Bootstrapping LLMs via Preference-Based Policy Optimization

Chen Jia

SI-TECH Information Technology
jiachenwestlake@gmail.com

Abstract

Bootstrapping large language models (LLMs) via preference-based policy optimization enables aligning model behavior with human preferences while reducing reliance on extensive manual annotations. We propose a novel preference-based policy optimization (PbPO) framework that formulates learning as a min-max game between the LLM policy and a reward model (RM). The RM is constrained within a confidence set derived from collected preferences to ensure reliable exploitation, while simultaneously promoting robust exploration. Our iterative online algorithm actively collects new preference data from the evolving policy, enabling continual self-improvement of both the policy and the RM. We provide theoretical guarantees, establishing high-probability regret bounds for both sequence-level and token-level RMs. Extensive experiments across five benchmark datasets demonstrate that PbPO consistently outperforms state-of-the-art preference optimization methods.

Extended version — https://arxiv.org/abs/2511.12867

Introduction

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017; Ziegler et al. 2019; Zhan et al. 2024a,b) has become a key paradigm for aligning machine learning models with human preferences (Stiennon et al. 2020; Ouyang et al. 2022; Rame et al. 2024; Rafailov et al. 2024), particularly for Large Language Models (LLMs) (Brown et al. 2020; Touvron et al. 2023; Achiam et al. 2023). By leveraging human annotations, RLHF trains a reward model (RM) that guides the policy to generate outputs that are helpful, truthful, and harmless (Ouyang et al. 2022; Bai et al. 2022; Casper et al. 2023).

Existing RLHF methods primarily rely on fixed preference datasets, separating preference data collection and RM pretraining (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022). Collecting large-scale, high-quality preference data is costly, often requiring human annotators (Bai et al. 2022) or LLMs (Cui et al. 2024). Moreover, reward misspecification and misgeneralization (Hong, Bhatia, and Dragan 2023; Gao, Schulman, and Hilton 2023) can lead to suboptimal policy updates.

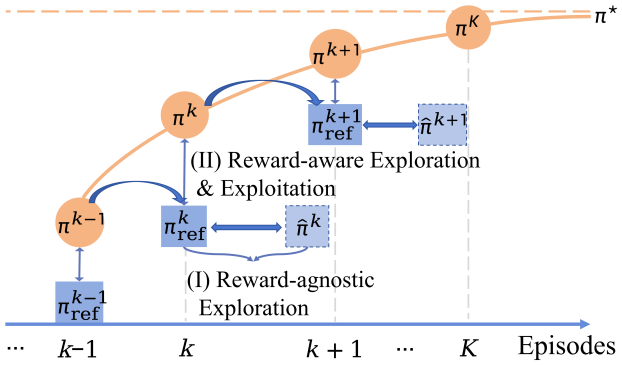


Figure 1: PbPO framework for bootstrapping LLMs. At each episode $k \in 1, 2, \ldots, K$: (I) **Reward-agnostic exploration**: collect new preference data using the current reference policy π^k_{ref} and an exploration-enhancing policy $\hat{\pi}^k$. (II) **Reward-aware exploration & exploitation**: update the main LLM policy π^k via a min-max objective using the reward model trained on collected preferences.

We adopt a bootstrapping approach in which LLMs iteratively improve themselves. Following online iterative RLHF, responses are sampled from the current LLM policy, feedback is collected to generate new preference data, and the RM is updated accordingly. Recent efforts (Zhan et al. 2024b; Xiong et al. 2024; Ye et al. 2024; Das et al. 2025) enhance this loop by incorporating reward-agnostic exploration strategies that maintain uncertainty of preference data collection, thereby improving data diversity to cover the whole preference data space. But standard online RLHF frameworks often optimize the RM solely for observed preferences, risking overfitting and premature convergence.

To address this, we propose **preference-based policy optimization (PbPO)**, a unified framework integrating reward-agnostic and reward-aware exploration. PbPO formulates a min-max game between the LLM policy and the RM. The RM is constrained within a confidence set derived from collected preferences to ensure reliable exploitation, while being optimized to minimize the performance gap with a reference policy, thereby promoting robust exploration. Iteratively collecting new preference data from the evolving policy enables effective bootstrapping of LLM alignment (Fig-

ure 1).

We provide theoretical guarantees for PbPO, establishing nearly optimal regret bounds for both sequence-level (Ouyang et al. 2022; Baheti et al. 2024; Xiong et al. 2024) and token-level RMs (Zeng et al. 2024; Cen et al. 2025; Zhong et al. 2025), as illustrated in Table 1. Extensive experiments on five benchmark datasets show that PbPO consistently outperforms existing state-of-the-art methods.

Our main contributions are summarized as follows:

1. We introduce PbPO, a unified framework combining reward-agnostic and reward-aware exploration within online RLHF.
2. We provide theoretical guarantees for iterative LLM self-improvement under PbPO, covering both sequence-level and token-level RMs.
3. We empirically validate PbPO on five benchmark datasets, achieving state-of-the-art performance.

Problem Formulation

We consider language modeling as a sequence-to-sequence decision-making task, where an input prompt x is first sampled, followed by generation of an output sequence $y = \{y_1, \ldots, y_H\}$. We formulate this process as an episodic *finite-horizon* Markov Decision Process (MDP), denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r^\star, \mu, H)$, where $\mathcal{S} = \{\{s_h\}_{1\le h\le H}\}$ denotes the state space consisting of the initial prompt state $\{s_1 = x\}$ and output prefixes $\{\{s_h = (x, y_{\le h-1})\}_{2\le h\le H}\}$, $\mathcal{A} = \{\{a_h\}_{1\le h\le H}\}$ denotes the action space corresponding to predicted tokens $a_h = y_h$, $\mu \in \Delta_{\mathcal{S}}$ is the initial state distribution, and $r^\star : \tau \mapsto r^\star(\tau) \in \mathbb{R}$ denotes the ground-truth reward model for a trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H) \in (\mathcal{S} \times \mathcal{A})^H$. Note that $r^\star$ is defined as a trajectory-wise reward, which is more general than the sequence-level and token-level rewards commonly used in RLHF. These commonly used rewards can be viewed as special cases and will be discussed in the following sections. Under our general MDP formulation, each of these cases corresponds to a specific sub-problem induced by a particular MDP structure. Additionally, we assume a deterministic state-transition function in sequence generation, where each action deterministically advances the state.

