# General Exploratory Bonus for Optimistic Exploration in RLHF

**Wendi Li  Changdae Oh  Sharon Li**
Department of Computer Sciences
University of Wisconsin-Madison
{wli679, changdae, sharonli}@cs.wisc.edu

## Abstract

Optimistic exploration is central to improving sample efficiency in reinforcement learning with human feedback, yet existing exploratory bonus methods to incentivize exploration often fail to realize optimism. We provide a theoretical analysis showing that current formulations, under KL or $\alpha$-divergence regularization, unintentionally bias exploration toward high-probability regions of the reference model, thereby reinforcing conservative behavior instead of promoting discovery of uncertain regions. To address this pitfall, we introduce the **General Exploratory Bonus** (**GEB**), a novel theoretical framework that provably satisfies the optimism principle. GEB counteracts divergence-induced bias via reference-dependent reward regulation and unifies prior heuristic bonuses as special cases, while extending naturally across the full $\alpha$-divergence family. Empirically, GEB consistently outperforms baselines on alignment tasks across multiple divergence settings and large language model backbones. These results demonstrate that GEB offers both a principled and practical solution for optimistic exploration in RLHF. Code is available here.

## 1 Introduction

Despite the acknowledged significance of online exploration for reinforcement learning with human feedback (RLHF) [1, 2, 3], there remains a paucity of theoretical frameworks governing *how to explore*. As shown in Fig. 1 (1, top), standard online RLHF algorithms [4, 5, 6] generally rely on passive exploration, *i.e.*, the stochasticity of the policy itself to generate responses, with no mechanism to incentivize novelty or diversity. As a result, this approach can be notoriously sample-inefficient. When the optimal behavior resides in low-probability regions, passive exploration is unlikely to discover it, leading to policies that remain trapped around local optima.

To address this, some works [7, 8, 9, 10, 11] have attempted to devise sample-efficient algorithms, inspired by the principle *optimism in the face of uncertainty*. As illustrated in Fig. 1 (2, top), the principle aims to generate responses for regions of high epistemic uncertainty, thus encouraging data collection in unexplored areas for further training. To operationalize this, recent attempts [12, 13, 14] encourage exploration by adding *exploratory bonuses* to reward modeling, which is practically optimizeable for large language models. These methods intend to artificially inflate rewards in underexplored regions, nudging the policy toward more informative data collection.

Unfortunately, our theoretical analysis in Section 3 reveals a fundamental pitfall: under the common KL-regularized RLHF, the existing theoretical framework of exploratory bonuses fails to satisfy optimism. In particular, we prove that existing bonus formulations can undesirably drive the policy $\pi$ toward the reference policy $\pi_{\text{ref}}$ due to the divergence regulation in the exploratory bonus, and the induced bonus actually biases exploration toward high-probability regions of the reference model. As illustrated in Fig. 1 (II, bottom), the bonus disproportionately amplifies rewards for regions already well-covered by $\pi_{\text{ref}}$, thereby reinforcing conservative behavior rather than driving exploration into

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

uncertain regions. This failure is not confined to KL-divergence; we further extend our analysis to the more general $\alpha$-divergence family and prove that the same collapse persists across a wide range of divergence-regularized objectives. Thus, while existing approaches appear to encourage exploration, they in fact undermine the very principle of optimism they aim to realize.

Motivated by these failures, we propose a new framework, **General Exploratory Bonus (GEB)**, which theoretically unifies existing approaches while provably satisfying optimism (Section 4). GEB corrects the failure modes of prior approaches by directly introducing a reference-dependent regulation into the reward. This adjustment offsets the undesired conservatism induced by divergence regularization, allowing the exploratory bonus to satisfy optimism—it increases the probability of responses rarely sampled to pursue potentially more preferred answers, as shown in Fig. 1 (III, bottom). Importantly, GEB provides a unified formulation: prior heuristic exploratory bonuses can be reinterpreted as special cases, and the framework naturally extends to the full class of $\alpha$-divergences. Beyond correcting the theoretical shortcomings, GEB remains practically implementable—it can be seamlessly integrated into the standard iterative RLHF loop without additional sampling cost.

We validate GEB on a large-scale alignment task across different divergences and model backbones. Empirically, GEB consistently yields stronger alignment compared to its counterpart of passive exploration. For example, the three GEB variants that we consider generally outperform the iterative f-DPO [15] across different divergence regulations, while the most performant variant surpasses several existing optimistic exploration methods that incorporate exploratory bonuses [12, 13, 14]. By analyzing the distribution of sampled responses, we validate that GEB can successfully encourage sampling in the region of small $\pi_{\text{ref}}$, thereby effectively achieving optimistic exploration.

We summarize our main contributions:

1. We formally prove that the existing theoretical framework of exploratory bonuses under KL and $\alpha$-divergence regularization fails to achieve optimistic exploration.
2. We introduce General Exploratory Bonus (GEB), a novel theoretical framework of optimistic exploration for RLHF that provably satisfies the optimism principle and unifies prior heuristic bonuses.
3. We empirically validate GEB on LLM alignment tasks, showing improved performance and broad applicability across multiple divergence families.

## 2 Preliminaries

**Iterative online RLHF.** Let $x$ be a prompt sampled from a distribution $\rho$ and $y$ be a response given $x$, which is sampled from a policy $\pi(\cdot|x)$ modeled by a language model. We denote by $r(x, y)$ a real-valued reward model. An iterative online RLHF proceeds for rounds $T$, where each round $t = 1, ..., T$ has the following three steps: (i) The reward model $r_t$ is trained on the human preference dataset $\mathcal{D}_t = \{(x, y^w, y^l)\}$, where $y^w, y^l$ denote the preferred and dispreferred response to $x$; (ii) The policy $\pi_t$ is updated to maximize the reward $r_t(x, y)$ for responses $y \sim \pi_t(\cdot|x)$ conditioned on prompt $x$; and (iii) using the updated policy, we sample $\tilde{x} \sim \rho$, and generate multiple response pairs $(\tilde{y}_1, \tilde{y}_2) \sim \pi_t(\cdot|\tilde{x})$. Human evaluators then annotate these pairs to produce preference-labeled data $\{(\tilde{x}, \tilde{y}^w, \tilde{y}^l)\}$. The dataset for the next round is formed by $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\tilde{x}, \tilde{y}^w, \tilde{y}^l)\}$. For reward modeling step (i), we typically adopt the Bradley-Terry objective [16]:

$$r_t = \arg\min_r \mathcal{L}_{BT}(\mathcal{D}_t, r) = \arg\min_r \mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_t} - \log[\sigma(r(x, y^w) - r(x, y^l))], \tag{1}$$

where $\sigma$ denotes the sigmoid function. Next, in each step (ii), given the learned reward function $r_t$, the policy $\pi_t$ is updated to maximize the expected reward, often with a KL-regularization as follows

$$\pi_t = \arg\max_\pi \mathcal{J}_{\beta,\text{KL}}(\pi, r_t) = \arg\max_\pi \mathbb{E}_{x\sim\rho,y\sim\pi(\cdot|x)} r_t(x, y) - \beta\mathbb{D}_{\text{KL}}(\pi\|\pi_{\text{ref}}), \tag{2}$$

where $\beta > 0$ is a hyperparameter and $\pi_{\text{ref}}$ is the reference model. The effectiveness of iterative online RLHF [17, 18] has been validated in various real-world systems such as Claude [19] and LLaMA-series [20, 21], but there is still much room for improvement in terms of sample-efficient exploration.

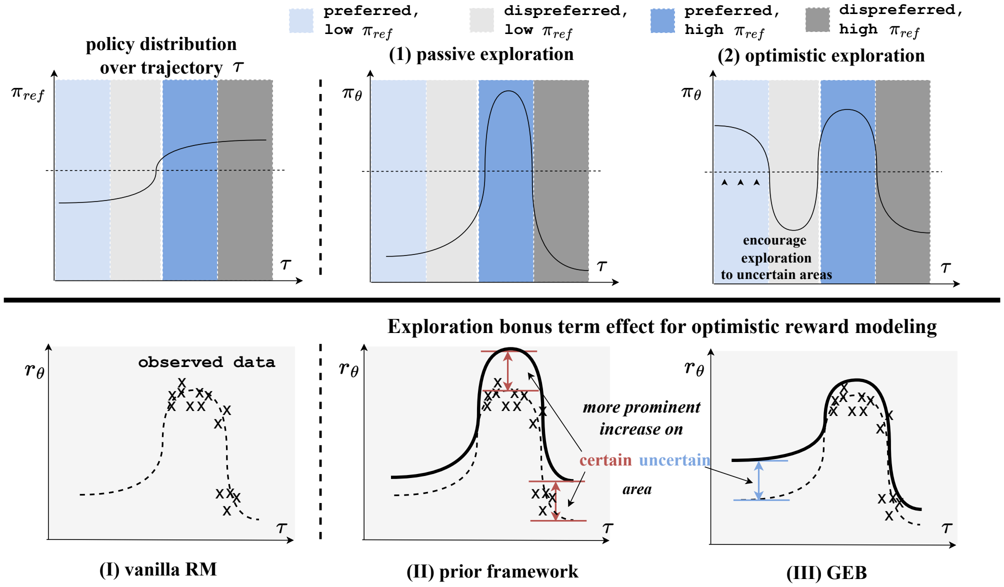
preferred, low $\pi_{ref}$
dispreferred, low $\pi_{ref}$
preferred, high $\pi_{ref}$
dispreferred, high $\pi_{ref}$
policy distribution over trajectory $\tau$
(1) passive exploration
(2) optimistic exploration
$\pi_{ref}$
$\pi_\theta$
$\tau$
encourage exploration to uncertain areas
Exploration bonus term effect for optimistic reward modeling
$r_\theta$
observed data
more prominent increase on
certain uncertain
area
(I) vanilla RM
(II) prior framework
(III) GEB

Figure 1: **The upper part** compares passive exploration and optimistic exploration. Optimistic exploration stimulates the trajectories $\tau$ of small $\pi_{\text{ref}}$ (seldom visited/uncertain), while passive exploration sticks to the high-$\pi_{\text{ref}}$ region, failing to approach global optima. The dashed line separates regions of high vs. low likelihood under the learning policy $\pi_\theta$. **The lower part** contrasts the effect of the exploration bonus term in optimistic reward modeling between prior works and our GEB. Prior works often emphasize rewards in frequently visited regions, which constrains exploration within certain areas. In contrast, our GEB amplifies rewards in seldom-visited regions, thereby encouraging further sampling in uncertain areas and successfully achieving optimistic exploration.

**Sample inefficiency of iterative online RLHF.** In online RLHF, standard online sampling is usually performed passively, relying solely on the LLM policy's inherent randomness. However, if the policy assigns a small probability to the optimal action, passive exploration may never explore it. Some recent theoretical analyses [22, 18] and empirical evidence [23, 14] present that the passive approach fails to sufficiently explore the prompt-response space. Particularly, Xie et al. [13] demonstrates that the sample complexity can be exponential in $1/\beta$ for passive exploration, which is unacceptable in the small-$\beta$ regime. After then, follow-up studies propose to implement the principle "optimism towards uncertainty" into RLHF algorithms, i.e., encourage exploration of uncertain trajectories. Several works on this try to estimate uncertainty by leveraging some uncertainty quantification techniques, such as elliptical potential [19], Bayesian modeling [10], and epistemic neural network training [11]. However, these methods are generally computationally prohibitive in LLM-scale settings. Therefore, recent works [13, 14, 12] propose *exploratory bonuses* for optimistic exploration, which can be computationally more tractable for LLM-based optimization.

