f_\theta\|_\infty W_1(\mu_{\mathrm{ref}}, \mu_\theta). \quad (12)$$

Implementation and Interpretation. In practice, F_{ref} is instantiated using a lagged policy (e.g., $F_{\text{ref}} = F_{\theta_{\text{old}}}$) and updated in a sliding-window manner akin to the target network trick adopted in deep RL algorithms [66, 53]. The gradient is taken with respect to the current pass rate $p_\theta(x)$ by fixing F_{ref} within each gradient update, analogous to a standard pointwise derivative. Theoretically, F_{ref} encodes the global information about the pass rate function $p_\theta(\cdot)$ and is thus policy-dependent. Consequently, although the gradient computation appears pointwise in implementation by fixing F_{ref}, the resulting utility functional $\mathcal{U}_\theta$ in Eq. (11) is inherently distribution-aware and cannot be reduced to a deterministic pointwise transformation $g(p_\theta(\cdot))$. From the perspective of distribution transport illustrated in Section 3.2, the introduced quantile-based representation via F_{ref} aligns with a Wasserstein-type geometry, rather than relying solely on the pointwise information like the KL divergence. We defer a detailed discussion about the induced geometry in Appendix F.5.

Our Method: CurveRL. In our main approach, we adopt $\psi(u) = \log u$ as the distortion risk function. This choice corresponds to a risk-seeking preference that emphasizes hard prompts, a preference well aligned with RLVR's empirical goal of improving worst-case reasoning performance. The log is preferable for designers who care about tail performance, but it is not universally optimal as discussed in Section 3.3. Consequently, a concise and interpretable weight is derived in the gradient form:

$$\nabla_\theta \mathcal{U}_\theta(F_{\text{ref}}) = \mathbb{E}_{x\sim d_0}\left[\frac{f_{\text{ref}}\left(p_\theta(x)\right)}{F_{\text{ref}}\left(p_\theta(x)\right)}\nabla_\theta p_\theta(x)\right], \tag{13}$$

where the weight above is interpreted as emphasizing prompts with the *lower-quantile* pass rates via $1/F_{\text{ref}}(p_\theta(x))$ and *higher density* via $f_{\text{ref}}(p_\theta(x))$. The logarithm transformation as ψ explicitly elicits the geometry of the pass rate function space of $p_\theta(\cdot)$, including the density and rank information of $p_\theta(x)$, to reflect the distributional structure. This weight has the same form of *reverse hazard rate* in probability and survival analysis [13, 31] albeit in a fundamentally different context.

4.2 Algorithm: CurveRL

In Algorithm 1, we elaborate on the update details of CurveRL at each training step t.

Sliding Window Estimate. Denote the input batch $\mathcal{B}$ and number of rollouts N. We initialize a queue W with the size $t_0 \times |\mathcal{B}|$, which stores the pass rates collected from the step $t-t_0$ to $t-1$. At step t, we estimate $F_{\text{ref}}(p)$ and $f_{\text{ref}}(p)$ via a histogram estimator from this lagged window for each $p \in [\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}]$. Therefore, F_{ref} is determined by old policies in the last t_0 steps.

Weight and Gradient Estimation. Next, for each prompt x in the batch $\mathcal{B}$, we perform N response attempts and estimate the pass rate $\hat{p}$. We evaluate $\hat{F}_{\text{ref}}(\hat{p})$ and $\hat{f}_{\text{ref}}(\hat{p})$ in the lagged window. We further use $\hat{p}$ as the group-level baseline, which reduces the variance of the policy gradient estimator. Finally, we maintain W as an active-pass-rate window, appending only prompts with non-vanishing gradients, i.e., $\hat{p} \in (0, 1)$. This active-prompt truncation also contributes to the stability in estimating the weight function. We defer the relevant discussion of this strategy in Appendix F.7. We also remove pass rates collected at step $t - t_0$, ensuring a length-t_0 sliding window estimation for F_{ref} in W. Thus, t_0 is the main hyperparameter in CurveRL.

Algorithm 1 CurveRL Update at Step t

Require: Batch $\mathcal{B}$ of inputs, number of rollouts N, sliding window W with the size $t_0 \times |\mathcal{B}|$

```
# Weight Estimation
for each p ∈ [1/N, 2/N, ..., (N-1)/N] do
    Evaluate density f̂_ref(p) and CDF F̂_ref(p)
end for
# Gradient Estimation
for each input x ∈ B do
    Sample y_1, ..., y_N ~ π_θ(· | x)
    for i = 1 to N do
        r_i ← r(x, y_i), S_i ← ∇_θ log π_θ(y_i | x)
    end for
    Estimate p̂ ← (1/N) Σ_{i=1}^N r_i
    if p̂ ∈ (0, 1) then
        w_t(p̂) ← f̂_ref(p̂) / F̂_ref(p̂)
        ĝ(x) ← (1/N) Σ_{i=1}^N w_t(p̂) (r_i − p̂) S_i
        Append p̂ to W
    end if
end for
Remove pass rates in t − t_0 from W
ĝ ← (1/|B|) Σ_{x∈B} ĝ(x)
return ĝ
```

4.3 Theoretical Interpretation and Advantages

A Unified Explanation of Prompt Reweighting as a Prior F_{ref}. The distribution-aware utility in Eq. (11) provides a unified interpretation of pointwise prompt reweighted algorithms. Specializing

to the log distortion $\psi(u) = \log u$ in Eq. (13), Theorem 1 shows that any pointwise weight $w_\theta(x)$, e.g., those in REINFORCE, GRPO, and MaxRL, can be equivalently represented by a prior distribution F_{ref} in the gradient update. In contrast, CurveRL estimates F_{ref} from the evolving pass-rate distribution and is thus adaptive and data-driven. The proof of Theorem 1 is given in Appendix C.4.

Theorem 1 (Pointwise Weight Induces a Prior F_{ref}). *Denote $p = p_\theta(x)$ and the pointwise weight $w_\theta(x) = g'(p_\theta(x)) := w(p)$ in the pass-rate space. Assume $\int_p^1 w(t)dt < \infty$ for any $p \in (0, 1]$. Under the distortion $\psi(u) = \log u$ in Eq. (13), $w(p) = f_{\text{ref}}(p) / F_{\text{ref}}(p)$ admits a unique F_{ref}:*

$$F_{\text{ref}}(p) = \exp\left(-\int_p^1 w(t)dt\right). \tag{14}$$

In Appendix F.3, we show that F_{ref} is Uniform$(0, 1)$ if and only if the gradients of distribution-aware and pointwise utility functionals coincide under the same distortion measure ψ, i.e., $\nabla_\theta \mathcal{U}_\theta(F_{\text{ref}}) = \nabla_\theta J_\psi(\theta)$. This identifies the uniform reference distribution as the boundary case in which the quantile transform does not change the gradient direction. Away from this boundary, the effect of distribution-aware reweighting can be characterized by a relative multiplier $R_\psi(p) = \psi'(F_{\text{ref}}(p))f_{\text{ref}}(p)/\psi'(p)$. In Appendix F.4, we further derive sufficient conditions based on $R_\psi(p)$: when $R'_\psi(p) < 0$, our distribution-aware method assigns relatively more weight to low-pass-rate prompts as opposed to the pointwise counterpart without F_{ref}.

Advantage 1: Distribution-aware and Data-Driven Reweighting. The classical weight function derived from the pointwise utility functional solely relies on the local information or raw value of $p_\theta(x)$ at each prompt. By contrast, our weight function is derived from a distribution-aware utility functional that characterizes the geometry (e.g., ranking and density in Eq. (13)) in the pass-rate function space based on the functional derivative defined in Definition 1. More importantly, as shown in Theorem 1, pointwise weights can be treated as special prior distribution of $F_{\text{ref}}(p_\theta(x))$, while our weight function of CurveRL in Eq. (13) is *adaptive and data-driven*, which is updated periodically via non-parametric statistical approaches, e.g., a histogram function (see Section 4.2). Our method is more adaptive than the exogenous prompt schedule in curriculum learning (see Appendix F.2).

Advantage 2: Mitigating Weight Collapse by Introducing Quantile Coordinate. As mentioned in Section 4.1, the pointwise utility function suffers from the weight collapse issue when $p_\theta(x) \approx 0$ or 1, leading to nearly indistinguishable learnability in the weights. By mapping the absolute value of pass rates to the CDF/quantile coordinate via F_{ref} in Eq. (11), CurveRL becomes robust to miscalibration of pass rates and enjoys a *monotone calibration invariance property* (see Appendix C.5). Unlike the evolving non-stationary pass-rate space in the learning dynamics, the CDF/quantile space via $F_{\text{ref}}(p_\theta(x))$ in CurveRL induces a more uniform difficulty spectrum and is thus more stationary as F_{ref} is close to F_θ as analyzed in Proposition 2. This potentially leads to more stable optimization.

5 Experiments

Experimental Setups. We train Qwen3-1.7B-Base and Qwen3-4B-Base on POLARIS-53K [2], a corpus of approximately 50K mathematical reasoning prompts, using the verl framework [80]. We compare CurveRL with GRPO [79] and MaxRL [88]. All methods use a shared training loop and differ only in the prompt-weighting rule. We adopt the batch size $|\mathcal{B}| = 256$, the number of rollouts $N = 8$ for all algorithms, and $t_0 = 10$ in Algorithm 1 for CurveRL. Training-time correctness is verified using MATH-VERIFY [48]. More implementation details are provided in Appendix D.

Evaluation. We evaluate trained models on eight challenging mathematical reasoning benchmarks: AIME 2025 [6], BeyondAIME [77], HMMT 02/25 [6], HMMT 02/26 [6], and MATH-500 [41] in the main content, and BRUMO 2025 [6], HMMT 11/25 [6], and Minerva [51] in Appendix E.

5.1 Main Results

Improved Trade-off Frontier. Table 1 reports evaluation results on five reasoning benchmarks, with similar results on the remaining three benchmarks deferred to Appendix E.1. Across both Qwen3-1.7B-Base and Qwen3-4B-Base, CurveRL achieves the highest pass@64 on all benchmarks, indicating stronger performance on hard prompts. Compared with MaxRL, the strongest baseline, it improves average pass@64 by $+\mathbf{5.9}\%$ and $+\mathbf{9.7}\%$, respectively. Importantly, CurveRL also

Table 1: **Main results on five math reasoning benchmarks.** We report pass@1 and pass@64 (%) for each method on Qwen3-1.7B-Base and Qwen3-4B-Base. The best per column is in bold.

Method	AIME 2025		BeyondAIME		HMMT 02/25		HMMT 02/26		MATH-500		Avg.	
	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
Qwen3-1.7B-Base												
Base	3.1	34.6	1.1	16.9	0.2	10.6	1.1	14.1	57.1	92.8	12.5	33.8
GRPO	4.8	29.3	1.1	15.6	0.5	10.0	**3.8**	15.5	70.5	92.6	16.1	32.6
MaxRL	6.7	36.5	1.3	21.8	1.3	18.4	3.2	17.7	**72.1**	94.0	16.9	37.7
CurveRL	**7.7**	**40.9**	**1.4**	**23.6**	**1.7**	**19.3**	3.0	**21.0**	71.6	**94.6**	**17.1**	**39.9**
Qwen3-4B-Base												
Base	6.9	44.3	4.0	23.8	1.0	22.5	3.0	20.6	67.5	95.7	16.5	41.4
GRPO	19.1	48.6	8.0	29.0	5.8	19.0	9.9	27.2	82.8	94.7	25.1	43.7
MaxRL	18.9	52.6	8.2	32.5	7.9	27.8	11.8	33.9	**85.1**	96.6	26.4	48.7
CurveRL	**19.6**	**55.6**	**8.3**	**40.4**	**9.0**	**32.4**	**11.9**	**40.9**	84.5	**97.7**	**26.7**	**53.4**

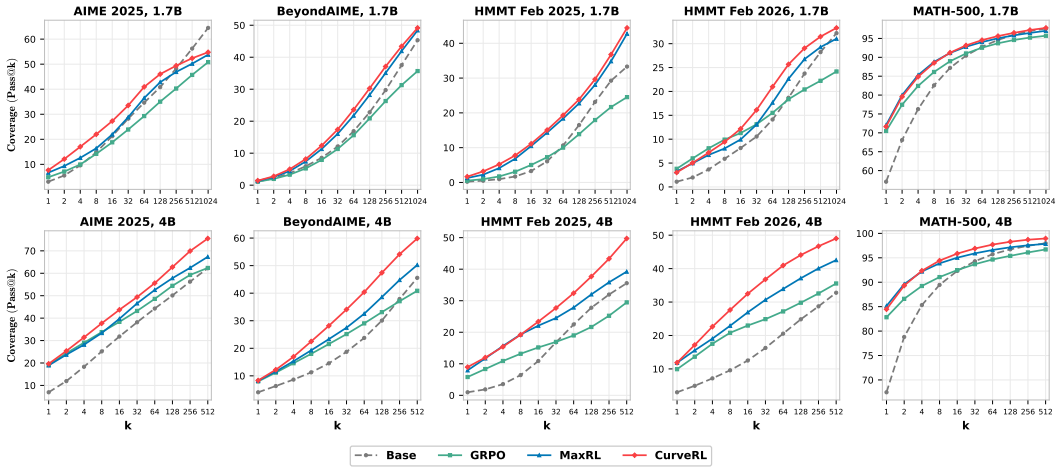


Figure 1: **Pass@k scaling on five representative benchmarks.** Top row: Qwen3-1.7B-Base, $k \in \{1, \ldots, 1024\}$. Bottom row: Qwen3-4B-Base, $k \in \{1, \ldots, 512\}$. CurveRL outperforms GRPO and MaxRL across the full range of k on both model sizes, and exceeds the pretrained base model on most panels.

achieves a higher average pass@1 on both model sizes, indicating that its improvement in Best-of-N performance does not come at the expense of single-shot accuracy. This mitigates the pass@k degradation issue observed in prior work [107, 20]. Overall, our results show that the distribution-aware reweighting of CurveRL improves the Pareto frontier between pass@1 and pass@k over the existing pointwise reweighting methods.

Analysis of Pass@k Scaling. Following prior work [107, 88, 98], we further examine the capability boundary of each algorithm through pass@k scaling over a wide range of k. Figure 1 reports results on five representative benchmarks, with the remaining three benchmarks in Figure 5 of Appendix E.2 following similar trends. CurveRL consistently outperforms the other algorithms across a wide range of pass@k in most settings, indicating a higher upper bound on reasoning capability. Consistent with prior observations [107], GRPO and MaxRL exhibit varying degrees of pass@k degradation relative to the pretrained base model, whereas CurveRL exceeds the pretrained base model in 9 out of 10 panels of Figure 1 across both model sizes. Meanwhile, the advantage of CurveRL over the strongest baseline, MaxRL, becomes larger as k increases on challenging reasoning benchmarks, reaching approximately $+\mathbf{7.3}\%$ on HMMT Feb 2026 at $k = 1024$ on Qwen3-1.7B-Base and $+\mathbf{26.8}\%$ on HMMT Feb 2025 at $k = 512$ on Qwen3-4B-Base. This widening advantage suggests that CurveRL effectively broadens the search space of reasoning trajectories, enabling the policy to discover rare correct solutions. This property may further benefit both reward-free and reward-guided test-time scaling methods, including scalable Best-of-N selection with strong statistical signals [45, 33], lightweight probing [38], or process reward models (PRMs) [52].

