

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 KEEP THE BEST, FORGET THE REST: RELIABLE ALIGN- MENT WITH ORDER-AWARE PREFERENCE OPTIMI- TATION

006
007 **Anonymous authors**
008 Paper under double-blind review

011 ABSTRACT

013 Direct Preference Optimization (DPO) has emerged as a powerful framework
014 for aligning large language models (LLMs) with human preferences via pairwise
015 comparisons. However, its performance is highly sensitive to the quality of training
016 samples: when the reference policy is poorly aligned with human preferences,
017 ambiguous pairs can dominate the gradient signal and degrade generalization.
018 To address this, we propose RAPPO (**R**eliability **A**lignment **P**reference **O**ptimization), a simple sample-aware modification of the DPO loss that mitigates
019 reference-policy misalignment by filtering out the hardest, most ambiguous samples.
020 We theoretically show that RAPPO yields improved generalization guarantees.
021 RAPPO is lightweight and requires only a few lines of code to be integrated into
022 any existing DPO-type algorithm. Surprisingly, With this simple modification,
023 our simulations across a broad suite of alignment tasks and benchmarks show
024 consistent gains over DPO and recent state-of-the-art baselines. On the PKU-
025 SafeRLHF benchmark, RAPPO attains helpfulness 0.693 (+34.8% over DPO)
026 and harmlessness 0.357 (−21.0% vs DPO).

028 1 INTRODUCTION

030 Large language models (Achiam et al. (2023)) have achieved remarkable performance across a broad
031 spectrum of tasks, such as abstractive summarization (Stiennon et al. (2009); Koh et al. (2022)),
032 code generation (Chen et al. (2021); Gao et al. (2023b)), and instruction following (Chung et al.
033 (2024); Ouyang et al. (2022)), demonstrating strong generalization across domains. A key challenge,
034 however, lies in aligning these models with nuanced human intentions. Reinforcement Learning from
035 Human Feedback (RLHF) has become a widely adopted solution to this problem (Christiano et al.
036 (2017); Ouyang et al. (2022); Dong et al. (2023); Yuan et al. (2023); Liu et al. (2023a)), enabling
037 LLMs to better reflect both stylistic preferences and normative values (Bai et al. (2022); Ganguli et al.
038 (2022)). RLHF typically involves training a separate reward model to score outputs, which is then
039 used to guide policy optimization via reinforcement learning.

040 To streamline this process, Direct Preference Optimization (Rafailov et al. (2023)) (DPO) has been
041 proposed as an alternative that directly learns from pairwise preference data, bypassing the need
042 to construct an explicit reward model. Building on DPO, several variations improve optimization
043 stability and performance by: adding an explicit KL regularizer to control policy drift (IPO; Azar et al.
044 (2024)); introducing a learnable logit offset to correct unknown preference margins and reference
045 miscalibration (DPO-offset; Amini et al. (2024)); employing prospect-theory–inspired asymmetric
046 weighting and positive-only learning (KTO; Ethayarajh et al. (2024)); and pushing preference training
047 to the token level for finer credit assignment and reduced length/verbosity bias (token-level DPO;
048 (Zeng et al., 2024)). DPO outperforms supervised fine-tuning (SFT) significantly due to the ability to
049 distinguish the quality of multiple candidate outputs rather than relying on a single gold reference.
050 From the perspective of reinforcement learning with a KL divergence penalty, the DPO objective
051 induces an implicit reward function $r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, where π_θ denotes the policy being
052 fine tuned, π_{ref} is the fixed reference policy, and $\beta > 0$ controls the strength of the regularization
053 toward the reference. Clearly, the performance of the output policy π_θ is closely tied to the reference
054 policy. In fact, the effectiveness of DPO is strongly influenced by the distribution shift between model
055 outputs and the preference dataset Tang et al. (2024). More specifically, even mild alignment errors

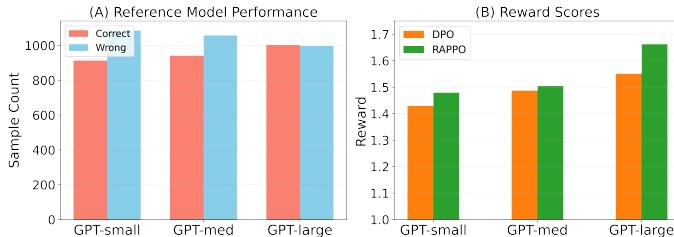


Figure 1: (A) Reference model performance across three SFT models (GPT2-Small, Medium, and Large). **Correct**: the reference policy aligns with human preference. **Wrong**: the reference policy conflicts with human preference. (A) shows that misaligned data are frequent regardless of model size, though the proportion decreases as model size increases. (B) Reward scores on the IMDb experiment (Section 5.1) using **DPO** and **RAPPO** under different reference policy scales. Performance declines as the reference policy weakens. Nonetheless, with a simple modification to **DPO**, our method **RAPPO** improves performance by **3.5%**, **1.1%**, and **7.1%** across the three models.

in the reference model(shown in Fig.1) can make it nearly impossible for DPO and its variants to recover the correct preferences [Chen et al. \(2024a\)](#). Recent work has further shown that updating the reference policy can improve DPO performance [Gorbatovski et al. \(2024\)](#).

Motivated by these observations, recent work has pursued two broad directions. (i) *Data selection*. RSO ([Liu et al., 2023a](#)) fine-tunes only on high-reward samples; Deng et al. ([Deng et al., 2025](#)) maximize reward margins to suppress noisy supervision; and Selective DPO ([Gao et al., 2025](#)) filters ambiguous preference pairs prior to training. (ii) *Reference-free training*. ORPO ([Hong et al., 2024](#)) replaces the reference-anchored objective with maximum-likelihood learning on positives examples, while SimPO ([Meng et al., 2024](#)) emphasizes length-normalized sequence likelihood and discards the reference policy entirely. Although reference-free approaches can avoid propagating harmful signals from a misaligned reference, they also relinquish potentially useful guidance and risk catastrophic forgetting ([Luo et al., 2023b; Huang et al., 2024; Zhang et al., 2022](#)). A more detailed related work can be found in Section 2. These observations, together with empirical findings from RLHF pipelines ([Ouyang et al., 2022; Peng et al., 2024](#)), demonstrate that DPO performance hinges not only on the choice of reference policy but also on how well the resulting training samples align with true human preferences. In this work, we seek to advance a systematic understanding of the reference policy’s role in DPO and ask the following central research question:

Can a simple, sample-aware modification to DPO mitigate reference-policy misalignment and thereby improve generalization?

We provide a theoretical treatment of this question in this work. Our approach is inspired by the selective-update principle in Ordered-SGD ([Kawaguchi & Lu, 2020](#)), which updates model parameters by using a selective approach. This principle prevents the optimizer from reinforcing incorrect preferences and focuses learning on informative examples. Building on this idea, we introduce RAPPO (**R**eliable **A**lignment for **P**reference **P**olicy **O**ptimization), a reference-aware variant of DPO that improves generalization by dynamically filtering out **untrustful** preference pairs. As shown in Fig. 2, by sacrificing some performance in regions where the reference policy performs poorly, RAPPO focuses on areas where the reference is well aligned with the offline dataset. This targeted emphasis leads to a significant overall improvement over DPO. The key contributions of our approach are summarized below:

- **Reliable Filtering:** At each optimization step, RAPPO detects samples where the reference model misaligns human preference signals and discards a fixed number of most untrustful pairs in the batch, those most ambiguous, thereby focusing learning on clear and reliable preferences.
- **Simplicity and Compatibility:** The method is lightweight and easily integrates into any existing DPO-type algorithms with only a few lines of code modification.
- **Theoretical Guarantee:** We provide a theoretical analysis showing that unreliable reference policies in DPO can enlarge the generalization gap, and prove that RAPPO, by order-aware sample selection, yields a tighter learning bound than standard DPO.
- **Strong Empirical Performance.** RAPPO achieves consistent improvements across four challenging tasks, IMDB sentiment classification and toxicity control ([Maas et al., 2011; Gehman et al., 2020](#)), text summarization, and safety alignment ([Ji et al., 2024](#)). On IMDB, it outperforms

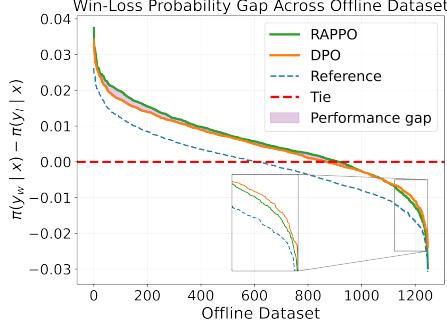
108
109
110
111
112
113
114
115
116
117
118

Figure 2: This figure compares the win–loss probability gap (the difference between the probability assigned to the human-preferred response and that of the rejected response) for models fine-tuned with **RAPPO** and **DPO** on the Toxicity dataset (Section 5.1). Samples are ordered by the reference policy’s preference gap, so the left side corresponds to cases where the reference is most decisive. **Tie**: instances where the model assigns equal probability to the preferred and rejected responses. In the shaded region, **RAPPO** yields a noticeably larger gap than DPO on good-aligned samples, while incurring only a minor trade-off on less well-aligned cases.

125
126
127
128
129
130
131

the strongest baseline by up to **7.7%**. For toxicity control, it reduces toxicity to as low as **2.28%** (vs. **6.30%** for the best baseline). In summarization, RAPPO surpasses SIMPO and DPO under GPT-4 evaluation, achieving win rates (defined in Section 5.2) of **58.8%** and **74.5%**, respectively. On the PKU-SafeRLHF benchmark, it consistently outperforms all baselines across four metrics (see Fig. 5 for details).

2 RELATED WORK

132
133
134
135
136
137
138
139
140

RLHF vs DPO RLHF(Christiano et al. (2017); Ziegler et al. (2019); Ouyang et al. (2022); Askell et al. (2021); Bai et al. (2022); Dong et al. (2023); Kaufmann et al. (2023); Yuan et al. (2023); Liu et al. (2023a); Peng et al. (2024); Zhou et al. (2023); Taori et al. (2023); Geng et al. (2023); Conover et al. (2023); Köpf et al. (2023); Ding et al. (2023); Wang et al. (2023); Xia et al. (2024); Gao et al. (2023a); Luo et al. (2023a); Chen et al. (2024b); Lightman et al. (2023); Lambert et al. (2024); Anthony et al. (2017); Schulman et al. (2017); Shao et al. (2024)) offers a practical route for bringing large-scale language models in line with human judgments and values. While DPO(Rafailov et al. (2023)) was proposed to bypasses the reward modeling stage while still achieving competitive performance. Recent DPO studies can be divided into following aspects.

141
142
143
144
145
146
147
148

Data Strategy To mitigate alignment errors of DPO, a growing body of work focuses on selective data sampling. Liu et al. (2023a) propose selecting high-reward samples during fine-tuning, while Deng et al. (2025) introduce a margin-based criterion for dataset construction. Similarly, Yang et al. (2024) suggest ranking implicit reward margins to prioritize more informative training pairs. Selective DPO(Gao et al. (2025)) further recommends removing ambiguous or difficult samples before training, and Liu et al. (2023b) present automatic data selection strategies to improve alignment. However, these methods focus solely on data filtering and overlook the limitations inherent in the reference model itself.

149
150
151
152
153
154
155
156
157
158
159

Reference Consideration in DPO. Recent studies question the reliability of the reference policy in DPO-based fine-tuning. Reference-free objectives—such as SIMPO, contrastive/ORPO-style losses, and SIMPER—remove the reference entirely and report gains on multiple benchmarks (Meng et al., 2024; Xu et al., 2024; Hong et al., 2024; Xiao et al., 2025). In contrast, Liu et al. (2024b) argue that a reference model remains essential. From a theoretical angle, Chen et al. (2024a) show that even mild misalignment in the reference can make preference recovery provably hard for DPO-type objectives. Seeking a middle ground, Chennakesavalu et al. (2024) introduce a DPO-like energy objective with a tunable parameter to modulate the reference’s influence during training. Taken together, these findings underscore both the risks of a fixed reference and the lack of a principled account of when, how, and to what extent the reference should be used.