We focus on optimizing a LLM policy for predicting the next action at a state s_h with a probability distribution $\pi(\cdot \mid s_h) \in \Delta_{\mathcal{A}}$. Starting from an initial state $s_1 = x \sim \mu$, a trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H)$ is generated with the probability distribution $\mathbb{P}^\pi(\tau) = \mu(x) \cdot \left(\prod_{h=1}^{H-1} \pi(a_h \mid s_h)\right) \cdot \pi(a_H \mid s_H)$. Given a LLM policy π, we define its performance as the expected cumulative reward over the episode as $J(\pi, r) := \mathbb{E}_{\tau\sim\pi}[r(\tau)]$, where $\mathbb{E}_{\tau\sim\pi}[\cdot]$ denotes the expectation w.r.t. τ drawn from the probabiliy distribution $\mathbb{P}^\pi(\tau)$.

Optimization objective. We consider an episodic learning objective, where the learning algorithm aims to minimize the following cumulative regret over K episodes:

$$\mathrm{Regret}(K) := \sum_{k=1}^{K} \left(J(\pi^\star, r^\star) - J(\pi^k, r^\star)\right), \qquad (1)$$

$\mathrm{Regret}(K)$	**Sequence-level RM**	**Token-level RM**
Upper Bound	$\widetilde{\mathcal{O}}(\kappa d\sqrt{K})$	$\widetilde{O}(\kappa d H^{3/2}\sqrt{K})$
Lower Bound	$\Omega\left(d\sqrt{K}\right)$	$\Omega\left(dH\sqrt{K}\right)$

Table 1: Regret analysis. Upper bounds relative to the PbPO algorithm and information-theoretic lower bounds. Parameters: d = feature dimension, H = sequence horizon, K = number of learning episodes, and κ is a coefficient satisfying $\sup_{r_{\min}\le x\le r_{\max}} |1/\sigma'(x)| \le \kappa$.

where the optimal LLM policy is defined as $\pi^\star = \arg\max_{\pi\in\Pi} J(\pi, r^\star)$.

Preference optimization. Standard preference alignment approaches for LLMs (Ouyang et al. 2022; Zhu, Jordan, and Jiao 2023) assume ground-truth RM $r^\star$, such that the trajectory-based preference probability over a binary feedback $o \in \{0, 1\}$ for a trajectory pair in precollected preference dataset $(\tau^0, \tau^1) \in \mathcal{D}^{\mathrm{pref}}$ is formulated by:

$$\begin{aligned}\mathbb{P}_{r^\star}(\tau^0 \succ \tau^1) =& \mathbb{P}_{r^\star}(o = 1 \mid \tau^0, \tau^1) \\ =& \sigma\left(r^\star(\tau^0) - r^\star(\tau^1)\right),\end{aligned} \qquad (2)$$

where $\sigma(x) = 1/(1+\exp(-x))$ leads to the Bradley-Terry-Luce (BTL) model (Christiano et al. 2017).

Classical preference optimization estimate the RM $r \in \mathcal{G}_r$ from precollected preference data to approximate $r^\star$. Our analysis will follow the assumption that the reward difference are bounded in the interval $[r_{\min}, r_{\max}]$ in $\mathbb{R}$ and leverage the quantity $\sup_{r_{\min}\le x\le r_{\max}} |1/\sigma'(x)| \le \kappa$ with some constant $\kappa \ge 0$.

Most previous work on preference optimization for LLMs focuses on the offline setting, where preference data are drawn from a fixed, predetermined distribution (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022; Bai et al. 2022). We extend this approach iteratively by updating the reference policy at each episode to be the policy optimized in the previous step, generating preference samples from this evolving policy while incorporating exploration strategies.

Notations. We use $\mathbb{E}_{\tau^0\sim\pi_0,\tau^1\sim\pi_1}[\cdot]$ to denote the expectation w.r.t. first drawing a prompt $x \sim \mu$, then draws $\tau^0 \sim \mathbb{P}^{\pi_0}(\cdot \mid x)$ and $\tau^1 \sim \mathbb{P}^{\pi_1}(\cdot \mid x)$. We use $\|z\|_\Sigma$ to denote the induced norm $\sqrt{z^\top \Sigma z}$ for some positive-definite matrix Σ. We write $\widetilde{\mathcal{O}}(\cdot)$ to omit logarithmic factors and constants in regret bound.

Preference-Based Policy Optimization

We illustrate the theoretical framework of preference-based policy optimization (PbPO) using a sequence-level reward model and a token-level reward model, respectively.

PbPO with Sequence-Level Reward Model

In this subsection, we follow Ouyang et al. (2022); Rafailov et al. (2024); Xiong et al. (2024) and compute the sequence-level reward based on the entire response sequence y given an input x. This formulation streamlines the reward model by providing a holistic assessment of the response quality. Note that under the assumption of a sequence-level RM, the

Algorithm 1: PbPO (Theoretical Version)

1: **Input:** Initial preference set $\mathcal{D}_0^{\text{pref}} = \emptyset$, LLM policy class Π, initial LLM policy π^0 (SFT pretrained), input distribution μ
2: **for** episodes $k = 1$ to K **do**
3: Set the reference policy as the optimized policy in the previous episode: $\pi_{\text{ref}}^k \leftarrow \pi^{k-1}$
4: **Step 1: Reward-agnostic exploration**:
5: Choose an enhancer policy $\hat{\pi}^k$ by maximizing the uncertainty measurement:
6: $\hat{\pi}^k = \arg\max_{\pi \in \Pi} \left\| \mathbb{E}_{\tau^0 \sim \pi, \tau^1 \sim \pi_{\text{ref}}^k} [\phi(\tau^0) - \phi(\tau^1)] \right\|_{\hat{\Sigma}_k^{-1}}^2$
7: s.t., $\hat{\Sigma}_k = \lambda I + \sum_{s=1}^{k-1} (\phi(\tau_s^0) - \phi(\tau_s^1))(\phi(\tau_s^0) - \phi(\tau_s^1))^\top$
8: Sample an input: $x \sim \mu$, trajectories: $\tau_k^0 \sim \hat{\pi}^k$, $\tau_k^1 \sim \pi_{\text{ref}}^k$
9: Observe preference label: $o_k \sim \mathbb{P}_{r^\star}(\cdot \mid \tau_k^0, \tau_k^1)$ with an oracle $r^\star$
10: Add to the preference dataset: $\mathcal{D}_k^{\text{pref}} \leftarrow \mathcal{D}_{k-1}^{\text{pref}} \cup \{(o_k; \tau_k^0, \tau_k^1)\}$
11: **Step 2: Reward-aware exploration & exploitation**:
12: Compute the best policy with a constrained min-max optimization problem:
13: $\pi^k = \arg\max_{\pi \in \Pi} \min_{r \in \mathcal{R}(\mathcal{D}_k^{\text{pref}})} J(\pi, r) - J(\pi_{\text{ref}}^k, r)$
14: s.t., $\mathcal{R}(\mathcal{D}_k^{\text{pref}}) = \left\{ r \in \mathcal{G}_r : \sum_{s=1}^k \log \mathbb{P}_r(o_s \mid \tau_s^0, \tau_s^1) \geq \max_{r' \in \mathcal{G}_r} \sum_{s=1}^k \log \mathbb{P}_{r'}(o_s \mid \tau_s^0, \tau_s^1) - \zeta_k \right\}$
15: **end for**
16: **Output:** Optimized policy sequence $\{\pi^k\}_{k=1}^K$

