

Article

Noise-Aware Direct Preference Optimization for RLAI

Alymzhan Toleu ^{1,2,*}, Gulmira Tolegen ^{1,2,*} , Alexandr Pak ^{1,*} and Assel Jaxylykova ¹ 

¹ School of Information Technology and Engineering, Kazakh-British Technical University, 050000 Almaty, Kazakhstan; alymzhan.toleu@gmail.com or a.toleu@satbayev.university (A.T.)

² AI Research Laboratory, Satbayev University, 050040 Almaty, Kazakhstan

* Correspondence: g.tolegen@satbayev.university or gulmira.tolegen.cs@gmail.com (G.T.); a.pak@kbtu.kz (A.P.)

Abstract

Reinforcement Learning from Human Feedback (RLHF) produces powerful instruction-following models but relies on a preference-labeling process that is both costly and slow. An effective alternative, Reinforcement Learning from AI Feedback (RLAIF), uses large language models as teachers for relabeling; however, this introduces substantial label noise. In our setting, we found that AI teachers flipped approximately 50% of the original human preferences on the dataset, a condition that degrades the performance of standard direct preference optimization (DPO). We propose noise-robust DPO (nrDPO) and nrDPO-gated, two drop-in variants that make DPO resilient to noisy preferences. nrDPO reweights each pair by (i) a margin-confidence term from a frozen reference policy (base or SFT), (ii) a context-stability term that penalizes preferences that change under truncated histories, and (iii) a length correction to curb verbosity bias. nrDPO-gated further filters low-confidence pairs via a simple threshold on the reference margin. On a dataset with heavy synthetic noise (30% flips), nrDPO-gated improves the preference accuracy by +3.8% over vanilla DPO; in a realistic RLAIF setting, nrDPO-gated is the only configuration that recovers competitive alignment, reaching $\approx 60\%$ on a 5k relabeled set (vs. $\approx 49\text{--}50\%$ for vanilla DPO) and approaching RLHF baselines.

Keywords: RLHF; RLAIF; DPO; LLM; noise robustness; preference optimization



Academic Editor: Luis Javier Garcia Villalba

Received: 4 September 2025

Revised: 20 September 2025

Accepted: 22 September 2025

Published: 23 September 2025

Citation: Toleu, A.; Tolegen, G.; Pak, A.; Jaxylykova, A. Noise-Aware Direct Preference Optimization for RLAI. *Appl. Sci.* **2025**, *15*, 10328. <https://doi.org/10.3390/app151910328>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Reinforcement Learning from Human Feedback (RLHF) has emerged as a pivotal technique for aligning large language models (LLMs) [1–3] with human values and preferences, enabling safer and more useful AI systems [4,5]. In RLHF, a reward model is trained on human-annotated preference data to approximate human judgments, which is then used to optimize the policy model via reinforcement learning algorithms, like proximal policy optimization (PPO). Direct preference optimization (DPO) was introduced as a simpler, reward-model-free alternative [6]. DPO directly optimizes the policy against preference pairs by reparameterizing the RLHF objective into a classification-like loss, achieving comparable alignment without explicit reward modeling or sampling.

Human preference labeling is the bottleneck in alignment for building a robust LLM: it is slow, costly, and hard to scale, especially for multi-turn prompts, low-resourced languages, and evolving safety policies. To reduce the human preference labeling cost, recent work explores Reinforcement Learning from AI Feedback (RLAIF), where one or more teacher LLMs act as preference judges. AI-generated preferences can be noisy, inconsistent, or biased, leading to label flips, ambiguous margins, or spurious correlations (e.g., length biases or context instability). These issues degrade downstream alignment,

particularly in small-scale or domain-specific datasets, like HH-RLHF [7], where noise amplifies overfitting and reduces the preference accuracy. In our setting, this divergence is not hypothetical; empirically, when we relabel human-annotated prompt–response pairs with teacher Llama-3.1 judges (8B and 70B LLMs), over half of the human preferences are flipped.

As shown in Figure 1, on a 2k subset, the 8B LLM teacher flips 1034/2000 pairs ($\approx 51.7\%$) and the 70B LLM teacher flips 1049/2000 ($\approx 52.5\%$); on a 5k subset, the 8B flips 2666/5000 ($\approx 53.3\%$) and the 70B flips 2693/5000 ($\approx 53.9\%$). These rates indicate that RLAIIF labels can be both noisy and biased relative to the human test distribution.

Motivated by these observations, this work investigates the following key research questions: (i) whether per-pair confidence weighting can improve robustness when labels are synthetically flipped at various levels; (ii) whether a simple sign-consistency gate helps recover accuracy under RLAIIF and narrows the gap to RLHF at matched sizes; (iii) how the choice of reference policy affects stability and accuracy across clean, noisy, and RLAIIF settings; (iv) on small clean subsets, whether robustness mechanisms hurt, match, or help standard DPO; and (v) what ranges of the hyper-parameters provide stable performance without extensive tuning.