## 3 Exploratory Bonus and How It Can Fail

In this section, we first provide the iterative online RLHF formulation with an exploratory bonus (Section 3.1). We then theoretically prove that the existing formulation can fail to achieve optimistic exploration under both KL-constrained RLHF (Section 3.2) and a more general $\alpha$-divergence-regularized RLHF (Section 3.3), motivating our proposed method in Section 4.

### 3.1 Exploratory Bonus

To improve the sample efficiency of iterative online RLHF, recent works [12, 14] introduce exploratory bonuses, which aim to encourage the policy model to explore the under-visited space given an optimistic reward estimation. These approaches modify the standard RLHF loop by adding an exploratory bonus term $\mathcal{L}_{\text{bonus}}$ in the reward modeling phase. Specifically, in the $t$-th iteration, the reward model $r_t$ and policy $\pi_t$ are optimized by

$$r_t = \arg\min_r \Big[\mathcal{L}_{BT}(\mathcal{D}_t, r) - \kappa \mathcal{L}_{\text{bonus}}(r)\Big], \tag{3}$$

$$\pi_t = \arg\max_\pi \mathcal{J}_{\beta,\text{KL}}(\pi, r_t) = \arg\max_\pi \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)} r_t(x,y) - \beta \mathbb{D}_{\text{KL}}(\pi \| \pi_{\text{ref}}), \tag{4}$$

where $\kappa > 0$ is a hyperparameter. By Eq. 3, the reward model $r_t$ should not only fit the observed data in $\mathcal{D}_t$, but also learn to maximize the bonus term $\mathcal{L}_{\text{bonus}}(r)$.

To boost exploration, the bonus term is designed to amplify the probability mass of policy more in underexplored areas rather than incentivizing it solely towards high empirical reward areas. As mentioned in § 2, early works on RLHF optimistic exploration are computationally prohibitive in the LLM fine-tuning regime. Thus, it is necessary to set a new approach that not only aligns with the principle of optimism in the face of uncertainty but is also cost-effective. For this, we derive a new condition for the exploration bonus to achieve optimism, avoiding direct uncertainty quantification:

**Definition 3.1 (Optimism condition for exploration bonus)** *Given an input prompt $x$ and a response $y$, when a reward model $r$ and a policy $\pi$ are computed with Eq. 3 and Eq. 4, respectively, the exploratory bonus $\mathcal{L}_{bonus}$ achieves optimism, if*

$$\frac{\partial}{\partial \pi_s(y|x)}\left(\frac{\partial \mathcal{L}_{bonus}(r(x,y))}{\partial \pi(y|x)}\right) < 0, \tag{5}$$

*where $\pi_s$ is a typical sampling policy, a joint policy on all iterations up to the current iteration.*

Specifically, at the $t$-th iteration, the typical sampling policy $\pi_s = \pi_1 \circ \pi_2 \circ \cdots \circ \pi_t$ is a joint distribution of all previous policies up to the current iteration. This distribution is not directly computable; rather, it serves as a theoretical construct describing how responses in $\mathcal{D}_t$ are generated. In Eq. 5, rather than characterizing $\mathcal{L}_{\text{bonus}}$ in its original function space, we define it by a condition on its partial derivatives with respect to two policies: current policy and typical policy. This new optimism condition not only enables us to flexibly define $\mathcal{L}_{\text{bonus}}$, but also serves as a core tool for theoretically analyzing existing methods. Although $\mathcal{L}_{\text{bonus}}$ appears unrelated to the current policy $\pi$—as it is defined in terms of the reward model $r(x, y)$—**the policy-reparameterized reward $r_\pi(x, y)$ allows us to express $r(x, y)$ directly in terms of $\pi$** as follows [24]:

$$r(x,y) := r_\pi(x,y) = \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x), \quad \text{where } Z(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}} \exp(r(x,y)/\beta).$$

This is derived from the closed-form solution of the proximal preference optimization (Eq. 2) as $\pi(y|x) = \frac{\exp(r(x,y)/\beta)}{Z(x)}$. Thanks to this alternative representation of the reward, we can express and interpret the bonus term with respect to our current policy $\pi$, which yields the following implication.

**Implication.** Our new optimism condition requires the derivative of the bonus term with respect to the current policy, $\partial\mathcal{L}_{\text{bonus}}(r_\pi(x, y))/\partial\pi(y|x)$, to be negatively correlated with the typical policy $\pi_s$. In other words, as a response $y$ is likely rare sample under $\pi_s$ (i.e., uncertain or underexplored responses), it should receive a larger ascending force in the policy distribution $\pi$, i.e., a higher $\partial\mathcal{L}_{\text{bonus}}(r_\pi(x, y))/\partial\pi(y|x)$. This new definition of the optimism principle, which is specified through the lens of partial derivative alignment, ensures the exploratory bonus nudges the exploration towards an uncertain response region without explicit uncertainty quantification. In practice, $\pi_s$ can be substituted by the reference model policy $\pi_{\text{ref}}$ or intermediate checkpoints of $\pi$ across iterations. We adopt the commonly used $\pi_{\text{ref}}$ as $\pi_s$ in our following demonstration.

### 3.2 Failure Under KL-constrained RLHF

Previous works, including Zhang et al. [12] and Cen et al. [14], formulate the exploratory bonus with $\mathcal{L}_{\text{bonus}}(r) = \max_\pi \mathcal{J}_{\beta,KL}(\pi, r)$. Under this formulation, optimizing the exploratory bonus in Eq. 3 yields a min–max bilevel objective: $\min_r -\kappa \max_\pi [\mathbb{E}_{x\sim\rho, y\sim\pi} r(x, y) - \beta\mathbb{D}_{\text{KL}}(\pi\|\pi_{\text{ref}})]$. Intuitively, this objective encourages $r$ not only to fit the observed data via $\mathcal{L}_{BT}$ but also to assign high reward to unobserved regions by maximizing $\max_\pi \mathbb{E}_{x\sim\rho, y\sim\pi} r(x, y)$ in $\mathcal{L}_{\text{bonus}}(r)$. Here, we theoretically show that such formulations can suffer from optimism failures under KL-regularized RLHF.

**Lemma 3.1 (Optimism failure under KL-divergence.)** *Let $r_1 = \arg\min_r \mathcal{L}_{BT}(\mathcal{D}, r)$ be a reward model trained with the vanilla BT loss, and let $r_2 = \arg\min_r[\mathcal{L}_{BT}(\mathcal{D}, r) - \kappa \max_\pi \mathcal{J}_{\beta,KL}(\pi, r)]$ be a reward model trained with an additional exploratory bonus. If the policy is optimized via Eq. 4, then $r_1$ and $r_2$ yield the same set of policies.*

See the proof in Appendix B.1. The lemma shows that incorporating the exploratory bonus $\mathcal{L}_{\text{bonus}}(r) = \max_\pi \mathcal{J}_{\beta,\text{KL}}(\pi, r)$ into the reward training objective ***fails to induce the policy to sample from low-$\pi_{ref}(y|x)$ regions, i.e., unexplored responses***. That is, $\mathcal{L}_{\text{bonus}}$ is ineffective for optimism. We next extend the result beyond KL divergence to a more general class of $\alpha$-divergence families.

### 3.3 Generalization to $\alpha$-divergence-constrained RLHF

In this subsection, we theoretically show that the failure of optimism can broadly be extended to the $\alpha$-divergence class. Many common divergences, such as reverse KL-divergence, Hellinger distance, and forward KL-divergence, are special cases of $\alpha$-divergence.

**Definition 3.2 ($\alpha$-divergence class)** *An $\alpha$-divergence is a certain type of function $D(p|q) = \int f(\frac{\mathrm{d}p}{\mathrm{d}q})\mathrm{d}q$ that measures the difference between two probability distributions $p$ and $q$, where*

$$f(x) = \frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(1-\alpha)},$$

*and $\alpha$ is a hyperparameter typically with $0 \leq \alpha \leq 1$.*

**Lemma 3.2 (Optimism failure under $\alpha$-divergence.)** *Consider an objective $\mathcal{J}_{\beta,f}(\pi, r) = \mathbb{E}_{x\sim\rho, y\sim\pi(y|x)} r(x,y) + \beta\mathbb{E}_{x\sim\rho, y\sim\pi_{ref}(y|x)} f(\frac{\pi(y|x)}{\pi_{ref}(y|x)})$, where $f$ belongs to $\alpha$-divergence class. If a reward is trained with $\hat{r} = \arg\min_r[\mathcal{L}_{BT}(\mathcal{D}, r) - \kappa\mathcal{L}_{bonus}]$ and a policy $\pi$ is updated by $\arg\max_\pi \mathcal{J}_{\beta,f}(\pi, \hat{r})$ with $\mathcal{L}_{bonus} = \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$, the gradient of the bonus satisfies $\frac{\partial^2 \mathcal{L}_{bonus}(r_\pi)}{\partial\pi_{ref}\partial\pi} \geq 0$, which means $\mathcal{L}_{bonus}$ encourage trajectories with large $\pi_{ref}$ more strongly, in contradiction to the optimism principle (Definition 3.1).*

*Proof* For a RL objective $\mathcal{J}_{\beta,f}(\pi, r)$, the relation between the optimal policy $\pi^*_f$ and the reward $r$ can be formulated as follows,

$$\pi^*_f(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)(f')^{-1}(r(x,y)/\beta), \quad r_\pi(x,y) = \beta f'(\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)}Z(x)), \tag{6}$$

where $Z(x)$ is a normalization term and $(f')^{-1}$ is the inverse function of $f'$. The bi-level objective can be similarly transformed to a single level one by canceling the inner maximization $\max_\pi$ by Eq. 6. The single-level objective can be written as $r_t = \arg\min_r \mathcal{L}_{BT}(\mathcal{D}, r) - \kappa\mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}} \frac{1}{Z(x)}(f')^{-1}(\frac{r(x,y)}{\beta}) \cdot r(x,y) - \beta f(\frac{1}{Z(x)}(f')^{-1}(\frac{r(x,y)}{\beta}))$. Since the policy is computed by $\arg\max_\pi \mathcal{J}_{\beta,f}(\pi, r)$, the reward can be reparameterized by the policy with Eq. 6, which fortunately cancels $Z(x)$. Then, the optimistic reward-modeling objective can be reparameterized as

$$\arg\min_\pi \mathcal{L}_{dpo}(\mathcal{D}, \pi) - \kappa\beta\mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}}\Big[\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} f'(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}) - \beta f(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)})\Big]. \tag{7}$$

Since for $\alpha$-divergence, $f(u) = \frac{u^\alpha - \alpha u - (1-\alpha)}{\alpha(\alpha-1)}$, the partial derivative of Eq. 7 is $(\frac{\pi_{\text{ref}}}{\pi})^{1-\alpha}$, which induces positively correlated gradients w.r.t. $\pi$ and $\pi_{\text{ref}}$ when $0 \leq \alpha < 1$, and is a constant when $\alpha = 1$, hence contradictory to the optimism defined in Definition 3.1. □

According to Lemma 3.2, we summarize several forms of exploratory bonus induced by different $\alpha$-divergences in Table 1. For clarity, these expressions are presented after removing constant coefficients and additive biases. In every case, the resulting bonus encourages the policy to place more probability mass on responses that the reference model already samples frequently, rather than on underexplored responses. Now, we further prove that it actually drives $\pi$ to collapse toward $\pi_{\text{ref}}$ and that the failure extends beyond $\alpha$-divergence to other $f$-divergences.