MDP problem we previously formulated degenerates into a sentence-wise bandit problem (Zhong et al. 2025). Given a ground-truth sequence-level RM $r^\star$, the trajectory-based preference probability over a binary feedback $o \in \{0, 1\}$ for a trajectory pair (τ^0, τ^1) is formulated by:

$$\mathbb{P}_{r^\star}(o = 1 \mid \tau^0, \tau^1) = \sigma\left(r^\star(x, y^0) - r^\star(x, y^1)\right).$$

We focus on linear approximation for the sequence-level RM and define the function class using the following assumption:

Assumption 1 (Linearity & boundedness of seq-level RM)**.** *We assume the RM is linearly parameterized as $r_\theta(x, y) = \langle \theta, \phi(x, y) \rangle$, where $\phi : \tau \mapsto \phi(x, y) \in \mathbb{R}^d$ is a fixed sequence-level feature extractor. For regularization, the features and parameters are bounded such that $\sup_{(x,y)} \|\phi(x, y)\|_2 \leq 1$ and $\|\theta\|_2 \leq B$ for some $B > 0$.*

We consider a function approximation approach for estimating the ground-truth RM $r^\star$. Specifically, we introduce a linear function class $\mathcal{G}_r^{\text{seq}}$ to approximate $r^\star$:

$$\mathcal{G}_r^{\text{seq}} := \Big\{ r(x, y) = \theta^\top \phi(x, y) : \|\theta\|_2 \leq B, \|\phi(x, y)\|_2 \leq 1 \Big\}. \tag{3}$$

We assume that the RM class satisfies realizability:

Assumption 2 (Realizability of sequence-level RM)**.** *We assume the reward class is realizable, i.e., the ground-truth sequence-level RM lies in the function class: $r^\star \in \mathcal{G}_r^{\text{seq}}$, i.e., it satisfies that $r^\star(x, y) = \langle \theta^\star, \phi(x, y) \rangle$ for some $\|\theta^\star\| \leq B$.*

Algorithm We introduce the PbPO algorithm in Algorithm 1. During training episodes $k \in \{1, 2, \ldots, K\}$, preference feedback can be adaptively collected. This enables the LLM policy to continuously refine itself based on up-to-date information, thereby achieving bootstrapping performance. At each episode, the reference policy π_{ref}^k is defined as the optimized policy from the previous round, then the algorithm mainly consists of two steps as follows.

Step 1: Reward-agnostic exploration by collecting trajectories with an enhancer policy (Lines 4–10). To learn the ground-truth RM, we collect exploratory trajectories that cover the space spanned by $\phi(\cdot)$ before collecting any human feedback. To achieve this we identify a set of explorative enhancer policy that are not covered by existing preference data from the previous episodes. We measure the extent to which the trajectory generated by $(\hat{\pi}^k, \pi_{\text{ref}}^k)$ can be covered by computing the norm of $\mathbb{E}_{\tau^0 \sim \hat{\pi}_k, \tau^1 \sim \pi_{\text{ref}}^k} [\phi(\tau^0) - \phi(\tau^1)]$ on the metric induced by the inverse empirical covariance matrix $\hat{\Sigma}_k$ at the k-th episode. This strategy encourages querying trajectory pairs (τ_k^0, τ_k^1) that highlight uncertain regions of the RM. The obtained preference-labeled pair $(o_k; \tau_k^0, \tau_k^1)$ is then added to the cumulative dataset $\mathcal{D}_k^{\text{pref}}$. The process can be viewed as reward-agnostic exploration.

Step 2: Reward-aware exploration and exploitation by solving a min–max optimization problem (Lines 11–14). Given the collected preference dataset $\mathcal{D}_k^{\text{pref}}$, the objective presented in Lines 13–14 constitutes a min-max optimization problem between the policy and an uncertain RM. The outer objective is to optimize a target LLM policy π^k by maximizing its performance gap with a reference policy π_k^{ref} $(= \pi^{k-1})$, while the inner minimization identifies the least favorable reward model $r \in \mathcal{R}(\mathcal{D}_k^{\text{pref}})$ that minimizes the performance gap. This ensures conservative yet guaranteed improvement relative to the reference policy, while accounting for the inherent uncertainty of the reward inference process. Specifically, the reward confidence set $\mathcal{R}(\mathcal{D}_k^{\text{pref}}) \subseteq \mathcal{G}_r$ (defined in Line 14) is constructed via the maximum likelihood estimation (MLE) on the preference dataset $\mathcal{D}_k^{\text{pref}}$ with $\zeta_k \geq 0$ being a slack parameter controlling the confidence radius, thereby encouraging reward-aware exploration for a distributionally robust formulation. Since the RM can be

constrained near the best empirical RM, the optimization of policy gains exploitation from the previous preference data.

Learning Guarantee We provide a theoretical guarantee for Algorithm 1 by establishing a cumulative regret bound of Eq. (1).

Theorem 1 (Regret bound with sequence-level RM)**.** *For any $\delta \in (0,1]$, let $\zeta_k = \mathcal{O}(d\log(Bk/\delta))$ for any $k \in \{1,2,\ldots,K\}$ with a maximum episode number of K, then under Assumptions 1 & 2, setting $\gamma = \sqrt{\log\left(1+4K/(c_2 d^2\log(K/\delta))\right)}$ we have with probability at least $1-3\delta$:*

$$\mathrm{Regret}(K) \le c_1 \gamma B\kappa d\sqrt{K}\log(BK/\delta), \tag{4}$$

where $c_1, c_2 > 0$ denote some universal constants.

Remark 1. *Theorem 1 establishes that Algorithm 1 achieves a regret bound of order $\widetilde{\mathcal{O}}(\kappa d\sqrt{K})$ with high probability, where K is the number of episodes and d is the feature dimension. If neglecting κ, this result is analogous to the regret bounds in standard linear bandit settings, such as those achieved by the LinUCB algorithm for exploration (Dani, Hayes, and Kakade 2008).*

Corollary 1 (Sample complexity with sequence-level RM)**.** *Under the conditions of Theorem 1, if the number of preference samples satisfies $K \ge \widetilde{\mathcal{O}}\big(\kappa^2 d^2/\epsilon^2\big)$, then Algorithm 1 achieves ϵ-approximate convergence with high probability; that is, there exists some $k_0 \in \{1,2,\ldots,K\}$ such that $J(\pi^\star, r^\star) - J(\pi^{k_0}, r^\star) \le \epsilon$ with high probability. Since the algorithm generates exactly one preference sample per episode, the sample complexity is equivalent to the number of episodes K.*

To show that the regret bound is nearly optimal, we use the follow information-theoretic lower bound.