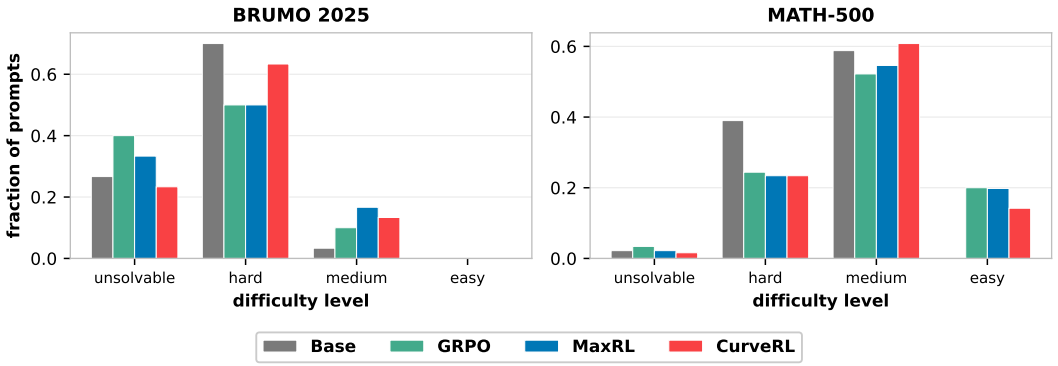


Figure 2: Prompt Distribution across Difficulty in Qwen3-1.7B-Base.

5.2 Mechanism Analysis of CurveRL

Difficulty Distribution Analysis. We investigate how the algorithm in the post-training changes the RLVR model's capability over prompts with distinct difficulty levels. We partition prompts by empirical pass rate $\hat{p}$ over 2048 rollouts into four groups with different prior difficulty levels: *unsolvable* ($\hat{p} = 0$), *hard* ($\hat{p} \in (0, 1/2]$), *medium* ($\hat{p} \in (1/2, 1)$), and *easy* ($\hat{p} = 1$). As illustrated in Figure 2 on two benchmarks BRUMO 2025 and MATH-500, CurveRL consistently reduces the fraction of *unsolvable* prompts, especially on BRUMO 2025. This verifies that CurveRL has a larger potential to expand the real solvability boundary by making more unsolvable problems tractable.

Moreover, CurveRL does not merely convert unsolvable prompts into barely solvable ones. It shifts a larger fraction of prompts into the *medium* regime, where $\hat{p} > 1/2$ and majority voting [95] becomes statistically reliable: repeated sampling can amplify individual correctness without requiring additional reward signals. This explains the stronger reward-free test-time scaling of CurveRL, as reflected by the higher majority@2048 scores on BRUMO 2025 and MATH-500 in Table 2.

Table 2: Majority@2048 on Qwen3-1.7B-Base.

Method	BRUMO 2025	MATH-500
Base	13.3	70.0
GRPO	16.7	74.6
MaxRL	16.7	77.2
CurveRL	**23.3**	**78.8**

Distribution-Aware and Data-Driven Weighting. Figure 3 visualizes how CurveRL differs from pointwise weighting algorithms from the perspective of prompt weights in the pass-rate space. **(1)** The left panel suggests the evolution of the empirical reference density $\hat{f}_{\text{ref}}(\hat{p})$ during training: the pass-rate distribution shifts from a sharp concentration near the lowest pass-rate bins toward higher-pass-rate regions, becoming smoother and more spread. This provides empirical evidence for the distribution-transport view in Eq. (7) of Section 3.2. **(2)** The center panel showcases the resulting weight dynamics $w_t(\hat{p}) = \hat{f}_{\text{ref}}(\hat{p})/\hat{F}_{\text{ref}}(\hat{p})$ in CurveRL, which adapts to the evolving estimates of reference distribution F_{ref} and f_{ref}. The CDF denominator preserves an emphasis on lower-quantile prompts, while the density numerator makes the allocation data-driven by assigning weights to pass-

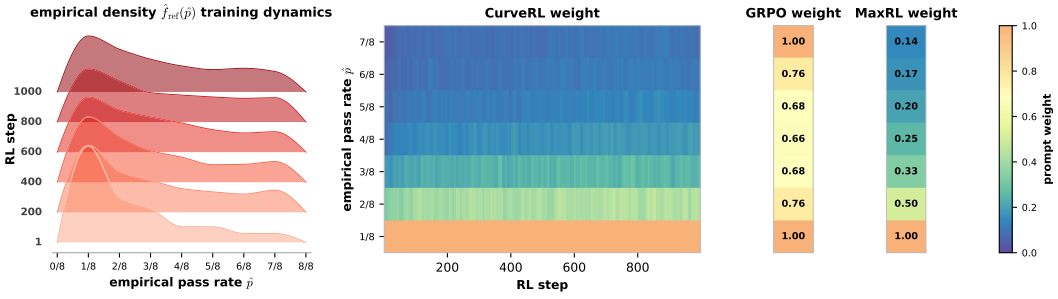


Figure 3: **Distribution-aware and data-driven weighting of CurveRL on Qwen3-4B-Base.** *Left*: dynamics of CurveRL's empirical pass-rate density $\hat{f}_{\text{ref}}(\hat{p})$, smoothed by interpolation. *Center*: dynamics of CurveRL's adaptive prompt weight $w_t(\hat{p}) = \hat{f}_{\text{ref}}(\hat{p})/\hat{F}_{\text{ref}}(\hat{p})$. *Right*: static GRPO weight $1/\sqrt{\hat{p}(1-\hat{p})}$ and MaxRL weight $1/\hat{p}$ rescaled to $[0, 1]$.

rate regions that are populated under the current policy. **(3)** The right panel contrasts this adaptive behavior with the static pointwise weights of GRPO and MaxRL, which can also be interpreted as a fixed prior in Theorem 1. With a symmetric weight $1/\sqrt{\hat{p}(1-\hat{p})}$, GRPO assigns large weights to both low- and high-pass-rate bins, while MaxRL employs $1/\hat{p}$ and monotonically emphasizes prompts with smaller pass rates.

Sensitivity Analysis of Sliding Window Size t_0. We further compare our selected sliding-window size $t_0 = 10$ batches in Algorithm 1 with two extreme choices: a single-batch window ($t_0 = 1$) and a much wider window ($t_0 = 50$). Table 3 suggests that CurveRL with $t_0 = 10$ achieves a better overall trade-off, obtaining the largest pass@64 on all benchmarks while remaining competitive on pass@1 relative to the alternatives. Specifically, a smaller window size can introduce larger variance when estimating the pass-rate distribution, while a larger window size relies more heavily on older policies and thus introduces bias in estimating the current pass-rate distributions. Notably, even with $t_0 = 1$ or 50, CurveRL still matches or exceeds the strongest pointwise baseline, MaxRL, on most benchmarks. Pass@k scaling curves for CurveRL under the two t_0 are provided in Figure 9 in Appendix E.5.

Table 3: **Sensitivity to sliding window size** t_0 regarding the pass@1 and pass@64 (%) performance on Qwen3-1.7B-Base.

t_0	BRUMO 2025		AIME 2025		HMMT 11/25		HMMT 02/26		MATH-500		Avg.	
	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
1	16.9	44.9	7.1	36.1	3.2	18.3	**3.5**	18.1	**72.0**	94.2	20.5	42.3
10 (default)	**18.0**	**45.3**	**7.7**	**40.9**	**3.9**	**20.6**	3.0	**21.0**	71.6	**94.6**	**20.8**	**44.5**
50	15.8	43.3	6.0	37.5	3.6	19.9	2.5	17.5	69.3	94.1	19.4	42.5

Ablation Study and Comparison with DAPO [104], We also perform an ablation study of our CurveRL, including the number of rollouts N, different components in the weight function of CurveRL, different density estimators, and comparison with DAPO in Appendix E.9.

6 Discussion and Conclusion

Our study suggests that the prompt reweighting in RLVR should be understood as principled context distribution control rather than heuristic modifications of weights in policy gradients. Under this view, there is no universally optimal prompt weight independent of the training objective. Instead, the optimal weight is utility-dependent and determined by the marginal value of improving each prompt's pass rate, which is formulated by the functional derivative. By using the quantile coordinate transform, CurveRL instantiates context distribution control with a distribution-aware utility functional in the pass-rate quantile space, allowing the prompt measure to adapt to the model's current competence profile through the distributional structure of pass rates independent of existing pointwise rules.

Limitations and Future Work. A natural extension of our method is to integrate pointwise and distribution-aware utility functionals into more expressive prompt weighting schemes. We explored two such integrated strategies in Appendix F.6, but did not observe empirical improvements. This suggests that the two utility functional classes may induce different geometries in the pass-rate space, and simply adding or multiplying them may not reconcile their optimization effects. A rigorous characterization of these geometries and their impact on learning dynamics remains an important direction for future work. Moreover, defining weight optimality directly in the prompt space, rather than in the pass-rate space, additionally captures the effect of parameter updates, although computing functional derivatives over high-dimensional prompt representations is generally intractable.

Acknowledgments

This work was supported in part by NIH grants R01EB036016, R01EB037101, and R01MH143267, NSF grant DMS-2310679, a Meta Faculty Research Award, and Wharton AI for Business. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of banking & finance*, 26(7):1505–1518, 2002.

[2] Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models. `https://hkunlp.github.io/blog/2025/Polaris`, 2025.

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[4] Sreejith Balakrishnan, Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Harold Soh. Efficient exploration of reward functions in inverse reinforcement learning via bayesian optimization. *Advances in Neural Information Processing Systems*, 33:4187–4198, 2020.

[5] Alejandro Balbás, José Garrido, and Silvia Mayoral. Properties of distortion risk measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009.

[6] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.

[7] Anas Barakat, Souradip Chakraborty, Khushbu Pahwa, and Amrit Singh Bedi. Why pass@ k optimization can degrade pass@ 1: Prompt interference in llm post-training. *arXiv preprint arXiv:2602.21189*, 2026.

[8] Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: an overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024.

[9] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[11] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[12] Tom Blau, Edwin V Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *International conference on machine learning*, pages 2107–2128. PMLR, 2022.

[13] Henry W Block, Thomas H Savits, and Harshinder Singh. The reversed hazard rate function. *Probability in the Engineering and informational Sciences*, 12(1):69–90, 1998.

[14] Haim Brezis and Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

[15] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[16] George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

[17] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[18] Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025.

[19] Xinzhu Chen, Xuesheng Li, Zhongxiang Sun, and Weijie Yu. Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning llms. *arXiv preprint arXiv:2512.00908*, 2025.

[20] Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

[21] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *International Conference on Learning Representations*, 2026.

[22] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

[23] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[24] Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, et al. CDE: Curiosity-driven exploration for efficient reinforcement learning in large language models. In *International Conference on Learning Representations*, 2026.

[25] Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv preprint arXiv:2510.13651*, 2025.

[26] Aniket Anand Deshmukh, Srinagesh Sharma, James W Cutler, Mark Moldwin, and Clayton Scott. Simple regret minimization for contextual bandits. *arXiv preprint arXiv:1810.07371*, 2018.

[27] Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2(2):319–328, 2012.

[28] Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

[29] Lawrence C Evans. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2025.

[30] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

[31] Maxim S Finkelstein. On the reversed hazard rate. *Reliability Engineering & System Safety*, 78(1):71–75, 2002.

[32] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[33] Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *The Fourteenth International Conference on Learning Representations*, 2026.

[34] Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why grpo needs normalization: A local-curvature perspective on adaptive gradients. *arXiv preprint arXiv:2601.23135*, 2026.

[35] Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

[36] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

[37] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[38] Jizhou Guo, Zhaomin Wu, Hanchen Yang, and Philip S Yu. Mining intrinsic rewards from llm hidden states for efficient best-of-n sampling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 371–382, 2026.

[39] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[40] Hangfeng He and Weijie J Su. The local elasticity of neural networks. *International Conference on Learning Representations*, 2019.

[41] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[42] Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[43] Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the learning dynamics of rlvr at the edge of competence. *arXiv preprint arXiv:2602.14872*, 2026.

[44] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[45] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *Advances in neural information processing systems*, 38:19720–19745, 2025.

[46] Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*, 2018.

[47] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[48] Hynek Kydlíček. Math-Verify. `https://github.com/huggingface/Math-Verify`, 2025. version 0.8.0.

[49] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

[50] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[51] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, and Theo Gutman-Solo. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[52] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2023.

[53] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[54] Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3):293–321, 1992.

[55] Jacques Louis Lions. *Optimal control of systems governed by partial differential equations*, volume 170. Springer, 1971.

[56] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022.

[57] Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

[58] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025.

[59] A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic temporal-difference learning. *arXiv preprint arXiv:1507.01569*, 2015.

[60] Anirudha Majumdar, Sumeet Singh, and Marco Pavone. Risk-sensitive inverse reinforcement learning via coherent risk models. In *Robotics: Science and Systems (RSS)*, volume 16, 2017.

[61] Yixiu Mao, Yun Qu, Qi Wang, Heming Zou, and Xiangyang Ji. Dynamics-predictive sampling for active rl finetuning of large reasoning models. In *International Conference on Learning Representations*, 2026.

[62] Viraj Mehta, Biswajit Paria, Jeff Schneider, Stefano Ermon, and Willie Neiswanger. An experimental design perspective on model-based reinforcement learning. In *International conference on learning representations*, 2022.

[63] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

[64] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2-3):267–290, 2002.

[65] Taywon Min, Haeone Lee, Yongchan Kwon, and Kimin Lee. Understanding impact of human feedback via influence functions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27471–27500, 2025.

[66] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[67] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

[68] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

[69] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[70] Yangchen Pan, Jincheng Mei, Amir-massoud Farahmand, Martha White, Hengshuai Yao, Mohsen Rohani, and Jun Luo. Understanding and mitigating the limitations of prioritized experience replay. In *Uncertainty in Artificial Intelligence*, pages 1561–1571. PMLR, 2022.

[71] Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

[72] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[73] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[74] Nived Rajaraman, Audrey Huang, Miro Dudik, Robert Schapire, Dylan J Foster, and Akshay Krishnamurthy. Learning to reason with curriculum i: Provable benefits of autocurriculum. *arXiv preprint arXiv:2603.18325*, 2026.

[75] Elvezio M Ronchetti and Peter J Huber. *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA, 2009.

[76] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[77] ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

[78] Jun Shao. *Mathematical statistics*. Springer, 1999.

[79] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[80] 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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[81] Kevin M Smith and Margaret P Chapman. On exponential utility and conditional value-at-risk as risk-averse performance criteria. *IEEE Transactions on Control Systems Technology*, 31(6):2555–2570, 2023.

[82] Weijie J Su. Envisioning future deep learning theories: some basic concepts and characteristics. *Science China Information Sciences*, 67(10):203101, 2024.

[83] Joe Suk and Yaqi Duan. On the optimization dynamics of rlvr: Gradient gap and step size thresholds, 2026.

[84] Ke Sun, Yingnan Zhao, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning with regularized wasserstein loss. *Advances in Neural Information Processing Systems*, 37:63184–63221, 2024.

[85] Ke Sun, Yingnan Zhao, Enze Shi, Yafei Wang, Xiaodong Yan, Bei Jiang, and Linglong Kong. Intrinsic benefits of categorical distributional loss: Uncertainty-aware regularized exploration in reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[86] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[87] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[88] Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. Maximum likelihood reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026.

[89] Yao Tang, Li Dong, Yaru Hao, Qingxiu Dong, Furu Wei, and Jiatao Gu. Multiplex thinking: Reasoning via token-wise branch-and-merge. *arXiv preprint arXiv:2601.08808*, 2026.

[90] Sebastian B Thrun. *Efficient exploration in reinforcement learning*. Carnegie Mellon University, 1992.

[91] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019.

[92] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[93] Christian Walder and Deep Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in neural information processing systems*, 2025.