Table 1: Realized exploratory bonus under different divergence classes when $\mathcal{L}_{\text{bonus}}(r) = \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$.

| $f$ | exploratory bonus |
|---|---|
| reverse KL | constant |
| forward KL | $\mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}} \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ |
| Hellinger distance | $\mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}} \sqrt{\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}}$ |

**Theorem 3.3 (Optimism failure beyond $\alpha$-divergence.)** *When $f$ belongs to $f$-divergence, and the reward function is obtained by $\hat{r} = \arg\min_r[\mathcal{L}_{BT}(\mathcal{D}_t, r) - \kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)]$ and the policy is updated by $\arg\max_\pi \mathcal{J}_{\beta,f}(\pi, \hat{r})$, the bonus term $-\kappa\max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ induces the policy model $\pi$ to coincide with $\pi_{ref}$ when $xf''(x)$ is a monotone function.*

The monotonic increase of $xf''(x)$ can be satisfied by a broader divergence class beyond $\alpha$-divergence, including JS-divergence and Pearson $\chi^2$. Please see the detailed proofs in Appendix B.2.

**Intuitive understanding.** The optimization of the exploratory bonus in Eq. 3 is a min-max bi-level objective, $\min_r -\kappa \max_\pi [\mathbb{E}_{x\sim\rho,y\sim\pi} r(x,y) - \beta \mathbb{D}_{\text{KL}}(\pi \| \pi_{\text{ref}})]$. Due to inner maximization $\max_\pi$, the divergence constraint implicitly makes $\pi$ close to $\pi_{\text{ref}}$ to reduce the KL divergence. Meanwhile, outer minimization $\min_r$ forces $r$ to provide high rewards in the region of high $\pi$ to maximize the expected reward. Their combination implicitly makes $r$ focus more on the region of high $\pi_{\text{ref}}$. As responses in high $\pi_{\text{ref}}$ area are easily sampled from scratch, prior exploratory bonuses just concentrate sampling on regions that are already frequently visited, *contradictory to the optimism principle, which requires encouraging exploration for responses $y$ rarely sampled by the reference model.*

## 4 General Exploratory Bonus with Optimism Principle

Motivated by the failure of the existing optimistic exploration works, we now propose a novel framework, *General Exploratory Bonus* (**GEB**), and prove that it achieves optimism. We further show that prior heuristic bonuses—and their broader variants—emerge as special cases of our formulation.

**Formulation of a novel exploratory bonus.** As shown in the previous section, existing bonus schemes fail because the divergence constraints in $\max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ force the optimal policy $\pi$ to remain close to $\pi_{\text{ref}}$, thereby biasing exploration toward regions where $\pi_{\text{ref}}$ is large. Achieving optimistic exploration requires the optimal $\pi$ to counteract this effect and deviate from $\pi_{\text{ref}}$. *Our key idea is therefore to incorporate an additional $\pi_{ref}$-dependent term into the reward that offsets the influence of the divergence regularization.* The resulting exploratory bonus takes the form

$$\mathcal{L}_{\text{bonus}} = \max_\pi J_{\beta,f}(\pi, R), \tag{8}$$

where our new reward formulation, $R$, now depends not only on the original reward model $r(x,y)$ but also on $\pi_{\text{ref}}(y|x)$. Now, the optimal policy of $\max_\pi J_{\beta,f}(\pi, R(x,y))$ can be obtained by replacing $r(x,y)$ with $R(x,y)$ in Eq 6. This yields $\pi^*(y|x) = \frac{1}{Z_R(x)} \pi_{\text{ref}} (f')^{-1}(\frac{R(x,y)}{\beta})$, where $Z_R(x)$ is a normalization term.

Following Lemma 3.2, we substitute $\pi(y|x)$ in $\max_\pi J_{\beta,f}(\pi, R(x,y))$ with its optimal form $\pi^*(y|x)$ and then apply the reward reparameterization trick for $\alpha$-divergences [25], i.e., $r(x,y) = f'(\pi^*(y|x)/\pi_{\text{ref}}(y|x))$, where $f$ specifies the divergence. Given this policy-reparameterized reward, we specify the exploratory bonus term in Eq 8 as follows:

$$\mathcal{L}_{\text{bonus}} = \beta \mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}(\cdot|x)} \Big[ \frac{u(x,y)}{Z_R(x)} f'(u(x,y)) - f(\frac{u(x,y)}{Z_R(x)}) \Big]. \tag{9}$$

In Eq. 9, we introduced $u(x,y)$ as an atomic function employed to construct the actual loss, and it is given by $u(x,y) = (f')^{-1}(R(x,y)/\beta)$ in this setup. Note that, after reward reparameterization, $u(x,y)$ can be expressed in terms of $\pi(y|x)$ and $\pi_{\text{ref}}(y|x)$. Moreover, as the functional form of $R(x,y)$ is not restricted to a specific class, $u(x,y)$ can be instantiated in many ways using these two distributions, while it must satisfy $u(x,y) > 0$ for all $x, y$ to ensure that the argument of $f'(\cdot)$ lies within its domain. See Table 2 for the example entries we are considering in this work.

**Equivalence to a practical objective.** In our proposed exploratory bonus, the normalization term $Z_R(x)$ in Eq. 9 cannot be eliminated. Fortunately, Lemma 4.1 (proved in Appendix B.4) shows that the training objectives with and without $Z_R(x)$ are equivalent. This equivalence allows us to convert the objective into a more concise form, facilitating both analysis and practical implementation.

**Lemma 4.1** *Denote two objectives as $h(u(x,y)) = \mathbb{E}_{x\sim\rho,y\sim\pi_{ref}} u(x,y) f'(u(x,y)) - f(u(x,y))$ and $\hat{h}(u(x,y)) = \mathbb{E}_{x\sim\rho,y\sim\pi_{ref}} \frac{u(x,y)}{Z_R(x)} f'(u) - f(\frac{u(x,y)}{Z_R(x)})$ where $u(x,y)$ is a function with $\pi(y|x)$ and $\pi_{ref}(y|x)$. If the ratio*

$$\frac{f'(u(x,y)) + u(x,y) f''(u(x,y)) - f'(\frac{u(x,y)}{Z_R(x)})}{Z_R(x) u(x,y) f''(u(x,y))} = \Lambda(x) \tag{10}$$

Table 2: GEB under different divergence classes and design of $u$. Note that (1) now the bonus term can be computed without the reference probability mass $\pi_{\text{ref}}$; (2) all of these $u$ instantiations meet the condition $u > \alpha$ when $0 < \pi < 1$. The presented bonuses are simplified by removing constants.

| $\mathcal{L}_{\text{bonus}}$ $u$ <br> $f$ | $1 + \alpha - \pi$ | $1/\pi$ | $\text{arctanh}(1-\pi) + \alpha$ |
|---|---|---|---|
| reverse KL | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} - \pi(y\|x)$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \frac{1}{\pi(y\|x)}$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \text{arctanh}(1-\pi(y\|x))$ |
| forward KL | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \log(1-\pi(y\|x))$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} - \log\pi(y\|x)$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \log\text{arctanh}(1-\pi(y\|x))$ |
| Hellinger Distance | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \sqrt{1.5-\pi(y\|x)}$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \frac{1}{\sqrt{\pi(y\|x)}}$ | $\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}} \sqrt{\text{arctanh}(1-\pi(y\|x)) + 0.5}$ |

*is independent of $y$ and $\Lambda(x) > 0$, then minimizing the two objectives, $\min_\pi -h(u(x,y))$ and $\min_\pi -\hat{h}(u(x,y))$, yields the same class of optimal policies.*

Note that the $\alpha$-divergence ($0 \le \alpha \le 1$; see Def. 3.2) naturally satisfies the condition in Eq. 10 whenever $u(x,y) > \alpha$. As we show in the next paragraph, enforcing $u(x,y) > \alpha$ is straightforward in practice, which grants our framework substantial flexibility and extensibility. Leveraging Lemma 4.1, we can thus rewrite our objective in Eq. 9 into a concise and analytically convenient form without the normalization term:

$$\mathcal{L}_{\text{bonus}} = \beta\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}}\Big[u(x,y)f'(u(x,y)) - f(u(x,y))\Big], \tag{11}$$

where $u(x,y)$ is flexibly formulated by $\pi(y|x)$ and $\pi_{\text{ref}}(y|x)$ satisfying $u(x,y) > \alpha$.

**GEB successfully achieves optimism.** Building on Lemma 4.1, we now show that our proposed framework achieves the optimism condition in Definition 3.1 (See Appendix B.5 for the proof).

**Theorem 4.2** *Consider an $\alpha$-divergence $f$ with $0 \le \alpha \le 1$, and the exploratory bonus $\mathcal{L}_{bonus} = \beta\mathbb{E}_{x\sim\rho,y\sim\pi_{ref}}\Big[u(x,y)f'(u(x,y)) - f(u(x,y))\Big]$, where $u(x,y)$ is a function dependent on $\pi(y|x)$ and $\pi_{ref}(y|x)$. For any $(x,y)$, if $\frac{\partial u}{\partial \pi} + \pi_{ref}\frac{\partial^2 u}{\partial\pi\partial\pi_{ref}} + \frac{(\alpha-1)\pi_{ref}}{u}\frac{\partial u}{\partial \pi}\frac{\partial u}{\partial \pi_{ref}} < 0$ and $u(x,y) > \alpha$, the optimism condition in Definition 3.1 is satisfied; that is, $\frac{\partial^2 \mathcal{L}_{bonus}}{\partial\pi\partial\pi_{ref}} \le 0$.*

In our formulation, $u(x,y)$ can be flexibly defined in terms of $\pi(y|x)$ and $\pi_{\text{ref}}(y|x)$ as long as it satisfies the derivative condition in Theorem 4.2 and $u(x,y) > \alpha$. In particular, when $u(x,y)$ depends solely on $\pi(y|x)$ and is independent of $\pi_{\text{ref}}(y|x)$, any function that is strictly decreasing in $\pi$ with $u(x,y) > \alpha$ constitutes a valid choice. This design flexibility underscores the extensibility of our framework. In Table 2, we list several such choices of $u$, along with their corresponding reparameterized exploratory bonus terms under three different $\alpha$-divergences. From a practical standpoint, since $\mathcal{L}_{\text{bonus}}$ is computed as an expectation over $\pi_{\text{ref}}(\cdot|x)$, it does not require additional sampling and can be seamlessly integrated into iterative online RLHF. Meanwhile, to avoid unintended decreases in the likelihood of preferred responses, we follow Chen et al. [23] and restrict the computation of the bonus on rejected responses to ensure that the probability of preferred responses continues to increase.