Theorem 2 (Regret lower bound with sequence-level RM)**.** *Under Assumptions 1 & 2, there exists a reward model $r^\star \in \mathcal{G}_r^{\mathrm{seq}}$ such that for any algorithm, the expected pseudo-regret over $K \ge \Omega(d^2/B^2)$ episodes is lower bounded as:*

$$\mathbb{E}_{\theta^\star}\left[\mathrm{Regret}(K)\right] \ge \Omega\left(d\sqrt{K}\right). \tag{5}$$

PbPO with Token-Level Reward Model

In this subsection, we focus on token-level RM, as discussed in recent RLHF advances based on token-level MDPs (Zeng et al. 2024; Cen et al. 2025; Zhong et al. 2025).

Given the ground-truth token-level RM $r^\star = \{r_h^\star\}_{h=1}^H$, the trajectory-based preference probability for trajectory pairs $\tau^0 = \{s_1, a_1^0, s_2^0, a_2^0, \ldots, s_H^0, a_H^0\}$ and $\tau^1 = \{s_1, a_1^1, s_2^1, a_2^1, \ldots, s_H^1, a_H^1\}$ over a binary feedback $o \in \{0,1\}$ is represented as:

$$\mathbb{P}_{r^\star}(o=1|\tau^0,\tau^1) = \sigma\left(\sum_{h=1}^H r_h^\star(s_h^0,a_h^0) - \sum_{h=1}^H r_h^\star(s_h^1,a_h^1)\right).$$

We focus on linear approximation for the token-level RM with the following assumption:

Assumption 3 (Linearity & boundedness of token-level RM)**.** *We assume that the token-level RM is parameterized by $\theta = (\theta_1, \theta_2, \ldots, \theta_H) \in \mathbb{R}^{dH}$ and defined as $r_\theta : \tau \mapsto \sum_{h=1}^H \theta_h^\top \phi(s_h, a_h) \in \mathbb{R}$, where H denotes the sequence length. For each $h \in [H]$, we assume the RM is represented as $r_{\theta_h}(s_h, a_h) = \langle \theta_h, \phi(s_h, a_h)\rangle$. For regularization, the step-wise parameters $\theta_h \in \mathbb{R}^d$ are bounded as $\|\theta_h\|_2 \le B$, and the features satisfy $\phi(s_h, a_h) \in \mathbb{R}^d$ with $\sup_{s_h,a_h} \|\phi(s_h,a_h)\|_2 \le 1$.*

We consider a function approximation approach for estimating the ground-truth RM $r^\star = \{r_h^\star\}_{h=1}^H$. Specifically, we introduce a linear function class $\mathcal{G}_r^{\mathrm{tok}}$ to approximate $r^\star$:

$$\mathcal{G}_r^{\mathrm{tok}} := \Big\{ \{r_h\}_{h=1}^H \,\big|\, r_h(s,a) = \theta_h^\top \phi(s,a) : \|\theta_h\|_2 \le B,\ \|\phi(s,a)\|_2 \le 1 \Big\}. \tag{6}$$

We define the realizability of RM with the following assumption:

Assumption 4 (Realizability of token-level RM)**.** *We assume the reward class is realizable, i.e., the ground-truth RM lies in the function class: $r^\star \in \mathcal{G}_r^{\mathrm{tok}}$, i.e., for each $h \in [H]$, it satisfies that $r_h^\star(s_h, a_h) = \langle \theta_h^\star, \phi(s_h, a_h)\rangle$ for some $\|\theta_h^\star\| \le B$.*

Algorithm The theoretical learning procedure is summarized in Algorithm 1, which follows the same structure as the sequence-level RM setting, with the only differences being the definition of the reward model and the feature projection: $\phi(\tau) = [\phi(s_1,a_1), \phi(s_2,a_2), \ldots, \phi(s_H,a_H)]$.

Learning Guarantee We present the regret bound and sample complexity of PbPO under the token-level RM setting as follows.

Theorem 3 (Regret bound with token-level RM)**.** *For any $\delta \in (0,1]$, let $\zeta_k = \mathcal{O}(dH\log(Bk\sqrt{H}/\delta))$ for any $k \in \{1,2,\ldots,K\}$ with a maximum episode number of K, then under Assumptions 3 & 4, setting $\gamma = \sqrt{\log\left(1+4K/(c_2 Hd^2\log(K/\delta))\right)}$, we have with probability at least $1-3\delta$:*

$$\mathrm{Regret}(K) \le c_1 \gamma B\kappa d H^{3/2}\sqrt{K}\log(B\sqrt{H}K/\delta), \tag{7}$$

where $c_1, c_2 > 0$ denote some universal constants.

Remark 2. *Theorem 3 establishes a high probabiliy regret bound of order $\widetilde{O}(\kappa d H^{3/2}\sqrt{K})$ for PbPO with token-level reward model.*

Based on the upper bound of regret, we can derive the following corollary on the sample complexity for the PbPO algorithm with token-level RM:

Corollary 2 (Sample complexity with token-level RM)**.** *To achieve an ϵ-approximate convergence, i.e., there exists some $k_0 \in \{1,2,\ldots,K\}$ such that $J(\pi^\star, r^\star) - J(\pi^{k_0}, r^\star) \le \epsilon$, it suffices to collect $K = \widetilde{\mathcal{O}}\big(\kappa^2 d^2 H^3/\epsilon^2\big)$ preference samples. Since our algorithm generates exactly one trajectory-based preference feedback per episode, the total number of episodes K coincides with the sample complexity.*

Algorithm 2: PbPO (Practical Version)