[94] Shaun Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1):71–92, 1996.

[95] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[96] Yuheng Wang and Margaret P Chapman. Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control. *Artificial Intelligence*, 311:103743, 2022.

[97] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[98] Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

[99] Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation: Stabilizing RLVR for LLM reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[100] Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework under non-linear rl objectives. *arXiv preprint arXiv:2510.04996*, 2025.

[101] Menahem E Yaari. The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, pages 95–115, 1987.

[102] 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.

[103] Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, et al. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

[104] 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.

[105] 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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

[106] Yurun Yuan, Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Trajectory bellman residual minimization: A simple value-based method for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[107] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, et al. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[108] Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025.

[109] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[110] Kaichen Zhang, Yuzhong Hong, Junwei Bao, Hongfei Jiang, Yang Song, Dingqian Hong, and Hui Xiong. Gvpo: Group variance policy optimization for large language model post-training. *Advances in neural information processing systems*, 2025.

[111] Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, An Jian, Kejiang Chen, and Xing Hu. Revisiting entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning. *arXiv preprint arXiv:2510.10959*, 2025.

[112] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[113] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning*, pages 11492–11502. PMLR, 2020.

[114] Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a u-statistic. *arXiv preprint arXiv:2603.01162*, 2026.

Appendix

Table of Contents

A Related Work **19**

B Details of Technical Background in Section 2 **20**

C Theoretical Results **21**
C.1 Functional derivative under Pointwise Utility Reduces to Partial Derivative . . . 21
C.2 Proof of Proposition 1 . . . 22
C.3 Proof of Proposition 2 . . . 23
C.4 Proof of Theorem 1 . . . 24
C.5 Monotone Calibration Invariance Property . . . 25

D Implementation Details **27**

E Supplemental Experimental Results **28**
E.1 Pass@1 and Pass@64 on Three Additional Benchmarks . . . 28
E.2 Pass@k Curves on the Additional Benchmarks . . . 29
E.3 Additional Difficulty Distribution . . . 29
E.4 Learning Signals Across Training Dynamics . . . 30
E.5 Supplemental Results on Sensitivity Analysis . . . 31
E.6 Stable Training Dynamics . . . 31
E.7 Validation Accuracy During Training . . . 32
E.8 Distribution-Aware Weighting on Qwen3-1.7B-Base . . . 32
E.9 Ablation Study and More Baselines . . . 32

F More Discussions **34**
F.1 Discussion about Policy-Reweighted Contextual Bandit . . . 34
F.2 Curriculum Learning under Our Framework . . . 34
F.3 When Gradients of Distribution-aware and Pointwise Utility Functionals Differ . . . 35
F.4 Sufficient Conditions for Prompt Weight Comparison and Risk Preference . . . 36
F.5 Induced Wasserstein-type Geometry in Our Method . . . 38
F.6 Integrating both Pointwise and Distribution-aware Utility Functionals . . . 38
F.7 An Active-prompt Truncation . . . 39

A Related Work

Action-level Exploration in Standard RL vs Context-level Distribution Control in RLVR. In standard RL, the state-visitation distribution is often reshaped indirectly through: (i) importance weighting in off-policy correction [86, 59], (ii) experiment-design-style transition selection [62, 12], (iii) goal-conditioning [56, 30], (iv) replay variants such as prioritized experience replay [54, 76, 3, 46, 70], all of which act on *previously collected* data. RLVR with prompt reweighting [79, 37, 104, 88, 100, 71] instead manipulates the *online* prompt distribution directly. Action-level exploration becomes more challenging in this setting, as the trajectory-level action space is combinatorially large and the contextual-bandit structure provides no state-visitation dynamics to leverage [22, 107, 24]. Therefore, the direct context distribution control by prompt reweighting supplies a new axis of information acquisition, which is exactly formulated by the PRCB framework we develop in Section 3.1.

Existing Strategies of Context-level Distribution Control in RLVR. There are mainly three classes of methods that exploit the freedom to shape the prompt distribution in RLVR. Our work casts all three mechanisms inside a single *effective context distribution* $d_\theta(x) \propto d_0(x)w_\theta(x)$ and supplies the optimality principle for w_θ that has been missing from this line of work.

- *Sample selection and dynamic filtering*: [104, 61] discard prompts whose empirical pass rate equals 0 or 1, serving as a heuristic yet effective way to select the most learnable prompts. [100] focuses on adaptively adjusting the number of samples for each prompt and prioritizes the harder prompts based on the prompt reweighting framework with non-linear RL objectives. We instead view reweighting as a direct context distribution control mechanism to improve the policy improvement by shaping learning dynamics, and establish the reweighting optimality from the perspective of utility functionals, which are risk-dependent.
- *Curriculum strategies*: Curriculum learning in RLVR, rooted in classical curriculum learning [10, 68], schedules prompts from easy to hard [71, 74, 18]. As analyzed in Appendix F.2, curriculum learning relies on an exogenous data schedule to help the policy optimization to chase the moving learnable window, while our method transforms the learnable window into a stationary and nearly fixed one by quantile reparameterization, thus enhancing the learning dynamics.
- *Prompt reweighting*: [25] first formalizes the prompt weighting objective for LLM reasoning. This motivates MaxRL [88] by introducing the maximum likelihood principle into the RL objective and the adaptive sampling framework of [100] through the lens of adaptive sample selection. This prompt reweighting framework is general to accommodate a flurry of RLVR algorithms, including the pass@K optimization in [93]. From the perspective of the model's edge of competence, [108, 43] also show that selecting prompts or manipulating the training distribution is crucial for effective learning. Our study focuses on this line of work in context distribution control of RLVR, as it allows more principled analysis than sample selection and curriculum learning, which can also be considered as special prompt reweighting in general.

Theoretical understanding of GRPO and Its Variants. A growing GRPO family modifies the algorithm primarily through the denominator and group-level normalization: Dr.GRPO [58] removes the standard-deviation factor to debias the gradient; GVPO [110], GPG [21], and follow-ups [34, 103, 19] explore alternative group-level rescalings. These modifications are typically motivated by variance reduction or empirical fixes, leaving the form of the denominator unjustified at the population level. A separate strand analyzes GRPO and REINFORCE through the lens of U-statistics [114] and optimization dynamics [83, 67, 106, 43]. In particular, [43] identifies a *replay effect* when the difficulty spectrum of $p_\theta(x)$ is sufficiently smooth. MaxRL [88] imports the maximum-likelihood principle into the policy gradient, which seemingly entails the optimality of maximum likelihood in statistics. However, we argue this is essentially a coincidence: maximum likelihood enjoys optimality under certain regularity conditions in statistical estimation, whereas policy optimization is mode-seeking instead of statistical estimation, where the optimality of $\log p_\theta$ does not transfer in a straightforward way. Our utility-functional view recovers MaxRL as one specific pointwise utility with a risk-seeking preference, and broadly provides a principled framework to explain the optimal prompt reweighting based on the choice of utility functional.

Distributional Structure in RLVR. The distributional structure between prompts and their rewards is crucial in algorithm analysis and design. [7, 93] reveals the conflict between improving pass@K and pass@1 induced by the prompt interference, which is also supported by the fundamental property

of deep nets called *local plasticity* [40, 82]. This trade-off also mirrors the well-known tension between adversarial robustness and clean accuracy [91, 109]: pushing harder prompts (akin to adversarial examples) trades off accuracy on easier ones. Therefore, it is beneficial to include the correlations between prompts in the algorithm design instead of treating them independently. Our distribution-aware method emerges across the prompt distribution in the pass-rate space by leveraging the rank/quantile and density information of $p_\theta(x)$, offering an explicit knob for navigating this trade-off between hard and easy prompts. [99] proposes quantile advantage estimation by replacing the mean with a group-wise K-quantile baseline to address entropy collapse and entropy explosion, which provides further evidence to demonstrate that leveraging the distributional structure of prompts can be beneficial.

Optimality of Utility Functional and Risk-sensitive Control. Risk-sensitive RL [42, 64, 60, 96, 81, 8] replaces the expectation of cumulative rewards with other utility functionals or risk measures, reflecting the principle that there is no universal optimality in this choice, which mainly depends on the designer's risk attitude. Distributional RL [23, 9, 84, 85] makes this dependence explicit by allowing arbitrary distortion functions. Our distribution-aware utility instead consumes the distributional structure of the pass rates, not its absolute magnitude. The CDF/quantile coordinate transform with the associated distortion function connects directly to dual utility and spectral risk measures [101, 94, 1, 5], which provides the risk-control rationale behind the quantile coordinate transform of Section 4.

B Details of Technical Background in Section 2

We provide a more detailed description of the technical background introduced in Section 2 for better clarity. The derivations below heavily depend on the insights from [25].

RL/REINFORCE Objective. In RLVR for LLM reasoning, the RL/REINFORCE objective can be rewritten as

$$J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim d_0, y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right] = \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\mathbb{I}\{y\in C(x)\}\right]\right] := \mathbb{E}_{x\sim d_0}\left[p_\theta(x)\right],$$

where $C(x)$ is a feasible set in the response, which is often determined by the domain-specific verifier. The score function trick is pivotal in REINFORCE [97] that offers a generic optimization tool to solve the optimization problem under decision-dependent sampling distributions for LLM-based generative model π_θ. As such, we have the following result:

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0}\left[\nabla_\theta p_\theta(x)\right] = \nabla_\theta J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim d_0}\left[\int_y r(x,y)\nabla\pi_\theta(y|x)dy\right] \\
&= \mathbb{E}_{x\sim d_0}\left[\int_y \pi_\theta(y|x) r(x,y)\nabla\log\pi_\theta(y|x)dy\right] = \mathbb{E}_{x\sim d_0, y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\nabla\log\pi_\theta(y|x)\right].
\end{aligned}$$

Group Relative Policy Optimization (GRPO). We next rewrite the gradient of GRPO in the same prompt-reweighting form. For each prompt x, GRPO performs n rollouts to generate responses $\{y_i\}_{i=1}^n$. We denote a reference policy π_{ref} and an old policy π_{old}. Under a contextual bandit or one-step decision abstraction analysis adopted in [25], GRPO can be written in a sequence-level form, which is closely related to Group Sequence Policy Optimization (GSPO) [112]. As such, GRPO maximizes the following objective:

$$\begin{aligned}
&J_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim d_0, \{y_i\}_{i=1}^n\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \\
&\left[\frac{1}{n}\sum_{i=1}^n\left(\min\left(\frac{\pi_\theta\left(y_i|x\right)}{\pi_{\theta_{\text{old}}}\left(y_i|x\right)}\hat{A}_i^x, \text{clip}\left(\frac{\pi_\theta\left(y_i|x\right)}{\pi_{\theta_{\text{old}}}\left(y_i|x\right)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_i^x\right) - \beta D_{KL}\left(\pi_\theta\|\pi_{\text{ref}}\right)\right)\right],
\end{aligned}$$

where ϵ is the clipping parameter and $\hat{A}_i^x$ is the advantage estimate. The clipped policy ratio $\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}$ prevents π_θ from deviating dramatically from the previous policy π_{old} and the regularization hyperparameter β penalizes the divergence from the reference policy π_{ref}. Notably, the group-based normalization scheme is employed to estimate the advantage by $\hat{A}_i^x$ in the gradient update rule:

$$\hat{A}_i^x = \frac{r\left(x,y_i\right) - \text{mean}\left(\left\{r\left(x,y_j\right)\right\}_{j=1}^n\right)}{\text{std}\left(\left\{r\left(x,y_j\right)\right\}_{j=1}^n\right)}.$$

Following [25], when we replace the group-wise empirical baseline and normalization by their population counterparts (i.e., an infinite group size) and ignore the clipping and the policy-ratio term, the gradient of GRPO can be approximated by the following simple form:

$$
\begin{aligned}
&\nabla_\theta J_{\mathrm{GRPO}}(\theta) \\
&\approx \mathbb{E}_{x\sim d_0, y\sim\pi_\theta(\cdot|x)}\left[\frac{r(x,y)-\mathbb{E}_{\tilde{y}\sim\pi_\theta(\cdot|x)}\left[r(x,\tilde{y})\right]}{\sqrt{\mathrm{Var}_{\tilde{y}\sim\pi_\theta(\cdot|x)}\left[r(x,\tilde{y})\right]}}\nabla_\theta \log \pi_\theta(y|x)\right] \\
&= \mathbb{E}_{x\sim d_0}\left[\frac{1}{\sqrt{p_\theta(x)(1-p_\theta(x))}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\left(r(x,y)-\mathbb{E}_{\tilde{y}\sim\pi_\theta(\cdot|x)}\left[r(x,\tilde{y})\right]\right)\nabla_\theta \log \pi_\theta(y|x)\right]\right] \\
&= \mathbb{E}_{x\sim d_0}\left[\frac{1}{\sqrt{p_\theta(x)(1-p_\theta(x))}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\nabla_\theta \log \pi_\theta(y|x)\right]\right] \\
&= \mathbb{E}_{x\sim d_0}\left[\frac{1}{\sqrt{p_\theta(x)(1-p_\theta(x))}}\nabla_\theta p_\theta(x)\right],
\end{aligned}
$$

which ideally corresponds to a population-level objective $J_{\mathrm{GRPO}} = \mathbb{E}_{x\sim d_0}\left[2\arcsin\sqrt{p_\theta(x)}\right]$ [25].

Advanced Prompt Reweighting RL Algorithms. We define pass@k as the probability of at least one correct example:

$$\mathrm{pass}\,@k(x) := \mathbb{P}\left(\exists i \in [k] \text{ s.t. } y_i \in C(x)\right).$$

An interesting property of MaxRL is that maximum likelihood optimizes an infinite harmonic mixture of pass@k gradients:

$$\nabla_\theta J_{\mathrm{ML}}(x) = \sum_{k=1}^{\infty}\frac{1}{k}\nabla_\theta\,\mathrm{pass}\,@k(x).$$

In other words, RL/REINFORCE can be viewed as optimizing a first-order approximation to the maximum-likelihood RL. In statistical estimation, maximum likelihood often enjoys estimation optimality. However, the learning in RLVR is inherently policy optimization rather than statistical estimation, and thus maximum likelihood does not generally entail optimality guarantees. Nevertheless, MaxRL achieves strong empirical performance and often surpasses GRPO across a range of benchmarks. One possible explanation is that the prompt weight $1/p_\theta(x)$ amplifies the learning signal from hard prompts, thereby accelerating training, particularly in the early stages.

In parallel, [100] studied a prompt reweighting framework that adaptively allocates the number of sampled responses, implicitly prioritizing difficult prompts to mitigate vanishing pass-rate estimates. In contrast, we focus on identifying what constitutes an optimal weighting scheme itself and characterizing its role in shaping the learning dynamics, rather than approaching the problem from the perspective of sampling budget allocation.

C Theoretical Results

C.1 Functional derivative under Pointwise Utility Reduces to Partial Derivative

If the utility function $\mathcal{U}_\theta$ equals to the pointwise utility function, functional derivative degenerates to the partial derivative, i.e.,

$$\mathcal{U}_\theta = \int_{\mathcal{X}} g\left(p_\theta(x)\right) d_0(x)dx \Rightarrow \frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) = g'\left(p_\theta(x)\right) = \frac{\partial g\left(p_\theta(x)\right)}{\partial p_\theta(x)}.$$

Even though it is a classical theoretical result in Calculus of Variations [35] and Partial Differential Equation [28], we provide a short version of the proof as a reference for completeness, as the remaining theoretical results in our paper all rely on this backbone.