**Prior exploratory bonuses are encompassed within GEB.** Although we have shown that existing theoretical formulations of $\mathcal{L}_{\text{bonus}}$ fail to guarantee optimism, many practical implementations have nevertheless been effective through various approximations and adaptations. These approximations and adaptations are generally inextensible beyond the reverse KL divergence (detailed in Appendix B.3). In this paragraph, we show that these practical implementations can be naturally subsumed into our GEB framework, and even broader objectives can be reinterpreted as instances of optimistic exploration. For example, Zhang et al. [12] and Xie et al. [13] finally implement their exploratory bonus as $\kappa\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{ref}}(y|x)}\log\pi(y|x)$, which belongs to GEB when $u = -\log\pi + 1$ and $f$ is KL-divergence. Similarly, Cen et al. [14] implement the exploratory bonus as $\kappa\mathbb{E}_{x\sim\rho,y\sim\pi_{\text{cal}}(\cdot|x)}\log\frac{\pi}{\pi_{\text{ref}}}$ where $\pi_{\text{cal}}$ is a fixed calibration distribution. This also falls under GEB by setting $u = -\frac{\pi_{\text{cal}}}{\pi_{\text{ref}}}\log\frac{\pi}{\pi_{\text{ref}}} - \frac{\pi_{\text{cal}}}{\pi_{\text{ref}}}\log\pi_{\text{ref}} + 1$ and $f$ is KL-divergence. Interestingly, even objectives not explicitly designed for exploration can be reinterpreted through our GEB framework. For instance, Chen et al. [23] augment the DPO loss with an additional term $\kappa\mathbb{E}_{x,y\sim\pi_{ref}}\sigma(-\beta\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)})$, which was originally introduced to control sample complexity. In our framework, this corresponds to optimistic exploration with $u = -\sigma(-\beta\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}) + 1$.

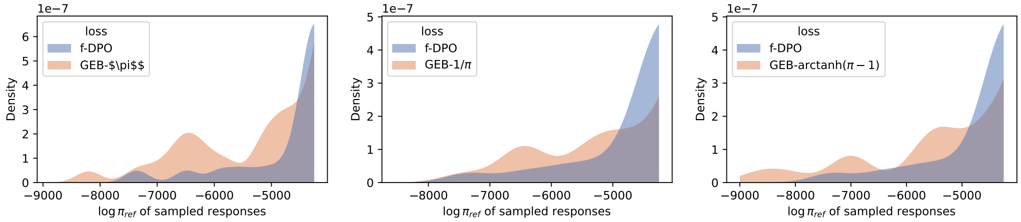

Figure 2: Comparison of $\log \pi_{\text{ref}}$ of sampled response in the last iteration between the general exploratory bonuses and vanilla iterative DPO. GEB-$\pi$, GEB-$1/\pi$, and GEB-$\operatorname{arctanh}(\pi - 1)$ corresponds to $1 + \alpha - \pi$, $1/\pi$, and $\operatorname{arctanh}(1 - \pi) + \alpha$ as in Table 2

**Takeaways.** In summary, we introduce a general formulation of the exploratory bonus term, GEB, which showcases the following unique strengths: (1) In contrast to the prior bonus terms, GEB meets the optimism condition theoretically, thus provably boosts exploration on relatively untapped region; (2) GEB spans a broad class of instantiations as shown in Table 2, specified by the combination of the $\alpha$-divergence class and the functional form of $u$, which provides a plenty options for practitioners to choose on demand; (3) GEB is practical as it does not incur additional sampling costs and can be seamlessly integrated into the existing iterative online RLHF framework (Appendix D.1); and (4) GEB offers an unified understanding of existing methods by generalizing them as special cases in a single flexible formulation.

## 5 Experiments

### 5.1 Experimental Settings

Following prior works [12, 13, 23], we adopt the same iterative online algorithm as in Algorithm 1 with three iterations, aiming to isolate the effects of different exploration bonuses. We adopt two LLM backbones: Llama-3-8B-SFT [18] following prior works, and Mistral-Instruct-v0.3 [26]. The training prompt set is RLHFlow-UltraFeedback [18] as in previous works. URM-LLaMa-3.1-8B [27] serves as the preference oracle. We evaluate the outcome policies on both in-domain and out-of-domain test sets. Specifically, for the in-domain test, we use a held-out test set from UltraFeedback [28], and sample 64 times per prompt with the outcome policy to compare the average reward and win-rates against the base model. We use length-controlled AlpacaEval2 benchmark [29] with GPT-4 as a judge for out-of-domain alignment test, and MATH-500 [30] to evaluate out-of-domain reasoning ability.

**Baselines.** We adopt f-DPO [25], which extends DPO to the f-divergence class, as the primary baseline. We further compare GEB with three optimistic-exploration methods that incorporate exploratory bonuses—SELM [12], XPO [13], and VPO [14]. Since the approximations or adaptations in their implementations do not extend beyond the KL divergence, we report their results only under KL. In contrast, we introduce a new baseline, Failed Exploratory Bonus (FEB), which removes these approximations or adaptations, i.e., Eq. 13.

### 5.2 Results & analyses

**GEB delivers robust improvements across different loss designs, divergence classes, and language model backbones.** The experimental results are shown in Table 3. Across both backbones, GEB generally outperforms f-DPO and FEB. Under the KL-divergence, GEB displays better or at least on-par performance compared to prior exploratory-bonus methods. Notably, the win-rate increases over 1.82% and 0.94% under the KL-divergence, over 2.36% and 1.29% under the Hellinger Distance, compared with their f-DPO counterpart. GPT-4 evaluation on the Alpaca benchmark also shows consistent performance gains on the out-of-domain alignment task. While GEB maintains on par, or usually better results in MATH, showing less performance degradation beyond alignment, known as alignment tax [31, 32].

**GEB effectively encourages exploration in small $\pi_{\text{ref}}$ region, yielding more diverse sampling.** In Figure 2, we visualize the distribution of $\log \pi_{\text{ref}}$ for sampled responses in the last iteration under the

Table 3: In-domain evaluation on different exploration bonuses. **Boldface** and underline indicate the best and the second-best results, respectively. GEB-$\pi$, GEB-$1/\pi$, and GEB-$\mathrm{arctanh}(\pi-1)$ corresponds to $1+\alpha-\pi$, $1/\pi$, and $\mathrm{arctanh}(1-\pi)+\alpha$ as in Table 2.

| | KL ($\alpha$=1) | | Hel. ($\alpha$=0.5) | | f-KL ($\alpha$=0) | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | WR | AvgR | WR | AvgR | WR | AvgR | WR | AvgR |
| *Mistral-Instruct-v0.3* | | | | | | | | |
| f-DPO | 78.42 | 0.7480 | 72.69 | 0.6536 | 51.11 | 0.5918 | 67.40 | 0.6645 |
| SELM | 77.56 | 0.7530 | - | - | - | - | - | - |
| XPO | 79.71 | 0.7492 | - | - | - | - | - | - |
| VPO | 78.57 | 0.7426 | - | - | - | - | - | - |
| FEB | 78.42 | 0.7480 | 71.54 | 0.6525 | 47.53 | 0.5928 | 65.83 | 0.6644 |
| GEB-$\pi$ | **81.00** | 0.7542 | 75.48 | **0.6641** | 51.68 | 0.5976 | **69.39** | 0.6720 |
| GEB-$1/\pi$ | 80.00 | 0.7554 | 73.97 | 0.6541 | 52.26 | **0.6051** | 68.74 | 0.6715 |
| GEB-$\mathrm{arctanh}(\pi-1)$ | 79.71 | **0.7559** | **75.69** | 0.6614 | **52.76** | 0.5989 | **69.39** | **0.6721** |
| *LLaMA-3-8B-SFT* | | | | | | | | |
| f-DPO | 73.11 | 0.8050 | 71.11 | 0.7859 | 67.38 | 0.7579 | 70.53 | 0.7829 |
| SELM | 74.19 | 0.8126 | - | - | - | - | - | - |
| XPO | 72.40 | 0.8119 | - | - | - | - | - | - |
| VPO | 71.61 | 0.7971 | - | - | - | - | - | - |
| FEB | 73.11 | 0.8050 | 68.17 | 0.7591 | 67.95 | 0.7611 | 69.74 | 0.7751 |
| GEB-$\pi$ | 74.34 | **0.8156** | 71.68 | 0.7840 | 67.67 | **0.7681** | 71.23 | **0.7892** |
| GEB-$1/\pi$ | 74.76 | 0.8102 | 72.25 | 0.7859 | 68.17 | 0.7591 | 71.73 | 0.7851 |
| GEB-$\mathrm{arctanh}(\pi-1)$ | **74.98** | 0.8080 | **73.26** | **0.7877** | **68.89** | 0.7569 | **72.38** | 0.7842 |

Table 4: Out-of-domain evaluation on different exploration bonuses with LLaMA-3-8B-SFT. **Boldface** and underline indicate the best and the second-best results, respectively. GEB-$\pi$, GEB-$1/\pi$, and GEB-$\mathrm{arctanh}(\pi-1)$ corresponds to $1+\alpha-\pi$, $1/\pi$, and $\mathrm{arctanh}(1-\pi)+\alpha$ as in Table 2.

| | KL($\alpha$=1) | | Hel.($\alpha$=0.5) | | f-KL($\alpha$=0) | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | Alpaca | Math | Alpaca | Math | Alpaca | Math | Alpaca | Math |
| f-DPO | 25.72 | 67.6 | 24.73 | 69.0 | 17.80 | 69.2 | 22.75 | 68.6 |
| FEB | 25.72 | 67.6 | 23.75 | 68.6 | 19.62 | 68.6 | 23.03 | 68.3 |
| GEB-$\pi$ | **28.27** | 69.2 | 25.87 | 69.6 | **20.05** | **71.6** | **24.73** | **70.1** |
| GEB-$1/\pi$ | 26.10 | 68.4 | 25.28 | **70.2** | 19.80 | 69.2 | 23.73 | 69.3 |
| GEB-$\mathrm{arctanh}(\pi-1)$ | 24.90 | **71.0** | **25.96** | 67.6 | 19.62 | 69.2 | 23.49 | 69.3 |

Table 5: Dist-n of the sampled corpus in the last iteration under the KL divergence.

| | dist-1 | dist-2 | dist-3 | dist-4 |
|---|---|---|---|---|
| f-DPO | 0.0189 | 0.2700 | 0.6349 | 0.8418 |
| GEB-$\pi$ | 0.0192 | 0.2694 | 0.6323 | 0.8420 |
| GEB-$1/\pi$ | 0.0191 | 0.2738 | 0.6401 | 0.8448 |
| GEB-$\mathrm{arctanh}(\pi-1)$ | 0.0192 | 0.2730 | 0.6391 | 0.8447 |

KL divergence. When trained with the GEB, the policy model consistently samples more trajectories with a smaller $\pi_{\text{ref}}$ compared to the policy trained by f-DPO loss. This validates our motivation that GEB can encourage sampling trajectories of small $\pi_{\text{ref}}$ for optimistic exploration. In Table 5, we further calculate the distinct-n ($n = 1, 2, 3, 4$) for the sampled responses in the last iterations under the KL divergence, which measures the diversity of a corpus. GEB generally has higher diversity scores, validating that GEB incentivizes qualitatively more diverse samples.