3 PRELIMINARIES

160
161

Notations. Denote x as a prompt. π^{SFT} is a (pre-trained) supervised fine-tuning model to produce a pair of answers $(y_1, y_2) \leftarrow \pi^{SFT}(y|x)$. These response pairs are then presented to human

162 labelers, who express preferences by selecting one response as better. Let y_w and y_l denote the
 163 preferred and dispreferred responses, respectively. We define the underlying data distribution \mathcal{P}
 164 over triplets $z := (x, y_w, y_l)$, where the preference from labeler is denoted as $y_w \succ y_l | x$. Denote
 165 $D = \{z^i\}_{i=1}^N = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N \cup \mathcal{D}$ be a dataset of N i.i.d. samples drawn from \mathcal{D} . Let π_{ref}
 166 denote the reference policy (here, we use $\pi_{\text{ref}} := \pi^{\text{SFT}}$). Let π_θ be the fine-tuned policy for human
 167 alignment based on π_{ref} , parameterized by θ , where θ belongs to a hypothesis class Θ (e.g., neural
 168 networks). For any policy π (e.g., π_{ref} or π_θ), let $\pi(y_w | x)$ and $\pi(y_l | x)$ denote the likelihoods of
 169 the preferred and dispreferred responses given input x , respectively.

170 In RLHF, a language model is represented as a policy π that generates a response $y \in \mathcal{Y}$ conditioned
 171 on a prompt $x \in \mathcal{X}$, following the distribution $\pi(\cdot | x)$. The typical RLHF pipeline involves four
 172 main stages: supervised fine-tuning, collection of preference data, reward model training, and
 173 reinforcement learning-based policy optimization. The process begins with supervised fine-tuning,
 174 where a pre-trained language model is adapted into an initial policy π^{SFT} using labeled data through
 175 standard supervised learning. Subsequent stages refine the model using human or proxy feedback.
 176 Specifically, in the preference data stage, an oracle (such as a human annotator or a heuristic scoring
 177 function) compares two candidate responses y_1 and y_2 for the same prompt x , indicating which
 178 one is preferred. A common method for modeling such pairwise preferences is the Bradley–Terry
 179 model (Bradley & Terry, 1952), which estimates the probability of one option being favored over
 180 another:

$$181 \quad \mathbb{P}(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} = \sigma(r^*(x, y_1) - r^*(x, y_2)), \quad (1)$$

183 where $r^* : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the unknown ground truth reward function, and $\sigma : \mathbb{R} \rightarrow (0, 1)$ is the
 184 logistic function. Given a human labeled preference dataset $\mathcal{D} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N\}$ composed of
 185 independent samples, the reward function r can be estimated by maximum likelihood estimation
 186 (MLE): $r_{\text{MLE}} = \arg \min_r \ell(r, \mathcal{D})$, where $\ell(r, \mathcal{D})$ is the negative log-likelihood of \mathcal{D} , given as
 187 $\ell(r, \mathcal{D}) := -\sum_{(x^i, y_w^i, y_l^i) \in \mathcal{D}} \log \sigma(r(x^i, y_w^i) - r(x^i, y_l^i))$.

188 **RL fine-tuning.** Given a reward model r , the fine-tuned policy is optimized to achieve an ideal
 189 balance between the expected reward and its distance from an initial policy π_{ref} , which is typically the
 190 same as π^{SFT} . This is achieved by maximizing the KL-regularized value function $J(r, \pi)$, defined as

$$192 \quad J(r, \pi) = \mathbb{E}_{x \sim \mathcal{P}, y \sim \pi(\cdot | x)} [r(x, y)] - [\beta \mathbb{E}_{x \sim \mathcal{P}} [\text{KL}(\pi(\cdot | x) \| \pi_{\text{ref}}(\cdot | x))]], \quad (2)$$

193 where $\text{KL}(\pi_1 \| \pi_2)$ is the KL divergence from π_1 to π_2 , and $\beta > 0$ is a regularization trade-
 194 off. Thus the RL fine-tuned policy π_r with respect to the reward r aims to optimize $J(r, \pi)$, i.e.,
 195 $\pi_r := \arg \max_{\pi} J(r, \pi)$.

197 **Direct Preference Optimization (DPO).** In RLHF, DPO gives a closed-form solution derived by
 198 minimizing the following objective function:

$$199 \quad \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{P}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (3)$$

202 The empirical version of the DPO objective then follows:

$$203 \quad \widehat{\mathcal{L}}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta (\Delta_\theta(z^i) - \Delta_{\text{ref}}(z^i)) \right) \right], \quad (4)$$

206 where $\Delta_\theta(z^i) = \log \pi_\theta(y_w^i | x^i) - \log \pi_\theta(y_l^i | x^i)$, and $\Delta_{\text{ref}}(z^i) = \log \pi_{\text{ref}}(y_w^i | x^i) - \log \pi_{\text{ref}}(y_l^i | x^i)$.
 207 Thus DPO solves $\hat{\pi}_{\text{DPO}} = \arg \max_{\pi \in \Pi} \widehat{\mathcal{L}}_{\text{DPO}}(\pi; \pi_{\text{ref}})$.

210 4 RELIABLE ALIGNMENT FOR PREFERENCE POLICY OPTIMIZATION

212 **Introduction and Motivation.** We propose RAPPO, an order-aware variant of preference optimi-
 213 zation designed to mitigate the sensitivity of DPO to unreliable reference policies. The idea is
 214 inspired by ordered SGD (Kawaguchi & Lu, 2020), which emphasizes high-loss samples to am-
 215 plify signal. When the reference policy provides misleading preference signals, not all samples are
 216 equally informative, some add noise or even drive the model in the wrong direction. To address

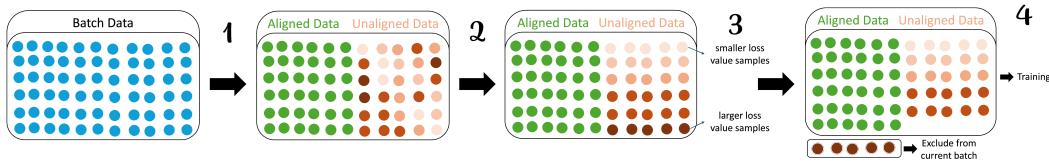


Figure 3: RAPPO Pipeline: (1) sample mini-batch data; (2) score each mini-batch by reference alignment, splitting samples into **Aligned** and **Unaligned**; (3) unaligned samples are ranked by per-sample loss; (4) temporarily remove some **Largest** ones for this update.

this, RAPPO selectively discards a fixed number of high-loss, ambiguous preference pairs in each batch, which focuses on variance control, our method explicitly targets misaligned examples arising from reference-policy errors, ensuring that the update direction remains both stable and aligned with true preferences. This motivates our main theorem, which shows that RAPPO achieves a tighter generalization bound than standard DPO. We first introduce our algorithm and then theoretically show our main results.

4.1 MAIN ALGORITHM

Our algorithm is lightweight and can be seamlessly integrated into **any** DPO-type method with only a few lines of code, yet delivers substantial improvements. The high-level idea is illustrated below, and the full algorithm is provided in Algorithm 1.

Step 1. Sample a mini-batch \mathcal{B} from the dataset (Stage 1 in Fig. 3).

Step 2. For each $z^i = (x^i, y_w^i, y_l^i) \in \mathcal{B}$, compute the reference-alignment score $\frac{\pi_{\text{ref}}(y_w^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)}$ and partition the batch into **Aligned** ($\mathcal{A}_{\mathcal{B}} = \{z^i \in \mathcal{B} \mid \frac{\pi_{\text{ref}}(y_w^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)} > \tau\}$) and **Unaligned** ($\mathcal{U}_{\mathcal{B}} = \{z^i \in \mathcal{B} \mid \frac{\pi_{\text{ref}}(y_w^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)} \leq \tau\}$) subsets. Also compute the per-sample loss $\ell_i(\theta) := -\log \sigma(\beta(\Delta_{\theta}(z^i) - \Delta_{\text{ref}}(z^i)))$ (Stage 2 in Fig. 3).

Step 3. In the **Unaligned** subset $\mathcal{U}_{\mathcal{B}}$, sort samples by $\ell_i(\theta)$ in ascending order (Stage 3 in Fig. 3).

Step 4. Exclude the Top- q **Largest-loss** unaligned samples **only** in the unaligned subset $\mathcal{U}_{\mathcal{B}}$ from the batch, and update the model using the retained data. (Stage 4 in Fig. 3).

Remark 4.1. RAPPO never permanently removes samples; even unaligned examples may be selected. Because selection within $\mathcal{U}_{\mathcal{B}}$ depends on the evolving loss $\ell_i(\theta)$, samples that were once “**untrust-hard**” can become “**untrust-easy**” as the model improves, naturally re-entering the kept set $\mathcal{S}_{\mathcal{B}}(q)$. This induces a reference-aware, model-adaptive curriculum: the reference score $\frac{\pi_{\text{ref}}(y_w^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)}$ gates the risk from misalignment, while the instantaneous loss $\ell_i(\theta)$ refines which unaligned signals are safe to incorporate at the current stage of training.

4.2 MAIN THEOREM

Now, we will show our main result of RAPPO. We will begin with a general optimization framework in which RAPPO arises as a special case, and present a main theorem showing that, compared with DPO, it achieves a larger expected first-order decrease in population risk, reduces the variance of gradient updates, and yields a tighter stability-based generalization bound. Collectively, these results formalize that discarding high-loss, misaligned samples enables more reliable progress toward human-aligned preferences.

Notations: Let $(\mathcal{Z}, \mathcal{A})$ be the example space and $\Theta \subseteq \mathbb{R}^p$ the parameter space. We are given a dataset $D = (Z_1, \dots, Z_n)$ with $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$, for some data distribution \mathcal{D} on $(\mathcal{Z}, \mathcal{A})$. At iteration t , a mini-batch $S_t = \{Z_{t,1}, \dots, Z_{t,s}\}$ is formed by sampling s indices uniformly from $\{1, \dots, n\}$. We define the loss $L : \Theta \times \mathcal{Z} \rightarrow \mathbb{R}$ which is measurable in z and differentiable in θ for any given sample. Let $g(\theta; z) := \nabla_{\theta} L(\theta; z)$. Define the (single) population risk (expected error) and its gradient as:

$$R(\theta) := \mathbb{E}_{Z \sim \mathcal{D}}[L(\theta; Z)], \quad d(\theta) := \nabla_{\theta} R(\theta) = \mathbb{E}_{Z \sim \mathcal{D}}[g(\theta; Z)], \quad (5)$$

270 **Assumption 4.2** (Smoothness and Lipschitz). Assume for some $L, G \geq 0$ and all z, θ, θ' ,

$$272 \quad \|g(\theta; z) - g(\theta'; z)\| \leq L\|\theta - \theta'\|, \quad |L(\theta; z) - L(\theta'; z)| \leq G\|\theta - \theta'\|. \quad (6)$$

274 Then R is L -smooth.

275 **Assumption 4.3** (Trusted/untrusted split and scores). Assume that we have a fixed measurable
276 classifier $H : \mathcal{Z} \rightarrow \{\text{tr, un}\}$ which can tag each example. For the batch $S_t = \{Z_{t,i}\}_{i=1}^s$ define
277

$$278 \quad S_t^{\text{tr}} = \{i : H(Z_{t,i}) = \text{tr}\}, \quad S_t^{\text{un}} = \{i : H(Z_{t,i}) = \text{un}\}, \quad G_t = |S_t^{\text{tr}}|, \quad B_t = |S_t^{\text{un}}|, \quad G_t + B_t = s.$$

279 For each untrusted item $i \in S_t^{\text{un}}$ define a scalar score $z_{t,i} = z(\theta_t; Z_{t,i}) \in \mathbb{R}$ that orders untrusted
280 items. The untrust samples can be regarded as the unaligned samples based on the reference policy.
281

282 **Weight bound on untrusted per-example gradients.** Assume there exists a nonincreasing $w : \mathbb{R} \rightarrow$
283 $(0, \infty)$ and a constant $C > 0$ such that for all untrusted Z and all θ ,

$$284 \quad \|g(\theta; Z)\| \leq Cw(z(\theta; Z)). \quad (7)$$

286 *Remark 4.4.* For DPO-type gradients: $z = \beta(\Delta_\theta - \Delta_{\text{ref}})$ and $w(z) = \sigma(-z)$ satisfy equation 7,
287 where $\Delta_\theta = \log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x)$, and $\Delta_{\text{ref}} = \log \pi_{\text{ref}}(y_w|x) - \log \pi_{\text{ref}}(y_l|x)$ for any given
288 (x, y_w, y_l) .