Proof. Firstly, after the perturbation, the perturbed utility function is

$$\mathcal{U}_\theta\left(p_\theta + \epsilon h\right) = \int_{\mathcal{X}} g\left(p_\theta(x) + \epsilon h(x)\right) d_0(x)dx.$$

By Taylor expansion at $t = p_\theta(x)$, we have

$$g\left(p_\theta(x) + \epsilon h(x)\right) = g\left(p_\theta(x)\right) + \epsilon \cdot g'\left(p_\theta(x)\right) \cdot h(x) + r_\epsilon(x),$$

where the reminder term satisfies $r_\epsilon(x)/\epsilon \to 0$ pointwise as $\epsilon \to 0$. Therefore, the utility difference is

$$\mathcal{U}_\theta\left(p_\theta + \epsilon h\right) - \mathcal{U}_\theta\left(p_\theta\right) = \epsilon \int_{\mathcal{X}} g'\left(p_\theta(x)\right) h(x) d_0(x) dx + \int_{\mathcal{X}} r_\epsilon(x) d_0(x) dx.$$

Next, by regularity condition, including the dominated convergence, the remainder term vanishes. Then we have

$$\begin{aligned}\frac{\mathcal{U}_\theta\left(p_\theta + \epsilon h\right) - \mathcal{U}_\theta\left(p_\theta\right)}{\epsilon} &= \int_{\mathcal{X}} g'\left(p_\theta(x)\right) h(x) d_0(x) dx + \int_{\mathcal{X}} \frac{r_\epsilon(x)}{\epsilon} d_0(x) dx \\ &\to \int_{\mathcal{X}} g'\left(p_\theta(x)\right) h(x) d_0(x) dx \quad \text{as } \epsilon \to 0.\end{aligned}$$

By Definition 1 and the uniqueness of Riesz representation, we obtain that

$$\int_{\mathcal{X}} \frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) \cdot h(x) d_0(x) dx = \int_{\mathcal{X}} g'\left(p_\theta(x)\right) \cdot h(x) d_0(x) dx, \forall h \in L^2(d_0).$$

Consequently, it results in the conclusion that $\frac{\delta \mathcal{U}_\theta}{\delta p_\theta}(x) = g'\left(p_\theta(x)\right)$ for d_0 almost everywhere for x.

□

C.2 Proof of Proposition 1

Proposition 1 (Entropic Risk RL Interpolates Classical RL and MaxRL.) Assume $p_\theta(x) > 0$ for d_0-almost every x and $\frac{\|\nabla_\theta p_\theta(x)\|}{p_\theta(x)} \in L^1\left(d_0\right)$, then

$$\lim_{\eta \to 0^+} \nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \nabla_\theta J_{\mathrm{RL}}(\theta), \ \lim_{\eta \to \infty} \eta \nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \nabla_\theta J_{\mathrm{ML}}(\theta). \tag{15}$$

Proof. As the reward $r(x, y)$ is a Bernoulli random variable, we have

$$\mathbb{E}_{y \sim \pi_\theta(\cdot|x)} e^{\eta r(x,y)} = \left(1 - p_\theta(x)\right) e^0 + p_\theta(x) e^\eta = 1 - p_\theta(x) + p_\theta(x) e^\eta.$$

Therefore, we have:

$$\mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \mathbb{E}_{x \sim d_0}\left[\frac{1}{\eta} \log\left(1 - p_\theta(x) + p_\theta(x) e^\eta\right)\right] := \mathbb{E}_{x \sim d_0}\left[h_\eta(p_\theta(x))\right],$$

where we denote the weight function as $w_\theta^\eta(x)$ and therefore we have

$$\nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) := \mathbb{E}_{x \sim d_0}\left[w_\theta^\eta(x) \nabla_\theta p_\theta(x)\right] = \mathbb{E}_{x \sim d_0}\left[\frac{e^\eta - 1}{\eta\left[1 + \left(e^\eta - 1\right) p_\theta(x)\right]} \nabla_\theta p_\theta(x)\right].$$

As $\eta > 0$, the weight function $w_\theta^\eta(x)$ above is always positive and thus h_η is an increasing function of $p_\theta(x)$, which leads to a meaningful utility function. Next, we discuss the two limiting behaviors of $\nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta)$.

Case 1: $\eta \to 0^+$**.** By Taylor expansion, $e^\eta = 1 + \eta + o(\eta)$ as $\eta \to 0^+$. Consequently, we have the pointwise limit:

$$\frac{e^\eta - 1}{\eta\left[1 + \left(e^\eta - 1\right) p_\theta(x)\right]} = \frac{\eta + o(\eta)}{\eta\left[1 + \left(\eta + o(\eta)\right) p_\theta(x)\right]} = \frac{1 + o(1)}{1 + \eta p_\theta(x) + o(\eta) p_\theta(x)} \to 1, \ (\eta \to 0^+).$$

Since $w_\eta(p_\theta(x)) \to 1$ and is uniformly bounded, the limit can be exchanged with expectation. This implies that $\lim_{\eta \to 0^+} \nabla_\theta \mathcal{U}_\theta^{\mathrm{risk}}(\eta) = \mathbb{E}_{x \sim d_0}\left[1 \cdot \nabla_\theta p_\theta(x)\right] = \nabla_\theta J_{\mathrm{RL}}(\theta)$.

Case 2: $\eta \to +\infty$**.** Denote $a_\eta \sim b_\eta$ when $\eta \to +\infty$ if $\frac{a_\eta}{b_\eta} \to 1$. Firstly, given each prompt x, we have the the pointwise limit:

$$\eta w_\theta^\eta(x) = \frac{e^\eta - 1}{1 + (e^\eta - 1)\, p_\theta(x)} = \frac{1}{\frac{1}{e^\eta - 1} + p_\theta(x)} \to \frac{1}{p_\theta(x)}. \; (\eta \to +\infty)$$

This implies $w_\theta^\eta(x) \sim \frac{1}{\eta p_\theta(x)}$ as $\eta \to +\infty$. However, the pointwise convergence does not directly ensure the limit of the integral is the integral of limit, i.e., $\lim \mathbb{E} = \mathbb{E} \lim$, which additionally requires some mild conditions (i.e., Dominated convergence theorem [29]). Since $\eta w_\theta^\eta(x) = \frac{1}{\frac{1}{e^\eta - 1} + p_\theta(x)} \leq \frac{1}{p_\theta(x)}$, according to our assumption in Proposition 1, we have

$$\left| \eta w_\theta^\eta \left(p_\theta(x) \right) \nabla_\theta p_\theta(x) \right| \leq \frac{\left| \nabla_\theta p_\theta(x) \right|}{p_\theta(x)} \in L^1 \left(d_0 \right),$$

where the function $g(x) \in L^1 \left(d_0 \right)$ indicates that $\mathbb{E}_{x \sim d_0}[|g(x)|] < \infty$. This inequality implies that a sequence of functions is bounded in absolute value by an integrable function. Therefore, the dominated convergence theorem [29] yields

$$\lim_{\eta \to \infty} \eta \nabla_\theta \mathcal{U}_\theta^{\text{risk}}(\eta) = \lim_{\eta \to \infty} \mathbb{E}_{x \sim d_0} \left[\eta w_\theta^\eta(x) \nabla_\theta p_\theta(x) \right] = \mathbb{E}_{x \sim d_0} \left[\frac{1}{p_\theta(x)} \nabla_\theta p_\theta(x) \right] = \nabla_\theta J_{\text{ML}}(\theta).$$

□

C.3 Proof of Proposition 2

Proposition 2. Denote W_1 as 1-Wasserstein distance. If ψ is L_ψ-Lipschitz and $\|f_\theta\|_\infty < \infty$, then

$$|\mathcal{U}_\theta(F_{\text{ref}}) - \mathcal{U}_\theta(F_\theta)| \leq L_\psi \|f_\theta\|_\infty W_1(\mu_{\text{ref}}, \mu_\theta). \tag{16}$$

Proof. As ψ is L_ψ-Lipschitz, we have $|\psi(a) - \psi(b)| \leq L_\psi \, |a - b|$. $\|f_\theta\|_\infty < \infty$ implies a bounded density function or a bounded the essential supremum of the density, i.e., $f_\theta(t) \leq \|f_\theta\|_\infty$ almost everywhere (a.e.), indicating that the inequality holds except on a set of measure zero. Recall the definition of 1-Wasserstein distance $W_1(\mu, \nu)$ to characterize the difference of two measure μ and ν:

$$W_1(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int \|x - y\|_1 d\gamma(x, y) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)| \, dt.$$

where γ is the coupling or the transport plan, and $\Pi(\mu, \nu)$ is the joint distribution space with the marginal distributions as μ and ν. The RHS above is the equivalent form of 1-Wasserstein distance, which is pivotal for our upper bound. In addition, we also write down the two utility functions:

$$\mathcal{U}_\theta(F_{\text{ref}}) = \mathbb{E}_{x \sim d_0} \left[\psi \left(F_{\text{ref}} \left(p_\theta(x) \right) \right) \right] = \int \psi(F_{\text{ref}}(t)) d\mu_\theta(t)$$

$$\mathcal{U}_\theta(F_\theta) = \mathbb{E}_{x \sim d_0} \left[\psi \left(F_\theta \left(p_\theta(x) \right) \right) \right] = \int \psi(F_\theta(t)) d\mu_\theta(t) = \int_0^1 \psi(z) dz = \text{const.}$$

Putting all together, we can derive

$$\begin{aligned}
|\mathcal{U}_\theta(F_{\text{ref}}) - \mathcal{U}_\theta(F_\theta)| &= \int \left[\psi(F_{\text{ref}}(t)) - \psi(F_\theta(t)) \right] d\mu_\theta(t) \\
&\leq \int \left| \psi(F_{\text{ref}}(t)) - \psi(F_\theta(t)) \right| d\mu_\theta(t) \\
&\overset{(a)}{\leq} L_\psi \int \left| F_{\text{ref}}(t) - F_\theta(t) \right| d\mu_\theta(t) \\
&= L_\psi \int \left| F_{\text{ref}}(t) - F_\theta(t) \right| f_\theta(t) dt \\
&\overset{(b)}{\leq} L_\psi \|f_\theta\|_\infty \int \left| F_{\text{ref}}(t) - F_\theta(t) \right| dt \\
&\overset{(c)}{=} L_\psi \|f_\theta\|_\infty W_1(\mu_{\text{ref}}, \mu_\theta),
\end{aligned}$$

where (a) holds by the L_ψ-Lipschitz condition of ψ, (b) holds by $\|f_\theta\|_\infty < \infty$, and (c) satisfies due to the definition of 1-Wasserstein distance.

Interpretation. The upper bound in Proposition 2 strengthens the connection of our distribution-aware utility and a distribution- or geometry-aware Wasserstein distance. Specifically, this bound shows that the discrepancy between the two utility functionals is controlled by the Wasserstein distance between the reference and current pass-rate distributions. In particular, it provides a formal justification that our method is distribution-aware: the utility depends on the global structure of the pass-rate distribution rather than pointwise values.

□

C.4 Proof of Theorem 1

Theorem 1 (Pointwise Weight Induces a Prior F_{ref}**)** Denote $p = p_\theta(x)$ and let the pointwise weight $w_\theta(x) = g'(p_\theta(x)) := w(p)$ be the per-prompt instantiation in the pass-rate space. Assume $\psi(t) = \log t$ in Eq. (13) and $\int_0^1 w(t)dt < \infty$, $w(p) = f_{\text{ref}}(p)/F_{\text{ref}}(p)$ admits a unique F_{ref}:

$$F_{\text{ref}}(p) = \exp\left(-\int_p^1 w(t)dt\right).$$

Proof. Recall that $w_\theta : (0,1] \to \mathbb{R}_{\geq 0}$ is a non-negative weight function that satisfies

$$\int_p^1 w(t)dt < \infty.$$

for any $p \in (0,1]$. The connection between the existing pointwise weight $w_\theta(\cdot)$ and the distribution-aware weight $\frac{f_{\text{ref}}(p)}{F_{\text{ref}}(p)}$ requires:

$$w(p) = \frac{f_{\text{ref}}(p)}{F_{\text{ref}}(p)}.$$

Example 1: REINFORCE with $w(p) = 1$. This requires $(\log F_{\text{ref}}(p))' = 1$ and $\log F_{\text{ref}}(p) = p + C$, i.e.,

$$F_{\text{ref}}(p) = e^{p+C}$$

Let $F_{\text{ref}}(1) = 1$, we have $C = -1$ and $f_{\text{ref}} = e^{p-1}$. Notably, there is a point mass at $p = 0$ as $F_{\text{ref}}(0) = e^{-1}$. This corresponds to a reflected and truncated exponential distribution. Assume $Z \sim \text{Exp}(1)$ with $p_Z(z) = e^{-1}$ for $z \in [0, +\infty)$. Define $Y = 1 - Z$ such that $y \in (-\infty, 1]$. $P_Y(Y \leq y) = P(1 - Z \leq y) = 1 - F_Z(1 - y) = e^{y-1}$. Furthermore, by putting all mass in $(-\infty, 0)$ on the point $z = 0$, the sub-distribution becomes $F_{\text{ref}}(p) = e^{p-1}$, $p \in [0,1]$, with the sub-density $f_{\text{ref}}(p) = e^{p-1}$.

Example 2: GRPO with $w(p) = 1/\sqrt{p(1-p)}$. This requires $(\log F_{\text{ref}})'(p) = 1/\sqrt{p(1-p)}$, implying

$$\log F_{\text{ref}}(p) = \int \frac{dp}{\sqrt{p(1-p)}} = 2\arcsin(\sqrt{p}) + C \Rightarrow F_{\text{ref}}(p) = \exp(2\arcsin(\sqrt{p}) + C).$$

As $\arcsin(1) = \frac{\pi}{2}$, we let $F_{\text{ref}}(1) = 1$. Then $C = -\pi$. Consequently, we have

$$F_{\text{ref}}(p) = \exp(2\arcsin(\sqrt{p}) - \pi), \quad f_{\text{ref}}(p) = \frac{\exp(2\arcsin(\sqrt{p}) - \pi)}{\sqrt{p(1-p)}}.$$

Note that this distribution is not common one, but a normalized arcsine distribution after an exponential transformation.

Example 3: MaxRL with $w(p) = 1/p$. It requires $(\log F_{\text{ref}})'(p) = 1/p$, which implies $\log F_{\text{ref}}(p) = \log p + C$ and $F_{\text{ref}}(p) = \exp(\log p + C)$. Let $F_{\text{ref}}(1) = 1$, we have $C = 0$. Therefore, we have $F_{\text{ref}}(p) = p$ and $f_{\text{ref}}(p) = 1$, which is a uniform distribution in $[0,1]$. We illustrate the density and CDF in Figure 4.

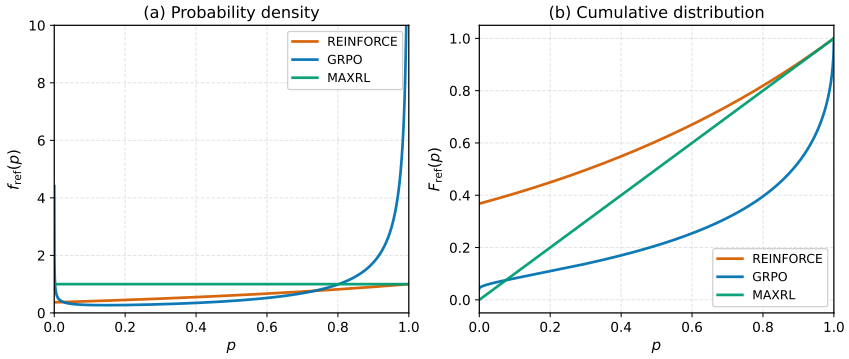


Figure 4: f_{ref} and F_{ref} for REINFORCE, GRPO, and MaxRL.