**The choice of $\kappa$.** Since the formulation of $u$ in Eq. 9 is flexible, the scale of the GEB term can differ substantially across designs, hence the absolute value of the bonus is less informative. Instead, we examine the relative ratio of the bonus term to the vanilla RL loss $|\kappa\mathcal{L}_{bonus}|/|\mathcal{L}_{RL}|$, which provides a more consistent basis for comparison and offers better practical guidance for tuning $\kappa$ across diverse settings. As shown in Fig. 3, performance remains stable when the ratio lies within a suitable range

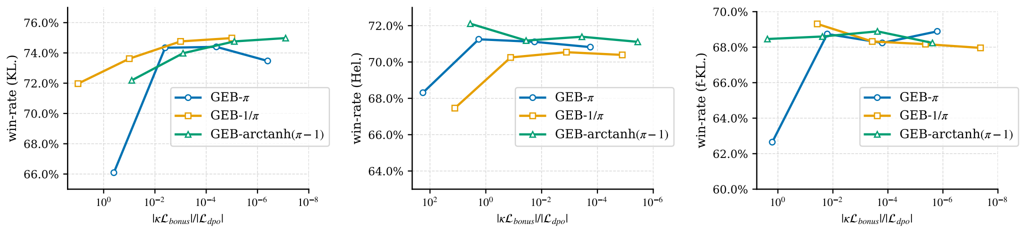

Figure 3: Experiments with different $\kappa$. The three graphs are under KL divergence, Hellinger Distance, and forward KL divergence from left to right, respectively. The p, f, tanh in the legends correspond to $1 + \alpha - \pi$, $1/\pi$, $\text{arctanh}(1 - \pi) + \alpha$ in Table 2 respectively.

Table 6: Ablation study on the sample responses used for the bonus term calculation. We compare the results on UltraFeedback (win rate) across three divergences with all responses (all) and with the rejected responses only (rejected-only).

| Method | KL all | KL rejected-only | Hel. all | Hel. rejected-only | fKL all | fKL rejected-only |
|---|---|---|---|---|---|---|
| GEB-$\pi$ | 79.78 | **81.00** | 73.40 | **75.48** | 51.32 | **51.68** |
| GEB-$\frac{1}{\pi}$ | 79.56 | **80.00** | 72.25 | **73.97** | 51.61 | **52.26** |
| GEB-arctan$\pi$ | 79.64 | **79.71** | 73.11 | **75.69** | 51.25 | **52.76** |

(1e-2 to 1e-6 in our case). However, if the ratio is too large, it impedes optimization of the RL objective and degrades performance; if too small, the exploration incentive in uncertain regions diminishes and performance reverts to the vanilla baseline.

**Restricting the exploratory bonus to rejected responses.** Restricting the bonus term to rejected responses is a common practice in prior works on exploratory bonuses [12, 23, 14]. Importantly, this restriction does not constitute a theoretical departure from our framework. The optimism guarantee in Theorem 4.2 hinges on increasing the probability of low $\pi_{\text{ref}}$ regions, i.e., underexplored regions. Because rejected samples lie precisely in these low-probability areas, applying the bonus only to rejected responses preserves the intended optimism direction. While preserving the theoretical guarantee on the optimism, it also shows a practical advantage as shown in Table 6.

**Semantic coherence of the sampled responses.** We use GPT-4 to evaluate whether a given sentence is coherent, nonsensical, or contains meaningless content. The scoring scale ranges from 0 (fully coherent) to 3 (complete nonsense with no coherent meaning). We apply this evaluation to the responses generated in the final training iteration of DPO and the three GEB variants. The resulting scores are as follows.

Table 7: Semantic coherence scores of responses produced by policy models trained using DPO and the GEB variants.

| | DPO | GEB-$\pi$ | GEB-$1/\pi$ | GEB-arctanh$\pi$ |
|---|---|---|---|---|
| KL. | 1.24 | 1.08 | 1.32 | 1.19 |
| Hel. | 1.33 | 1.42 | 1.12 | 1.23 |
| fKL. | 1.32 | 1.38 | 1.32 | 1.30 |

As shown in Table 7, the responses produced by the GEB variants exhibit semantic coherence comparable to those generated by DPO. This indicates that, in practice, GEB promotes exploration into moderately underrepresented yet still semantically meaningful regions of the output space.

## 6 Conclusion

While recent work proposes exploratory bonuses to operationalize the "optimism in the face of uncertainty" principle, our work shows that the existing theoretical frameworks of exploratory bonuses fail under KL and $\alpha$-divergence regularization. To address prior theoretical pitfalls, we introduce General Exploratory Bonus (GEB), a novel theoretical framework for sample-efficient RLHF. Our approach provably satisfies the optimism principle and unifies prior heuristic bonuses. We empirically validate GEB on LLM alignment tasks with diverse bonus designs and LLM backbones, showing improved performance and broad applicability across multiple divergence families.

## References

[1] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[2] Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms. *CoRR*, abs/2405.08448, 2024.

[3] Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[4] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Ramé, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online AI feedback. *CoRR*, abs/2402.04792, 2024.

[5] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *CoRR*, abs/2404.19733, 2024.

[6] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 12248–12267, 2024.

[7] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024.

[8] Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *CoRR*, abs/2402.09401, 2024.

[9] Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff G. Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *CoRR*, abs/2312.00267, 2023.

[10] Han Qi, Haochen Yang, Qiaosheng Zhang, and Zhuoran Yang. Sample-efficient reinforcement learning from human feedback via information-directed sampling. *arXiv preprint arXiv:2502.05434*, 2025.

[11] Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, and Min Lin. Sample-efficient alignment for llms. *arXiv preprint arXiv:2411.01493*, 2024.

[12] Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *CoRR*, abs/2405.19332, 2024.

[13] Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient RLHF. *CoRR*, abs/2405.21046, 2024.

[14] Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. 2025.

[15] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[16] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[17] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[18] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *CoRR*, abs/2405.07863, 2024.

[19] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock,

Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[22] William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[23] Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. Avoiding $\exp(r_{max})$ scaling in rlhf through preference-based exploration. *arXiv preprint arXiv:2502.00666*, 2025.

[24] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2024.

[25] Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

[27] Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024.

[28] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2024.

[29] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475, 2024.

[30] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[31] Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C. Courville. Language model alignment with elastic reset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*, 2023.

[32] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 580–606, 2024.

[33] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *CoRR*, abs/2109.13916, 2021.

[34] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871, 2018.

[35] Xuanming Zhang, Yuxuan Chen, Yiming Zheng, Zhexin Zhang, Yuan Yuan, and Minlie Huang. Seeker: Towards exception safety code generation with intermediate language agents framework. *arXiv preprint arXiv:2412.11713*, 2024.

[36] Xuanming Zhang, Yuxuan Chen, Min-Hsuan Yeh, and Yixuan Li. Metamind: Modeling human social thoughts with metacognitive multi-agent systems. *CoRR*, abs/2505.18943, 2025.

[37] Min-Hsuan Yeh, Jeffrey Wang, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, and Yixuan Li. Position: Challenges and future directions of data-centric ai alignment. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

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

[39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[40] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[41] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[42] Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for llms. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[43] Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: segment-level direct preference optimization for

social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 12409–12423, 2025.

[44] Yueqin Yin, Shentao Yang, Yujia Xie, Ziyi Yang, Yuting Sun, Hany Hassan Awadalla, Weizhu Chen, and Mingyuan Zhou. Segmenting text and learning their rewards for improved RLHF in language model. *CoRR*, abs/2501.02790, 2025.

[45] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

[46] Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. Reinforcement learning with token-level feedback for controllable text generation. In *Findings of the Association for Computational Linguistics: NAACL*, pages 1704–1719, 2024.

[47] 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 *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

[48] Wendi Li and Yixuan Li. Process reward model with q-value rankings. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

[49] Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kai Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *CoRR*, abs/2412.01981, 2024.

[50] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758, 2025.

[51] Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. First return, entropy-eliciting explore. *CoRR*, abs/2507.07017, 2025.

[52] Fang Wu, Weihao Xuan, Ximing Lu, Zaïd Harchaoui, and Yejin Choi. The invisible leash: Why RLVR may not escape its origin. *CoRR*, abs/2507.14843, 2025.

[53] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939, 2025.

[54] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A unified framework for interactive decision making in mdp, pomdp, and beyond. *CoRR*, abs/2211.01962, 2022.

# A Related Works

**Alignment & RLHF.** Alignment [33, 34, 35, 36, 37] aims to ensure AI systems act in accordance with human values, preferences, and goals; and it has become a critical field in AI research. To steer language models to match human preferences, Reinforcement Learning from Human Feedback (RLHF) [38, 39] achieves great success and has become the standard alignment pipeline. However, its computational complexity has motivated a family of Direct Preference Optimization (DPO) [24, 40, 41] that forgoes explicit reward modeling. Despite their efficiency, recent researchers [3, 1, 17] reemphasize the significance of online sampling.

**Optimistic exploration of RLHF.** To address the computational overheads of passive exploration in RLHF, which samples trajectories just based on randomness, some existing attempts have been devoted to sample-efficient RL algorithms. Most of the works [7, 8, 42, 9, 22] adhere to the principle of optimism, proposing specialized prompt or response selection strategies to emphasize uncertain samples. While some research [11, 27] propose uncertainty-aware reward models with epistemic neural networks or bootstrap ensembles, these methods introduce additional cost. Some research also addresses the sample efficiency with different theoretical foundations, such as information theory [10], preference-incentive exploration [23]. Notably, several works [12, 13, 14] introduce different exploratory bonuses, which can implement optimism toward uncertainty without additional computations. However, they only focus on KL-divergence and their theoretical framework cannot result in real optimism as shown in Section 3.2.

**Efficient RL for LLM.** Beyond optimistic exploration, some research proposes fine-grained signals for RL learning. For instance, several studies propose segment-level [43, 44] or token-level [45, 46] reward functions for alignment or text control. Notably, for reasoning tasks, the process reward model [47, 48, 49], which provides step-wise feedback for solutions, has shown promising effectiveness. On the other hand, recent research [50, 51, 52] on LLM reasoning reveals that high-entropy tokens guide the model toward diverse reasoning paths. Training with only high-entropy tokens is more beneficial for reasoning performance [53]. While our approach is highly extensible, we believe the orthogonal methods can be further incorporated with our general exploratory bonus.

# B Optimism Failure of previous works

## B.1 Optimism failure under KL-divergence

We start by proving how the existing exploratory bonus term under KL-divergence instantiation fails to achieve optimistic exploration.