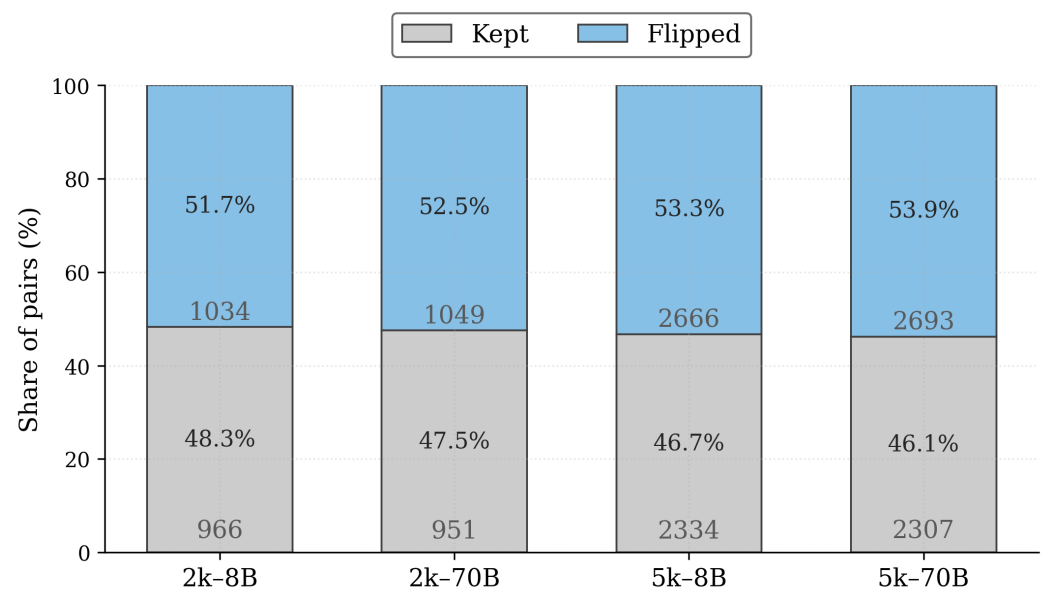


Figure 1. RLAIIF relabeling—kept vs. flipped (stacked % bars). Flipped is highlighted in light blue; kept is gray.

To address these questions, in this paper, we propose noise-robust DPO (nrDPO), an extension of DPO that incorporates per-sample weighting to prioritize reliable pairs during training. nrDPO computes weights based on reference margins (for confidence), context stability (via prompt truncation), and optional length regularization, dynamically down-weighting noisy examples. We further enhance it with nrDPO-gated, which adds optional hard gating to filter sign-inconsistent pairs (where the reference model prefers the rejected response), providing targeted robustness against flips common in RLAIIF or noisy data.

We structure the following experiments directly around the following:

1. Hyper-parameter analysis: we tune nrDPO's intrinsic-confidence knobs—margin slope γ , context-stability window K , and length weight λ —with small grids to find stable regimes rather than cherry-picked peaks.
2. Small clean data: we train SFT, vanilla DPO, nrDPO, and nrDPO-gated on harmless-base subsets (2k/5k/10k) to check whether nr-regularization harms clean-data performance, performing an evaluation on the human-labeled test split.
3. Noisy training dataset: we inject synthetic pairwise flips at 10/20/30% in the 5k train set and measure how preferred accuracy degrades on the clean test split, isolating label-noise robustness.
4. Real RLAIIF scenario: we relabel harmless-base 2k/5k with Llama-3.1 (8B, 70B) teachers via a likelihood comparison and train the same variants, reflecting practical AI-feedback scenarios; we perform an evaluation on the human test set to carry out a direct comparison against human-labeled targets.
5. Reference policy analysis: across clean/noisy/RLAIIF settings, we compare $\text{ref} = \text{base}$ versus a light SFT reference to quantify how a stronger anchor affects DPO and nrDPO (with and without gating).
6. RLAIIF vs. RLHF: we report head-to-head results on matched sizes (2k/5k), highlighting when nrDPO-gated closes the gap to RLHF.

The results demonstrate nrDPO's efficacy: on clean data, it marginally outperforms vanilla DPO (e.g., 65.53% vs. 64.79% at 5k), while nrDPO-gated provides a significant improvement over baselines in noisy settings, achieving 2–4% gains at high noise (e.g., 62.28% at 30% flips on a harmless-base dataset and 66.27% on a helpful-base dataset). For RLAIIF, gated nrDPO yields 7–9% improvements over baselines (e.g., 52.81% on 2k–8B relabels), confirming its noise resilience.

The contributions of this work can be summarized as follows: (i) nrDPO—a weighted variant of DPO that improves robustness to noisy preferences without complicating the DPO loss function. (ii) nrDPO-gated—a sign-consistency filter on top of nrDPO that down-weights (or zeroes) pairs whose reference margin disagrees with the given label beyond a tunable threshold. This one-line gate is especially effective when labels are systematically corrupted, as in RLAIIF scenarios where teacher relabeling flips a large fraction of pairs. (iii) A systematic empirical study across clean small data (2k/5k/10k), synthetic noise (10/20/30% flips), and real RLAIIF (2k/5k relabeled by Llama-3.1-8B/70B). On clean data, nrDPO matches or slightly improves over vanilla DPO. Under noise, nrDPO degrades more slowly. Under severe noise and in RLAIIF, nrDPO-gated consistently outperforms ungated baselines and narrows the gap to RLHF, with improvements supported by 95% confidence intervals.

The remainder of this article is structured as follows. Section 2 reviews related work on RLHF, DPO, and robust variants. Section 3 describes the nrDPO methodology, including weighting and gating. Section 4 presents the experimental results. Section 5 discusses the results. Section 6 outlines the limitations of the approach and future directions. Finally, Section 7 concludes this article.