1: **Input:** Input dataset $\mathcal{D}_{\text{in}}$, number of episodes K, batch size $\mathcal{B}$, outer step size T_{out}, inner step size T_{in}. Initialize $\hat{\Sigma}_0 = \lambda I$, $\mathcal{D}_0^{\text{pref}} = \emptyset$, $\pi^0 = \pi_{\text{SFT}}$
2: **for** episodes $k = 1$ to $\lfloor K/\mathcal{B} \rfloor$ **do**
3: Set the reference policy as the optimized policy in the previous episode: $\pi_{\text{ref}}^k \leftarrow \pi^{k-1}$
4: **Step 1: Reward-agnostic exploration**:
5: Update $\hat{\Sigma}_k \leftarrow \hat{\Sigma}_{k-1} + \sum_{j=1}^{\mathcal{B}} (\phi(\tau_{k-1,j}^0) - \phi(\tau_{k-1,j}^1))(\phi(\tau_{k-1,j}^0) - \phi(\tau_{k-1,j}^1))^\top$
6: Maximize enhancer policy:
7: $\hat{\pi}^k \leftarrow \arg\max_{\pi \in \Pi} \left\| \mathbb{E}_{\tau^0 \sim \pi, \tau^1 \sim \pi_{\text{ref}}^k} [\phi(\tau^0) - \phi(\tau^1)] \right\|_{\hat{\Sigma}_k^{-1}}^2$
8: **for** $j = 1$ to $\mathcal{B}$ **do**
9: Sample input $x_j \in \mathcal{D}_{\text{in}}$, trajectories $\tau_{k,j}^0 \sim \hat{\pi}^k$, $\tau_{k,j}^1 \sim \pi_{\text{ref}}^k$, observe preference label $o_{k,j} \sim \mathbb{P}_{r^\star}(\cdot \mid \tau_{k,j}^0, \tau_{k,j}^1)$
10: Add $\mathcal{D}_k^{\text{pref}} \leftarrow \mathcal{D}_{k-1}^{\text{pref}} \cup \{(o_{k,j}; \tau_{k,j}^0, \tau_{k,j}^1)\}$
11: **end for**
12: **Step 2: Reward-aware exploration & exploitation**:
13: **for** outer steps $t = 1$ to T_{out} **do**
14: $\pi^k \leftarrow \text{SGA}\left(J(\pi^k, \hat{r})\right)$ ▷ Eq. (9)
15: **for** inner steps $t' = 1$ to T_{in} **do**
16: $\hat{r} \leftarrow \text{SGD}\left(J(\pi, \hat{r}) - J(\pi_{\text{ref}}^k, \hat{r}) - \beta \sum_{n=1}^{|\mathcal{D}_k^{\text{pref}}|} \log \mathbb{P}_{\hat{r}}(o_n \mid \tau_n^0, \tau_n^1)\right)$ ▷ Eq. (10)
17: **end for**
18: **end for**
19: **end for**
20: **Output:** Optimized policy sequence $\{\pi^k\}_{k=1}^K$

Theorem 4 (Regret lower bound with token-level RM)**.** *Under Assumptions 3 & 4, there exists a reward model $r^\star \in \mathcal{G}_r^{\text{tok}}$ such that for any algorithm, the expected pseudo-regret over $K \geq \Omega(d^2/B^2)$ episodes is lower bounded as follows:*

$$\mathbb{E}_{\theta^\star}\left[\text{Regret}(K)\right] \geq \Omega\left(dH\sqrt{K}\right). \tag{8}$$

This theorem demonstrates that our regret bound is nearly optimal, with only a small gap of $\widetilde{\mathcal{O}}(\kappa\sqrt{H})$.

Approximate Policy Optimization for PbPO

In this section, we describe how to implement the theoretical PbPO algorithms in practice.

Stackelberg Game Formulation of the Min-Max Objective The original objective in Algorithm 1 (Lines 13–14) defines a constrained min-max optimization problem over the LLM policy and the RM. However, solving this problem is generally intractable when employing flexible function approximators such as neural networks. To address this issue, we reformulate the objective as a two-player Stackelberg game (Von Stackelberg 2010) between the LLM policy (leader) and the RM (follower).

To circumvent the difficulty of directly optimizing under the RM's confidence-set constraint, we apply a Lagrangian relaxation. Specifically, we introduce a Lagrange multiplier $\beta \geq 0$ and convert the constrained min-max problem into an unconstrained bi-level optimization problem:

$$\hat{\pi} \in \arg\max_{\pi \in \Pi} \; J(\pi, r^\pi) - J(\pi_{\text{ref}}, r^\pi), \tag{9}$$

such that

$$r^\pi \in \arg\min_{r \in \mathcal{G}_r} \Big\{ J(\pi, r) - J(\pi_{\text{ref}}, r) - \beta \sum_{n=1}^{|\mathcal{D}^{\text{pref}}|} \log \mathbb{P}_r(o_n \mid \tau_n^0, \tau_n^1) \Big\}. \tag{10}$$

Here, the RM r is optimized to maximize the likelihood of the observed preferences via maximum likelihood estimation (MLE) based on the preference dataset. This dataset, denoted by $\mathcal{D}_k^{\text{pref}}$, is iteratively updated at each online training episode $k \in \{1, 2, \ldots, K\}$ in Algorithm 1.

Gradient-based Policy Optimization Following the two-player game formulation, we adopt a gradient-based adversarial training procedure for policy optimization, as illustrated in Algorithm 2. At the beginning of each online episode k, after optimizing the enhancer policy, input samples are sampled from the input dataset. Preference data are collected by sampling trajectories from both the the enhancer policy and the reference policy, which are then aggregated with previous preference data (Lines 8–11).

The two-player game is solved via gradient-based adversarial training for policy optimization (Lines 13–18). Specifically, the main LLM policy is updated via stochastic gradient ascent (SGA) to maximize the objective in Eq. (9) (Line 14), while the reward model (RM) is updated via stochastic gradient descent (SGD) to minimize the objective in Eq. (10) (Line 16).

Experiments

We evaluate the PbPO approach on five benchmarks.

Experimental Setup

Benchmarks. We evaluate on a suite of benchmarks including the complex reasoning dataset BBH (Suzgun et al. 2023), knowledge-based benchmarks AGIEval (Zhong et al. 2024), ARC-C (Clark et al. 2018), and MMLU (Hendrycks et al. 2021), as well as the math reasoning benchmark GSM8K (Cobbe et al. 2021). All reported accuracy results are averaged over three random seeds $\{10, 20, 30\}$ using zero-shot decoding.

Preference data collection. In each online learning episode, we randomly sample a batch of prompts from the training set to generate preference pairs using outputs from the reference policy and the enhancer policy. The corresponding preference labels are derived from GPT-4 feedback. To construct the offline preference dataset, we utilize models from the LLaMA2 (Touvron et al. 2023) and Qwen2 (Yang et al. 2024) families, including LLaMA2-7B/13B/70B-Chat and Qwen2-1.5B/7B/72B to generate candidate response pairs.