**Lemma 3.1** *Let $r_1 = \arg\min_r \mathcal{L}_{BT}(\mathcal{D}, r)$ be a reward model trained with the vanilla BT loss, and let $r_2 = \arg\min_r [\mathcal{L}_{BT}(\mathcal{D}, r) - \kappa \max_\pi \mathcal{J}_{\beta,KL}(\pi, r)]$ be a reward model trained with an additional exploratory bonus. If the policy is optimized via Eq. 4, then $r_1$ and $r_2$ yield the same set of policies.*

*Proof* First, the inner maximization of the bonus term admits a closed-form solution, $\pi^*(y|x) = \pi_{\text{ref}}(y|x)e^{\frac{r(x,y)}{\beta}}/Z(x)$ where $Z(x) = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}e^{\frac{r(x,y)}{\beta}}$ is a normalization term. Substituting this solution for the bi-level objective of $r_2$ reduces it to a single-level form:

$$r_2 = \arg\min_r \left[\mathcal{L}_{BT}(\mathcal{D}, r) - \kappa \mathbb{E}_{x\sim\rho}\beta \log \mathbb{E}_{y\sim\pi_{\text{ref}}} e^{\frac{r(x,y)}{\beta}}\right]. \tag{12}$$

As shown in Rafailov et al. [24], the log-ratio $\beta \log \pi_\theta(y|x) - \beta \log \pi_{\text{ref}}(y|x)$ represents the same class of the original reward function $r$ through Eq. 4, thus all $r$ in the reward modeling objectives can be reparameterized by the log-ratio. Plugging this into Eq. 12 yields

$$\arg\min_\pi \mathcal{L}_{dpo}(\mathcal{D}, \pi) - \kappa \mathbb{E}_{x\sim\rho}\beta \log \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)} \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}. \tag{13}$$

Since the second term equals 0, the reparameterized Eq. 12 is exactly the vanilla DPO loss, that is, the reparameterized training objective of $r_1$. Thus, the exploratory bonus in the reward training objective does not affect the final policy set. □

### B.2 Extension beyond $\alpha$-divergence

The following theorem formally proves that the exploratory bonus $-\kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ cannot encourage optimism for a more general divergence class.

**Theorem 3.3** *When $f$ belongs to $f$-divergence, the reward function is obtained via $\hat{r} = \arg\min_r[\mathcal{L}_{BT}(\mathcal{D}_t, r) - \kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)]$, and the policy is updated by $\arg\max_\pi \mathcal{J}_{\beta,f}(\pi, \hat{r})$, the bonus term $-\kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ induces the policy model $\pi$ to coincide with $\pi_{ref}$ when $xf''(x)$ is a monotone function.*

*Proof* By Lemma 3.2, we can reparameterize the bonus term used for optimistic reward modeling into Eq. 6. Denote $h(u) = uf'(u) - f(u)$. For a fixed prompt $x$, the training step can then be written as the following constrained optimization problem:

$$\arg\max_\pi \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)} h(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}) \quad s.t. \quad \sum_y \pi(y|x) = 1 \quad \text{and} \quad \forall y, \pi(y|x) > 0. \tag{14}$$

Then we can apply the Lagrange multiplier as

$$\mathcal{L} = \mathbb{E}_{y\sim\pi_{\text{ref}}} h(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}) - \mu\Big(\sum_y \pi(y|x) - 1\Big) - \sum_y \eta(y)\pi(y|x), \tag{15}$$

where $\mu, \eta$ are the dual variables. Then we utilize the Karush-Kuhn-Tucker (KKT) conditions for the given optimization problem. The complementary slackness gives that $\forall y, \eta(y)\pi(y|x) = 0$. The stationary condition requires

$$\frac{\partial \mathcal{L}}{\partial \pi(y|x)} = h'(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}) - \mu - \eta(y) = 0. \tag{16}$$

Let $S_y = \{y|\pi(y|x) > 0\}$. For all $y \in S_y$, complementary slackness implies $\eta(y) = 0$. Since $h'(u) = uf''(u)$ is a monotone function, we can obtain $\forall y \in S_y$, $\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ is a constant. Then applying the normalization constraint, $\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)} \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} = 1$. Hence, we conclude that the unique interior optimum is $\pi^*(y|x) = \pi_{\text{ref}}(y|x)$. □

The theorem implies that the reparameterized exploratory bonus attains its maximum only when $\pi$ and $\pi_{\text{ref}}$ coincide. The condition that $xf''(x)$ is a monotone function is satisfied by $\alpha$-divergence and beyond, e.g. Pearson $\chi^2$. Hence, the exploratory bonus $-\kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ in the reward training objective generally contradicts the optimism, since it cannot encourage trajectories with small initialized possibility.

### B.3 Prior adaptions & approximations cannot generalize

Although the theoretical justification for the previously proposed exploratory bonus does not hold, its empirically implemented loss remains effective due to various adaptations and approximations. In this subsection, however, we show that these adaptations and approximations cannot be extended beyond the KL-divergence class.

Zhang et al. [12] modify the formulation of $\mathcal{J}_{\beta,f}(\pi, r)$ in the optimistic exploratory bonus $-\kappa \max_\pi \mathcal{J}_{\beta,f}(\pi, r)$ as

$$\mathcal{J}'_{\beta,f}(\pi, r) = E_{x,y\sim\pi,y'\sim\pi_{\text{ref}}}[r(x,y) - r(x,y')] - \beta D_{KL}(\pi|\pi_{\text{ref}}), \tag{17}$$

which adds a bias in the reward expectation term. Under KL-divergence, the original $\mathcal{J}_{\beta,f}(\pi, r)$ will be zero after re-parameterization as shown in Lemma 3.1, thus the sole reparameterized bias term will remain as $-\mathbb{E}_{y'\sim\pi_{\text{ref}}} \log \pi(y'|x)$. Since $\mathcal{J}_{\beta,f}(\pi, r)$ cannot be reparameterized to zero except KL-divergence, this adaptation cannot generalize then.

In the derivations of Cen et al. [14], an idealized calibration distribution $\pi_{cal}$ is assumed to satisfy $\mathbb{E}_{y\sim\pi_{cal}} r(x,y) = 0$. Since $\pi_{cal}$ is not directly accessible in practice, the method substitutes rejected responses to approximate expectations under $\mathbb{E}_{\pi_{cal}}$. These rejected samples, however, do not satisfy the defining property of $\pi_{cal}$, making the approximation theoretically inconsistent.

The theoretical framework of Xie et al. [13] is based on Implicit Q-Approximation (refer to Lemma C.3 in the original paper) as follows,

$$\beta \frac{\log \pi(y|x)}{\log \pi_{\text{ref}}(y|x)} = r(x, y) - V^*(x) \tag{18}$$

where $V(x)^*$ is the KL-regularized value function. Because this relation depends critically on the logarithmic structure of the KL divergence, the framework cannot be extended to more general divergence families either.

In contrast, our general exploratory bonus integrates seamlessly with iterative online RLHF algorithms and extends naturally to the entire $\alpha$-divergence family. As discussed in §4, all these heuristic bonus terms can be encompassed by our unified theoretical framework.

### B.4 Equivalence between a sophisticated objective and a simple one.

**Lemma 4.1** *Denote two objectives as $h(u(x,y)) = \mathbb{E}_{x\sim\rho, y\sim\pi_{ref}} u(x,y)f'(u(x,y)) - f(u(x,y))$ and $\hat{h}(u(x,y)) = \mathbb{E}_{x\sim\rho, y\sim\pi_{ref}} \frac{u(x,y)}{Z(x)} f'(u) - f(\frac{u(x,y)}{Z(x)})$ where $u(x,y)$ is a function with $\pi(y|x)$ and $\pi_{ref}(y|x)$. If the ratio*

$$\frac{f'(u(x,y)) + u(x,y)f''(u(x,y)) - f'(\frac{u(x,y)}{Z(x)})}{Z(x)u(x,y)f''(u(x,y))} = \Lambda(x) \tag{19}$$

*is independent of $y$ and $\Lambda(x) > 0$, then minimizing the two objectives, $\min_\pi -h(u(x,y))$ and $\min_\pi -\hat{h}(u(x,y))$, yields the same class of optimal policies.*

*Proof* Similar to Lemma 4.1, for a fixed $x$, we can write a similar formulation of the Lagrange multipliers as follows,

$$\mathcal{L} = \mathbb{E}_{y\sim\pi_{\text{ref}}} h(u(x,y)) - \mu_1(\sum_y \pi(y|x) - 1) - \sum_y \eta_1 \pi(y|x), \tag{20}$$

where $\mu, \eta$ are the dual variables. Then, we utilize the KKT conditions formulation for the given optimization problem. Similar to Lemma 4.1, when $\pi(x,y) > 0$, we obtain

$$\frac{\partial h}{\partial \pi(y|x)} = u(x,y)f''(u(x,y)) \cdot \pi_{\text{ref}}(y|x) \cdot \frac{\partial u(x,y)}{\partial \pi(y|x)} = \mu_1. \tag{21}$$

Similarly, we can obtain the KKT conditions for $\hat{h}(u(x,y))$ as follows,

$$\frac{\partial \hat{h}}{\partial \pi(y|x)} = \frac{1}{Z(x)}(f'(u(x,y)) + u(x,y)f''(u(x,y)) - f'(\frac{u(x,y)}{Z(x)})) \cdot \pi_{\text{ref}}(y|x) \cdot \frac{\partial u(x,y)}{\partial \pi(y|x)} = \mu_2, \tag{22}$$

where $\mu_2$ is another dual variable. With the condition of Eq. 20, these two partial derivatives in Eq. 21 and Eq. 22 are equivalent when $\mu_2 = \mu_1 \Lambda(x)$. Hence, every policy that satisfies the stationary condition for $h$ also satisfies it for $\hat{h}$. Since $\Lambda(x) > 0$, the second-order derivative $\frac{\partial^2 h}{\partial^2 \pi(y|x)}$ and $\frac{\partial^2 \hat{h}}{\partial^2 \pi(y|x)}$ have the same sign, which indicates they share the same local minima. Hence, minimizing the two objectives $\min_\pi -h(u)$ and $\min_\pi -\hat{h}(u)$ induces the same class of policies.

□

### B.5 GEB enables optimistic exploration

In this subsection, we prove how GEB meets the optimism principle specified by Definition 3.1.

**Theorem 4.2** *Consider an $\alpha$-divergence $f$ with $0 \le \alpha \le 1$, and the exploratory bonus $\mathcal{L}_{bonus} = \beta \mathbb{E}_{x\sim\rho, y\sim\pi_{ref}} \Big[u(x,y)f'(u(x,y)) - f(u(x,y))\Big]$, where $u(x,y)$ is a function dependent on $\pi(y|x)$ and $\pi_{ref}(y|x)$. For any $(x,y)$, if $\frac{\partial u}{\partial \pi} + \pi_{ref} \frac{\partial^2 u}{\partial \pi \partial \pi_{ref}} + \frac{(\alpha-1)\pi_{ref}}{u} \frac{\partial u}{\partial \pi} \frac{\partial u}{\partial \pi_{ref}} < 0$ and $u(x,y) > \alpha$, the optimism condition in Definition 3.1 is satisfied; that is, $\frac{\partial^2 \mathcal{L}_{bonus}}{\partial \pi \partial \pi_{ref}} \le 0$.*

*Proof* First, substituting the optimal solution of the inner maximization and utilizing reward reparameterization, we obtain the training objective as in Eq. 9. By Lemma 4.1, this can be equivalently expressed as

$$\mathcal{L}_{\text{bonus}} = \beta \mathbb{E}_{x\sim\rho, y\sim\pi_{\text{ref}}(\cdot|x)} \Big[ u(x,y) f'(u(x,y)) - f(u(x,y)) \Big]. \tag{23}$$

For $\alpha$-divergences, the conditions in Lemma 4.1 are satisfied. Since we have $f'(u) + uf''(u) - f'(\frac{u}{Z}) = u^{\alpha-1}(Z^{1-\alpha} - \alpha)/(1-\alpha)$ and $uf''(u) = u^{\alpha-1}$, the fraction $\Lambda(x) = (Z^{1-\alpha} - \alpha)/Z(1-\alpha)$ in Lemma 4.1 is independent of $y$. Since $u > \alpha$, $Z_R = \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)} u > 0$, thus $\Lambda(x) > 0$ is also satisfied. Finally, the mixed second-order derivative of Eq. 11 is computed as

$$\frac{\partial^2 \mathcal{L}_{\text{bonus}}}{\partial\pi\partial\pi_{\text{ref}}} = \beta \mathbb{E}_{x\sim\rho} \sum_y u^{\alpha-1} (\frac{\partial u}{\partial \pi} + \pi_{\text{ref}} \frac{\partial^2 u}{\partial\pi\partial\pi_{\text{ref}}} + \frac{(\alpha-1)\pi_{\text{ref}}}{u} \frac{\partial u}{\partial \pi} \frac{\partial u}{\partial \pi_{\text{ref}}}) < 0, \tag{24}$$

which achieves the optimism defined in Definition 3.1. □

# C Regret Bound

In derivations, we utilize the theoretical tools in [12, 13, 14]. First, we make some standard statistical assumptions following Cen et al. [14].