2. Related Work

Reinforcement Learning from Human Feedback (RLHF) emerges from studies on learning reward functions from preference comparisons and optimizing policies against those learned rewards [4,8]. Early demonstrations showed that collecting pairwise human preferences enables practical reward modeling and subsequent policy optimization, laying the conceptual groundwork for modern RLHF in language models [4]. Subsequent work adapted these ideas to language models (LMs): a reward model is trained from human com-

parisons and the policy is optimized under a Kullback–Leibler (KL) divergence regularizer toward a reference model (often supervised fine-tuning), typically with proximal policy optimization (PPO) [9,10]. Landmark applications include summarization with human feedback and instruction following, establishing the now standard three-stage pipeline of (i) supervised fine-tuning (SFT) on demonstrations, (ii) preference data collection and reward modeling, and (iii) RL fine-tuning under KL control [11,12].

Because RLHF can be brittle and compute-intensive, a rich line of work proposes direct preference optimization objectives that avoid explicit online RL while targeting the same KL-regularized optimum. Direct preference optimization provides a closed-form link from Bradley–Terry-style preferences to an optimal KL-constrained policy, enabling stable, classification-style training [6]. Other ranking/odds-ratio formulations similarly optimize preferences without an explicit reward model or PPO, e.g., RRHF [13], PRO [14], and ORPO [15]. Collectively, these methods simplify post-training, reduce hyper-parameter sensitivity, and often match or exceed PPO-based RLHF on standard alignment evaluations.

Reinforcement Learning from AI Feedback (RLAIF) extends the RLHF pipeline by using a strong model as the preference annotator, often alongside AI-generated demonstrations for SFT (distillation) [16,17]. Anthropic’s Constitutional AI popularized this framing: a constitution of principles guides self-critique/revision in the supervised stage and preference labeling for the RL stage, scaling oversight and reducing human labeling for safety [16]. Since then, RLAIF has been widely adopted in open-source post-training recipes, including variants that use the same (or stronger) model as both a teacher (for SFT) and critic (for preferences), and even direct-RLAIF (d-RLAIF) that queries an LLM judge for rewards online [17,18].

Recent work explores automatic preference construction and self-reward without human labels. Self-Rewarding LMs (SRLM) [19] use LLM-as-a-judge prompting to score their own outputs and iterate preference optimization (e.g., iterative DPO), achieving strong results on MT-Bench and AlpacaEval-style evaluations. DLMA [20] proposes contrastive prompt pairs and a probability-based self-reward that integrates directly with DPO, reporting improvements over RLAIF in Llama-2 7 B/13 B settings. RLCD [21] constructs preference pairs by generating outputs under positive vs. negative prompts for a target principle (e.g., harmlessness), yielding clean, synthetic comparisons without humans and outperforming RLAIF baselines in several alignment tasks.

Using an LLM as a judge scales preference collection but introduces systematic evaluation noise, e.g., position, verbosity, and self-enhancement biases documented by MT-Bench/Chatbot Arena [22] and by broader audits of LLM-as-a-judge bias [23]. Length confounding is particularly salient, and Length-Controlled AlpacaEval shows that simple regression-style adjustment can substantially reduce this bias [24]. Beyond measurement, a critical study of AI feedback argues that some reported RLAIF gains arise from a teacher-critic mismatch (e.g., GPT-3.5 as teacher vs. GPT-4 as critic), implying that raw AI-judged preferences can be misleading unless denoised [25].

On the objective/optimization side, several works make preference learning robust to noisy labels. Provably Robust DPO (rDPO) explicitly models random preference flips and derives a de-biased DPO loss that maintains performance as noise increases [26]. ROPO complements this with iterative quality-aware weighting that down-weights uncertain pairs, improving tolerance to corrupted comparisons [27]. These results suggest that swapping standard DPO/PPO for noise-aware objectives can directly harden the post-training loop against AI-judge noise.

A parallel thread makes the RLHF/RLAIF RL stage itself robust to noisy reward models. R^3M treats corrupted preferences as sparse outliers and uses l_1 -regularized reward learning, with theory and gains in both RL control and LLM generation [28]. Other

formulations design reward-robust RLHF that trades off performance and robustness to avoid over-optimizing noisy reward signals (e.g., reward-robust RLHF) [29]. Additional policy-side strategies filter samples whose rewards are unreliable before PPO updates, improving signal-to-noise without changing the objective [30].

Outside preference learning, there is a broader trend toward robustness-oriented optimization under noise and uncertainty. In machine learning, distributionally robust optimization prepares models for worst-case distribution shifts, and other approaches inject uncertainty during training to improve real-world reliability. In parallel, hybrid nature-inspired metaheuristics address complex, dynamic objectives in systems like cloud load balancing, e.g., Kookaburra–Pelican hybrids [31]. Together, these lines of work illustrate a cross-domain move toward task-aware, robustness-focused optimization.

3. Model

We begin by formalizing the learning setting and notation used throughout the method and experiments. In Sections 3.2–3.4, we introduce the vanilla DPO and its noise-robust variants.

3.1. Problem Setup and Notation

Each training instance is a preference triple (x, y^+, y^-) consisting of a prompt x , a preferred response y^+ , and a dispreferred response y^- . We fine-tune a parametric policy $\pi_\theta(\cdot | x)$ starting from a pretrained LLM, together with a fixed reference policy $\pi_{\text{ref}}(\cdot | x)$ (either the same base model or a small SFT model trained on the same distribution).

For any response y , we compute the average per-token log-likelihood over answer tokens.

$$\ell_\pi(x, y) = \frac{1}{|A|} \sum_{t \in A} \log \pi_\theta(y_t | x, y_{<t}), \quad (1)$$

where A indexes answer tokens and prompt tokens are masked out. For the reference policy, define $\ell_{\text{ref}}(x, y)$ analogously by replacing π_θ with π_{ref} in (1).