General Case. In general, there is a key connection equation:

$$\int_p^1 w(t)dt = \int_p^1 \frac{f_{\text{ref}}(t)}{F_{\text{ref}}(t)}dt = \log F_{\text{ref}}(1) - \log F_{\text{ref}}(p) = -\log F_{\text{ref}}(p).$$

This implies that

$$F_{\text{ref}}(p) = \exp\left(-\int_p^1 w(t)dt\right).$$

Since w is non-negative, we can easily verify that F_{ref} is non-decreasing function as $F_{\text{ref}}(p_1) < F_{\text{ref}}(p_2)$ if $p_1 < p_2$ and $F_{\text{ref}}(1) = 1$. Therefore, F_{ref} is a valid CDF. Given the fact that $\frac{d}{dp}\left(-\int_p^1 w(t)dt\right) = w(p)$, by the chain rule, we can derive

$$f_{\text{ref}}(p) = \frac{d}{dp}\left(\exp\left(-\int_p^1 w(t)dt\right)\right) = F_{\text{ref}}(p)w(p),$$

which is the exact goal we want to prove in the beginning.

Uniqueness. Assume $\tilde{F}$ also satisfies $\frac{\tilde{f}(p)}{\tilde{F}(p)} = w(p)$. This implies that

$$\frac{d}{dp}\log\tilde{F}(p) = w(p).$$

By taking the integral between $[p, 1]$ on both sides, we derive

$$\log\tilde{F}(1) - \log\tilde{F}(p) = \int_p^1 w(t)dt \Rightarrow \log\tilde{F}(p) = -\int_p^1 w(t)dt \Rightarrow \tilde{F}(p) = \exp\left(-\int_p^1 w(t)dt\right).$$

Therefore, we have $\tilde{F} = F_{\text{ref}}$ and F_{ref} is unique.

Remark. The induced F_{ref} only depends on the pointwise weight $w(p)$ with the pass rate value p in the pass-rate space, which is independent of the prompt $x \sim d_0$ and the policy parameter θ.

□

C.5 Monotone Calibration Invariance Property

For any increasing function $G : [0, 1] \to [0, 1]$ on the absolute value of pass rate $p_\theta(x)$, the gradient of our distribution-aware method keeps the same form, which we call a monotone calibration invariance property in Proposition 3:

Proposition 3 (Monotone Calibration Invariance). *Assume $F_{\text{ref}} = F_{\theta_{old}}$ and denote $\tilde{p}_\theta(x) = G(p_\theta(x))$, i.e., $\tilde{p}_\theta = G \circ p_\theta$. Denote the induced reference CDF and density as $\tilde{F}_{\text{ref}}$ and $\tilde{f}_{\text{ref}}$, then*

$$\nabla_\theta J_{curve}\left(\theta; p_\theta, F_{\text{ref}}\right) = \nabla_\theta J_{curve}\left(\theta; \tilde{p}_\theta, \tilde{F}_{\text{ref}}\right).$$

Proof. After the transformation with $\tilde{p}_\theta$, we have

$$\tilde{F}_{\text{ref}}(u) = \tilde{F}_{\theta_{\text{old}}}(u) = \mathbb{P}_{x\sim d_0}(\tilde{p}_{\theta_{\text{old}}}(x) \leq u) = \mathbb{P}_{x\sim d_0}(p_{\theta_{\text{old}}}(x) \leq G^{-1}(u)) = F_{\text{ref}}(G^{-1}(u)).$$

Then, we can derive

$$\tilde{f}_{\text{ref}}(u) = \frac{d\tilde{F}_{\text{ref}}(u)}{du} = f_{\text{ref}}(G^{-1}(u))\frac{dG^{-1}(u)}{du}.$$

Based on the derivative equation of the inverse function and let $u = G(v)$, we know $(G^{-1}(u))' = \frac{1}{G'(v)}$. Therefore, we have

$$\tilde{f}_{\text{ref}}(u) = f_{\text{ref}}(G^{-1}(u))\frac{dG^{-1}(u)}{du} = f_{\text{ref}}(G^{-1}(u))\frac{1}{G'(G^{-1}(u))}.$$

Recall the gradients before and after the transformation have the following form:

$$\nabla_\theta J_{\text{curve}}(\theta; p_\theta, F_{\text{ref}}) = \mathbb{E}_{x\sim d_0}\left[\frac{f_{\text{ref}}\left(p_\theta(x)\right)}{F_{\text{ref}}\left(p_\theta(x)\right)}\nabla_\theta p_\theta(x)\right]$$
$$\nabla_\theta J_{\text{curve}}(\theta; \tilde{p}_\theta, \tilde{F}_{\text{ref}}) = \mathbb{E}_{x\sim d_0}\left[\frac{\tilde{f}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}{\tilde{F}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}\nabla_\theta \tilde{p}_\theta(x)\right].$$

Next, we know that $\nabla\tilde{p}_\theta(x) = G'(p_\theta(x))\nabla p_\theta(x)$. By putting all together, we have the following key result:

$$\begin{aligned}
\frac{\tilde{f}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}{\tilde{F}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}\nabla_\theta\tilde{p}_\theta(x) &= \frac{f_{\text{ref}}(G^{-1}(\tilde{p}_\theta(x)))\frac{1}{G'(G^{-1}(\tilde{p}_\theta(x)))}}{F_{\text{ref}}(G^{-1}(\tilde{p}_\theta(x)))}\nabla_\theta\tilde{p}_\theta(x) \\
&= \frac{f_{\text{ref}}(G^{-1}(G(p_\theta(x))))\frac{1}{G'(G^{-1}(G(p_\theta(x))))}}{F_{\text{ref}}(G^{-1}(G(p_\theta(x))))}\nabla_\theta\tilde{p}_\theta(x) \\
&= \frac{f_{\text{ref}}(p_\theta(x))\frac{1}{G'(p_\theta(x))}}{F_{\text{ref}}(p_\theta(x))}G'(p_\theta(x))\nabla p_\theta(x) \\
&= \frac{f_{\text{ref}}(p_\theta(x))}{F_{\text{ref}}(p_\theta(x))}\nabla p_\theta(x).
\end{aligned}$$

This implies that

$$\begin{aligned}
\nabla_\theta J_{\text{curve}}(\theta; \tilde{p}_\theta, \tilde{F}_{\text{ref}}) &= \mathbb{E}_{x\sim d_0}\left[\frac{\tilde{f}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}{\tilde{F}_{\text{ref}}\left(\tilde{p}_\theta(x)\right)}\nabla_\theta\tilde{p}_\theta(x)\right] \\
&= \mathbb{E}_{x\sim d_0}\left[\frac{f_{\text{ref}}\left(p_\theta(x)\right)}{F_{\text{ref}}\left(p_\theta(x)\right)}\nabla_\theta p_\theta(x)\right] = \nabla_\theta J_{\text{curve}}\left(\theta; p_\theta, F_{\text{ref}}\right).
\end{aligned}$$

□

For the pointwise objective function, this monotone calibration invariance property does not hold. Immediately, we have the following corollary.

Corollary 1. *Denote* $J_g(\theta; p_\theta) = \mathbb{E}_{x\sim d_0}\left[g\left(p_\theta(x)\right)\right]$ *and* $J_g(\theta; \tilde{p}_\theta) = \mathbb{E}_{x\sim d_0}\left[g\left(\tilde{p}_\theta(x)\right)\right]$ *with* $\tilde{p}_\theta = G \circ p_\theta$. *There exists some* g *such that*

$$\nabla_\theta J_g(\theta; p_\theta) \neq \nabla_\theta J_g(\theta; \tilde{p}_\theta).$$

Proof. Firstly, we have

$$\nabla_\theta J_g\left(\theta; \tilde{p}_\theta\right) = \mathbb{E}_{x\sim d_0}\left[g'\left(\tilde{p}_\theta(x)\right)\nabla_\theta\tilde{p}_\theta(x)\right] = \mathbb{E}_{x\sim d_0}\left[g'\left(G\left(p_\theta(x)\right)\right)\cdot G'\left(p_\theta(x)\right)\cdot\nabla_\theta p_\theta(x)\right].$$

If we let

$$\nabla_\theta J_g(\theta) = \mathbb{E}_{x\sim d_0}\left[g'\left(p_\theta(x)\right)\nabla_\theta p_\theta(x)\right] = \mathbb{E}_{x\sim d_0}\left[g'\left(G\left(p_\theta(x)\right)\right)\cdot G'\left(p_\theta(x)\right)\cdot\nabla_\theta p_\theta(x)\right],$$

this implies a functional equation for all $p_\theta(x)$ and G.

$$g'\left(p_\theta(x)\right) = g'\left(G\left(p_\theta(x)\right)\right) G'\left(p_\theta(x)\right),$$

A counterexample is when $g'(t) = \frac{1}{t}$ in MaxRL and $G(t) = t^2$, we have:

$$\frac{1}{p_\theta(x)} \neq \frac{1}{p_\theta(x)^2} 2p_\theta(x) = \frac{2}{p_\theta}.$$

In summary, our distribution-aware prompt reweighting enjoys the monotone calibration invariance and is thus robust to the miscalibration for the absolute value of pass rates. By contrast, the pointwise prompt reweighting does not satisfy this property in general. □

D Implementation Details

More Details of Experimental Setup. Based on verl framework, our training applies a fully on-policy policy gradient update, meaning that there is no importance ratio or associated clipping. We also disable both the KL penalty and the entropy bonus. As such, the effect of the prompt-weighting function is not entangled with auxiliary regularizers, thereby avoiding complex tuning. This practice is also adopted in prior work on RLVR for LLM reasoning [88, 69, 89, 111]. During training across 1000 RL steps, we employ the temperature 1.0 and the learning rate 1×10^{-6}. The maximum prompt length is capped at 1024 tokens and the maximum response length at 4096 tokens for both Qwen3-1.7B-Base and Qwen3-4B-Base. In evaluation, we use $\mathrm{pass@}k$ [17] as the primary metric, with $k \in \{1, 2, 4, \ldots, 1024\}$ for Qwen3-1.7B-Base and $k \in \{1, 2, 4, \ldots, 512\}$ for Qwen3-4B-Base. $\mathrm{pass@}k$ measures the probability that at least one of k independently sampled responses is correct, serving as a proxy for a model's exploration capability under a fixed sampling budget. In practice, $\mathrm{pass@}1$ is the raw mean accuracy, while $\mathrm{pass@}k$ for $k \geq 2$ uses 1000 best-of-k bootstrap resamples (with replacement) of size k, averaged across prompts. We sample 2048 rollouts per prompt for Qwen3-1.7B-Base and 1024 for Qwen3-4B-Base to reduce estimation variance.

RL Stack and Computational Devices. All experiments are run within the verl post-training framework [80], which couples a policy-gradient trainer with a vLLM-backed rollout engine. Using a single inference stack for both training rollouts and downstream evaluation removes a common source of distribution shift between the two stages. Both backbones are optimized with the AdamW optimizer using its default first- and second-moment coefficients, under bfloat16 mixed precision and FSDP-style sharding of parameters, gradients, and optimizer states. Each experiment runs on a single node of $8\times$ NVIDIA B200 GPUs interconnected by NVLink.

Qwen-math prompt template

```
<|im_start|>system
Please reason step by step and put the final answer in \boxed{}.
<|im_end|>
<|im_start|>user
{problem statement}
Let's think step by step and put the final answer within \boxed{}.
<|im_end|>
<|im_start|>assistant
```

Prompt Template. Every training and evaluation prompt is rendered with the Qwen-math chat template [102]. We append the instruction "`Let's think step by step and put the final answer within \boxed{}.`" to the problem statement and route the result through the tokenizer's chat template, yielding inputs of the form shown below. At inference time the verifier extracts the final boxed expression from the generated chain of thought and grades it with MATH-VERIFY.

On-policy Update. Each RL step generates a fresh batch of rollouts from the current policy and consumes it with a single optimizer update. Therefore, the data-generating policy and the target policy coincide. The token-level importance ratio is identically one along the entire response,

$$\rho_{i,t}(\theta) \;=\; \frac{\pi_\theta(y_{i,t} \mid x,\, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x,\, y_{i,<t})} \;\equiv\; 1,$$

Table 4: Training hyperparameters shared by both Qwen3 backbones.

Parameter	Value	Parameter	Value
Base model	Qwen3-{1.7B, 4B}-Base	Prompts per batch	256
Rollouts per prompt	8	Grad updates per step	1
Max prompt length	1024	Max response length	4096
Learning rate	1×10^{-6}	Training steps	1000
KL coefficient	0	Entropy coefficient	0
Rollout temperature	1.0	Eval temperature	0.6
Eval top-p	0.95	Device	8× NVIDIA B200

which implies that the PPO clipping range $[1-\varepsilon,\, 1+\varepsilon]$ is never active and the threshold ε has no effect on the optimization trajectory. The KL penalty and the entropy bonus are also disabled, isolating the contribution of the prompt-weighting rule from auxiliary regularizers and matching the simplified objective adopted by [88]. This supports the concise formulations in the preliminaries of Section 2.

Evaluation Decoding. Following [88], evaluation generations are sampled at temperature 0.6 with top-p 0.95, and both top-k and min-p truncation are disabled. We do not employ adaptive sampling and do not apply any logit correction for the mismatch between inference and training runtimes. Reported metrics are computed on the final checkpoint of each run.

Hyperparameter Summary. Table 4 consolidates the training-side hyperparameters, which are shared across both Qwen3-1.7B-Base and Qwen3-4B-Base.

E Supplemental Experimental Results

This appendix contains experimental results that are not shown in the main text due to the space limit. The training setup, evaluation protocol, and bootstrap procedure are identical to those described in Section 5.

E.1 Pass@1 and Pass@64 on Three Additional Benchmarks

Table 5: **Supplemental results on three additional math reasoning benchmarks.** pass@1 and pass@64 (%), best per column in bold.

Method	**BRUMO 2025**		**HMMT 11/25**		**Minerva**		**Avg.**	
	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
Qwen3-1.7B-Base								
Base	8.0	40.1	2.4	18.7	20.8	56.9	10.4	38.6
GRPO	14.2	40.2	3.0	14.7	**31.4**	57.7	16.2	37.5
MaxRL	15.9	42.5	3.7	18.4	31.2	59.5	16.9	40.2
CurveRL	**18.0**	**45.3**	**3.9**	**20.6**	30.3	**60.3**	**17.4**	**42.1**
Qwen3-4B-Base								
Base	18.2	46.7	3.5	21.3	25.9	61.3	15.9	43.1
GRPO	27.3	47.1	5.7	27.8	40.8	58.3	24.6	44.4
MaxRL	**30.4**	**66.5**	**8.1**	35.6	**42.8**	62.3	**27.1**	54.8
CurveRL	27.6	66.2	7.2	**39.1**	39.9	**62.9**	24.9	**56.1**

Table 5 reports pass@1 and pass@64 on three additional benchmarks: **BRUMO 2025** [6], **HMMT 11/25** [6], and **Minerva Math** [51]. The qualitative picture mirrors Table 1: CurveRL matches or exceeds MaxRL at pass@64 on every benchmark on both model sizes.