Hyperparameters. For both the initial policy and the enhancer policy, we employ a pretrained LLaMA2-7B (Touvron et al. 2023) or Qwen2-7B (Yang et al. 2024) model,

Method	LLaMA2-7B						Qwen2-7B					
	BBH	AGIEval	ARC-C	MMLU	GSM8K	**Avg.**	BBH	AGIEval	ARC-C	MMLU	GSM8K	**Avg.**
SFT	$43.6_{\pm.3}$	$34.3_{\pm.3}$	$66.7_{\pm.5}$	$48.3_{\pm.4}$	$49.3_{\pm.2}$	48.4	$61.4_{\pm.2}$	$50.0_{\pm.2}$	$61.5_{\pm.4}$	$69.5_{\pm.3}$	$75.5_{\pm.3}$	63.6
DPO	$45.2_{\pm.4}$	$35.3_{\pm.3}$	$67.6_{\pm.5}$	$49.5_{\pm.4}$	$51.0_{\pm.4}$	49.7	$65.3_{\pm.3}$	$52.7_{\pm.4}$	$63.2_{\pm.5}$	$70.5_{\pm.3}$	$77.8_{\pm.3}$	65.9
PPO	$44.7_{\pm.5}$	$33.7_{\pm.2}$	$68.0_{\pm.4}$	$50.2_{\pm.4}$	$49.5_{\pm.2}$	49.2	$64.7_{\pm.5}$	$54.9_{\pm.4}$	$65.3_{\pm.3}$	$72.7_{\pm.3}$	$78.2_{\pm.5}$	67.2
Online DPO (300 eps)	$45.6_{\pm.5}$	$35.6_{\pm.3}$	$68.7_{\pm.4}$	$52.3_{\pm.4}$	$51.7_{\pm.4}$	50.8	$65.6_{\pm.3}$	$55.7_{\pm.2}$	$68.2_{\pm.5}$	$70.2_{\pm.4}$	$76.5_{\pm.4}$	67.2
Online PPO (300 eps)	$46.2_{\pm.5}$	$35.5_{\pm.4}$	$66.7_{\pm.6}$	$50.0_{\pm.7}$	$51.0_{\pm.6}$	49.9	$66.2_{\pm.6}$	$54.2_{\pm.4}$	$65.2_{\pm.6}$	$73.2_{\pm.3}$	$79.2_{\pm.6}$	67.6
Best-of-N Distill (300 eps)	$46.8_{\pm.3}$	$33.8_{\pm.4}$	$67.5_{\pm.3}$	$49.2_{\pm.2}$	$50.7_{\pm.4}$	49.6	**68.2**$_{\pm.4}$	$53.2_{\pm.5}$	$65.4_{\pm.3}$	$72.7_{\pm.2}$	$78.2_{\pm.4}$	67.5
PbPO w/ seq RM (100 eps)	$49.4_{\pm.4}$	$40.5_{\pm.3}$	$68.3_{\pm.5}$	$53.2_{\pm.3}$	$52.1_{\pm.6}$	52.7	$65.7_{\pm.3}$	$57.8_{\pm.4}$	$66.7_{\pm.4}$	$73.4_{\pm.5}$	$81.4_{\pm.4}$	69.0
PbPO w/ seq RM (200 eps)	$52.4_{\pm.3}$	**44.0**$_{\pm.2}$	$71.6_{\pm.4}$	$55.4_{\pm.5}$	$55.2_{\pm.4}$	55.7	$66.3_{\pm.4}$	$\underline{59.2}_{\pm.5}$	$68.7_{\pm.5}$	$75.0_{\pm.3}$	$82.5_{\pm.3}$	70.3
PbPO w/ seq RM (300 eps)	$\underline{53.9}_{\pm.6}$	$\underline{43.7}_{\pm.3}$	$\underline{72.9}_{\pm.5}$	**57.1**$_{\pm.6}$	$\underline{58.1}_{\pm.4}$	$\underline{57.1}$	$67.8_{\pm.4}$	$58.2_{\pm.4}$	$\underline{70.5}_{\pm.4}$	**77.2**$_{\pm.4}$	$85.2_{\pm.3}$	$\underline{71.8}$
PbPO w/ tok RM (100 eps)	$49.0_{\pm.5}$	$39.2_{\pm.5}$	$68.8_{\pm.4}$	$52.0_{\pm.4}$	$54.5_{\pm.4}$	52.7	$66.3_{\pm.4}$	$57.2_{\pm.6}$	$67.5_{\pm.3}$	$71.5_{\pm.3}$	$83.6_{\pm.3}$	69.2
PbPO w/ tok RM (200 eps)	$52.7_{\pm.4}$	$41.7_{\pm.3}$	$72.0_{\pm.3}$	$54.8_{\pm.5}$	$57.2_{\pm.5}$	55.7	$67.2_{\pm.3}$	$58.1_{\pm.4}$	$69.5_{\pm.4}$	$74.6_{\pm.2}$	$\underline{85.8}_{\pm.2}$	71.0
PbPO w/ tok RM (300 eps)	**54.3**$_{\pm.5}$	$43.5_{\pm.4}$	**73.5**$_{\pm.3}$	$\underline{56.3}_{\pm.4}$	**60.2**$_{\pm.4}$	**57.6**	$\underline{68.0}_{\pm.3}$	**59.5**$_{\pm.4}$	**71.0**$_{\pm.3}$	$\underline{76.4}_{\pm.6}$	**87.2**$_{\pm.4}$	**72.4**

Table 2: Main results on LLaMA2-7B and Qwen2-7B backbones. We format **the best** and the second best results.

fine-tuned on the supervised training dataset. The reward model is constructed by appending a linear output layer to a frozen LLaMA2-7B or Qwen2-7B backbone. For the main experiments, we conduct a total of 300 episodes of online PbPO. In each episode, preference data are sampled with a batch size of 32.

Baselines. We compare our approach against several strong offline and online preference optimization baselines:

- **DPO:** Performing Direct Preference Optimization (DPO) (Rafailov et al. 2024) using the offline preference dataset.
- **PPO:** Training a reward model using the offline preference dataset, followed by PPO fine-tuning (Ouyang et al. 2022).
- **Online DPO:** Similar to Self-Rewarding LM (Yuan et al. 2024), where the main LLM policy generates new preference pairs for iterative fine-tuning via DPO.
- **Online PPO:** Following Xiong et al. (2024), iterative RLHF training where new preference pairs are generated and added to the dataset for reward model learning at each episode.
- **Best-of-N Distill:** Following (Sessa et al. 2024), we perform iterative self-distillation. In each episode, $N = 10$ candidate completions are generated using the teacher policy via nucleus sampling (top-p = 0.9, top-k = 40, temperature = 0.8). The response with the highest reward score is used as the training signal for knowledge distillation with KL divergence.