Define the policy and reference margins:

$$\Delta_\pi = \ell_{\pi_\theta}(x, y^+) - \ell_{\pi_\theta}(x, y^-), \quad \Delta_{\text{ref}} = \ell_{\pi_{\text{ref}}}(x, y^+) - \ell_{\pi_{\text{ref}}}(x, y^-). \quad (2)$$

The sign of Δ encodes which response is preferred, and the magnitude $|\Delta|$ reflects confidence.

3.2. Vanilla DPO

Direct preference optimization (DPO) [6] optimizes the per-pair logistic loss

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma(\beta[\Delta_\pi - \Delta_{\text{ref}}]) = \text{softplus}(-\beta[\Delta_\pi - \Delta_{\text{ref}}]), \quad (3)$$

with temperature $\beta > 0$ and $\sigma(\cdot)$ being the logistic function. Intuitively, DPO pushes Δ_π to exceed Δ_{ref} on pairs where the reference prefers y^+ to y^- . Vanilla DPO treats all preference pairs equally, making it sensitive to dataset noise, such as ambiguous pairs (small margins), unstable preferences (sensitive to context), sign-inconsistent labels (e.g., flipped preferences), or biases (e.g., response length discrepancies).

3.3. Noise-Robust DPO (nrDPO)

To address these limitations, nrDPO introduces per-sample weights into the DPO loss, dynamically prioritizing reliable pairs during training. We reweigh each pair by a non-negative weight w , normalized to the unit mean within the batch:

$$\mathcal{L}_{\text{nrDPO}}(\theta) = \mathbb{E} \left[\underbrace{\frac{w}{\mathbb{E}[w]}}_{\text{normalized}} \text{softplus}(-\beta[\Delta_{\pi} - \Delta_{\text{ref}}]) \right]. \quad (4)$$

The weight is the product of three interpretable factors,

$$w = w_{\text{margin}} \cdot w_{\text{stab}} \cdot w_{\text{len}}, \quad w \geq 0, \quad (5)$$

designed to discount low-confidence, unstable, or length-biased pairs:

(i) Margin-confidence weight.

Larger $|\Delta_{\text{ref}}|$ means the reference is more confident about the ordering. We therefore set

$$w_{\text{margin}} = \sigma(\gamma |\Delta_{\text{ref}}|), \quad \gamma \geq 0, \quad (6)$$

where γ controls how aggressively we emphasize high-confidence pairs.

(ii) Context-stability weight.

Let $\Delta_{\text{ref}}^{(K)}$ be the margin computed after keeping only the last K tokens (or turns) of the prompt before recomputing (1). Pairs whose ordering flips under small context changes are down-weighted:

$$w_{\text{stab}} = \exp\left(-|\Delta_{\text{ref}} - \Delta_{\text{ref}}^{(K)}|\right), \quad K \in \mathbb{N}_{\geq 0}. \quad (7)$$

(iii) Length-correction weight.

To attenuate the known length bias of auto-regressive LMs, we penalize large answer-length gaps:

$$w_{\text{len}} = \exp\left(-\lambda ||y^+| - |y^-||\right), \quad \lambda \geq 0. \quad (8)$$

Setting $\gamma=0$, $K=0$ (with $\Delta_{\text{ref}}^{(K)}=\Delta_{\text{ref}}$), and $\lambda=0$ collapses nrDPO back to vanilla DPO in (3).

3.4. Gated nrDPO

When label noise is directional, for example, under RLAIIF relabeling with a teacher that sometimes disagrees systematically with human preferences, even a small fraction of sign-flipped pairs can dominate the update. We therefore introduce a simple sign gate using the same reference model:

$$w_{\text{gate}} = 1\{\Delta_{\text{ref}} > \tau\}, \quad \tau \geq 0, \quad w^* = \frac{1\{\Delta_{\text{ref}} > \tau\} w}{\mathbb{E}[1\{\Delta_{\text{ref}} > \tau\} w]}. \quad (9)$$

We use a hard gate with $\tau = 0$ by default and refer to the resulting objective as *nrDPO-gated*. The sign of Δ_{ref} supplies a directional prior from a trusted SFT model; gating prevents the optimizer from amplifying pairs that conflict with this prior.

For completeness, Algorithm 1 summarizes one nrDPO training step (with an optional hard sign gate). Batch of pairs $\{(x_i, y_i^+, y_i^-)\}_{i=1}^B$; policy π_{θ} ; reference π_{ref} ; hyper-parameters $\beta, \gamma, K, \lambda$; optional gate threshold $\tau \geq 0$.

Algorithm 1: One nrDPO training step; the optional hard gate keeps pairs with $\Delta_{\text{ref}} > \tau$

```

for  $i = 1, \dots, B$ :
     $\Delta_{\pi}^{(i)} \leftarrow \ell_{\pi}(x_i, y_i^+) - \ell_{\pi}(x_i, y_i^-)$ 
     $\Delta_{\text{ref}}^{(i)} \leftarrow \ell_{\text{ref}}(x_i, y_i^+) - \ell_{\text{ref}}(x_i, y_i^-)$ 
     $\Delta_{\text{ref},K}^{(i)} \leftarrow \ell_{\text{ref}}(\text{keep\_last}_K(x_i), y_i^+) - \ell_{\text{ref}}(\text{keep\_last}_K(x_i), y_i^-)$ 
     $w_{\text{margin}}^{(i)} \leftarrow \sigma(\gamma |\Delta_{\text{ref}}^{(i)}|)$ 
     $w_{\text{stab}}^{(i)} \leftarrow \exp(-|\Delta_{\text{ref}}^{(i)} - \Delta_{\text{ref},K}^{(i)}|)$ 
     $w_{\text{len}}^{(i)} \leftarrow \exp(-\lambda ||y_i^+| - |y_i^-||)$ 
     $w^{(i)} \leftarrow w_{\text{margin}}^{(i)} w_{\text{stab}}^{(i)} w_{\text{len}}^{(i)}$ 
    if gate:  $w^{(i)} \leftarrow 1\{\Delta_{\text{ref}}^{(i)} > \tau\} w^{(i)}$ 
normalize  $\bar{w}^{(i)} \leftarrow \frac{w^{(i)}}{\frac{1}{B} \sum_{j=1}^B w^{(j)}}$ 

 $\mathcal{L} \leftarrow \frac{1}{B} \sum_{i=1}^B \bar{w}^{(i)} \cdot \text{softplus}(-\beta[\Delta_{\pi}^{(i)} - \Delta_{\text{ref}}^{(i)}])$ 
 $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$ 