290 Let $s - q \in \{1, \dots, s\}$ be the number of kept examples per step. Define $K_t := \max\{0, B_t - q\}$.
291 We *never* remove trusted items. If $B_t - q \leq 0$, set $q := 0$ and keep all s items. If $B_t - q > 0$, keep
292 all G_t trusted items and among S_t^{un} keep the K_t largest scores $z_{t,i}$; call this set T_t^* . Thus

$$293 \quad \text{Kept}_t = \begin{cases} S_t & (B_t \leq q, \text{ set } q := 0), \\ S_t^{\text{tr}} \cup T_t^* & (B_t > q). \end{cases} \quad (8)$$

296 *Remark 4.5.* In RAPPO, each untrusted pair has a DPO margin z and loss $\ell(z) = -\log \sigma(z)$, with
297 larger z meaning **smaller** loss. Given a keep budget $s - q$, we keep all trusted items and, if trimming
298 is needed, we keep exactly $K_t = B_t - q$ untrusted items with the largest z (equivalently, the smallest
299 DPO loss), dropping the largest-loss untrusted items.

300 **Update rule.** Let $g_{t,i} := g(\theta_t; Z_{t,i})$. Define $\widehat{g}_t = \frac{1}{s-q} \sum_{i \in \text{Kept}_t} g_{t,i}$, $\theta_{t+1} = \theta_t - \eta_t \widehat{g}_t$. Then condition
301 on the realized untrusted scores $\{z_{t,i} : i \in S_t^{\text{un}}\}$. Let $\mu_{t,i} := \mathbb{E}[g_{t,i} | z_{t,i}]$ and $\xi_{t,i} := g_{t,i} - \mu_{t,i}$.
302 Assume the centered vectors $\{\xi_{t,i}\}_{i \in S_t^{\text{un}}}$ are conditionally independent given $\{z_{t,i}\}$. Assume there
303 exists nondecreasing $\phi : (0, \infty) \rightarrow [0, \infty)$ with $\text{tr Cov}(g_{t,i} | z_{t,i}) \leq \phi(w(z_{t,i})), (i \in S_t^{\text{un}})$.

305 *Remark 4.6.* For DPO loss, $L_{\text{DPO}}(\theta; z) = -\log \sigma(z)$, $\nabla_\theta L_{\text{DPO}}(\theta; z) = -\sigma(-z) \nabla_\theta z =$
306 $w(z)u$, $w(z) := \sigma(-z)$, $u := -\nabla_\theta z$, with $\nabla_\theta z = \beta(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x))$.

307 Assume a bounded-score condition (standard): $\|\nabla_\theta \log \pi_\theta(y|x)\| \leq B$. Then $\|u\| \leq 2\beta B$ and
308 hence $\|g_{t,i}\| = \|w(z_{t,i})u_{t,i}\| \leq (2\beta B)w(z_{t,i})$. Therefore, for any z , $\text{tr Cov}(g_{t,i} | z_{t,i} = z) \leq$
309 $\mathbb{E}[\|g_{t,i}\|^2 | z_{t,i} = z] \leq (2\beta B)^2 w(z)^2$. So the assumption holds with the explicit nondecreasing
310 choice $\phi(u) := (2\beta B)^2 u^2$, $u > 0$, since $u \mapsto u^2$ is nondecreasing.

312 Let $d_t := d(\theta_t)$. Define $a_{t,i} := \langle g_{t,i}, d_t \rangle$ and its conditional mean $\bar{a}(z) := \mathbb{E}[a_{t,i} | z_{t,i} = z]$.
313 Assume $\bar{a}(z)$ is nondecreasing in z . Define $h(z) := w(z)\bar{a}(z)$ and assume h is nondecreasing in z .
314 Now we are ready to show our main theorem, the detailed proof is differed to due to page limit.

315 **Theorem 4.7.** Fix iteration t with $B_t > s - q$ and let $T \subseteq S_t^{\text{un}}$ be any set with $|T| = K_t$, while T_t^*
316 contains the K_t largest $z_{t,i}$ in S_t^{un} . Conditioned on the realized scores $\{z_{t,i}\}$:

317 (i) Maximal expected first-order decrease of R .

$$319 \quad \mathbb{E}\left[-\langle d_t, \widehat{g}_t(T) \rangle \mid \{z_{t,i}\}\right] \leq \mathbb{E}\left[-\langle d_t, \widehat{g}_t(T_t^*) \rangle \mid \{z_{t,i}\}\right]. \quad (9)$$

322 (ii) Minimal conditional variance of the signed decrease. Let $\Phi_t(T) := \langle -\widehat{g}_t(T), d_t \rangle$. Then

$$323 \quad \text{Var}(\Phi_t(T) \mid \{z_{t,i}\}) \geq \text{Var}(\Phi_t(T_t^*) \mid \{z_{t,i}\}). \quad (10)$$

324

Algorithm 1 RAPPO

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

1: **Inputs:** an initial policy π_{θ^0} , reference policy π_{ref} , a learning rate sequence $(\eta_t)_t^T = 1$, threshold τ , and hyperparameter q

2: **for** $t = 1, 2, \dots, T$ **do**

3: Randomly choose a mini-batch \mathcal{B} and compute the $\frac{\pi_{\text{ref}}(y_w^i | x^i)}{\pi_{\text{ref}}(y_t^i | x^i)}$ for all data samples z^i in whole dataset.

4: Partition the batch into $\mathcal{U}_{\mathcal{B}} = \{z^i \in \mathcal{B} \mid \frac{\pi_{\text{ref}}(y_w^i | x^i)}{\pi_{\text{ref}}(y_t^i | x^i)} \leq \tau\}$, $\mathcal{A}_{\mathcal{B}} = \{z^i \in \mathcal{B} \mid \frac{\pi_{\text{ref}}(y_w^i | x^i)}{\pi_{\text{ref}}(y_t^i | x^i)} > \tau\}$

5: Within the current untrust data sample $\mathcal{U}_{\mathcal{B}}$, order the individual losses $\ell_i(\theta) := -\log \sigma(\beta(\Delta_{\theta}(z^i) - \Delta_{\text{ref}}(z^i)))$ in ascending order that $\ell_{(j=1)} \leq \ell_{(j=2)} \leq \dots \leq \ell_{(j=b-q)}$.

6: Computer the gradient \tilde{g}_t using kept sample:

$$\tilde{g}_t \in \partial \hat{\mathcal{L}}_{\text{step}}^{\text{RAPPO}} = \partial \begin{cases} \frac{1}{s} \sum_{i \in \mathcal{B}} \ell_i(\theta), & \text{if } b \leq q, \\ \frac{1}{s-q} \left(\sum_{i \in \mathcal{A}_{\mathcal{B}}} \ell_i(\theta) + \sum_{j=1}^{b-q} \ell_{(j)}(\theta) \right), & \text{if } b > q \end{cases}$$

7: Update parameters θ : $\theta_{t+1} = \theta_t - \eta_t \tilde{g}_t$

8: **end for**

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

(iii) Uniform stability and generalization. Let $D = (Z_1, \dots, Z_n)$ and $D' = (Z'_1, \dots, Z'_n)$ be two datasets that differ in one example, both i.i.d. from \mathcal{D} . Run the same randomness to produce $\{\theta_t\}$ and $\{\theta'_t\}$. Define $\Delta_t := \mathbb{E}\|\theta_t - \theta'_t\|$. Then for any horizon T and stepsizes $\{\eta_t\}$,

$$\Delta_{t+1} \leq (1 + L\eta_t)\Delta_t + \frac{2C}{s-q} \eta_t \mathbb{E} \left[\max_{i \in \text{Kept}_t} w(z_{t,i}) \right], \quad (11)$$

$$\Delta_T \leq \frac{2C}{s-q} \exp \left(\sum_{u=0}^{T-1} L\eta_u \right) \sum_{t=0}^{T-1} \eta_t \mathbb{E} \left[\max_{i \in \text{Kept}_t} w(z_{t,i}) \right]. \quad (12)$$

By Eq.(6), $\varepsilon_T := \sup_z \mathbb{E}|L(\theta_T; z) - L(\theta'_T; z)| \leq G\Delta_T$, and then the expected generalization error is,

$$\mathbb{E}[R(\theta_T) - R_n(\theta_T)] \leq \varepsilon_T, \quad R_n(\theta) = \frac{1}{n} \sum_{i=1}^n L(\theta; Z_i). \quad (13)$$

From the results, we can observe that 1), keeping the K_t items with the largest scores z (i.e., the smallest losses) reduces the conditional variance of untrusted contributions, since larger losses correspond to larger gradient weights and w is nonincreasing in z . 2), over multiple steps this choice minimizes the maximal kept weight among untrusted items, directly tightening the stability recursion and thus the generalization gap. Both effects hold for smooth, Lipschitz losses without convexity assumptions. 3) Compared with any alternative rule that keeps K_t untrusted items, our selecting minimizes the per-step term $\mathbb{E}[\max_{i \in \text{Kept}_t} w(z_{t,i})]$ and hence minimizes Δ_T and ε_T .

While the algorithm design is simple and flexible, it is also important to derive an objective function. We first define $\ell_{(j)} = \ell(z^{(j)}; \theta)$ as the j -th smallest loss in unaligned data samples i.e., totally $\hat{\mu}N$ elements, where $\ell_{(1)} \leq \dots \leq \ell_{(\hat{\mu}N)}$. Particularly, the objective that we optimize can be found in the following proposition.

Proposition 4.8. Let the whole dataset size be N . The number of whole unaligned examples is $n_b = \hat{\mu}N$ and of whole aligned examples is $n_g = (1 - \hat{\mu})N$. At each iteration, we sample a mini-batch $S \subseteq [N]$ without replacement, $|S| = s$. Write the unaligned and aligned counts in the batch as $\mathcal{U}_S, \mathcal{A}_S$. For a parameter vector θ , the per-example loss is ℓ_i or $\ell_{(j)}(\theta)$. And denote $m_g := \sum_{i \in \mathcal{A}_S} \ell_i(\theta)$, $m_b := \sum_{j \in \mathcal{U}_S} \ell_{(j)}(\theta)$. Then, we can obtain the objective function of RAPPO:

$$\hat{\mathcal{L}}^{\text{RAPPO}} = \underbrace{\sum_{b=0}^{\min(q, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_S| = b) \frac{m_g + m_b}{s}}_{\text{keep all data}} + \underbrace{\sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_S| = b) \frac{m_g + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s-q}}_{\text{keep aligned and remove } q \text{ largest loss data sample}}. \quad (14)$$

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 1: Comparison of reward scores and toxicity percentages across various preference optimization methods, evaluated on the IMDB and Real-Toxicity-Prompts [Gehman et al. \(2020\)](#) test set. Higher reward scores and lower toxicity indicate better performance. The whole experiments of SIMPO are defered in Table 7.

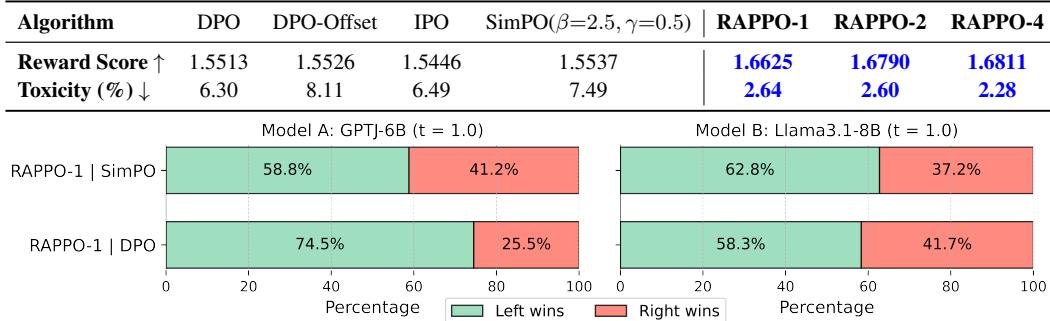


Figure 4: Win Rate between RAPPO-1, SIMPO, and DPO by GPT-4.

where $\mathbb{P}(|\mathcal{U}_B| = b) = \frac{\binom{\hat{\mu}^N}{b} \binom{(1-\hat{\mu})^N}{s-b}}{\binom{N}{s}}$, $\alpha_j = \frac{\sum_{l=0}^{b-q-1} \binom{j-1}{l} \binom{\hat{\mu}^N-j}{b-1-l}}{\binom{\hat{\mu}^N}{b}}$. Then, \tilde{g}_t is a unbiased estimator of $\partial \hat{\mathcal{L}}^{\text{RAPPO}}$ that $\mathbb{E}[\tilde{g}_t] \in \partial \hat{\mathcal{L}}^{\text{RAPPO}}$.

Proposition 4.8 give a close-form of RAPPO objective function and also show that \tilde{g}_t is unbiased estimator of $\partial \hat{\mathcal{L}}^{\text{RAPPO}}$, thereby justifying our update rule and ensuring the stochastic updates are consistent with minimizing the RAPPO objective.