**Assumption C.1** *For any reward function $r$, and any trajectory $\tau$, we have $-R_{max} < r(\tau) < R_{max}$, where $R_{max}$ is a constant.*

This is an assumption generally made for theoretical analyses of RLHF. Note that $R_{max}$ is measurable and controllable in practice. Then we introduce the assumption of the reward class proposed in Cen et al. [14], which offers a regularization mechanism for the subsequent derivations.

**Assumption C.2** *We assume that $r^* \in \mathcal{R}$, where*

$$\mathcal{R} = \{r : \mathbb{E}_{x\sim\rho, \tau\sim\pi_{cal}(\cdot|x)} r(x,y) = 0\}, \tag{25}$$

*where $\rho$ is the prompt distribution and $\pi_{cal}$ is a fixed calibration distribution independent of the algorithm.*

We also introduce the preference generalized eluder coefficient proposed in Zhang et al. [12], an extension of the generalized eluder coefficient [54], which connects prediction error and in-sample estimation error.

**Definition C.1** *Let $f_r(x,y,y') = r_t(x,y) - r^*(x,y)$. For a reward function class $\mathcal{R}$, we define the preference generalized eluder coefficient as the smallest $d_{\text{PGEC}}$ as*

$$\sum_{t=1}^{T} \mathbb{E}_{x\sim\rho, y\sim\pi^t(\cdot|x), y'\sim\pi_{cal}} [f_{r_t}(x,y,y') - f_{r^*}(x,y,y')]$$
$$\leq \sqrt{d_{\text{PGEC}} \sum_{t=1}^{T} \mathbb{E}_{x\sim\rho, y\sim\widetilde{\pi}_t(\cdot|x), y'\sim\pi_{cal}} [f_{r_t}(x,y,y') - f_{r^*}(x,y,y')]^2} + 4\sqrt{d_{\text{PGEC}} T}, \tag{26}$$

With the above assumptions and the theoretical tool, we can have the following regret boundary.

**Theorem C.1** *Let $\mathcal{J}_{\beta,f}(\pi, r) = \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)} r(x,y) - \beta \mathbb{E}_{x\sim\rho, y\sim\pi_{ref}(\cdot|x)} f(\frac{\pi(y|x)}{\pi_{ref}(y|x)})$. In the $t$-th iteration of the iterative online RLHF, the reward function and the policy are updated via*

$$r_t = \arg\min_r \Big[ \mathcal{L}_{BT}(\mathcal{D}_t, r) - \kappa \mathcal{L}_{bonus} \Big], \tag{27}$$

$$\pi_t = \arg\max_\pi \mathcal{J}_{\beta,KL}(\pi, r_t) = \arg\max_\pi \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)} r_t(x,y) - \beta \mathbb{D}_{KL}(\pi \| \pi_{ref}), . \tag{28}$$

*When the exploratory bonus is $\mathcal{L}_{bonus} = \max_\pi J_{\beta,f}(\pi, R)$, and the hyperparameter $\kappa$ satisfies $\kappa = \sqrt{\frac{\log(T|\mathcal{R}|\delta^{-1})}{(\gamma d_{\text{PGEC}} T)}} (32 R_{max} e^{4R_{max}})^{-1}$, with probability at least $1-\delta$, the regret can be bounded* *as follows,*

$$\sum_{t=1}^{T} \mathcal{J}_{\beta,f}(\pi^*, r^*) - \mathcal{J}_{\beta,f}(\pi^t, r^*) \le \mathcal{O}(R_{max} e^{4R_{max}} T \sqrt{d_{\mathrm{PGEC}} \gamma \log(|\mathcal{R}|\delta^{-1})}), \tag{29}$$

*where $\gamma = \sup_{x,y} \frac{\pi}{\pi_{cal}}$, $r^*$ and $\pi^*$ are ground-truth reward function and corresponding optimal policy with $\pi^* = \arg\max_\pi \mathcal{J}_{\beta,f}(\pi, r^*)$.*

*Proof* First, we can decompose the regret function as in Cen et al. [14] as follows,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathcal{J}_{\beta,f}(\pi^*, r^*) - \mathcal{J}_{\beta,f}(\pi^t, r^*) \\
&\qquad = \underbrace{\sum_{t=1}^{T} [\mathcal{J}_{\beta,f}(\pi^*, r^*) - \mathcal{J}_{\beta,f}(\pi^t, r^t)]}_{\text{Term 1}} + \underbrace{\sum_{t=1}^{T} [\mathcal{J}_{\beta,f}(\pi^t, r^t) - \mathcal{J}_{\beta,f}(\pi^t, r^*)]}_{\text{Term 2}}.
\end{aligned} \tag{30}$$

Then, we will bound term 1 and term 2 individually and combine them at last.

**Bound term 1.** Since the function $R(x, y)$ is dependent on $r(x, y)$, we denote it as $R(r)$ for this part. First, we connect the term 1 with $\max_\pi \mathcal{J}_{\beta,f}(\pi, R(r_t)) - \max_\pi \mathcal{J}_{\beta,f}(\pi, R(r^*))$. When $\pi^*$ is the optimal $\pi$ for $\max_\pi J_{\beta,f}(\pi, r^*)$, we have

$$\text{Term 1} \le J_{\beta,f}(\pi^*, r^*) - J_{\beta,f}(\pi^*, r_t) \le \sup_{x,y} \frac{\pi^*}{\pi_t} \mathbb{E}_{x\sim\rho, y\sim\pi_t}(r^* - r^t). \tag{31}$$

Similarly, we can obtain its lower bound as $-2R_{max}$. Then, we have

$$\text{Term 1} \le \sum_{t=1}^{T} \Big[ \max_\pi \mathcal{J}_{\beta,f}(\pi, R(r_t)) - \max_\pi \mathcal{J}_{\beta,f}(\pi, R(r^*)) \Big] + 4R_{max}T. \tag{32}$$

**Bound term 2.** First, we utilize the preference generalized eluder coefficient to connect the prediction error to in-sample error.

$$\text{Term 2} = \sum_{t=1}^{T} \mathbb{E}_{x\sim\rho, y\sim\pi^t(\cdot|x), y'\sim\pi_{cal}} [f_{r_t}(x, y, y') - f_{r^*}(x, y, y')] \tag{33}$$

$$\le \frac{\eta T d_{\mathrm{PGEC}}}{4} + 4\sqrt{d_{\mathrm{PGEC}} T} + \frac{1}{\eta} \sum_{t=1}^{T} \mathbb{E}_{x\sim\rho, y\sim\widetilde{\pi}_t(\cdot|x), y'\sim\pi_{cal}} [f_{r_t}(x, y, y') - f_{r^*}(x, y, y')]^2, \tag{34}$$

where the first equality uses the property of the reward class in Assumption C.2, and the inequality follows Definition C.1 with Cauchy–Schwarz inequality. Then, we bound the squared in-sample error as follows.

$$\begin{aligned}
\mathbb{E}_{x\sim\rho, \tau\sim\widetilde{\pi}_t} [f_{r_t}(x, y, y') - f_{r^*}(x, y, y')]^2 &\le \gamma \mathbb{E}_{x\sim\rho, \tau, \tau'\sim\widetilde{\pi}_t} [f_{r_t}(x, y, y') - f_{r^*}(x, y, y')]^2 & (35)\\
&\le \gamma (32 R_{max} e^{4R_{max}})^2 \mathbb{E}_{x\sim\rho, \tau, \tau'\sim\widetilde{\pi}_t} [\sigma(f_{r_t}(x, y, y')) - \sigma(f_{r^*}(x, y, y'))]^2 & (36)\\
&\le 8\gamma (32 R_{max} e^{4R_{max}})^2 \mathbb{E}_{x\sim\rho, \tau, \tau'\sim\widetilde{\pi}_t(\cdot|x)} D_H^2 (P_{r_t}(\cdot|\tau, \tau') \| P_{r^*}(\cdot|\tau, \tau')), & (37)
\end{aligned}$$

where $\gamma = \sup_{x,y} \frac{\widetilde{\pi}_t}{\pi_{cal}}$, and the second inequality utilizes the Lemma C.8 in Xie et al. [13], the third inequality uses $(x - y)^2 < 4(x + y)(\sqrt{x} - \sqrt{y})$. Refer to the Lemma C.6 in Xie et al. [13], we have

$$\sum_{i<t} \mathbb{E}_{x\sim\rho, \tau, \tau'\sim\widetilde{\pi}_t(\cdot|x)} D_H^2 (P_{r_t}(\cdot|\tau, \tau') \| P_{r^*}(\cdot|\tau, \tau')) \le L_{BT}^{(t)}(r_t) - L_{BT}^{(t)}(r^*) + 2\log(|\mathcal{R}|\delta^{-1}) \tag{38}$$

where $L_{BT}^{(t)}(r) = \sum_{i<t} \mathbb{E}_{y,y'\sim\mathcal{D}_\sqcup} - \log\sigma(f_r(x, y, y'))$ is the vanilla BT loss for reward modeling. Finally, the term 2 can be bounded by

$$\begin{aligned}
\text{Term 2} \le 4\sqrt{d_{\mathrm{PGEC}} T} + \frac{\eta T d_{\mathrm{PGEC}}}{4} +& \\
&\frac{8\gamma}{\eta} (32 R_{max} e^{4R_{max}})^2 (L_{BT}^{(t)}(r_t) - L_{BT}^{(t)}(r^*) + 2T\log(|\mathcal{R}|\delta^{-1})).
\end{aligned} \tag{39}$$

**Bound the regret.** Since the $r_t$ is optimized by $L_{BT}^{(t)}(r_t) - \sum_{i=1}^{T} \kappa \max_{\pi} \mathcal{J}_{\beta,f}(R(r_t),\pi)$, we have $r_t = \arg\min_{r\in\mathcal{R}} L_{BT}^{(t)}(r_t) - \sum_{i=1}^{T} \kappa \max_{\pi} \mathcal{J}_{\beta,f}(R(r_t),\pi)$ Therefore, when $\eta = 4\sqrt{\frac{2\gamma\log(T|\mathcal{R}|\delta^{-1})}{Td_{\mathrm{PGEC}}}}(32R_{max}e^{4R_{max}})$ and $\kappa = \frac{\eta}{4\gamma}(32R_{max}e^{4R_{max}})^{-2}$, the regret can be bounded by

$$\text{Regret} \leq 4\sqrt{d_{\mathrm{PGEC}}T} + \sqrt{2^3\gamma d_{\mathrm{PGEC}}\log(|\mathcal{R}|\delta^{-1})}(32R_{max}e^{4R_{max}}T) + 4R_{max}T. \tag{40}$$

□

# D Experiments

## D.1 Implementation Details

**Algorithm.** Following prior work [12, 13, 23], we adopt the same algorithmic backbone for empirical validation in order to isolate and compare the effects of different exploratory bonuses in the loss function. This backbone bypasses reward modeling at each iteration through reward reparameterization, a procedure commonly referred to as iterative DPO [18]. Previous methods of exploratory bonuses further reparameterize their bonus terms to make them compatible with this framework. To extend the original framework to $\alpha$-divergence, we extend the iterative DPO with f-DPO [25] loss, and likewise reparameterize our generalized exploratory bonus accordingly. The full procedure is summarized in Algorithm 1. Since the only difference is the loss function in Line 5 of Algorithm 1, our GEB does not induce any additional compute costs.