```

4. Experiment

The following experiments address key questions:

1. The impact of nrDPO's hyper-parameters (e.g., margin scaling γ , context retention K) on robustness;
2. Whether nrDPO and nrDPO-gated maintain or enhance preference accuracy on small clean human-labeled subsets (2k/5k/10k pairs);
3. Their patterns under synthetic noise injection (10/20/30% label flips);
4. Their performance in practical RLAIIF scenarios with 8B and 70B teacher models for relabeling;
5. The role of reference policy choice (base vs. SFT) in influencing training stability and overall alignment quality.

4.1. Dataset

All data come from the Anthropic HH-RLHF corpus [7]. Unless otherwise stated, we perform our evaluations on the standard test data containing 2312 prompt-pair examples (for harmless-base test data) and 2354 prompt-pair examples (for helpful-base test data). Each example has a chosen and a rejected response. For each example we recover the common prompt prefix and form a triple (x, y^+, y^-) ; prompt tokens are masked during likelihood computation. Subsets are drawn uniformly at random with a fixed seed (seed = 42). Noise injection is performed by uniformly flipping labels (chosen $\leftrightarrow$ rejected) for a given proportion $p \in \{10\%, 20\%, 30\%\}$ of the training pairs.

Table 1 summarizes the splits we use. We denote noise-injected sets with the suffix np (e.g., H-5k-n20 refers to use harmless-base, 5k, 20% random label flips).

To apply the approach in a real RLAIIF setting, we build a relabeled split by obtaining AI feedback from a stronger teacher model and using the teacher's pairwise preference as supervision. Concretely, starting from the harmless-base 2k and 5k training subsets, we query Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct to decide which response it prefers, and we flip the pair when the teacher's preference opposes the original label. We then train with these AI-derived preferences (shown in Table 2) and evaluate on the standard harmless-base test split (2312 pairs), exactly mirroring an RLAIIF training pipeline.

Table 1. Training subsets used in our experiments. All evaluations use the corresponding HH-RLHF test split with 2312 prompt-pair examples (H stands for harmless-base) and 2354 prompt-pair examples (He stands for helpful-base).

Corpus	ID	#Pairs	Clean/Noisy	Construction
harmless-base	H-2k	2000	clean	random sample from train
	H-5k	5000	clean	random sample from train
	H-10k	10,000	clean	random sample from train
	H-5k-n10	5000	noisy (10%)	H-5k with 10% random label flips
	H-5k-n20	5000	noisy (20%)	H-5k with 20% random label flips
	H-5k-n30	5000	noisy (30%)	H-5k with 30% random label flips
helpful-base	He-5k	5000	clean	random sample from train
	He-5k-n10	5000	noisy (10%)	He-5k with 10% random label flips
	He-5k-n20	5000	noisy (20%)	He-5k with 20% random label flips
	He-5k-n30	5000	noisy (30%)	He-5k with 30% random label flips

Table 2. RLAIF split used to apply the proposed method in a realistic AI-feedback setting. Training uses teacher preferences; evaluation uses the harmless-base test set (2312 pairs). We use # to denote the count of pairs.

ID	Base Corpus	#Pairs	Teacher (AI Feedback)
RLAIF-2k-8B	harmless-base	2000	Llama-3.1-8B-Instruct
RLAIF-2k-70B	harmless-base	2000	Llama-3.1-70B-Instruct
RLAIF-5k-8B	harmless-base	5000	Llama-3.1-8B-Instruct
RLAIF-5k-70B	harmless-base	5000	Llama-3.1-70B-Instruct

Table 3 summarizes teacher-driven relabeling on harmless-base. For both training sizes (2k and 5k) and both teachers (8B, 70B), the teacher flips slightly more than half of the pairs, indicating a mild anti-correlation with the original HH choices. At 2k, flip rates are approximately 52% (8B: 51.7%; 70B: 52.5%); at 5k, they rise to approximately 53–54% (8B: 53.3%; 70B: 53.9%). Ties are zero in all settings. The small gap between 8B and 70B suggests the phenomenon is teacher-agnostic. Practically, these rates mean RLAIF introduces systematic label disagreement relative to human preferences, helping explain why ungated DPO variants degrade and why margin-based gating (which down-weights low-confidence or misaligned pairs) yields substantial robustness gains in our RLAIF experiments.

Table 3. RLAIF relabeling outcomes on harmless-base. “Kept” means that the teacher agrees with the HH chosen. “Flipped” means that the pair is inverted.