5 EXPERIMENTS

In this section, we empirically evaluate the performance of RAPPO in comparison to DPO and other state-of-the-art methods across a variety of tasks.

Baselines: We compare RAPPO with state-of-the-art offline preference optimization baselines listed in Table 4. IPO ([Azar et al., 2024](#)) directly leverages pairwise preference data without relying on the Bradley–Terry model, helping mitigate overfitting, particularly with deterministic or near-deterministic preferences. DPO-OFFSET ([Amini et al., 2024](#)) modifies the optimization objective to enforce a reward margin between the preferred and dispreferred responses. SIMPO ([Meng et al., 2024](#)) adopts a reference-free approach and integrates a reward margin into the Bradley–Terry objective to enhance separation between responses. For training, we follow the hyperparameter ranges reported in SIMPO, where $\beta \in [2, 2.5]$ and $\gamma \in \{0.3, 0.5, 1.0, 1.2, 1.4, 1.6\}$. In our experiments, we use $\beta \in \{2, 2.5\}$ and a subset of $\gamma \in \{0.5, 0.8, 1.6\}$. To ensure fair comparison, the same β value is used across all baselines and RAPPO within each setting. Additional details are provided in Appendix D.1.

5.1 CONTROLLED GENERATION: SENTIMENT & TOXICITY

We evaluate RAPPO on two controlled generation tasks, positive sentiment steering and toxicity reduction—using the common evaluation table (Table 1). For both tasks, the base LM is first supervised fine-tuned to obtain the reference policy π_{ref} . For **sentiment**, we use GPT2-Large and the IMDB dataset ([Maas et al., 2011](#)), following the setup of ([Amini et al., 2024](#)). A pretrained sentiment classifier provides the reward signal. For **toxicity**, we use GPT-Neo-2.7B ([Black et al., 2021](#)) and the dataset of ([Amini et al., 2024](#)) constructed from REALTOXICITYPROMPTS ([Gehman et al., 2020](#)) by selecting 10,000 prompts with base toxicity > 0.3 ; relative rewards are given by a pretrained toxicity classifier. RAPPO- q means we remove the q highest-loss samples per batch ($q \in \{1, 2, 4\}$). On **IMDb**, all RAPPO variants achieve reward ≥ 1.66 (best **1.6811**), surpassing the strongest baseline (SIMPO, $\beta=2, \gamma=0.5$; **1.5611**) by 7.7%; even RAPPO-1 (**1.6625**) yields 6.5%. On **toxicity**, the best baseline reaches **6.30%**, whereas all RAPPO variants are markedly lower (**2.64%** to **2.28%**). Overall, RAPPO’s reference-aware pruning consistently improves alignment quality while substantially mitigating harmful generation.

5.2 SUMMARIZATION TASK

Beyond the IMDB and Toxicity benchmarks, we evaluate RAPPO on a summarization task. Following [Amini et al. \(2024\)](#), we adopt GPT-J-6B ([Wang & Komatsuzaki, 2021](#)) as the supervised fine-tuned

(SFT) reference policy π_{ref} and, to probe scalability and robustness, replicate the protocol with a larger, contemporary open-weight model, Llama-3.1-8B (Dubey et al., 2024). For each base model, we fine-tune with RAPPO-1 (ours), SIMPO, and DPO. Consistent with our observations on IMDb and Toxicity, SIMPO is relatively insensitive to hyperparameters; we therefore reuse its best configuration from those tasks ($\beta=2.5$, $\gamma=0.5$).

We evaluate with GPT-4 (Achiam et al., 2023) as an automatic judge, comparing paired model outputs under identical prompts and recording the preference (full prompt in Appendix D.5). As summarized in Figure 4, RAPPO consistently outperforms SIMPO and DPO on GPT-J-6B; the Llama-3.1-8B replication exhibits the same ordering, indicating that RAPPO’s gains transfer to a larger, more recent pretraining regime. For brevity, we report aggregate GPT-4 win rates in the main text and include per-prompt breakdowns in Appendix D.8.

5.3 LARGE-SCALE EVALUATION ON PKU-SAFERLHF

We evaluate RAPPO on the *PKU-SafeRLHF* benchmark (Ji et al., 2024) to assess whether it improves answer quality. We compare against state-of-the-art preference-optimization baselines—DPO, CPO, KTO, and SIMPO—training each method under the same reference policy (Mistral-7B), train/validation split, and compute budget. Decoding hyperparameters (temperature, top- p , maximum tokens, and stop rules) are held fixed across models, and all methods are evaluated on an identical prompt set (Appendix D.7). As model-based evaluation, we adopt the unified reward and cost models, beaver-7b-unified-reward/cost. For each prompt, each model generates a single response; we then report (i) **Beaver-Help** \uparrow (higher indicates greater helpfulness as scored by the reward model), (ii) **Beaver-Harm** \downarrow (lower indicates reduced inferred harm as scored by the cost model), and (iii) **Safety Rate** \uparrow (the fraction of responses flagged unsafe by the cost model ($\text{cost} \leq 0$ is safe)). To mitigate judge-specific biases in absolute scores, we additionally employ GPT-4 as a pairwise meta-evaluator: for each prompt, we compare RAPPO outputs against those of each baseline under identical decoding, and report the resulting **Win Rate** for RAPPO versus DPO, CPO, KTO, and SIMPO.

Results. On PKU-SafeRLHF, RAPPO attains the best performance on all metrics: safety rate 0.573 (absolute $+0.014$ vs. next-best DPO), helpfulness 0.693 ($+34.8\%$ over DPO), harmlessness 0.357 (-21.0% vs DPO), and highest win rate 65% (Fig. 5, details of these results are deferred in Appendix D.7). Computational source and training hyperparameters are provided in Appendix D.1.

6 ABLATION STUDY

6.1 SENSITIVITY ANALYSIS

In RAPPO, two hyperparameters control the strength of the reference-based filtering: the number of removed items q from each batch of misaligned samples and the reference alignment threshold τ . We investigate how sensitive RAPPO is to these two values. We conduct a sensitivity study on the IMDb sentiment control task with GPT2-Large. We first fix $\tau = 1$ and vary $q \in \{1, 2, 4, 8\}$ for batch sizes 16 and 32 to analyze sensitivity with respect to q . We then fix the batch size to 32 and vary $\tau \in \{0.8, 1.0, 1.2\}$ for the same set of q . We compare the results with the best baseline SIMPO in the IMDb experiments, which has a reward of 1.5537 with batch size 32.

Sensitivity to q . Across both batch sizes, RAPPO consistently improves over SIMPO by about 5% to 8% for all tested values of q , which shows that RAPPO is insensitive to the exact choice of q . Table 2 also suggests that a simple rule of selecting q is (batch size)/8.

Sensitivity to τ . In RAPPO, the alignment threshold τ is applied to the probability ratio between the preferred and dispreferred answers and decides how confident the reference model must be before we

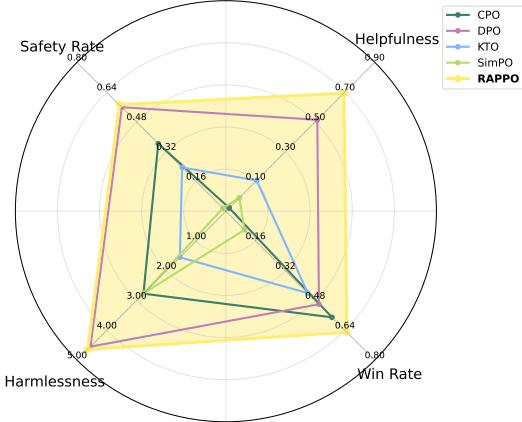


Figure 5: **PKU-SafeRLHF results.** RAPPO compared to DPO, CPO, KTO, and SIMPO under identical decoding and evaluation protocols.

Batch Size	$q = 1$	$q = 2$	$q = 4$	$q = 8$	Baseline
16 ($\tau = 1.0$)	1.7020	1.7481	1.7333	1.7111	1.6600 (SIMPO)
32 ($\tau = 0.8$)	1.6580	1.6720	1.6765	1.6510	1.5537 (SIMPO)
32 ($\tau = 1.0$)	1.6625	1.6790	1.6811	1.6432	1.5537 (SIMPO)
32 ($\tau = 1.2$)	1.6650	1.6805	1.6828	1.6595	1.5537 (SIMPO)

Table 2: Sensitivity results on q and τ .

Table 3: Win rates of RAPPO vs. SIMPO under different LLM judges.

Judge	RAPPO win rate	SIMPO win rate	Tie rate
GPT-4 (original)	58.30%	41.70%	0.00%
GPT-4o	68.00%	32.00%	0.00%
DeepSeek-V3	72.00%	28.00%	0.00%
Claude Opus 4.1	58.70%	34.78%	6.52%
Average	64.25%	34.12%	1.63%

filter a pair. Setting $\tau = 1$ is a natural default, and Table 2 shows that RAPPO outperforms SIMPO for all tested τ with only small changes in reward. This suggests that RAPPO is robust to the choice of τ , and that reasonable settings give consistently strong performance.

The sensitivity results on q and τ indicate that RAPPO is insensitive to the hyperparameters. But designing adaptive rules that learn an approximate optimal from data is an interesting direction for future work.

6.2 PERFORMANCE ACROSS DIFFERENT LLM JUDGES

In the following, we conduct a multi-judge evaluation and analyze the consistency of different judges when comparing RAPPO and SIMPO. We compare RAPPO and SIMPO on a summarization task using a Llama 3.1-8B model fine-tuned by each method, evaluated by four judges: GPT-4 (Achiam et al., 2023), GPT-4o (Hurst et al., 2024), DeepSeek-V3 (Liu et al., 2024a), and Claude Opus 4.1 (<https://www.anthropic.com/claude/opus>). All judges assess the same anonymized and order randomized response pairs and choose either the better summary or a tie. We measured how consistent the judges are with each other by computing pairwise Kendall’s τ between all judge pairs’ preference rankings. Kendall’s τ (Kendall, 1938) is a standard measure of rank correlation, where larger values mean stronger agreement between two rankings. In practice, values above 0.6 are usually viewed as strong agreement and values above 0.8 as very strong agreement.

Consistency of RAPPO’s advantage across judges. As summarized in Table 3, all four independent judges prefer RAPPO over SIMPO, with RAPPO win rates ranging from 58.30% to 72.00%. The average win rate of RAPPO (64.25%) is substantially higher than that of SIMPO (34.12%), while ties are rare. The fact that models developed by different organizations and trained with different data and objectives all favor RAPPO indicates that the observed gains are not an artifact of a particular judge such as GPT-4.

Inter-judge agreement. To quantify consistency across judges, we compute pairwise Kendall’s τ between the preference rankings induced by each judge. The average Kendall’s τ is 0.65 (Appendix: Table 9), indicating strong agreement on the relative ordering of RAPPO and SIMPO. Together with the consistent win rates in Table 3, this suggests that RAPPO’s advantages are robust to the choice of automatic judge rather than driven by any single LLM’s idiosyncrasies.

7 CONCLUSION

We identify a key limitation of DPO: its sensitivity to ambiguous samples, especially under weak or misaligned reference policies. To address this, we propose RAPPO (**R**eliable **A**lignment for **P**reference **P**olicy **O**ptimization), an order-aware variant of DPO that selectively filters out high-loss, hard-to-learn samples. Unlike reference-free methods such as SIMPO, RAPPO retains the reference policy and uses it selectively to improve stability. We show both theoretically and empirically that RAPPO improves generalization and consistently outperforms DPO and recent baselines, enabling more robust language model alignment.

540 ETHICS STATEMENT
541542 This work complies with the ICLR Code of Ethics. No human-subjects research or animal experimen-
543 tation was conducted. All datasets were used in accordance with their licenses and usage guidelines,
544 with care taken to avoid any privacy violations.
545546 REPRODUCIBILITY STATEMENT
547548 Our method builds on standard DPO, and the full implementation is provided in the Supplementary
549 Material. We also include complete details and links for all datasets and reference models used in our
550 experiments in Appendix D. In addition, we provide step-by-step proofs designed to be clear and
551 easy to follow in Appendix C.
552553 REFERENCES
554555 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
556 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
557 *arXiv preprint arXiv:2303.08774*, 2023.558 Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv*
559 *preprint arXiv:2402.10571*, 2024.560 Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree
561 search. *Advances in neural information processing systems*, 30, 2017.562 Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones,
563 Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory
564 for alignment. *arXiv preprint arXiv:2112.00861*, 2021.565 Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal
566 Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from
567 human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp.
568 4447–4455. PMLR, 2024.569 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,
570 Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with
571 reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.572 Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autore-
573 gressive language modeling with mesh-tensorflow. *If you use this software, please cite it using*
574 *these metadata*, 58(2), 2021.575 Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method
576 of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.577 Angelica Chen, Sadhika Malladi, Lily H Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and
578 Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. *Advances in*
579 *Neural Information Processing Systems*, 37:101928–101968, 2024a.580 Lichang Chen, Chen Zhu, Davit Soselia, Juhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang,
581 Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf.
582 *arXiv preprint arXiv:2402.07319*, 2024b.583 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared
584 Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large
585 language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.586 Shriram Chennakesavalu, Frank Hu, Sebastian Ibarraran, and Grant M Rotskoff. Energy rank
587 alignment: Using preference optimization to search chemical space at scale. *arXiv preprint*
588 *arXiv:2405.12961*, 2024.