**Algorithm 1** Iterative Online Algorithm with Exploratory Bonus

```
Input: Reference model $\pi_{\text{ref}}$, iteration number $T$, prompt set for each interation $\mathcal{D}_1, \ldots, \mathcal{D}_T$, reward
    function $r$;
Output: Trained model $\pi_T$;
1: for iteration t = 1, 2, . . . , T do
2:     for $x \in \mathcal{D}_t$ do
3:         $y_1, y_2 \sim \pi_{\text{ref}}(\cdot|x)$ and obtain the rewards $r(y_1), r(y_2)$;
4:         Rank the reward and denote $y^+, y^-$ as the preferred and dispreferred response between
           $y_1, y_2$ and update $D_t = \{x, y^w, y^l\}$;
5:         $\pi_t = \arg\min_\pi \mathcal{L}_{\text{DPO}} - \kappa\mathcal{L}_{bonus}(\pi)$
6:         update $\pi_{\text{ref}}$ with $\pi_t$ (optional)
7:     end for
8: end for
```

**Hyperparameter settings and environments.** All experiments are conducted on two NVIDIA H200 GPUs. When training and sampling, the max length is set to 2048. For training, the batch size per device is set to 2; we enable gradient checkpointing, and the gradient accumulation step is set to 64; the learning rate is 5e-7 with a cosine scheduler, and the warm-up ratio is 0.03. In the main experiments, we use the best performance with $\kappa$ with a suitable ratio range to f-dpo loss across $1, 1e-2, 1e-4, 1e-6, 1e-8$. For sampling, the temperature is set to 1. For in-domain evaluation and MATH evaluation, we set temperature to 0.6 and top-p to 0.9; we use the default setting of alpaca-eval.

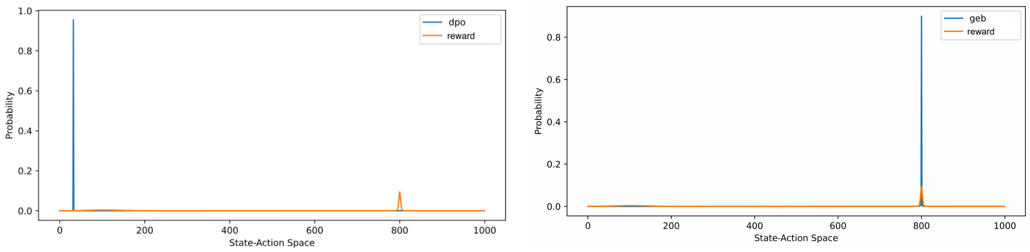

Figure 4: Comparison on the bandit policy distributions trained with DPO (left) and GEB (right). The DPO policy collapses to a local optimum, while the GEB policy continues to explore and ultimately chooses the globally preferred action.

# E Toy Experiments

To illustrate a setting in which GEB yields substantial improvement, we construct a toy example in which the most preferred action lies in a rarely visited region.

## E.1 Experimental setting

We consider a 1000-arm bandit with 1000 parameters, each parameter corresponding to a distinct arm. As shown in Fig. 5, the most preferred action lies in a rarely visited region, making it unlikely to be sampled under pure passive exploration, as in f-DPO. Each experiment is run for 5000 iterations. At each iteration, the bandit policy generates 64 rollouts to form a batch of 32 preference pairs. The learning rate is set to 1e-2 with no warm-up phase.

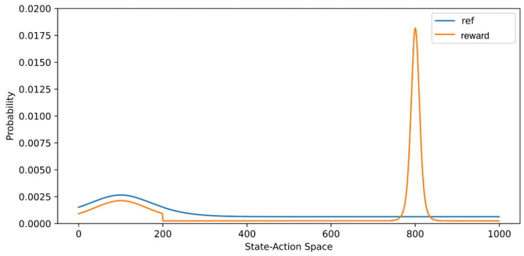

Figure 5: Initial reference bandit distribution ("ref") and the reward distribution. Because the most preferred action lies in a low-probability region, it is rarely visited under purely passive exploration.

## E.2 Results and Analyses

As shown in Fig. 4, the bandit policy trained with DPO becomes trapped in a local optimum: its probability mass collapses onto a suboptimal action because the policy never encounters the truly preferred action during training. In contrast, all three GEB variants successfully recover the desired distribution (right panel of Fig. 4), concentrating probability on the most preferred action. This demonstrates that GEB effectively promotes exploration into low-probability regions, enabling the policy to discover and select the optimal action despite its small initial likelihood.

# F Supplement results of main experiments

## F.1 Repeated run of the main experiment

To assess the statistical significance of GEB's performance gains, we repeat each experiment three times and report the mean and standard deviation. As shown in Table 8, GEB achieves consistently higher performance than baseline methods, with statistically significant improvements.

## F.2 The choice of $u$

The three variants of the exploration bonuses in Table 2 represent different instantiations of the GEB framework. Each satisfies the optimism condition in Definition 3.1 and exhibits consistent performance improvements on the alignment task.

Nonetheless, the curvature of $u$ with respect to $\pi$ meaningfully affects the optimization dynamics of the exploratory bonus. For instance, when $u = 1 - \pi + \alpha$, the function is linear, thus the gradient of $u$ to $\pi$ is a constant. Therefore, the per-trajectory incentive to decrease $\pi$ is constant. In contrast, when $u$ is convex—such as $u = 1/\pi$—the gradient magnitude diminishes as $\pi$ becomes larger. This results in a more conservative reduction of $\pi$ compared to the linear case.

Table 8: Repeated in-domain evaluation on different exploration bonuses. **Boldface** and underline indicate the best and the second-best results, respectively. GEB-$\pi$, GEB-$1/\pi$, and GEB-$\operatorname{arctanh}(\pi - 1)$ corresponds to $1 + \alpha - \pi$, $1/\pi$, and $\operatorname{arctanh}(1 - \pi) + \alpha$ as in Table 2.

| | KL ($\alpha$=1) | | Hel. ($\alpha$=0.5) | | f-KL ($\alpha$=0) | |
|---|---|---|---|---|---|---|
| | WR | AvgR | WR | AvgR | WR | AvgR |
| | | | *Mistral-Instruct-v0.3* | | | |
| f-DPO | 78.39 ±0.54 | 0.7480 ±0.0171 | 72.61 ±0.97 | 0.6536±0.0291 | 51.56±1.52 | 0.5918±0.0540 |
| FEB | 78.39 ±0.54 | 0.7480±0.0171 | 70.18±1.77 | 0.6428 ±0.1002 | 48.22±1.72 | 0.5910 ±0.0093 |
| GEB-$\pi$ | **80.64**±1.71 | 0.7523±0.0270 | $\underline{74.79}$±1.32 | **0.6710**±0.0231 | 52.23±0.89 | $\underline{0.5990}$±0.0102 |
| GEB-$1/\pi$ | $\underline{79.47}$±0.63 | **0.7602**±0.0036 | 72.97±1.17 | 0.6561±0.0054 | **53.40**±1.22 | **0.6030**±0.0093 |
| GEB-$\operatorname{arctanh}(\pi - 1)$ | 79.15±0.89 | $\underline{0.7567}$±0.0231 | **75.12**±1.65 | $\underline{0.6602}$±0.0171 | $\underline{52.59}$±0.77 | 0.5974±0.0372 |
| | | | *LLaMA-3-8B-SFT* | | | |
| f-DPO | 73.35 ±0.68 | 0.7984±0.0270 | 71.73±1.33 | $\underline{0.7902}$±0.0177 | 67.04±2.39 | 0.7579±0.0090 |
| FEB | 73.35 ±0.68 | 0.7984±0.0270 | 68.36±2.17 | 0.7560±0.0312 | 66.44±1.36 | 0.7598 ±0.0063 |
| GEB-$\pi$ | 74.52±1.80 | **0.8096**±0.0136 | 72.02±1.11 | **0.7911** ±0.0171 | 66.97±1.45 | **0.7702**±0.0048 |
| GEB-$1/\pi$ | **75.07**±1.42 | $\underline{0.8092}$±0.0063 | $\underline{72.40}$±0.80 | 0.7709±0.0242 | $\underline{67.93}$ ±0.25 | $\underline{0.7588}$±0.0033 |
| GEB-$\operatorname{arctanh}(\pi - 1)$ | $\underline{74.97}$±1.44 | 0.8055±0.0021 | **73.78**±2.40 | 0.7877±0.0054 | **68.34**±0.49 | 0.7602±0.0372 |

Takeaway: The curvature of $u$ with respect to $\pi$ governs the behavior of the exploration bonus. Greater convexity leads to a more conservative shift of probability mass toward underexplored regions.

## 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: The failure of existing theoretical framework is demonstrated in §3.2. The theoretical framework is introduced in §4. The empirical studies are in §**??**

   Guidelines:

   - The answer NA 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 NA 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: The limitations are discussed in §**??**.

   Guidelines:

   - The answer NA 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 generally provide proofs directly after the lemma or theorem, while some of proofs are supplemented in Appendix.

Guidelines:

- The answer NA 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: The implementation details can be found in §**??** and §D.1. Moreover, we provide the reproducible code and scripts here.

   Guidelines:

   - The answer NA 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: The datasets we use are all open-sourced, and we have provided the reproducible code and scripts here..

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The implementation details can be found in §**??** and §D.1. Moreover, we provide the reproducible code and scripts here.

Guidelines:

- The answer NA 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: [No]

Justification: The repeated training on large-scale LLMs are computation-costly, but we use multiple variants of GEB and experiments of different hyper-parameters to validate the stability of GEB.

Guidelines:

- The answer NA 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: The implemented details are in §D.1.

   Guidelines:

   - The answer NA 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: The research is conducted according to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: [Yes]

    Justification: In §**??**.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: A aim of RLHF is to reduce the occurrence of jailbreaking behaviors. Existing safeguard strategies can be generally applied to our outcome policies.

    Guidelines:

    - The answer NA 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 creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited.

    Guidelines:

    - The answer NA 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: [Yes]

    Justification: The usage of outcome policies follows the standard usage of LLMs in the huggingface package.

    Guidelines:

    - The answer NA 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: [NA]

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

    Guidelines:

    - The answer NA 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: [NA]

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

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.