Train Size	Teacher	Kept	(%)	Flipped	(%)
2k	Llama-3.1-8B	966/2000	48.3	1034/2000	51.7
2k	Llama-3.1-70B	951/2000	47.6	1049/2000	52.5
5k	Llama-3.1-8B	2334/5000	46.7	2666/5000	53.3
5k	Llama-3.1-70B	2307/5000	46.1	2693/5000	53.9

4.2. Training Setup

We adopt $(\gamma, K, \lambda) = (1.0, 64, 2 \times 10^{-3})$ as a robust starting point; ablations sweep $\gamma \in \{0.8, 1.0, 1.2, 1.5\}$, $K \in \{0, 32, 64, 128\}$, and $\lambda \in \{0, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$. Unless stated otherwise, we fine-tune Llama-3.2-1B-Instruct as a base model with AdamW, $\beta = 0.1$, one epoch, a batch size of approx. 1, and a learning rate of 5×10^{-6} . The SFT reference is trained for one epoch with a learning rate of 5×10^{-5} and batch size of 1 using prompt masking. We run a compact grid of 26 configurations across γ , K , λ , and gating on/off only at 20% label flips, and we use the selected configuration that remains

unchanged for 10%/30%, clean, and RLAIIF experiments, which saves time and computing resources while preventing hyper-parameter overfitting to a particular corruption level.

4.3. Evaluation Metric

We assess alignment performance using preference accuracy (PrefAcc). For a test set of preference triples (x, y_w, y_l) , PrefAcc is the fraction of pairs where the model assigns a higher length-normalized answer log-likelihood to the winning response y_w than to the losing response y_l (prompt tokens are masked):

$$\text{PrefAcc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left(\frac{1}{|y_w|} \log \pi(y_w | x) > \frac{1}{|y_l|} \log \pi(y_l | x) \right). \quad (10)$$

Evaluations use the official HH-RLHF test splits: 2312 pairs for harmless-base and 2354 for helpful-base. To capture the uncertainty of our estimates, we report 95% confidence intervals (95% CI). Intuitively, this interval gives a range of values around the observed PrefAcc within which the true accuracy would fall in 95 out of 100 repeated samples of the same size. The width of the interval reflects how reliable the estimate is: narrower intervals indicate more certainty, while wider intervals indicate greater variability. In Appendix A, we report confidence intervals at multiple levels (90%, 95%, and 99%).

4.4. Results

4.4.1. Hyper-Parameter Sensitivity

Figure 2 plots the preference accuracy (PrefAcc) on the HH test split while sweeping three nrDPO hyper-parameters, with two training conditions shown in every panel: clean (solid blue) and 20% label-noise (dashed orange).

- Margin weight γ . Accuracy is fairly flat and robust on clean data (0.646–0.650). Under 20% noise, performance improves as γ increases—from 0.612 at $\gamma = 1.0$ and up to 0.623–0.626 near $\gamma = 1.5$ —indicating that putting more weight on large reference margins helps denoise preferences.
- Context keep window K . Best results occur around $K = 32$ for both settings. Larger truncation ($K = 64, 128$) negatively affects performance, especially on clean data (drops 1–1.5 pts from the peak). Keeping a small recent history seems sufficient; retaining too much context dilutes the stability signal.
- Length regularization λ . A small positive λ improves or matches the baseline; on clean data, the curve peaks ($\lambda = 1 \times 10^{-3}$) and then softens. With 20% noise, λ reduces variance, and the strongest value among those tried is $\lambda = 2 \times 10^{-3}$, which brings accuracy back up after a dip at 1×10^{-3} .

4.4.2. Results with Clean Data

Table 4 presents results on clean (human-labeled) subsets of varying sizes: 2k, 5k and 10k. SFT establishes a baseline around 55–57%, improving modestly as the data scale increases. Vanilla DPO consistently outperforms SFT by 3–12% (e.g., 66.87% at 10k vs. SFT’s 54.80%), demonstrating DPO’s effectiveness in preference alignment. nrDPO provides marginal gains over vanilla DPO (e.g., +0.74% at 5k, +0.13% at 2k), with non-overlapping CIs at 5k suggesting statistical significance; this highlights the value of its weighting scheme in emphasizing high-confidence and stable pairs even on clean data.

However, nrDPO-gated underperforms in comparison to non-gated variants, dropping to 51.73% at 2k and 62.37% at 10k. This is expected; in clean datasets with fewer inconsistencies, gating (which down-weights sign-inconsistent margins) discards useful pairs, leading to reduced performance. Gains scale with data size for all DPO methods, underscoring sample efficiency in harmless alignment.