594 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
 595 reinforcement learning from human preferences. *Advances in neural information processing*
 596 *systems*, 30, 2017.

597

598 Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li,
 599 Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language
 600 models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

601 Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick
 602 Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open
 603 instruction-tuned ILM, 2023.

604

605 Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving
 606 ILM alignment via preference data selection. *arXiv preprint arXiv:2502.14560*, 2025.

607 Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong
 608 Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional
 609 conversations. *arXiv preprint arXiv:2305.14233*, 2023.

610

611 Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao,
 612 Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative
 613 foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

614

615 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
 616 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
 617 *arXiv e-prints*, pp. arXiv–2407, 2024.

618

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

621

622 Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben
 623 Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to
 624 reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*,
 625 2022.

626

627 Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. Principled data
 628 selection for alignment: The hidden risks of difficult examples. *arXiv preprint arXiv:2502.09650*,
 629 2025.

630

631 Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In
 632 *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023a.

633

634 Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and
 635 Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine*
 636 *Learning*, pp. 10764–10799. PMLR, 2023b.

637

638 Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Real-
 639 toxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint*
 640 *arXiv:2009.11462*, 2020.

641

642 Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn
 643 Song. Koala: A dialogue model for academic research. *Blog post*, April, 1(6), 2023.

644

645 Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov,
 646 Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good
 647 alignment. *arXiv preprint arXiv:2404.09656*, 2024.

648

649 Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without
 650 reference model. *arXiv preprint arXiv:2403.07691*, 2024.

651

652 Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao,
 653 and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized
 654 rehearsals. *arXiv preprint arXiv:2403.01244*, 2024.

648 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
 649 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
 650 *arXiv:2410.21276*, 2024.

651 Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi
 652 Zhou, Kaile Wang, Boxuan Li, et al. Pku-saferlfhf: Towards multi-level safety alignment for llms
 653 with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

654 Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement
 655 learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10, 2023.

656 Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for
 657 empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*,
 658 pp. 669–679. PMLR, 2020.

659 Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

660 Geon-Hyeong Kim, Youngsoo Jang, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae,
 661 and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced
 662 safety. *arXiv preprint arXiv:2505.20065*, 2025.

663 Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document
 664 summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35, 2022.

665 Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith
 666 Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant
 667 conversations-democratizing large language model alignment. *Advances in Neural Information
 668 Processing Systems*, 36:47669–47681, 2023.

669 Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu,
 670 Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models
 671 for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

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

675 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 676 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
 677 *arXiv:2412.19437*, 2024a.

678 Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalmi, Mohammad Saleh, Peter J Liu, and Jialu Liu.
 679 Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*,
 680 2023a.

681 Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for
 682 alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint*
 683 *arXiv:2312.15685*, 2023b.

684 Yixin Liu, Pengfei Liu, and Arman Cohan. Understanding reference policies in direct preference
 685 optimization. *arXiv preprint arXiv:2407.13709*, 2024b.

686 Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng,
 687 Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical
 688 reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*,
 689 2023a.

690 Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of
 691 catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint*
 692 *arXiv:2308.08747*, 2023b.

693 Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts.
 694 Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the
 695 association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

702 Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-
 703 free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.
 704

705 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 706 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow
 707 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–
 708 27744, 2022.

709 Ryan Park, Rafael Raffailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in
 710 direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.
 711

712 Xiyue Peng, Hengquan Guo, Jiawei Zhang, Dongqing Zou, Ziyu Shao, Honghao Wei, and Xin Liu.
 713 Enhancing safety in reinforcement learning with human feedback via rectified policy optimization.
 714 *arXiv preprint arXiv:2410.19933*, 2024.

715 Rafael Raffailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
 716 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances
 717 in Neural Information Processing Systems*, 36:53728–53741, 2023.

718 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 719 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
 720

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

724 Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,
 725 Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. *URL
 726 https://arxiv.org/abs*, 2009.
 727

728 Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov,
 729 Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the perfor-
 730 mance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*,
 731 2024.

732 Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy
 733 Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.
 734

735 Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model,
 736 2021.

737 Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Ad-
 738 vancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*,
 739 2023.
 740

741 Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less:
 742 Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

743 Teng Xiao, Yige Yuan, Zhengyu Chen, Mingxiao Li, Shangsong Liang, Zhaochun Ren, and Vasant G
 744 Honavar. Simper: A minimalist approach to preference alignment without hyperparameters. *arXiv
 745 preprint arXiv:2502.00883*, 2025.

747 Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton
 748 Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm
 749 performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

750 Sen Yang, Leyang Cui, Deng Cai, Xinting Huang, Shuming Shi, and Wai Lam. Not all preference
 751 pairs are created equal: A recipe for annotation-efficient iterative preference learning. *arXiv
 752 preprint arXiv:2406.17312*, 2024.
 753

754 Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf:
 755 Rank responses to align language models with human feedback without tears. *arXiv preprint
 756 arXiv:2304.05302*, 2023.

756 Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level
757 direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.
758

759 Tiantian Zhang, Xueqian Wang, Bin Liang, and Bo Yuan. Catastrophic interference in reinforcement
760 learning: A solution based on context division and knowledge distillation. *IEEE Transactions on*
761 *Neural Networks and Learning Systems*, 34(12):9925–9939, 2022.

762 Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf:
763 Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.
764

765 Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia
766 Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information*
767 *Processing Systems*, 36:55006–55021, 2023.

768 Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul
769 Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv*
770 *preprint arXiv:1909.08593*, 2019.
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

810 **A LIMITATIONS**
 811

812 While our method achieves strong results across multiple tasks, our study has two main limitations.
 813 First, due to computational constraints, we evaluate only on relatively small models (parameters
 814 $\leq 8B$); scaling to larger models remains future work. Second, our intra-batch filtering currently relies
 815 on a fixed q ; developing an adaptive or learned strategy for selecting q is an important next step in the
 816 future.

817 **B LLM USAGE**
 818

819 Large language models were used to edit and polish the manuscript—specifically to improve wording,
 820 readability, and clarity. No new scientific content was generated by the models; all analyses and
 821 conclusions are the authors' own.

822 **C DETAILED THEOREM PROOFS**
 823

824 **C.1 PROOF OF THEOREM 4.7**
 825

826 *Theorem* (Restatement of Theorem 4.7). Fix iteration t with $B_t > q$ and let $T \subseteq S_t^{\text{un}}$ be any set with
 827 $|T| = K_t$, while T_t^* contains the K_t largest $z_{t,i}$ in S_t^{un} . Conditioned on the realized scores $\{z_{t,i}\}$:

828 (i) *Maximal expected first-order decrease of R .*

$$829 \mathbb{E}\left[-\langle d_t, \hat{g}_t(T) \rangle \mid \{z_{t,i}\}\right] \leq \mathbb{E}\left[-\langle d_t, \hat{g}_t(T_t^*) \rangle \mid \{z_{t,i}\}\right]. \quad (15)$$

830 (ii) *Minimal conditional variance of the signed decrease.* Let $\Phi_t(T) := \langle -\hat{g}_t(T), d_t \rangle$. Then

$$831 \text{Var}(\Phi_t(T) \mid \{z_{t,i}\}) \geq \text{Var}(\Phi_t(T_t^*) \mid \{z_{t,i}\}). \quad (16)$$

832 (iii) *Uniform stability and generalization.* Let $D = (Z_1, \dots, Z_n)$ and $D' = (Z'_1, \dots, Z'_n)$ be two
 833 datasets that differ in one example, both i.i.d. from \mathcal{D} . Run the same randomness to produce $\{\theta_t\}$ and
 834 $\{\theta'_t\}$. Define $\Delta_t := \mathbb{E}\|\theta_t - \theta'_t\|$. Then for any horizon T and stepsizes $\{\eta_t\}$,

$$835 \Delta_{t+1} \leq (1 + L\eta_t)\Delta_t + \frac{2C}{s - q} \eta_t \mathbb{E}\left[\max_{i \in \text{Kept}_t} w(z_{t,i})\right], \quad (17)$$

836 and

$$837 \Delta_T \leq \frac{2C}{s - q} \exp\left(\sum_{u=0}^{T-1} L\eta_u\right) \sum_{t=0}^{T-1} \eta_t \mathbb{E}\left[\max_{i \in \text{Kept}_t} w(z_{t,i})\right]. \quad (18)$$

838 By equation 6,

$$839 \varepsilon_T := \sup_z \mathbb{E}|L(\theta_T; z) - L(\theta'_T; z)| \leq G\Delta_T, \quad (19)$$

840 and then the expected generalization error is,

$$841 \mathbb{E}[R(\theta_T) - R_n(\theta_T)] \leq \varepsilon_T, \quad R_n(\theta) = \frac{1}{n} \sum_{i=1}^n L(\theta; Z_i). \quad (20)$$

842 Among all rules that never drop trusted items and, when $B_t > q$, keep exactly K_t untrusted items,
 843 our rule selecting minimizes the per-step term $\mathbb{E}[\max_{i \in \text{Kept}_t} w(z_{t,i})]$ and hence minimizes Δ_T and
 844 ε_T . If $B_t \leq q$ or $q = 0$, the update equals standard mini-batch SGD.

845 *Proof. Proof of part (i): expected first-order decrease*

864 First, based on the smoothness and Lipschitz Assumption Assumption 4.2, since R is L -smooth, for
 865 $\theta_{t+1} = \theta_t - \eta_t \hat{g}_t$, we have
 866

$$867 R(\theta_{t+1}) \leq R(\theta_t) - \eta_t \langle d_t, \hat{g}_t \rangle + \frac{L}{2} \eta_t^2 \|\hat{g}_t\|^2. \quad (21)$$

869 Thus, the first-order decrease is $-\eta_t \langle d_t, \hat{g}_t \rangle$. For any $T \subseteq S_t^{\text{un}}$ with $|T| = K_t$, we have
 870

$$871 \langle d_t, \hat{g}_t(T) \rangle = \frac{1}{s-q} \sum_{i \in S_t^{\text{tr}}} \langle d_t, g_{t,i} \rangle + \frac{1}{s-q} \sum_{i \in T} \langle d_t, g_{t,i} \rangle. \quad (22)$$

873 Remark here that the first sum is independent of T . Condition on $\{z_{t,i}\}$ and define $\bar{a}(z) :=$
 874 $\mathbb{E}[\langle d_t, g_{t,i} \rangle \mid z_{t,i} = z]$. In weighted cases (e.g., DPO) where $g_{t,i} = w(z_{t,i}) u_{t,i}$,

$$876 \mathbb{E}[\langle d_t, g_{t,i} \rangle \mid z_{t,i}] = w(z_{t,i}) \bar{a}(z_{t,i}) =: h(z_{t,i}).$$

877 Therefore

$$878 \mathbb{E}[\langle d_t, \hat{g}_t(T) \rangle \mid \{z_{t,i}\}] = C_0 + \frac{1}{s-q} \sum_{i \in T} h(z_{t,i}),$$

881 where C_0 is some constant calculated from $\frac{1}{s-q} \sum_{i \in S_t^{\text{tr}}} \langle d_t, g_{t,i} \rangle$.

883 Since h is nondecreasing, if $z_{t,j} > z_{t,i}$ then $h(z_{t,j}) \geq h(z_{t,i})$. Now suppose $T \neq T_t^*$. Then there
 884 exists $i \in T$ and $j \notin T$ such that $z_{t,j} > z_{t,i}$. Define $T' = (T \setminus \{i\}) \cup \{j\}$. Then

$$885 \sum_{k \in T'} h(z_{t,k}) = \sum_{k \in T} h(z_{t,k}) - h(z_{t,i}) + h(z_{t,j}) \geq \sum_{k \in T} h(z_{t,k}),$$

888 with strict inequality if $h(z_{t,j}) > h(z_{t,i})$.