E.2 Pass@k Curves on the Additional Benchmarks

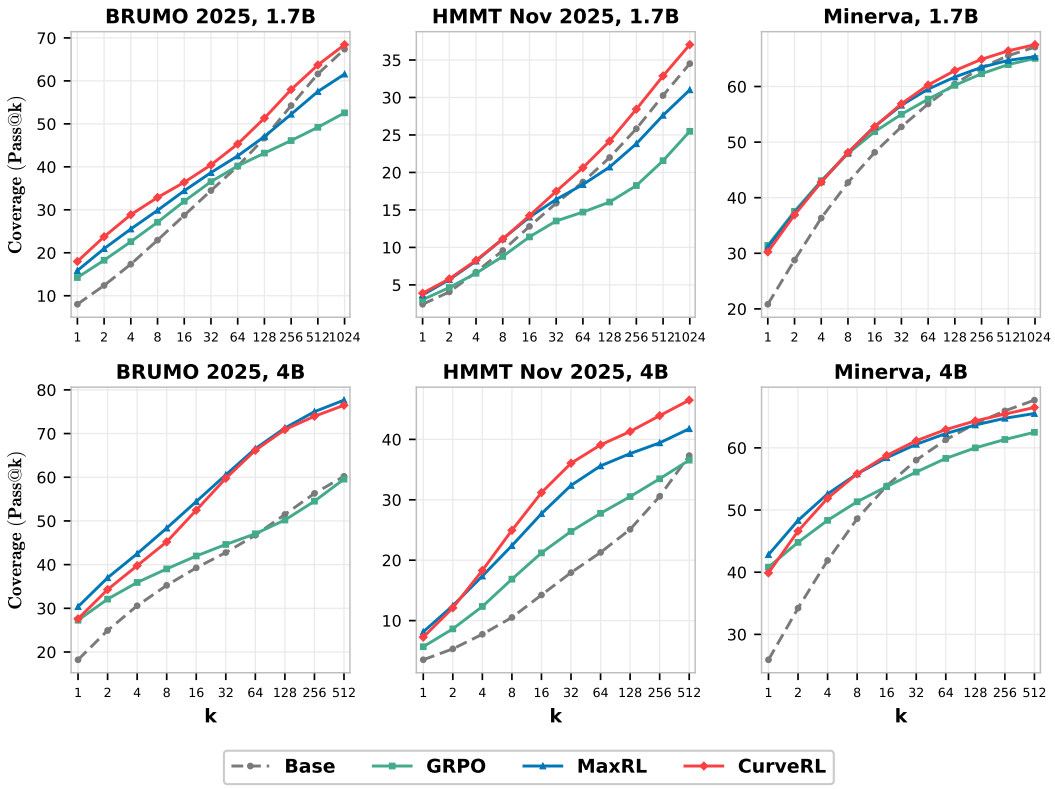


Figure 5: **Pass@k scaling on the additional three benchmarks.** Top row: Qwen3-1.7B-Base, $k \in \{1, \ldots, 1024\}$. Bottom row: Qwen3-4B-Base, $k \in \{1, \ldots, 512\}$. CurveRL tracks or dominates the GRPO and MaxRL baselines on most benchmarks.

E.3 Additional Difficulty Distribution

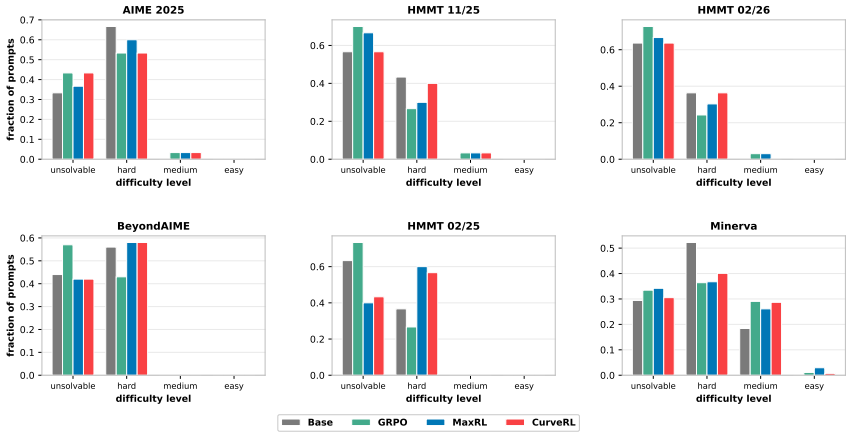


Figure 6: **Qwen3-1.7B-Base post-training prompt-difficulty distribution on the remaining six benchmarks.** Across benchmarks, CurveRL's *unsolvable* fraction is mostly no higher than that of the strongest pointwise-weighted baseline, while the *hard/medium* mass is preserved or enlarged.

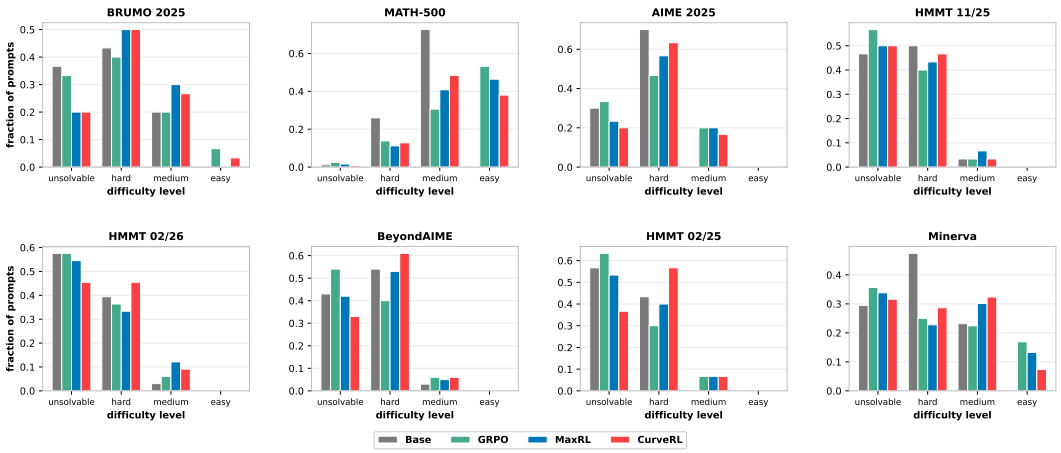


Figure 7: Qwen3-1.7B-Base post-training prompt-difficulty distribution on all benchmarks.

E.4 Learning Signals Across Training Dynamics

Figure 8 tracks the fraction of training prompts with strictly positive empirical pass rate $\hat{p}$ at each RL step on Qwen3-1.7B-Base and Qwen3-4B-Base. Larger values indicate a richer pool of prompts that provide non-vanishing gradient signals. While all three methods start from a similar pool, CurveRL and MaxRL remain consistently above GRPO throughout the training, with the gap widening after the early phase. This suggests that they slow down the collapse of prompts into the zero-gradient $\hat{p} = 0$ region, preserving useful learning signals for more steps and broadening the reasoning boundary.

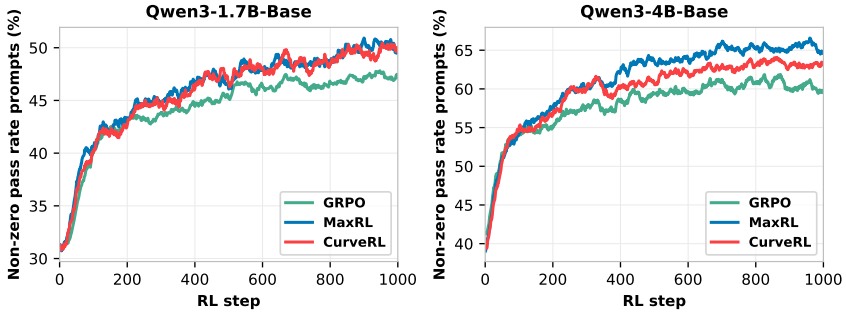


Figure 8: Fraction of prompts where the model generates at least one correct rollout out of 8 samples.

E.5 Supplemental Results on Sensitivity Analysis

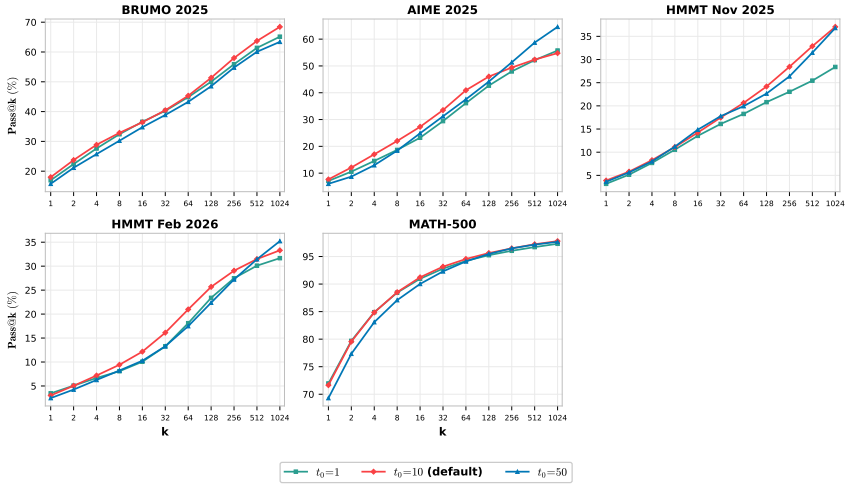


Figure 9: **Pass@k scaling for the three sliding-window sizes.** Five benchmarks from Table 3; each panel overlays CurveRL with $t_0 \in \{1, 10, 50\}$ batches on Qwen3-1.7B-Base. The default $t_0 = 10$ tracks or improves on both alternatives at every k on the most benchmarks.

E.6 Stable Training Dynamics

Figure 10 reports additional training dynamics on Qwen3-1.7B-Base (top row) and Qwen3-4B-Base (bottom row), including mean response length, policy entropy, and gradient norm. CurveRL generally produces longer chains of thought and maintains higher actor entropy, supporting greater reasoning diversity and a wider capability boundary, as shown in Section 5. CurveRL also exhibits a flatter gradient-norm profile, indicating more stable training.

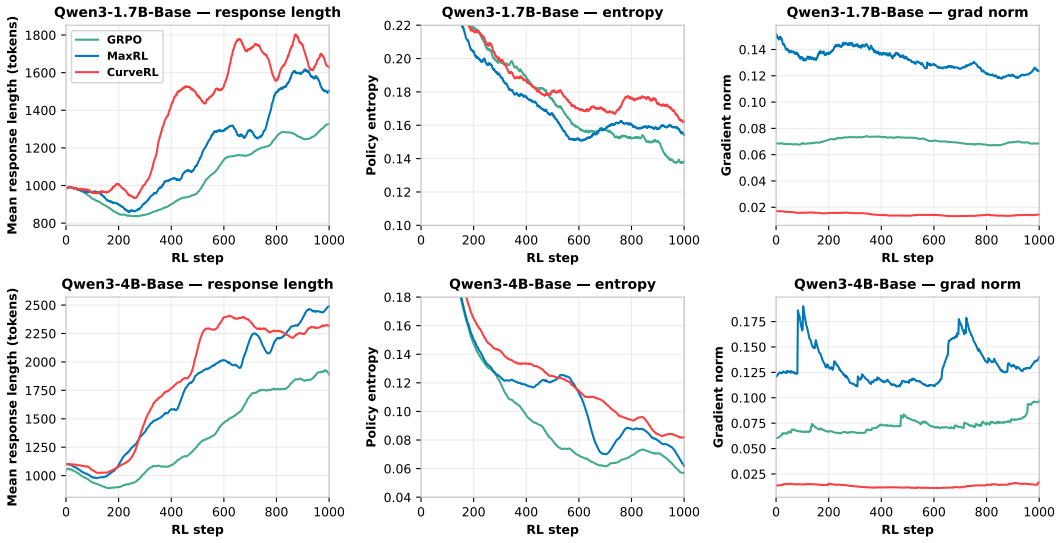


Figure 10: **Additional training dynamics metrics for Qwen3-1.7B-Base (top row) and Qwen3-4B-Base (bottom row).** Mean response length, policy entropy, and gradient norm over RL steps for three RLVR algorithms.

E.7 Validation Accuracy During Training

Figures 11 and 12 report the validation accuracy on the MATH-500 during training. CurveRL generally outperforms other baselines on both pass@1 and pass@k during training.

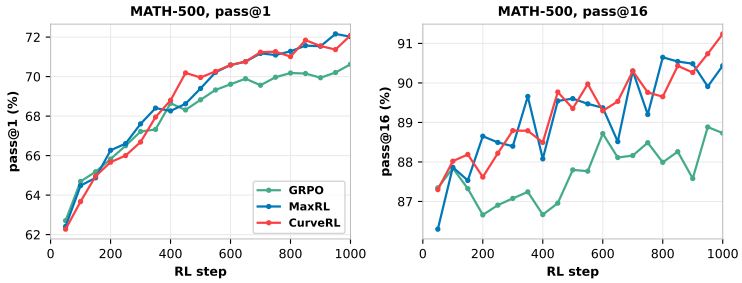


Figure 11: **Qwen3-1.7B-Base validation accuracy during training.**

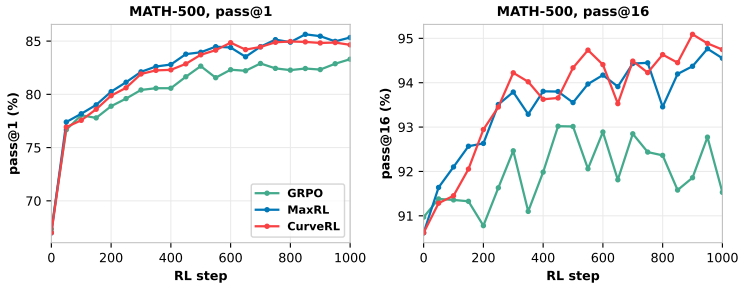


Figure 12: **Qwen3-4B-Base validation accuracy during training.**

E.8 Distribution-Aware Weighting on Qwen3-1.7B-Base

Results shown in Figure 13 on Qwen3-1.7B-Base are similar to Figure 3 on Qwen3-4B-Base. The qualitative pattern matches the 4B case in the main text: the empirical pass-rate density $\hat{f}_{\text{ref}}$ drifts toward higher pass-rate bins as training proceeds, while the data-driven weight $w_t = \hat{f}_{\text{ref}}/\hat{F}_{\text{ref}}$ continues to emphasize the low-pass-rate prompts. The static GRPO and MaxRL weights are model-size invariant and therefore identical to the 4B panels.

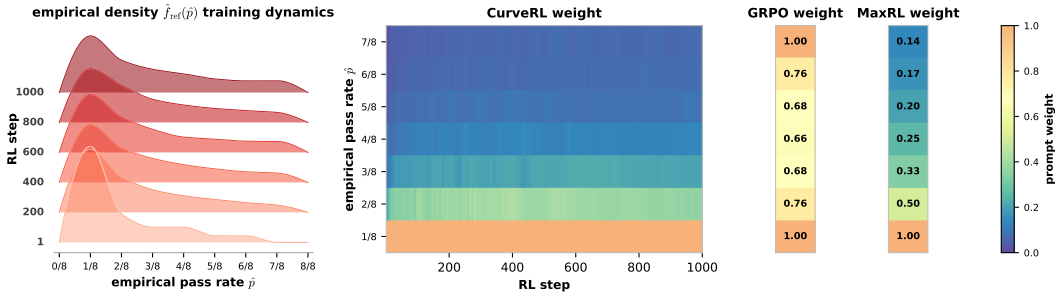


Figure 13: **Weighting schemes on Qwen3-1.7B-Base.** Same layout and scale as Figure 3. Only CurveRL's panels (left two) reflect 1.7B training, since the GRPO and MaxRL weights are static.