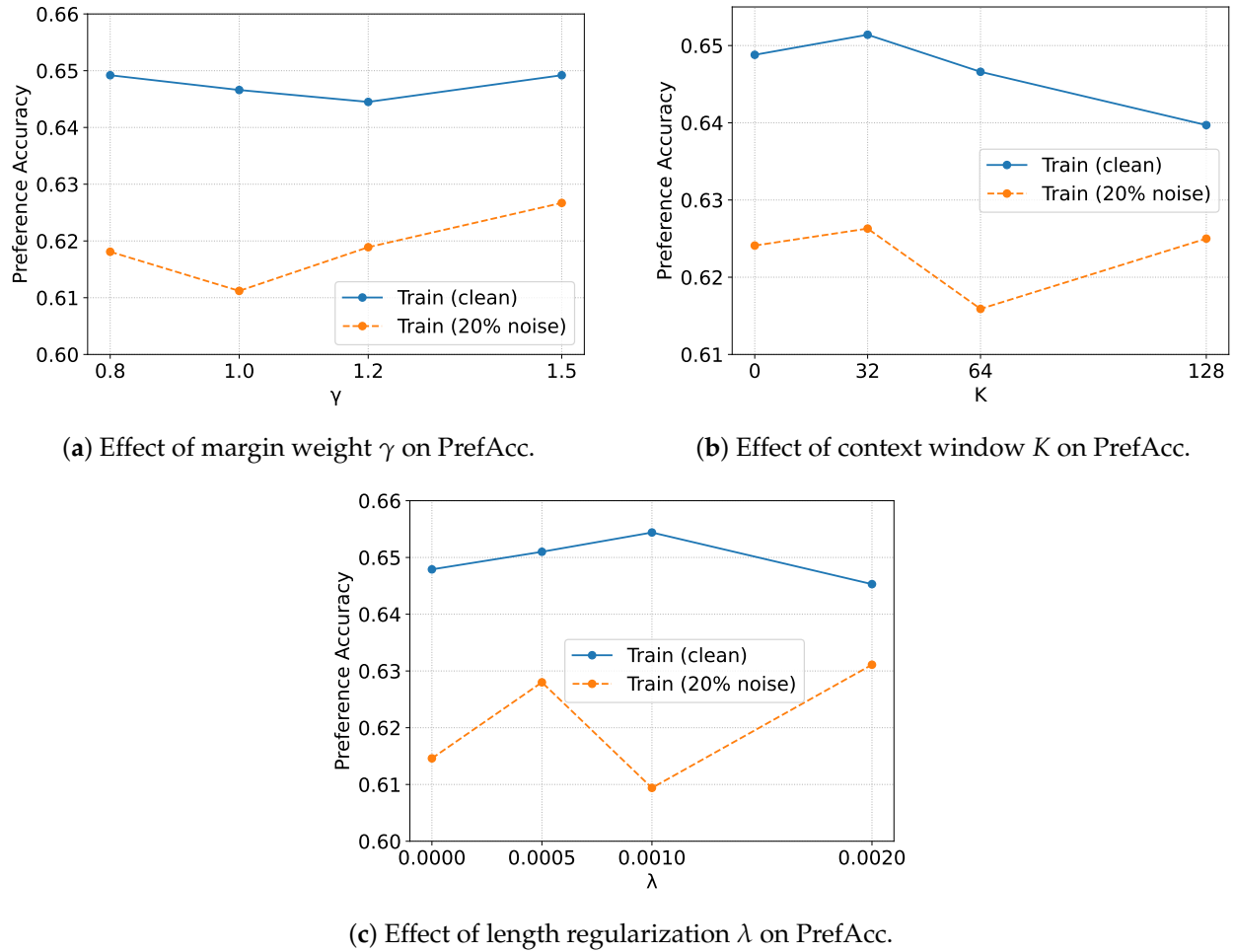


Figure 2. Hyper-parameter sensitivity of nrDPO on the H-5k dataset. The figure shows PrefAcc as a function of (a) margin weight γ , (b) context keep window K , and (c) length regularization λ .

Table 4. Results for clean training subsets of harmless-base. Test split: 2312 pairs.

Models	H-2k		H-5k		H-10k	
	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI
SFT	56.96	(55.02, 59.00)	56.06	(54.07, 58.04)	54.80	(52.85, 56.83)
Vanilla DPO	59.47	(57.48, 61.46)	64.79	(62.85, 66.74)	66.87	(64.92, 68.81)
nrDPO	59.60	(57.61, 61.59)	65.53	(63.54, 67.43)	66.44	(64.49, 68.38)
nrDPO-gated	51.73	(49.74, 53.81)	60.81	(58.82, 62.80)	62.37	(60.38, 64.36)

4.4.3. Results with Noisy Data

To assess robustness, we inject noise into the 5k harmless subset (Table 5). SFT remains stable (approx. 56%) as noise increases, and it is unaffected by preferences. Vanilla DPO starts strong at 10% noise (62.98%) but degrades progressively (to 58.52% at 30%), with widening CIs reflecting increased variance from noisy labels. nrDPO mitigates this somewhat, outperforming vanilla by 1–2% at lower noise (e.g., 64.14% vs. 62.98% at 10%), but it also declines at higher levels (to 58.43% at 30%). In contrast, nrDPO-gated exhibits superior resilience, maintaining approx. 63% across noise levels and achieving the highest score at 30% noise (62.28%, CI: 60.29–64.23). Non-overlapping CIs with baselines at 20–30% noise confirm its significance. This validates the gating mechanism’s role in filtering sign-inconsistent (flipped) pairs, complementing the core weights to preserve alignment under corruption.

Table 5. Results for 5k noise-injected training subsets of harmless-base). Test split: 2312 pairs. Bold indicates the highest PrefAcc in each column.

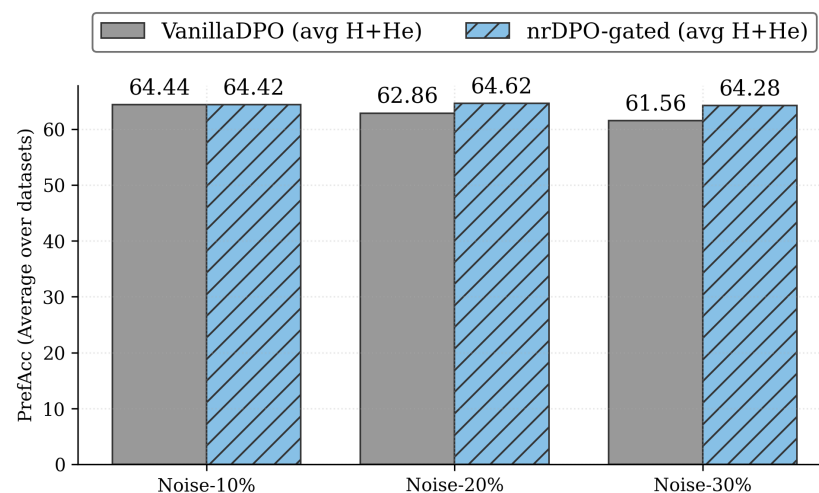
Models	H-5k-n10		H-5k-n20		H-5k-n30	
	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI
SFT	56.36	(54.33, 58.39)	56.62	(54.58, 58.65)	56.23	(54.20, 58.26)
Vanilla DPO	62.98	(61.03, 64.92)	60.99	(59.56, 62.37)	58.52	(56.49, 60.51)
nrDPO	64.14	(62.15, 66.13)	61.25	(59.26, 63.24)	58.43	(56.44, 60.42)
nrDPO-gated	63.24	(61.29, 65.27)	63.15	(61.25, 65.10)	62.28	(60.29, 64.23)