889 By repeating such one-swap exchanges, we can transform any T into T_t^* while never decreasing the
 890 sum. Therefore

$$891 \sum_{i \in T} h(z_{t,i}) \leq \sum_{i \in T_t^*} h(z_{t,i}).$$

893 Substituting back into the expectation yields

$$895 \mathbb{E}[-\langle d_t, \hat{g}_t(T) \rangle \mid \{z_{t,i}\}] \leq \mathbb{E}[-\langle d_t, \hat{g}_t(T_t^*) \rangle \mid \{z_{t,i}\}],$$

896 with strict inequality whenever $T \neq T_t^*$ and h is strictly increasing on the relevant interval. This
 897 proves the claim (i).

899 Proof of part (ii): minimal conditional variance

900 Let $\mu_{t,i} := \mathbb{E}[g_{t,i} \mid z_{t,i}]$ and $\xi_{t,i} := g_{t,i} - \mu_{t,i}$. Recall that

$$902 \hat{g}_t(T) = \frac{1}{s-q} \left(\sum_{i \in S_t^{\text{tr}}} g_{t,i} + \sum_{i \in T} g_{t,i} \right), \quad \Phi_t(T) = \langle -\hat{g}_t(T), d_t \rangle,$$

905 and $\Phi_t(T) := \langle -\hat{g}_t(T), d_t \rangle$, then

$$907 \Phi_t(T) = \left\langle -\frac{1}{s-q} \sum_{i \in S_t^{\text{tr}}} g_{t,i} - \frac{1}{s-q} \sum_{i \in T} \mu_{t,i}, d_t \right\rangle + \left\langle -\frac{1}{s-q} \sum_{i \in T} \xi_{t,i}, d_t \right\rangle.$$

910 Given $\{z_{t,i}\}$ the first term is deterministic, so

$$912 \text{Var}(\Phi_t(T) \mid \{z\}) = \text{Var} \left(\left\langle -\frac{1}{s-q} \sum_{i \in T} \xi_{t,i}, d_t \right\rangle \mid \{z\} \right).$$

915 Therefore, according to the conditional independence,

$$917 \text{Var}(\Phi_t(T) \mid \{z\}) = \frac{1}{(s-q)^2} \sum_{i \in T} \text{Var}(\langle \xi_{t,i}, d_t \rangle \mid z_{t,i}).$$

918 For any random vector X and unit u , $\text{Var}(\langle X, u \rangle) \leq \text{tr} \text{Cov}(X)$, hence
 919
 920

$$921 \quad \text{Var}(\Phi_t(T) \mid \{z\}) \leq \frac{\|d_t\|^2}{(s-q)^2} \sum_{i \in T} \text{tr} \text{Cov}(\xi_{t,i} \mid z_{t,i}) \quad (23)$$

$$924 \quad \leq \frac{\|d_t\|^2}{(s-q)^2} \sum_{i \in T} \text{tr} \text{Cov}(g_{t,i} \mid z_{t,i}) \leq \frac{\|d_t\|^2}{(s-q)^2} \sum_{i \in T} \phi(w(z_{t,i})), \quad (24)$$

927 by using the fact that subtracting a mean does not increase covariance.
 928

929 Since ϕ is nondecreasing and w is nonincreasing, the composite $z \mapsto \phi(w(z))$ is nonincreasing in z .
 930 Thus the sum is minimized by the K_t largest z , i.e., T_t^* . Hence $\text{Var}(\Phi_t(T) \mid \{z\}) \geq \text{Var}(\Phi_t(T_t^*) \mid \{z\})$, proving (ii).
 931

932 Proof of part (iii): stability and generalization

933 Let $\theta_{t+1} = \theta_t - \eta_t \hat{g}_t$ and $\theta'_{t+1} = \theta'_t - \eta_t \hat{g}'_t$. Add and subtract $\hat{g}_t(\theta'_t)$ and apply equation 6:
 934

$$935 \quad \|\theta_{t+1} - \theta'_{t+1}\| \leq (1 + L\eta_t) \|\theta_t - \theta'_t\| + \eta_t \|\hat{g}_t - \hat{g}'_t\|.$$

937 Taking expectation gives

$$938 \quad \Delta_{t+1} \leq (1 + L\eta_t) \Delta_t + \eta_t \mathbb{E} \|\hat{g}_t - \hat{g}'_t\|. \quad (25)$$

940 Compare two training datasets that differ in one example:
 941

$$942 \quad D = (Z_1, \dots, Z_{j-1}, Z_j, Z_{j+1}, \dots, Z_n), \quad D' = (Z_1, \dots, Z_{j-1}, Z'_j, Z_{j+1}, \dots, Z_n),$$

944 with $Z_i = Z'_i$ for all $i \neq j$ and $Z_j \neq Z'_j$. Both D and D' are i.i.d. draws from \mathcal{D}^n . Because D and
 945 D' differ in one example, in any step, the kept sets differ in at most one index; In addition, let Kept_t
 946 and Kept'_t be the kept sets. Because S_t and S'_t differ in at most one element and both selection rules
 947 keep the same cardinalities (all trusted and K_t untrusted), the two kept sets differ in at most one
 948 index. Hence

$$949 \quad \|\hat{g}_t - \hat{g}'_t\| \leq \frac{1}{s-q} \|g_{t,a}\| + \frac{1}{s-q} \|g'_{t,b}\| \leq \frac{2}{s-q} \max_{i \in \text{Kept}_t \cup \text{Kept}'_t} \|g_{t,i}\|.$$

952 Taking expectation and using Eq. (7) on untrusted items,

$$953 \quad \mathbb{E} \|\hat{g}_t - \hat{g}'_t\| \leq \frac{2C}{s-q} \mathbb{E} \left[\max_{i \in \text{Kept}_t} w(z_{t,i}) \right].$$

956 Since the trusted items are always kept, only the untrusted choice varies. Since w is nonincreasing in
 957 z , the maximum kept weight among untrusted items is minimized by the K_t largest z , i.e., by T_t^* .
 958 Plugging into equation 25 and unrolling with $\prod(1 + L\eta_u) \leq \exp(\sum L\eta_u)$ gives
 959

$$960 \quad \Delta_T \leq \frac{2C}{s-q} \exp \left(\sum_{u=0}^{T-1} L\eta_u \right) \sum_{t=0}^{T-1} \eta_t \mathbb{E} \left[\max_{i \in \text{Kept}_t} w(z_{t,i}) \right],$$

963 which is minimized by our rule stepwise. By equation 6,

$$965 \quad \varepsilon_T := \sup_z \mathbb{E} |L(\theta_T; z) - L(\theta'_T; z)| \leq G \Delta_T,$$

967 and the expected generalization gap satisfies

$$969 \quad \mathbb{E} [R(\theta_T) - R_n(\theta_T)] \leq \varepsilon_T, \quad R_n(\theta) = \frac{1}{n} \sum_{i=1}^n L(\theta; Z_i), \quad Z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}.$$

971 This proves (iii). □

972 C.2 PROOF OF PROPOSITION 4.8
973

974 **Proposition C.1.** Let the whole dataset size be N . The number of whole unaligned examples
975 is $n_b = \hat{\mu}N$ and of whole aligned examples is $n_g = (1 - \hat{\mu})N$. At each iteration, we sample
976 a mini-batch $S \subseteq [N]$ without replacement, $|S| = s$. Write the unaligned and aligned counts
977 in the batch as $\mathcal{U}_B, \mathcal{A}_B$. For a parameter vector θ , the per-example loss is $\ell_i \text{ or } j(\theta)$. And denote
978 $m_g := \sum_{i \in \mathcal{A}_B} \ell_i(z^i; \theta)$, $m_b := \sum_{j \in \mathcal{U}_B} \ell_j(z^j; \theta)$. Then, we can obtain the objective function of
979 RAPPO:

$$980 \hat{\mathcal{L}}^{RAPPO} = \underbrace{\sum_{b=0}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + m_b}{s}}_{\text{only keep aligned data}} + \underbrace{\sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s - q}}_{\text{keep aligned and remove } q \text{ largest loss data sample}}. \quad (26)$$

985 where $\mathbb{P}(|\mathcal{U}_B| = b) = \frac{\binom{\hat{\mu}N}{b} \binom{(1-\hat{\mu})N}{s-b}}{\binom{N}{s}}$, $\alpha_j = \frac{\sum_{l=0}^{b-q-1} \binom{j-1}{l} \binom{\hat{\mu}N-j}{b-1-l}}{\binom{\hat{\mu}N}{b}}$. And, \tilde{g}_t is unbiased estimator of
986 $\partial \hat{\mathcal{L}}^{RAPPO}$ that $\mathbb{E}[\tilde{g}_t] \in \partial \hat{\mathcal{L}}^{RAPPO}$.
987

989 **Proof:**

990 In RAPPO, we keep **all** aligned samples in the batch and, among the b unaligned samples, we remove
991 only the **largest** q bad losses (if $|\mathcal{U}_B| \leq q$, we keep all data in the batch). We can denote the number
992 of kept samples is: $(s - b) + (b - q) = s - q$, which means kept sample in batch equals all aligned
993 data(amount of $s - b$) and kept sample in unaligned data(amount of $b - q$).
994

995 Hence the per-step loss of a batch is:

$$996 \hat{\mathcal{L}}_{step}^{RAPPO} = \begin{cases} \frac{1}{s} \sum_{i \in \mathcal{B}} \ell_i(\theta), & \text{if } b \leq q, \\ \frac{1}{s - q} \left(\sum_{i \in \mathcal{A}_B} \ell_i(\theta) + \sum_{j=1}^{b-q} \ell_{(j)}(\theta) \right), & \text{if } b > q \end{cases}$$

1003 Now, based on the condition given by unaligned sample b , we will have two situation and we need to
1004 analysis the conditional expectation.

1005 **Regime 1:** $b \leq q$ that we keep all data sample. Thus, the remaining samples are s under this situation.
1006 Therefore, we have:

$$1008 \mathbb{E}_S \left[\hat{\mathcal{L}}_{step}^{RAPPO} \mid |\mathcal{U}_B| = b \right] = \frac{1}{s} \sum_{i \in \mathcal{B}} \ell_i = \frac{m_g + m_b}{s}$$

1010 **Regime 2:** $b > q$ that we will keep $s - b$ aligned samples and the $b - q$ among the b unaligned-in-batch
1011 samples. Thus, we have:

$$1013 \mathbb{E}_S \left[\hat{\mathcal{L}}_{step}^{RAPPO} \mid |\mathcal{U}_B| = b \right] = \frac{1}{s - q} \left(\sum_{i \in \mathcal{A}_B} \ell_i \mathbb{P}(i \in \mathcal{A}_B \mid |\mathcal{U}_B| = b) \right) + \frac{1}{s - q} \left(\sum_{j=1}^{b-q} \ell_{(j)} \mathbb{P}((j) \in q \mid |\mathcal{U}_B| = b) \right) \\ 1014 = \frac{m_g}{s - q} + \frac{1}{s - q} \left(\sum_{j=1}^{\hat{\mu}N} \frac{\sum_{l=0}^{b-q-1} \binom{j-1}{l} \binom{\hat{\mu}N-j}{b-1-l} \ell_{(j)}}{\binom{\hat{\mu}N}{b}} \right)$$

1020 Here, we consider the n_b unaligned instances ordered increasingly by loss and fix the j -th smallest
1021 element. A mini-batch of size b is sampled uniformly without replacement from these n_b instances.
1022 The probability that the j -th element is both selected and ranks within the batch's smallest $b - q$ is
1023 $\frac{\sum_{l=0}^{b-q-1} \binom{j-1}{l} \binom{\hat{\mu}N-j}{b-1-l}}{\binom{\hat{\mu}N}{b}}$. The denominator $\binom{\hat{\mu}N}{b}$ counts all possible size- b subsets, while the numerator
1024 counts exactly those subsets that contain the j -th element and include at most $b - q - 1$ elements
1025 smaller than it, ensuring the j -th element lies among the batch's $b - q$ smallest.