E.9 Ablation Study and More Baselines

Number of Rollouts N and A Smoothed Estimator. We have added an ablation on the number of rollouts N across five benchmarks and the pass@1 and pass@64 metrics. As shown in Table 6, reducing N from $N = 8$ (default) to $N = 4$ worsens both average pass@1 and average pass@64 of CurveRL, indicating that a small N produces an overly coarse pass-rate grid for estimating reliable

and precise prompt weights. By contrast, performance appears to plateau as N increases from $N = 8$ (default) to $N = 16$, with marginal and inconsistent gains: a limited improvement in average pass@64 accompanied by a drop in average pass@1. Therefore, we select $N = 8$ as the default, as it achieves the best overall performance–compute trade-off among the evaluated settings and is also commonly used in representative RLVR algorithms, such as GRPO.

Table 6: Ablation study of CurveRL with different rollout counts N and a smoothed estimator on Qwen3-1.7B-Base. The best result in each column is in bold.

	AIME 2025		BeyondAIME		HMMT 02/25		HMMT 11/25		BRUMO 2025		Average	
Method	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
CurveRL ($N = 4$)	6.1	35.0	1.1	19.4	1.0	16.2	3.3	18.6	15.7	42.1	5.5	26.3
CurveRL ($N = 8$, default)	**7.7**	**40.9**	1.4	23.6	**1.7**	19.3	**3.9**	**20.6**	**18.0**	45.3	**6.5**	29.9
CurveRL ($N = 16$)	7.2	39.0	**1.5**	**25.2**	1.3	**22.4**	3.6	19.9	16.6	**46.1**	6.0	**30.5**
CurveRL ($N = 8$, smoothed estimator)	5.8	38.1	**1.5**	23.2	1.2	20.7	3.5	**20.6**	16.7	45.9	5.7	29.7

In addition to the standard histogram estimator used in CurveRL, we evaluate a smoothed variant for comparison. This variant preserves the original f/F-based prompt weighting while dividing each $\hat{p}$ bin into two sub-bins. Prompts within the same original bin are randomly assigned to these sub-bins before their weights are computed. Thus, prompts with the same $\hat{p}$ can have distinct density estimates, resulting in a finer-grained density estimator. Table 6 shows that this variant achieves a similar average pass@64 to standard CurveRL, but a lower average pass@1. Although more advanced density estimators could also be explored, we view them as extensions of the proposed CurveRL algorithm.

Different Components in CurveRL. We have added the corresponding ablation study across five benchmarks using both pass@1 and pass@64 metrics in Table 7. The results show that each component captures only part of the distributional structure and therefore accounts for only part of CurveRL's overall performance gains.

Table 7: Ablation study of CurveRL with different components on Qwen3-1.7B-Base. The best result in each column is in bold.

	AIME 2025		BeyondAIME		HMMT 02/25		HMMT 11/25		BRUMO 2025		Average	
Method	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
CurveRL (rank)	5.7	36.3	**1.6**	21.9	0.9	17.2	3.7	**21.4**	14.2	40.6	5.2	27.5
CurveRL (density)	6.3	38.3	1.5	20.5	1.0	17.1	3.6	20.1	15.0	42.1	5.5	27.6
CurveRL (rank+density, default)	**7.7**	**40.9**	1.4	**23.6**	**1.7**	**19.3**	**3.9**	20.6	**18.0**	**45.3**	**6.5**	**29.9**

Comparison with DAPO [105]. We have added a comparison with DAPO in Table 8. While DAPO achieves a slightly higher average pass@1 than MaxRL, it yields a lower average pass@64. Our distribution-aware CurveRL outperforms DAPO on both average metrics. Moreover, DAPO incurs more than twice the wall-clock training time of the other prompt-reweighting methods in our experiments. This additional overhead arises from its dynamic sampling procedure, which generates additional rollout groups and retains prompts with empirical pass rates strictly between zero and one until the effective batch is filled.

Table 8: Overall performance comparison including DAPO on Qwen3-1.7B-Base. The best result in each column is in bold.

	AIME 2025		BeyondAIME		HMMT 02/25		HMMT 11/25		BRUMO 2025		Average	
Method	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64	pass@1	pass@64
Base	3.1	34.6	1.1	16.9	0.2	10.6	2.4	18.7	8.0	40.1	3.0	24.2
GRPO	4.8	29.3	1.1	15.6	0.5	10.0	3.0	14.7	14.2	40.2	4.7	22.0
MaxRL	6.7	36.5	1.3	21.8	1.3	18.4	3.7	18.4	15.9	42.5	5.8	27.5
DAPO	7.3	35.7	**1.5**	16.5	**1.7**	12.7	3.4	16.2	15.5	38.4	5.9	23.9
CurveRL	**7.7**	**40.9**	1.4	**23.6**	**1.7**	**19.3**	**3.9**	**20.6**	**18.0**	**45.3**	**6.5**	**29.9**

F More Discussions

F.1 Discussion about Policy-Reweighted Contextual Bandit

New Information Acquisition in RLVR. In standard RL with an external environment, information acquisition is primarily achieved through action-level exploration [90, 49], where the policy affects future observations only indirectly through the environment dynamics. In contrast, RLVR under the PRCB formulation introduces a fundamentally new control axis: the learner can directly reshape the effective context distribution through policy-dependent prompt reweighting. This mechanism is closely related in spirit to active learning [57, 63], adaptive experimental design [62, 12], and Bayesian optimization [32, 15, 4], where learning efficiency is improved by adaptively allocating sampling effort to more informative inputs. With a slight notational abuse, we let $d_t(x)$ denote the context distribution and $d_t(x, y)$ the joint context-response distribution at time t. The following schematic contrast highlights the distinction between the two paradigms:



$$\text{Standard RL: } \cdots \to d_t(x) \overset{\pi_{\theta_t}}{\to} d_t(x,y) \overset{\text{update}}{\to} \pi_{\theta_{t+1}} \overset{\textbf{exploration}}{\to} d_{t+1}(x) \to \cdots,$$

$$\text{RLVR under PRCB: } \cdots \to d_t(x) \overset{\pi_{\theta_t}}{\to} d_t(x,y) \overset{\text{update}}{\to} \pi_{\theta_{t+1}} \overset{\boldsymbol{w_{\theta_{t+1}}(x)}}{\to} d_{t+1}(x) \to \cdots.$$

This perspective highlights that RLVR differs fundamentally from standard RL by enabling direct control over the state or context distribution, rather than relying solely on action-level exploration.

New Challenges under PRCB. The PRCB viewpoint also introduces two conceptual challenges. *(1) Coupled decision-dependence sampling structure*. RLVR already involves decision-dependent sampling in the response space, where the data distribution in the optimization depends on π_θ. Under PRCB, this dependence further extends to the context space through $d_\theta(x)$. As such, the learning dynamics are shaped jointly by response-level policy dependence via π_θ and context-level reweighting via w_θ. *(2) Breakdown of classical optimality principles.* Many optimality criteria for contextual bandits, such as simple and cumulative regret minimization [26, 113, 50], are formulated under a fixed context distribution d_0. In PRCB, however, the effective training measure changes with the policy through w_θ. This policy-dependent measure adaptivity falls outside the standard analysis and motivates our utility-based formulation as one principled optimality in Section 3.2.

F.2 Curriculum Learning under Our Framework

Curriculum Learning is a Time-Varying Pointwise Prompt Reweighting Method. The curriculum learning strategy in RLVR, such as [71], develops different data scheduling from easy to hard prompts, therefore improving the policy improvement. Within our framework in Definition 1, it indeed changes the measure from a fixed d_0 to a time-varying measure $d_0^{(t)}$ that is manually developed, which is orthogonal to the policy-dependent prompt reweighting $w_\theta(x)$. The utility function $\mathcal{U}_\theta$ based on the pointwise objective $J_g(\theta)$ in Eq. (6) is changed to a time-varying version:

$$\mathcal{U}_\theta^{(t)} = \mathbb{E}_{x\sim d_0^{(t)}}\left[g\left(p_\theta(x)\right)\right] = \mathbb{E}_{x\sim d_0}\left[\frac{d_0^{(t)}(x)}{d_0(x)} g\left(p_\theta(x)\right)\right].$$

By the functional derivative defined in Definition 1, for any perturbation $h \in L^2(d_0^{(t)})$, according to the Riesz representation, the first variation of $\mathcal{U}_\theta^{(t)}$ is given by

$$\lim_{\epsilon\to 0} \frac{\mathcal{U}_\theta^{(t)}(p_\theta + \epsilon h) - \mathcal{U}_\theta^{(t)}(p_\theta)}{\epsilon} = \int w_\theta^*(x)\, h(x)\, d_0^{(t)}(x)dx = \int w_\theta^*(x)\, \frac{d_0^{(t)}(x)}{d_0(x)}\, h(x)\, d_0(x)dx.$$

where $\frac{d_0^{(t)}(x)}{d_0(x)}$ is the Radon-Nikodym derivative. Immediately, we can derive a time-varying optimal weight function for the pointwise objective $J_g(\theta)$ as

$$w_\theta^*(x,t) = \frac{\delta \mathcal{U}_\theta^{(t)}}{\delta p_\theta}(x) = \frac{d_0^{(t)}(x)}{d_0(x)} g'(p_\theta(x)).$$

This illuminates the role of curriculum learning in RLVR: it incorporates a pointwise time-varying prompt reweighting by manually developing $d_0^{(t)}$ in a curriculum way, which is on top of the policy-dependent prompt reweighting function $w_\theta(x)$.

Similarity and Difference Between Curriculum Learning and Our Distribution-aware Method. Given the aforementioned mechanism of curriculum learning in RLVR, a natural question is: *why curriculum learning and our distribution-aware method are both effective in RLVR and what the fundamental distinction is*. We argue that both methods aim to reshape the learning dynamics of RLVR to focus on prompts near the learning frontier or in the current learnable window at each training time by prompt reweighting, even though they are in different ways:

- The easy-to-hard schedule employed in curriculum learning incorporates a non-stationary sample allocation mechanism to help the policy optimization to **chase** the moving learnable window. However, the endogenous easy-to-hard schedule is typically conducted by leveraging prior knowledge and based on the absolute values of pass rates, to the best of our knowledge.
- By introducing the quantile coordinate transform, i.e., quantile reparameterization, our distribution-aware method flattens the difficulty distribution in a nearly uniform and stationary spectrum, transforming the moving learnable windows to a stationary and fixed one. This strategy implicitly yet adaptively helps the policy to concentrate on the prompts in the learning frontier, and at the same time it takes advantage of the distributional information of pass rates in the pass rate function space.

F.3 When Gradients of Distribution-aware and Pointwise Utility Functionals Differ

Under the same distortion function, the distribution-aware and pointwise utility functions are

$$\mathcal{U}_\theta(F_{\text{ref}}, \psi) = \mathbb{E}_{x\sim d_0}\left[\psi\left(F_{\text{ref}}\ \left(p_\theta(x)\right)\right)\right],\ \mathcal{U}_\theta := J_\psi(\theta) = \mathbb{E}_{x\sim d_0}\left[\psi\left(p_\theta(x)\right)\right],$$

with their respective gradients:

$$\begin{aligned}\nabla_\theta \mathcal{U}_\theta(F_{\text{ref}}, \psi) &= \mathbb{E}_{x\sim d_0}\left[\psi'\left(F_{\text{ref}}\ \left(p_\theta(x)\right)\right) f_{\text{ref}}\ \left(p_\theta(x)\right) \nabla_\theta p_\theta(x)\right],\\ \nabla_\theta J_\psi(\theta) &= \mathbb{E}_{x\sim d_0}\left[\psi'\left(p_\theta(x)\right) \nabla_\theta p_\theta(x)\right].\end{aligned}$$

Denote $p = p_\theta(x)$. For the general increasing function ψ, the equivalence of the two gradients implies

$$\psi'(p) = \psi'(F_{\text{ref}}(p)) f_{\text{ref}}(p) \Rightarrow \frac{d}{dp}\psi\left(F_{\text{ref}}(p)\right) = \frac{d}{dp}\psi(p).$$

By taking the integral on both sides between p and 1, we have

$$\psi(F_{\text{ref}}(1)) - \psi(F_{\text{ref}}(p)) = \psi(1) - \psi(p).$$

As we know $F_{\text{ref}}(1) = 1$, this implies that

$$\psi(F_{\text{ref}}(p)) = \psi(p),$$

which holds for each p. As ψ is an increasing function, we can derive

$$F_{\text{ref}}(p) = p,$$

which implies that F_{ref} is Uniform$(0, 1)$ for the pass rate distributions of $p_\theta(x)$.

Remark. This connection with the uniform distribution reveals a fundamental source of differences between distribution-aware prompt reweighting and pointwise counterpart: the mismatch between model capability and data difficulty. Specifically, in the general learning dynamics, the pass rate distributions can be more concentrated around 0 if the model struggles to achieve high pass rates for most prompts, especially in the early training phase. That's when the *weight collapse* issue happens and distribution-aware prompt reweighting differs from the pointwise counterpart. For example, if we increase the model size on a fixed training dataset, the advantage may diminish at the pass rate distribution can be more spread and closer to a uniform distribution. However, this does not happen in the general learning dynamics. Therefore, the fundamental source of the advantage of distribution-aware prompt reweighting over the pointwise version depends on capability-difficulty match between the model and the dataset.

F.4 Sufficient Conditions for Prompt Weight Comparison and Risk Preference

Under the same risk measure ψ, we define the pointwise and distribution-aware prompt weight by

$$w_0(p) = \psi'(p),\ w_F(p) = \psi'(F_{\text{ref}}(p))f_{\text{ref}}(p).$$

Following the analysis in Appendix F.3, we hope to know when $w_0(p) > w_F(p)$ for the concerned pass rates, e.g., low pass rate in the early training. Answering this question recovers a fundamental relationship between model capability and data difficulty.

For $\psi(t) = \log(t)$. $w_0(p) = \frac{1}{p}$ and $w_F(p) = \frac{f_{\text{ref}}(p)}{F_{\text{ref}}(p)}$. As the two weights have different magnitudes, comparing the weight requires us to perform normalization. For N number of rollouts, the pass rate satisfies $p \in [0, \frac{1}{N}, \ldots, \frac{N}{N}]$. Denote $p_i = \frac{i}{N}$ and thus the normalized weights are

$$\bar{w}_0(p_i) = \frac{w_0(p_i)}{\sum_{i=0}^{N} w_0(p_i)},\ \bar{w}_F(p_i) = \frac{w_F(p_i)}{\sum_{i=0}^{N} w_F(p_i)}.$$

We now analyze the scenario when the distribution-aware weight $\bar{w}_F$ is more aggressive than the pointwise weight $\bar{w}_0$. We first define:

$$r(p) = \frac{w_F(p)}{w_0(p)} = \frac{p f_{\text{ref}}(p)}{F_{\text{ref}}(p)} = \frac{d\log F_{\text{ref}}(p)}{d\log p} = \frac{f_{\text{ref}}(p)}{\frac{1}{p}\int_0^p f_{\text{ref}}(t)dt}.$$

Given any $p_l < p_u$, being more aggressive for $\bar{w}_F$ indicates that

$$\frac{\bar{w}_F(p_l)}{\bar{w}_F(p_u)} > \frac{\bar{w}_0(p_l)}{\bar{w}_0(p_u)} \Leftrightarrow \frac{w_F(p_l)}{w_F(p_u)} > \frac{w_0(p_l)}{w_0(p_u)} \Leftrightarrow r(p_l) > r(p_u).$$

Therefore, **the sufficient condition** is

$$r'(p) < 0,$$

in the considered range of p, which can guarantee $r(p_l) > r(p_u)$ if $p_l < p_u$. Let $r'(p) = 0$, we have:

$$\frac{d\log F_{\text{ref}}(p)}{d\log p} = \alpha \Rightarrow \log F_{\text{ref}}(p) = \alpha \log p + C,$$

where α is a constant. Let $F_{\text{ref}}(1) = 1$, we have $C = 0$ and

$$F_{\text{ref}}(p) = p^{\alpha},\ f_{\text{ref}}(p) = \alpha p^{\alpha-1},\ w_F(p) = \frac{\alpha}{p}.$$

which is a power function. In our case, $\alpha = 1$. This result indicates that if $r(p)$ is a decreasing function regarding p or $w_F(p)$ decreases faster than $w_0(p) = \frac{1}{p}$, we will observe a more aggressive prompt reweighting on the low-pass-rate prompts over MaxRL.