Main Results

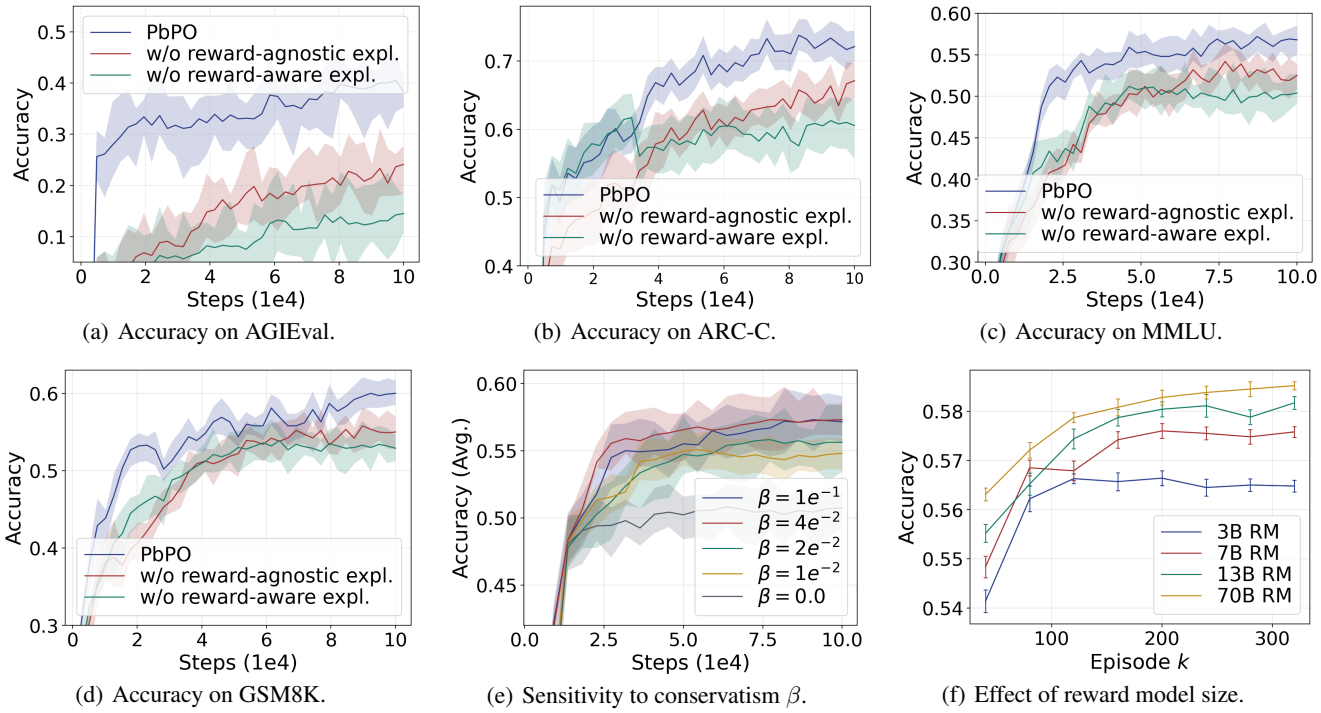
Table 2 presents the main results. Compared to the SFT baseline, all preference-based optimization methods yield noticeable improvements, confirming the effectiveness of leveraging preference data for LLM fine-tuning. However, existing online approaches—including Online DPO, Online PPO, and Best-of-N Distillation—only marginally outperform the offline RLHF baselines. In contrast, our proposed method consistently surpasses both offline and online baselines across all five evaluation datasets, demonstrating the superiority of our preference-based policy optimization (PbPO) framework in enhancing policy optimization. We conduct an ablation study on the number of online PbPO episodes. The results indicate that online PbPO reliably outperforms its offline variant, underscoring the advantage of iterative preference alignment. Performance continues to improve with more episodes, highlighting the effectiveness of bootstrapping LLMs through repeated preference-based refinement. Additionally, we observe that PbPO based on token-level RM achieves superior final performance compared to its sequence-level counterpart, particularly on datasets that require multi-step reasoning such as BBH, ARC-C, and GSM8K. This shows that token-level RM has an advantage for mastering complex, long-horizon tasks.

Experimental Analysis

Ablation study on exploration strategies. We evaluate PbPO against two ablations: (i) **PbPO w/o reward-agnostic explor.**, which removes the optimized enhancer and instead uses only the reference policy to generate trajectory pairs for preference data; and (ii) PbPO w/o reward-aware exploration, which follows a pipeline approach: it first trains the reward model using the online preference data and then performs policy optimization in each episode (similar to Xiong et al. (2024)). Figure 2(a–d) shows that our full method converges to a higher plateau across all tasks. The ablation without reward-agnostic exploration achieves lower final performance, while the one without reward-aware exploration performs significantly worse, confirming that both exploration components are critical.

Sensitivity of the conservatism regularizer β. Figure 2(e) shows the effect of β, which balances adversarial and preference losses in Eq. (10). Any $\beta > 0$ leads to steady improvement and higher stability compared to $\beta = 0$, where the RM fails to learn effectively. With $\beta \in \{0, 0.01, 0.02, 0.04, 0.1\}$, larger values ($0.04$, 0.1) achieve similar top performance, indicating robustness to β selection without fine-tuning.

Effect of RM size. We test RM sizes from OpenLLaMA-3B to LLaMA2-70B (Figure 2(f)). All sizes benefit from more episodes, confirming the advantage of online learning. Smaller RMs (e.g., 3B) converge faster but plateau



(a) Accuracy on AGIEval. (b) Accuracy on ARC-C. (c) Accuracy on MMLU.

(d) Accuracy on GSM8K. (e) Sensitivity to conservatism β. (f) Effect of reward model size.

Figure 2: Experimental analysis based on LLaMA2-7B backbone.

lower; larger RMs achieve better final performance, suggesting greater RM capacity improves preference alignment.

Related Work

RLHF. Reinforcement learning from human feedback (RLHF) has become a pivotal approach for fine-tuning large language models (LLMs) to produce text better aligned with human preferences. Early methods in this area rely on reward-based RLHF, where human preferences are used to train a reward model that guides reinforcement learning to optimize the LLM's outputs (Christiano et al. 2017; Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022). In contrast, reward-free methods such as Direct Preference Optimization (DPO) (Rafailov et al. 2024) and its variants (Azar et al. 2024; Ethayarajh et al. 2024; Park et al. 2024; Meng, Xia, and Chen 2024) bypass the explicit learning of a reward model. These approaches directly optimize the model based on pre-collected human preference data through pairwise comparisons, avoiding the construction of an explicit reward function. This often leads to improved data efficiency and scalability for LLM fine-tuning. Moreover, recent studies (Guo et al. 2024; Pang et al. 2024; Xiong et al. 2024; Ye et al. 2024; Shani et al. 2024; Cen et al. 2025; Zhang et al. 2025; Das et al. 2025) propose online variants of RLHF, where preference data are collected interactively from LLM annotators to iteratively evaluate and update the current policy. These works can be viewed as reward-based preference optimization relying on online collective preference datasets.