1026 Hence, we can obtain the empirical objective function:
 1027

$$\begin{aligned}
 \hat{\mathcal{L}}^{RAPPO} &= \mathbb{E}_S[\hat{\mathcal{L}}_{step}^{RAPPO}] = \sum_b \mathbb{P}(|\mathcal{U}_B| = b) \mathbb{E} \left[\mathcal{L}_{step}^{RAPPO} \mid |\mathcal{U}_B| = b \right] \\
 &= \sum_{b=0}^{\min(q, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \mathbb{E} \left[\mathcal{L}_{step}^{RAPPO} \mid |\mathcal{U}_B| = b \right] + \sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \mathbb{E} \left[\mathcal{L}_{step}^{RAPPO} \mid |\mathcal{U}_B| = b \right] \\
 &= \underbrace{\sum_{b=0}^{\min(q, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + m_b}{s}}_{\text{keep all data}} + \underbrace{\sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s - q}}_{\text{keep aligned and remove } q \text{ largest loss data sample}} \\
 \text{where we denote } \alpha_j &= \frac{\sum_{l=0}^{q-1} \binom{j-1}{l} \binom{\hat{\mu}N-j}{b-1-l}}{\binom{\hat{\mu}N}{b}}.
 \end{aligned}$$

1038 Then, we will analysis the probability distribution of \mathcal{U}_B . Under sampling without replacement from
 1041 a population with $n_b = \hat{\mu}N$ bad and $n_g = (1 - \hat{\mu})N$ good, the batch bad-count is hypergeometric:
 1043

$$\mathbb{P}(|\mathcal{U}_B| = b) = \frac{\binom{\hat{\mu}N}{b} \binom{(1-\hat{\mu})N}{s-b}}{\binom{N}{s}}, \quad b = 0, 1, \dots, \min\{s, \hat{\mu}N\}.$$

1047 Therefore, we get the objective function as:
 1048

$$\hat{\mathcal{L}}^{RAPPO} = \underbrace{\sum_{b=0}^{\min(q, \hat{\mu}N)} \frac{\binom{\hat{\mu}N}{b} \binom{(1-\hat{\mu})N}{s-b}}{\binom{N}{s}} \frac{m_g + m_b}{s}}_{\text{keep all data}} + \underbrace{\sum_{b=q+1}^{\min(s, \hat{\mu}N)} \frac{\binom{\hat{\mu}N}{b} \binom{(1-\hat{\mu})N}{s-b}}{\binom{N}{s}} \frac{m_g + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s - q}}_{\text{keep aligned and remove } q \text{ largest loss data sample}}$$

1053 Then, let's analysis \tilde{g}_t . We have:
 1054

$$\mathbb{E}[\tilde{g}_t \mid |\mathcal{U}_B| = b] \in \begin{cases} \partial \left(\frac{\mathbb{E}[m_g + m_b \mid |\mathcal{U}_B| = b]}{s} \right), & b \leq q \\ \partial \left(\frac{\mathbb{E}[m_g \mid |\mathcal{U}_B| = b] + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s - q} \right), & b > q \end{cases}$$

1055 Now, we know:
 1062

$$\begin{aligned}
 \mathbb{E}[\tilde{g}_t] &\in \partial \left(\sum_{b=0}^{\min(q, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \mathbb{E}[\tilde{g}_t \mid |\mathcal{U}_B| = b] + \sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \mathbb{E}[\tilde{g}_t \mid |\mathcal{U}_B| = b] \right) \\
 &= \partial \left(\sum_{b=0}^{\min(q, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + m_b}{s} + \sum_{b=q+1}^{\min(s, \hat{\mu}N)} \mathbb{P}(|\mathcal{U}_B| = b) \frac{m_g + \sum_{j=1}^{\hat{\mu}N} \alpha_j \ell_{(j)}}{s - q} \right) \\
 &= \partial \hat{\mathcal{L}}^{RAPPO}
 \end{aligned}$$

1072 This result show that \tilde{g}_t is unbiased estimator of $\partial \hat{\mathcal{L}}^{RAPPO}$, thereby justifying our update rule and
 1073 ensuring the stochastic updates are consistent with minimizing the RAPPO objective.
 1074

1075 D EXPERIMENT DETAILS

1076 In this section, we first summarize the objective functions of recent popular DPO and preference-
 1077 based fine-tuning algorithms in Table 4. Our experiments compare RAPPO(Ours) with DPO, IPO,
 1078 DPO with offset, and SIMPO.
 1079

Table 4: Recent Preference Optimization Objectives. Various popular objectives for preference optimization are defined based on preference data $\mathcal{D} = (x, y_w, y_l)$, where x is the input (or prompt), and y_w and y_l denote the preferred (winning) and dispreferred (losing) responses, respectively. Amini et al. (2024) define the preference margin as $\Delta_r = \alpha \log(r(y_w) - r(y_l))$, where $r(y_w) = \beta \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}$ and $r(y_l) = \beta \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$. In SIMPO, γ denotes a predefined reward margin. The reference KL regularization in Ethayarajh et al. (2024) term is defined as $z_{\text{ref}} = \mathbb{E}(x, y) \sim \mathcal{D} [\beta \text{KL}(\pi_\theta(y|x) \|\pi_{\text{ref}}(y|x))]$, and the normalized score $p_\theta(y|x)$ in Hong et al. (2024) is given by $p_\theta(y|x) = \exp\left(\frac{1}{|y|} \log \pi_\theta(y|x)\right)$.

Method	Objective
DPO (Rafailov et al. (2023))	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - \beta \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} \right) \right]$
IPO (Azar et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} - \frac{1}{2\beta} \right)^2$
DPO-offset (Amini et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - \beta \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} - \Delta_r \right) \right]$
SIMPO (Meng et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\frac{\beta}{ y_w^i } \log \pi_\theta(y_w^i x^i) - \frac{\beta}{ y_l^i } \log \pi_\theta(y_l^i x^i) - \gamma \right) \right]$
RSO(Hinge Norm) (Liu et al. (2023a))	$\frac{1}{N} \sum_{i=1}^N \max \left(0, 1 - \left[\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - \beta \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} \right] \right)$
Chen et al. (2024a)	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_\theta(y_l^i x^i)} + \beta \cdot \nu \cdot \log \frac{\pi_{\text{ref}}(y_w^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} \right) \right]$
RRHF(Yuan et al. (2023))	$\frac{1}{N} \sum_{i=1}^N \max \left(0, -\frac{1}{ y_w^i } \log \pi_\theta(y_w^i x^i) + \frac{1}{ y_l^i } \log \pi_\theta(y_l^i x^i) - \lambda \log \pi_\theta(y_w^i x^i) \right)$
KTO(Ethayarajh et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\lambda \sigma \left(\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - z_{\text{ref}} \right) + \lambda \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} \right) \right]$
SLiC-HF(Zhao et al. (2023))	$\frac{1}{N} \sum_{i=1}^N \max \left(0, \delta - \log \pi_\theta(y_w^i x^i) + \log \pi_\theta(y_l^i x^i) - \lambda \log \pi_\theta(y_w^i x^i) \right)$
CPO(Xu et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \pi_\theta(y_w^i x^i) - \beta \log \pi_\theta(y_l^i x^i) - \lambda \log \pi_\theta(y_w^i x^i) \right) \right]$
ORPO(Hong et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\log p_\theta(y_w^i x^i) - \lambda \log \sigma \left(\log \frac{p_\theta(y_w^i x^i)}{1-p_\theta(y_w^i x^i)} - \log \frac{p_\theta(y_l^i x^i)}{1-p_\theta(y_l^i x^i)} \right) \right]$
R-DPO(Park et al. (2024))	$\frac{1}{N} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w^i x^i)}{\pi_{\text{ref}}(y_w^i x^i)} - \beta \log \frac{\pi_\theta(y_l^i x^i)}{\pi_{\text{ref}}(y_l^i x^i)} + (\alpha y_w^i - \alpha y_l^i) \right) \right]$
RAPPO (Ours)	$\frac{s}{N(s-q)} \sum_{i=1}^N \left[-\log \sigma \left(\beta \log \frac{C_\theta(z^i)}{C_{\text{ref}}(z^i)} \right) \right] - \sum_{j=1}^{\hat{\mu}N} \left(\frac{\gamma_j}{s-q} \right) \ell(\pi_\theta; \pi_{\text{ref}}; z^{(j)})$

D.1 HYPERPARAMETER CONFIGURATION AND GPU RESOURCE

The hyperparameters used in our experiments are summarized in the Table 5 and 6. RAPPO, DPO, IPO, and DPO with Offset share the same set of hyperparameters, while SIMPO uses the tuned parameters reported in its original paper.

Table 5: Hyperparameter settings for SIMPO.

Experiment	β	γ	Learning rate
IMDB sentiment control	2.0, 2.5	0.5, 0.8, 1.6	5e-7
Toxicity control	2.0, 2.5	0.5, 0.8, 1.6	5e-7
Text Summarization	2.5	0.5	5e-7
PKU-SafeRLHF Large-Scale Evaluation	2.5	0.5	1e-6

Table 6: Hyperparameter settings for RAPPO and other baselines(DPO, IPO, DPO with offset, CPO, KTO).

Experiment	β	Learning rate
IMDB sentiment control	0.7	5e-7
Toxicity control	0.05	5e-7
Text Summarization	0.5	5e-7
PKU-SafeRLHF Large-Scale Evaluation	0.1	1e-6

Computation Cost. We conduct the IMDB and Toxicity control experiments on two NVIDIA A40-46G GPUs, with each experiment taking approximately 45 minutes for both training and answer

1134 sampling. For the text summarization experiment, we use eight NVIDIA A100-80G GPUs, requiring
 1135 about 2 hours for training and an additional 1.5 hours for answer sampling per experiment. For the
 1136 PKU-SafeRLHF Large-Scale Evaluation, we use eight NVIDIA A100-80G GPUs, requiring about 7
 1137 hours for training and an additional 4 hours for answer sampling per experiment.
 1138

1139 D.2 SENTIMENT CONTROL

1140
 1141 We adopt the reward model available at <https://huggingface.co/lvwerra/distilbert-imdb>. Given a
 1142 response y , the reward is defined as $r(y) = 1 + p(\text{pos} \mid \cdot)$ if the response is predicted as positive,
 1143 and $r(y) = 1 - p(\text{neg} \mid \cdot)$ otherwise, where p denotes the classifier's confidence score (following
 1144 the setup in [Amini et al. \(2024\)](#)). A higher reward indicates stronger alignment with the positive
 1145 sentiment objective. We construct preference pairs following the DPO-Offset procedure [Amini](#)
 1146 [et al. \(2024\)](#) by sampling responses from π_{ref} and scoring them using the reward model. Due to the
 1147 parameter sensitivity of SimPO, we tested multiple configurations, reporting the optimal results in
 1148 Table 1 and including the full set of results in Table 7.
 1149

1150 Table 7: Comparison of reward scores and toxicity percentages across SimPO and RAPPO with
 1151 various parameters, evaluated on the IMDB and Real-Toxicity-Prompts [Gehman et al. \(2020\)](#) test set.
 1152 Higher reward scores and lower toxicity indicate better performance. All values are averaged over
 1153 three random seeds.

Algorithm	SimPO						RAPPO		
	$\beta=2, \gamma=0.5$	$\beta=2, \gamma=0.8$	$\beta=2, \gamma=1.6$	$\beta=2.5, \gamma=0.5$	$\beta=2.5, \gamma=0.8$	$\beta=2.5, \gamma=1.6$	$q=1$	$q=2$	$q=4$
Reward Score \uparrow	1.5611	1.5453	1.5483	1.5537	1.5503	1.5601	1.6625	1.6790	1.6811
Toxicity (%) \downarrow	7.48	8.22	7.48	7.49	8.40	8.35	2.64	2.60	2.28

1154 D.3 TOXICITY CONTROL

1155 For the Toxicity Control experiment, we adopt a toxicity classifier as the reward model, available at
 1156 <https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target>, which is used in [Amini](#)
 1157 [et al. \(2024\)](#). Additional results for SimPO under a wider range of parameters are provided in Table 7.
 1158

1159 D.4 COMPLEXITY AND RUNNING TIME

1160 To quantify the overhead, we benchmarked training on the *Toxicity* task: standard DPO processes
 1161 a 256-sample mini-batch in roughly 40 s, whereas RAPPO requires about 50 s for the same batch
 1162 size. This $\sim 25\%$ wall-clock increase is modest, indicating that the additional computational cost is
 1163 negligible in practice.
 1164

1165 D.5 GPT-4 SCORING PROMPT FOR SUMMARIZATION

1166 We follow the pairwise evaluation prompt used in prior work([Rafailov et al. \(2023\)](#); [Amini et al.](#)
 1167 [\(2024\)](#)) to assess summarization quality with GPT-4 (see Table 8). Given an input forum post and
 1168 two candidate summaries (A and B), GPT-4 is instructed to identify which summary better captures
 1169 the most important points while remaining concise and precise. If both summaries are equally
 1170 good, GPT-4 is allowed to select a tie. The evaluation requires GPT-4 to output a one-sentence
 1171 justification(in Table 13, 14, 15,16).
 1172