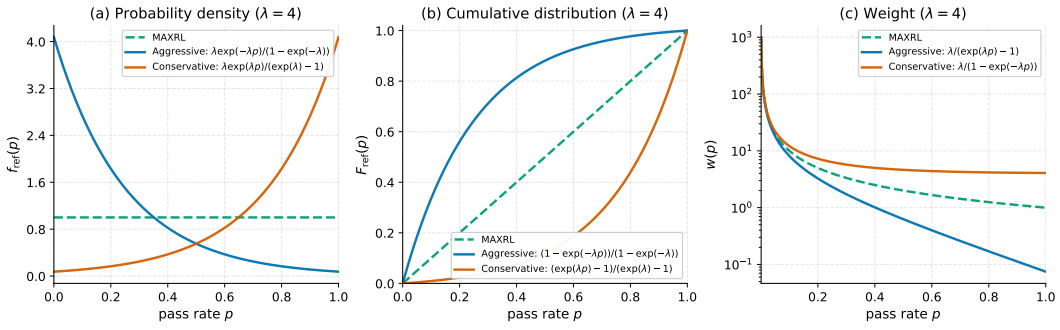
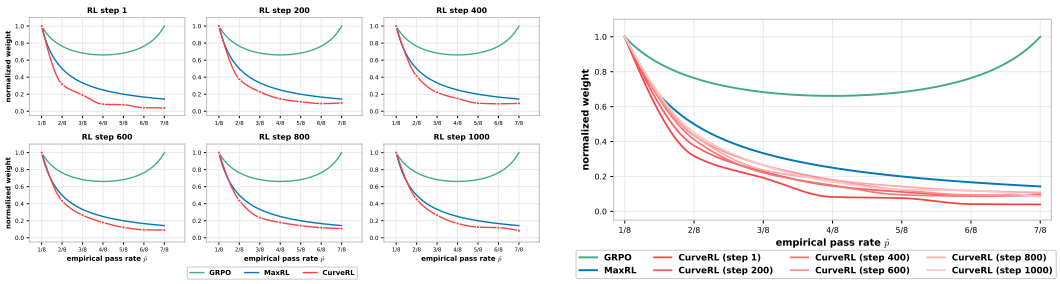


Figure 14: The probability density function, CDF and weight function in terms of the pass rate p among MaxRL and the two considered examples. The blue line shows a more aggressive weight update on the low-pass-rate prompts, while the red line suggests a more conservative update relative to MaxRL.



(a) Normalized Weight Dynamics on Six Steps.

(b) Combined Normalized Weight Dynamics.

Figure 15: Dynamics of normalized weights among GRPO, MaxRL, and CurveRL on Qwen3-4B-Base.

For example, consider a random variable Z with a truncated exponential distribution in $[0, 1]$:

$$f_Z(p) = \frac{\lambda e^{-\lambda p}}{1 - e^{-\lambda}},\ F_Z(p) = \frac{1 - e^{-\lambda p}}{1 - e^{-\lambda}},\ p \in [0, 1].$$

If $w_F(p)$ follows the truncated exponential distribution above, we have

$$w_F(p) = \frac{\lambda e^{-\lambda p}}{1 - e^{-\lambda p}} = \frac{\lambda}{e^{\lambda p} - 1}, r(p) = \frac{\lambda p}{e^{\lambda p} - 1} \Rightarrow r'(p) = \lambda \frac{e^{\lambda p} - 1 - \lambda p e^{\lambda p}}{(e^{\lambda p} - 1)^2} < 0,$$

as we can easily prove that $g(z) = e^z - 1 - ze^z$ with $z = \lambda p > 0$ is a decreasing function with a negative derivative and $g(z) < g(0) = 0$. Another example is a symmetric random variable Y relative to Z in $[0, 1]$, i.e., $Y = 1 - Z$. Therefore, we have

$$f_Y(p) = \frac{\lambda e^{\lambda p}}{e^{\lambda} - 1},\ F_Y(p) = \frac{e^{\lambda p} - 1}{e^{\lambda} - 1},\ p \in [0, 1].$$

Similarly, this distribution implies that

$$w_F(p) = \frac{\lambda}{1 - e^{-\lambda p}}, r(p) = \frac{p\lambda}{1 - e^{-\lambda p}} \Rightarrow r'(p) = \lambda \frac{1 - e^{-\lambda p} - \lambda p e^{-\lambda p}}{(1 - e^{-\lambda p})^2} > 0,$$

which indicates that this strategy is more conservative than MaxRL with a slower decreasing speed of $W_F(p)$ than $\frac{1}{p}$. Figure 14 showcases the density function, CDF, and weight functions in terms of the pass rate p of MaxRL and two considered examples, with different aggressiveness of weights in the gradient update.

Empirically, we compare the dynamics of normalized weights among GRPO, MaxRL, and CurveRL in Figure 15. It suggests that CurveRL behaves more aggressively and puts larger weights on low-pass-rate prompts than MaxRL with a faster decreasing speed when we increase p. We argue that the current model capability is still less sufficient for the dataset POLARIS-53K, such that a more risk-seeking utility function is preferable, which is guided in principle by the relationship between data difficulty and the model's capability. This result demonstrates that our method is data-driven with an adaptive weight function relative to the static ones in GRPO and MaxRL. Conversely, a conservative weight should be observed when we deploy a larger model on an easier dataset. This is an interesting investigation for empirical demonstration, which we leave as future work.

For the General ψ. Our sufficient condition can be extended to the general risk function ψ, where the condition quantity is

$$R_\psi(p) := \frac{w_F(p)}{w_0(p)} = \frac{\psi'\left(F_{\text{ref}}(p)\right) f_{\text{ref}}(p)}{\psi'(p)}.$$

If $R'_\psi(p) < 0$, our distribution-aware weight function $w_F(p)$ is more aggressive over the low-pass-rate prompts than the pointwise counterpart in the learning dynamics.

Relationship between Model Capacity and Data Difficulty. A key observation is that our distribution-aware weight is data-driven and depends on both the model π_θ and the dataset. The

mismatch between the model capability and data difficulty fundamentally determines the dynamics of the distribution-aware weights, including whether the weight is more aggressive than the pointwise version or not. In particular, if a simpler model is employed on a hard dataset, more pass rates of prompts will be more concentrated around 0, leading to more aggressive weight on the low-pass-rate prompts. On the other hand, if a large model is employed on an easy dataset, the weight distribution is more spread and tends to be more uniform. Our distribution-aware weight likely behaves similarly as the pointwise counterpart, e.g., MaxRL, or even more conservatively on low-pass-rate prompts. We leave more empirical demonstration as future work.

F.5 Induced Wasserstein-type Geometry in Our Method

The transform F_{ref} provides a quantile-based difficulty coordinate: it measures the relative position of the pass rate $p_\theta(x)$ under a fixed reference distribution. This representation depends on the ordering structure of pass rates rather than their raw scale. In this sense, it is closer in spirit to one-dimensional Wasserstein-type quantile representations than to pointwise density-ratio-based objectives such as Kullback–Leibler (KL) divergence.

Kullback–Leibler (KL) divergence with pointwise density geometry. When we compare the distances of two one-dimensional distributions μ and ν, if they are absolutely continuous with respect to Lebesgue measure, the densities p_μ and p_ν exist. Consequently, the Kullback–Leibler (KL) divergence is defined by using the pointwise probability ratio:

$$D_{\text{KL}}(\mu,\nu) = \int_{\mathbb{R}} p_\mu(x) \log \frac{p_\mu(x)}{p_\nu(x)} dx,$$

which induces a pointwise geometry in the density space.

Optimal Transport (OT)-based Distance with expressive geometry. As a typical instance of optimal transport, p-Wasserstein distance on one-dimensional random variables is defined by

$$W_p^p(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int \|x-y\|^p d\gamma(x,y),$$

where γ is the coupling or the transport plan, and $\Pi(\mu,\nu)$ is the joint distribution space with the marginal distributions as μ and ν. By definition, Wasserstein distance explicitly accounts for the underlying geometry of the data space, as opposed to the KL divergence. In the one-dimensional space, Wasserstein distance is equivalent to the **quantile matching**:

$$W_p^p(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^p du,$$

which evaluates the distance in the quantile space with respect to the quantile function F^{-1}. When $p=1$, 1-Wasserstein distance has an equivalent form based on the CDF:

$$W_1(\mu,\nu) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|\, dt,$$

which induces a transport or ordering geometry.

In summary, our method operates $F_{\text{ref}}(p_\theta(x))$ in the distribution-aware utility functional, a surrogate of the one from the perspective of distribution transport (moving the mass from $p_\theta(x)$ to $p^*(x)$) as introduced in Section 3.2. Since optimal transport in the one-dimensional space reduces to quantile matching and CDF matching, our method effectively operates in the quantile space (equivalent to the CDF space), inducing a Wasserstein-type geometry.

F.6 Integrating both Pointwise and Distribution-aware Utility Functionals

Strategy 1: Single Function Transformation. We integrate both $p_\theta(x)$ and $F_{\text{ref}}\left(p_\theta(x)\right)$ into a single function transformation ϕ:

$$\mathcal{U}_\theta = \mathbb{E}_{x\sim d_0}\left[\phi\left(p_\theta(x), F_{\text{ref}}\left(p_\theta(x)\right)\right)\right].$$

We choose $\phi(a,b) = \log(a^{1-\lambda}b^\lambda)$ for computational convenience, which leads to

$$\mathcal{U}_\theta = \mathbb{E}_{x\sim d_0}\left[\log\left(p_\theta(x)^{1-\lambda} F_{\text{ref}}\left(p_\theta(x)\right)^\lambda\right)\right] = (1-\lambda)\mathbb{E}_{x\sim d_0}\left[\log p_\theta(x)\right] + \lambda\mathbb{E}_{x\sim d_0}\left[\log F_{\text{ref}}\left(p_\theta(x)\right)\right].$$

Notably, this utility functional is naturally meaningful as it is a monotonically increasing function of $p_\theta(x)$. Therefore, by applying the functional derivative in Definition 1, the optimal weight function and gradient update rule is:

$$\nabla_\theta J_{\text{integrated}}(\theta) = \mathbb{E}_{x\sim d_0}\left[\left((1-\lambda)\frac{1}{p_\theta} + \lambda\frac{f_{\text{ref}}\left(p_\theta(x)\right)}{F_{\text{ref}}\left(p_\theta(x)\right)}\right)\nabla_\theta p_\theta(x)\right].$$

Strategy 2: Multiplication. Alternatively, we can directly multiply the two function transformations ψ and g under a monotonic constraint to ensure a meaningful utility function over $p_\theta(x)$:

$$\begin{aligned}&\mathcal{U}_\theta = \mathbb{E}_{x\sim d_0}\left[\psi\left(F_{\text{ref}}\left(p_\theta(x)\right)\right) g\left(p_\theta(x)\right)\right],\\ &\text{s.t. } \frac{\delta U_\theta}{\delta p_\theta}(x) = \psi\left(F_{\text{ref}}\left(p_\theta(x)\right)\right) g'(p_\theta(x)) + \psi'\left(F_{\text{ref}}\left(p_\theta(x)\right)\right) f_{\text{ref}}\left(p_\theta(x)\right) g(p_\theta(x)) > 0.\end{aligned}$$

where the monotonic constraint is derived by ensuring a positive first-order derivative in terms of $p_\theta(x)$. When we employ $\psi(\mu) = -\log(\mu)$ and $g(\nu) = \log(\nu)$, it guarantees $\frac{\delta U_\theta}{\delta p_\theta}(x) > 0$. Consequently, the integrated gradient update rule is:

$$\nabla_\theta J_{\text{integrated}}(\theta) = \mathbb{E}_{x\sim d_0}\left[-\left(\frac{\log F_{\text{ref}}\left(p_\theta(x)\right)}{p_\theta(x)} + \frac{f_{\text{ref}}\left(p_\theta(x)\right)\log p_\theta(x)}{F_{\text{ref}}\left(p_\theta(x)\right)}\right)\nabla_\theta p_\theta(x)\right]. \tag{17}$$

Remark. Although the above integrated formulations provide a natural way to combine local sensitivity and distributional information, we do not observe empirical improvements in practice. This suggests that simply combining pointwise and distribution-aware terms does not necessarily lead to more effective learning signals due to the *geometry mismatch or gradient conflicts*. This observation demonstrates that the benefit of distribution-aware objectives does not arise from a direct additive or multiplicative fusion with pointwise terms, but rather from explicitly modeling the geometry of the pass-rate distribution. Therefore, it further helps to posit our contribution: purely distribution-aware utility functionals with clearer geometric interpretation are not considered as a regularization or supplement to pointwise ones, but correspond to a fundamentally different objective class that derives a new separate weight function family.

F.7 An Active-prompt Truncation

Firstly, active prompts should in principle be used to form the effective pass-rate distribution to evaluate the weight function. The non-active prompts have zero gradients and thus do not contribute to policy optimization. Therefore, in evaluating the reference pass-rate distribution, we focus on prompts with non-zero contribution to the policy optimization to form the effective pass-rate distribution. We also emphasize that the truncation is only applied to the reference pass-rate distribution to evaluate the weight function rather than serving as a prompt-filtering strategy within one batch like DAPO.

Secondly, as revealed in the nonparametric statistics literature, histogram-based density estimation can be unstable near the boundaries when the underlying empirical distribution contains substantial endpoint mass. This issue is particularly relevant in the early stage of RLVR, when many prompts have low population pass rates with $p_\theta(x) \approx 0$. Consequently, the weight computed in the first bin can become artificially large and dominate the normalization of the remaining active bins, resulting in unstable updates. The active-prompt truncation mitigates this finite-sample boundary instability. Its motivation is related to conventional boundary treatments in density estimators in nonparametric statistics, such as beta-smoothing in kernel density estimators, which are designed to improve stability near the endpoints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We first provide a framework to justify the principle and optimality of choosing prompt reweighting in Section 3. Next, we propose a distribution-aware weighing method in Section 4. Experiments are performed in Section 5 to substantiate the claims made in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide one separate paragraph in Section 6 that discusses the limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide detailed proof of each proposition and theorem in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided the full implementation details in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use the open-source implementation of the baseline method, MaxRL, and evaluate our algorithms on widely adopted benchmarks.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the implementation details in Appendix D and include details of the experimental setup for each empirical result in Section 5 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Following the standard in RLVR, we train our algorithm once in the standard training dataset, but we employ as many evaluation benchmarks as possible for reliable evaluation of the algorithm's performance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the information of computer resources in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses the pre-trained Qwen3 models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We mainly use LLM for editing (e.g., grammar, spelling, word choice), checking the proof, drafting the paragraph, and visualizing results for submission.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.