Self-improvement of LLMs. Self-improvement of large language models is an emerging paradigm aiming to enhance model capabilities by leveraging the model's own outputs as training signals. A prominent line of work is based on RLHF: starting from a supervised fine-tuned (SFT) model, Sun et al. (2024) prompt the SFT model to generate preference labels by selecting preferred responses according to certain principles, then train a principle-driven reward model and optimize the policy via PPO. Yuan et al. (2024) build a preference dataset from their own SFT model fine-tuned on instruction-following and evaluation data, followed by DPO training. Chen et al. (2025) focus on further enhancing a DPO-tuned model through bootstrapping with implicit rewards. Another related direction is Best-of-N distillation (Sessa et al. 2024; Yang et al. 2025), where a model learns from its own Best-of-N sampled outputs to improve consistency and generalization.

Conclusion

We propose Preference-based Policy Optimization (PbPO), a framework for bootstrapping large language models (LLMs) by iteratively refining both the policy and reward model via a min-max optimization game. Unlike conventional RLHF methods, which can suffer from reward misspecification and premature convergence, PbPO leverages confidence set constraints to ensure robust and reliable policy improvement. Guided exploration further enhances data collection by maintaining uncertainty, enabling continual self-improvement. Our theoretical analysis provides high-probability guarantees, and extensive experiments on multiple benchmarks show that PbPO consistently outperforms state-of-the-art preference optimization methods.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by SITECH Information Technology Co., Ltd.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, 4447–4455.

Baheti, A.; Lu, X.; Brahman, F.; Le Bras, R.; Sap, M.; and Riedl, M. 2024. Leftover Lunch: Advantage-based Offline Reinforcement Learning for Language Models. In *The Twelfth International Conference on Learning Representations*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Cen, S.; Mei, J.; Goshvadi, K.; Dai, H.; Yang, T.; Yang, S.; Schuurmans, D.; Chi, Y.; and Dai, B. 2025. Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF. *The Thirteenth International Conference on Learning Representations*.

Chen, C.; Liu, Z.; Du, C.; Pang, T.; Liu, Q.; Sinha, A.; Varakantham, P.; and Lin, M. 2025. Bootstrapping Language Models with DPO Implicit Rewards. In *The Thirteenth International Conference on Learning Representations*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; et al. 2024. Ultrafeedback: Boosting Language Models with Scaled AI Feedback. In *International Conference on Machine Learning*, 9722–9744. PMLR.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, 101, 355–366.

Das, N.; Chakraborty, S.; Pacchiano, A.; and Chowdhury, S. R. 2025. Active preference optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 96–112. Springer.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Model Alignment as Prospect Theoretic Optimization. In *International Conference on Machine Learning*, 12634–12651. PMLR.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 10835–10866. PMLR.

Guo, S.; Zhang, B.; Liu, T.; Liu, T.; Khalman, M.; Llinares, F.; Rame, A.; Mesnard, T.; Zhao, Y.; Piot, B.; et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hong, J.; Bhatia, K.; and Dragan, A. 2023. On the Sensitivity of Reward Inference to Misspecified Human Models. In *The Eleventh International Conference on Learning Representations*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Pang, R. Y.; Yuan, W.; He, H.; Cho, K.; Sukhbaatar, S.; and Weston, J. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37: 116617–116637.

Park, R.; Rafailov, R.; Ermon, S.; and Finn, C. 2024. Disentangling Length from Quality in Direct Preference Optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, 4998–5017.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36.

Rame, A.; Vieillard, N.; Hussenot, L.; Dadashi-Tazehozi, R.; Cideron, G.; Bachem, O.; and Ferret, J. 2024. WARM: On the Benefits of Weight Averaged Reward Models. In *International Conference on Machine Learning*, 42048–42073. PMLR.

Sessa, P. G.; Dadashi-Tazehozi, R.; Hussenot, L.; Ferret, J.; Vieillard, N.; Rame, A.; Shahriari, B.; Perrin, S.; Friesen, A. L.; Cideron, G.; et al. 2024. BOND: Aligning LLMs with Best-of-N Distillation. In *The Thirteenth International Conference on Learning Representations*.

Shani, L.; Rosenberg, A.; Cassel, A.; Lang, O.; Calandriello, D.; Zipori, A.; Noga, H.; Keller, O.; Piot, B.; Szpektor, I.; et al. 2024. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37: 118953–118993.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *NeurIPS*, 33: 3008–3021.

Sun, Z.; Shen, Y.; Zhang, H.; Zhou, Q.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2024. Salmon: Self-alignment with principle-following reward models. In *International Conference on Learning Representations*.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; et al. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13003–13051.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Von Stackelberg, H. 2010. *Market structure and equilibrium*. Springer Science & Business Media.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, T.; Mei, J.; Dai, H.; Wen, Z.; Cen, S.; Schuurmans, D.; Chi, Y.; and Dai, B. 2025. Faster WIND: Accelerating Iterative Best-of-N Distillation for LLM Alignment. In *International Conference on Artificial Intelligence and Statistics*, 4537–4545. PMLR.

Ye, C.; Xiong, W.; Zhang, Y.; Dong, H.; Jiang, N.; and Zhang, T. 2024. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37: 81773–81807.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. E. 2024. Self-Rewarding Language Models. In *International Conference on Machine Learning*, 57905–57923. PMLR.

Zeng, Y.; Liu, G.; Ma, W.; Yang, N.; Zhang, H.; and Wang, J. 2024. Token-level Direct Preference Optimization. In *International Conference on Machine Learning*, 58348–58365. PMLR.

Zhan, W.; Uehara, M.; Kallus, N.; Lee, J. D.; and Sun, W. 2024a. Provable Offline Preference-Based Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Zhan, W.; Uehara, M.; Sun, W.; and Lee, J. D. 2024b. Provable Reward-Agnostic Preference-Based Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Zhang, S.; Yu, D.; Sharma, H.; Zhong, H.; Liu, Z.; Yang, Z.; Wang, S.; Awadalla, H. H.; and Wang, Z. 2025. Self-Exploring Language Models: Active Preference Elicitation for Online Alignment. *Transactions on Machine Learning Research*.

Zhong, H.; Shan, Z.; Feng, G.; Xiong, W.; Cheng, X.; Zhao, L.; He, D.; Bian, J.; and Wang, L. 2025. DPO Meets PPO: Reinforced Token Optimization for RLHF. In *Forty-second International Conference on Machine Learning*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314.

Zhu, B.; Jordan, M.; and Jiao, J. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, 43037–43067. PMLR.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.