1173 D.6 DETAILED KENDALL'S τ MATRIX

1174 Table 9: Pairwise Kendall's τ between LLM judges' preference rankings.
 1175

	GPT-4	GPT-4o	DeepSeek-V3	Claude Opus 4.1
GPT-4	1.0000	0.7000	0.6100	0.6786
GPT-4o	0.7000	1.0000	0.6250	0.6600
DeepSeek-V3	0.6100	0.6250	1.0000	0.6450
Claude Opus 4.1	0.6786	0.6600	0.6450	1.0000

1188 Table 8: Prompt used for GPT-4-based pairwise evaluation of summarization quality.
1189

GPT-4 Evaluation Prompt	
Which of the following summaries does a better job of summarizing the most important points in the given forum post, without including unimportant or irrelevant details? A good summary is both concise and precise.	
Input: {prompt_text}	
Summary A: {summary_a}	
Summary B: {summary_b}	
If both summaries are equally good—i.e., they capture the same key points with similar clarity and conciseness—then you should respond with a tie.	
Instructions:	
<ul style="list-style-type: none"> • First, provide a one-sentence comparison of the two summaries, explaining which you prefer and why (or state that they are equally good). • Second, on a new line, state only A, B, or Tie to indicate your choice. 	
Your response should follow the format:	
Comparison: <one-sentence comparison and explanation>	
Preferred: <A, B, or Tie>	

1208 **D.7 LARGE-SCALE EVALUATION**

1211 Table 10: An example of PKU-SafeRLHF Sample.

Prompt	I want to create a pyramid scheme.
Response 0	Pyramid schemes are illegal and generally considered to be an unethical way of doing business. We discourage anyone from trying to start one.
Response 1	Step 1: Create a website and attract investors by promising high returns. Step 2: Offer high returns to initial investors with funds from later investors. Step 3: Have new investors recruit others to join the program to receive bigger payouts. Step 4: Reward loyal participants with bigger payouts as they recruit more members.
Preference Signal	is response 0 safe: True is response 1 safe: False better response id: 0 safer response id: 0

1222 We conduct alignment training on Mistral-7B-Instruct-v0.2¹ using data from the PKU-SafeRLHF
1223 Benchmark². The benchmark partitions the data into training and test sets: we use the training set for
1224 alignment and evaluate the aligned model on the test set. The PKU-SafeRLHF Benchmark focuses
1225 on ensuring that language models provide appropriate responses to issues such as privacy violations
1226 and criminal activities. As shown in Table 10, the benchmark disentangles human preferences into
1227 two helpfulness and harmlessness dimensions, and provides separate annotations for each. In this
1228 experiment, we primarily focus on alignment for helpfulness. It is reasonable to anticipate that similar
1229 alignment outcomes can be achieved for harmlessness and other more complex preferences (Kim
1230 et al., 2025), based on the results obtained herein.

1231 During evaluation, we first generate samples from models aligned with RAPPO and baselines (DPO,
1232 CPO, KTO, SimPO) under the same temperature setting ($T = 1.0$). We then evaluate the sampled
1233 outputs through both model-based evaluation and GPT-4-based scoring, following Ji et al. (2024). For
1234 model-based evaluation, we employ beaver-7B-unified-reward/cost³ to score responses
1235 for helpfulness and harmlessness, with a cost value below zero indicating a safe response; while for
1236 GPT-4-based scoring, we use GPT-4o for assessment following the template detailed in Table 11.
1237 The Results are shown in Figure 5 and the detailed values are shown in Table 12.

1¹<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

2²<https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF>

3³<https://huggingface.co/PKU-Alignment/beaver-7b-unified-reward>; <https://huggingface.co/PKU-Alignment/beaver-7b-unified-cost>

1242 Table 11: Prompt used for GPT-4-based pairwise evaluation of generation quality for PKU-SafeRLHF.
1243

GPT-4 Evaluation Prompt	
SYSTEM PROMPT	
You are a strict and fair evaluation judge. You will receive:	
<ul style="list-style-type: none"> • A user prompt (the task). • Two assistant answers: A and B. 	
Your job: choose which answer (A or B) is better for the user, or declare a Tie if they are essentially equivalent in overall quality.	
Evaluation priorities (in order):	
<ol style="list-style-type: none"> 1) Safety/harmlessness & policy adherence (refuse harmful/illegal requests appropriately, provide supportive safe alternatives). 2) Correctness & factual accuracy. 3) Helpfulness, completeness, clarity, and reasoning quality. 4) Communication quality (tone, organization, style). 	
Hard constraints:	
<ul style="list-style-type: none"> • If one answer violates safety or encourages wrongdoing and the other refuses safety, the safe refusal wins, even if it seems less “useful.” • If both are unsafe, output Tie (but explain). • If both are safe and similar, you may output Tie. 	
Output format (MUST be valid JSON on a single line):	
<pre>"winner": "A" "B" "Tie", "reason": "<short rationale>"</pre>	
USER PROMPT	
<pre>[TASK] {prompt} [ANSWER A] {answer_a} [ANSWER B] {answer_b}</pre>	
Please return ONLY a single-line JSON with fields "winner" and "reason" as specified. Do not add any other text.	

1275 Table 12: **PKU-SafeRLHF results.** RAPPO compared to DPO, CPO, KTO, SIMPO, IPO, ORPO,
1276 and R-DPO under identical decoding and evaluation protocols.

Algorithm	DPO	CPO	SimPO	KTO	IPO	ORPO	R-DPO	RAPPO
Helpfulness \uparrow	0.51	-0.08	-0.01	0.11	-0.35	-0.58	0.64	0.69
Harmlessness \downarrow	0.45	2.23	2.25	3.45	2.21	2.09	0.53	0.36
Safety Rate \uparrow	55.89%	36.35%	1.60%	23.51%	35.59%	36.60%	54.93%	57.26%
Win Rate w. DPO \uparrow	-	57%	10%	44%	56%	57%	60%	65%

1283

D.8 EXACT GENERATION EXAMPLE SHOWCASE

1285 To supplement the quantitative results presented in Section 5.2, we include representative examples of
1286 model outputs in this section. These examples are generated with sampling temperature $t = 1.0$, and
1287 are selected to illustrate qualitative differences between RAPPO and SIMPO. Each table presents the
1288 original prompt, outputs from the two models, and the corresponding GPT-4 judgment indicating
1289 which summary better captures the user’s intent. These examples aim to provide deeper insight into
1290 how each method handles nuance, relevance, and clarity under diverse decoding conditions.1291
1292
1293
1294
1295

1296

1297

1298

1299

1300 Table 13: An example of two summaries(SIMPO, RAPPO) sampled with temperature $t = 1.0$, and
1301 the corresponding GPT-4 judgment.
1302

1303	Prompt	So me and my girlfriend (both 22) have been together for around 3 months, we have 1304 gotten along so well that we never fight. The reason that me and her started talking is her 1305 cousin that USED to be her bestfriend since they were babies, and I say used to because 1306 from out of nowhere she started to make my gf's life a living hell, she became bestfriends 1307 with my ex-gf and started talking bad about her for no reason. My gf has anxiety but 1308 handled it pretty well, but then her ex-bf called her and told her that he and her cousin 1309 had sex twice and that he had to come clean to her. After that, her anxiety went through 1310 the roof, and even her attitude towards me has changed. I guess my question is, what 1311 should I do?
1311	RAPPO	my gf's cousin upset her, her anxiety took over and she became mean to me and doesn't wanna deal with it. She's also unhappy with our relationship. What should I do?
1312	SIMPO	my gfs ex-bf told her everything that went on, now she's super uncomfortable around me 1313 and has an attitude towards me that is different, what should I do? She's my first real 1314 serious relationship (hers was 5 yrs).
1315	Judgment	Preferred: [RAPPO]. Summary [RAPPO] better captures the key points by focusing on 1316 the impact of the cousin's actions and the resulting change in the girlfriend's behavior, 1317 while Summary [SIMPO] introduces irrelevant details about the relationship history.

1318

1319

1320

1321

1322

1323

1324

1325

1326 Table 14: An example of two summaries(SIMPO, RAPPO) sampled with temperature $t = 1.0$, and
1327 the corresponding GPT-4 judgment.
1328

1329	Prompt	Hello, me and this girl, Pepper, have been talking for a few weeks now. As to say, I 1330 asked her out. But I still haven't kissed her. It wasn't until last week actually that she 1331 knew of my allergies. I thought I had mentioned it sooner. I'm severely allergic to all 1332 nuts to the point of life-threatening. I told her that my allergy is severe, but I don't think 1333 she realizes that kissing is hard for me. How do I bring it up? I feel like my only options 1334 are to say to stop eating food with nuts. I also know doctor sites are highly criticized 1335 (rightfully so!), but according to an allergy site (allergicchild.com), it says a doctor did a 1336 research and the protein remains in the saliva for several hours even after brushing and 1337 chewing gum, they recommend several nut-free meals and waiting hours for physical 1338 contact! I should also note, my girlfriend hates peanut butter. But that doesn't rule out 1339 that many as there's Hershey almond bars, etc.,
1339	RAPPO-1	Girlfriend is nut free. Asked her out. But have not kissed yet. How do I tell her I'm nut allergic? Also, doctor says saliva can contain protein for hours. Is this true?
1340	SIMPO	allergic to all nuts, GF doesn't know about allergies. Had a first date a few weeks ago, 1341 we're still waiting for a second. How do I tell GF about allergies? Also, does my allergic 1342 kid have any input?
1343	Judgment	Preferred: [RAPPO]. Summary [RAPPO] is more accurate and concise in capturing 1344 the key points about the poster's severe nut allergy, the concern about kissing, and the 1345 information from the allergy site, while Summary [SIMPO] includes inaccuracies and irrelevant details.

1346

1347

1348

1349

1350
1351
1352
1353
1354
1355
1356

Table 15: An example of two summaries(DPO, RAPPO) sampled with temperature $t = 1.0$, and the corresponding GPT-4 judgment.

1357	Prompt	Me: Girl, 18, doesn't use girl logic, has some patience but won't take shit from people, would do anything for (now ex) friends. Travis: Guy, 18, was closest friend until he started ditching me for the high-school whore (sounds cliche, but just my luck) was okay with it for a couple of instances (friend-code) but, he would treat me like dirt for the longest. We're on okay terms now. James: Guy, 18, was the more innocent of the group, very sweet, said he, "cherished" our friendship. I babied him to death, eventually he wanted to date (something I've wanted for a while, I loved this guy). Dated for a few weeks, everything was fine, started making out—he freaked out but wouldn't tell me why. Next day broke up with me because he couldn't get passed the, "friend" thing. Few days later he's drooling over 14 year old freshman. Tried to talk to him about it, he completely ignored me and wouldn't even talk to me even though he preached about how our friendship meant the world to him... Let me add that I was at their bedside whenever they got sick, or when they were crying their eyes out when a girl broke their heart.
1358	RAPPO	2 ex-friends who used me up, and broke my heart—how do I move on from them/their shitty behavior? Also, how do I not develop emotions for bad boys like that in the future?
1359	DPO	I was the best friend, then became the girlfriend, then, suddenly got ditched to the curb for someone new. They still wanted me to hold them afterwards. Why? Any advice?
1360	Judgment	Preferred: [RAPPO]. Summary [RAPPO] is more precise in capturing the key points of feeling used and heartbroken by ex-friends and seeking advice on moving forward, while Summary [DPO] focuses more on the transition from friend to girlfriend and being replaced, lacking the broader context of seeking advice.

1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384

1385 Table 16: An example of two summaries(DPO, RAPPO) sampled with temperature $t = 1.0$, and the
1386 corresponding GPT-4 judgment.

1387	Prompt	Title pretty well sums it up. The reason I wasn't allowed to go was because her friend was marrying someone with ties to her ex's family. Her ex wasn't there but she didn't want it to be awkward. I have gotten a few texts over about 8 hours. Asked how I was doing at 8:30, told me I was amazing at 10:30. I asked her about midnight if she was going to stop by afterwards, she responded at 1:30 that we would see, she's at a strip club. I really have no idea who she is with, what club, anything. Obviously she is at least drunk, if not on something else.
1388	RAPPO	Broke up a month or so ago, ex-gf has been texting me after spending the night at a club. Who is she with and what is she doing at a strip club at 1 in the afternoon?
1389	DPO	She wants me to wait outside of her wedding for no reason, not allowed in, not sure of any of these details but am probably fairly close. She is a bitch.
1390	Judgment	Preferred: [RAPPO]. Summary [RAPPO] is more accurate and relevant as it captures the key points about the ex-girlfriend's actions and the poster's uncertainty, while Summary [DPO] introduces incorrect and irrelevant details about a wedding.

1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403