Table 6 extends the noise analysis to the helpfulness split. SFT hovers near (49–50%), as expected for unaligned baselines. Vanilla DPO and nrDPO both yield strong gains at 10% noise (approx. 66%), with nrDPO slightly ahead (66.44% vs. 65.89%). As noise rises, vanilla DPO holds steady (approx. 64–65%), while nrDPO shows minor degradation at 30% (64.95%). nrDPO-gated performs competitively at low noise (65.59%) and excels at higher levels, reaching 66.27% at 30% (CI: 64.32–68.22), outperforming others with non-overlapping CIs.

Table 6. Results for 5k noise-injected training subsets of helpful-base set. Test split: 2354 pairs. Bold indicates the highest PrefAcc in each column.

Models	He-5k-n10		He-5k-n20		He-5k-n30	
	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI
SFT	50.25	(48.26, 52.29)	50.13	(48.13, 52.12)	49.58	(47.58, 51.57)
Vanilla DPO	65.89	(63.98, 67.76)	64.74	(62.83, 66.69)	64.61	(62.66, 66.53)
nrDPO	66.44	(64.53, 68.39)	65.17	(63.21, 67.08)	64.95	(63.08, 66.91)
nrDPO-gated	65.59	(63.68, 67.50)	66.10	(64.19, 68.05)	66.27	(64.32, 68.22)

Aggregating harmless+helpful 5k (Figure 3), nrDPO-gated is almost equal to vanilla DPO at 10% noise, but surpasses vanilla DPO by approx. +1.8 pp at 20% and approx. +2.7 pp at 30%. On average, across datasets, nrDPO-gated is robust to label noise, matching vanilla DPO at low noise and outperforming it substantially at medium/high noise.

**Figure 3.** Overall average across datasets (harmless-base + helpful-base, 5k): PrefAcc vs. injected noise (10%, 20%, 30%) comparing vanilla DPO (gray) and nrDPO-gated (light-blue).

4.4.4. Results with RLAIF

We evaluate RLAIF on the harmless-base corpus by relabeling a 2k and 5k subset with two teacher policies (Llama-3.1-8B and 70B) and training several DPO variants in Table 7. Across both teachers, for 2k AI-labeled dataset, nrDPO-gated is the only configuration that consistently lifts performance: with the 8B teacher, it reaches 52.81% (95% CI 50.74–54.84), compared to 43.99% (41.91–46.02) for vanilla DPO and 45.59% (43.51–47.62) for the standard nrDPO. With the 70B teacher, it remains best at 50.17% (48.18–52.21), again exceeding vanilla DPO (43.47%, 41.48–45.50) and nrDPO (42.99%, 40.96–45.03). In both cases, the confidence intervals of nrDPO-gated do not overlap with those of the baselines, indicating statistically robust improvements. Scaling RLAIF to a 5k relabeled subset preserves—and amplifies—this trend: nrDPO-gated attains 60.12% (58.13–62.11) with the 8B teacher and 57.18% (55.19–59.21) with the 70B teacher, outperforming vanilla DPO (49.39%/47.66%) and nrDPO (50.26%/47.02%) by approximately +9–11 points. In all settings, the gated model’s confidence intervals do not overlap with the baselines, indicating statistically robust gains from intrinsic-confidence gating under RLAIF.

Table 7. Results for RLAIF on harmless-base with 2k and 5k relabeled training subsets and two teachers (Llama-3.1-8B/70B). Test split: 2312 pairs. Bold indicates the highest PrefAcc in each column.

Model	RLAIF-2k-8B		RLAIF-2k-70B		RLAIF-5k-8B		RLAIF-5k-70B	
	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI
Vanilla DPO	43.99	(41.91, 46.02)	43.47	(41.48, 45.50)	49.39	(47.36, 51.43)	47.66	(45.63, 49.70)
Vanilla DPO (ref = SFT)	43.60	(41.57, 45.59)	43.17	(41.18, 45.20)	50.09	(48.05, 52.12)	46.76	(44.72, 48.83)
nrDPO	45.59	(43.51, 47.62)	42.99	(40.96, 45.03)	50.26	(48.27, 52.29)	47.02	(45.03, 49.05)
nrDPO (ref = SFT)	44.16	(42.17, 46.15)	42.95	(40.92, 44.98)	50.09	(48.10, 52.08)	48.14	(46.15, 50.17)
nrDPO-gated	52.81	(50.74, 54.84)	50.17	(48.18, 52.21)	60.12	(58.13, 62.11)	57.18	(55.19, 59.21)

Figure 4 clearly shows on a harmless-base dataset with 2k and 5k RLAIF labels that nrDPO-gated outperforms vanilla DPO under both teachers.

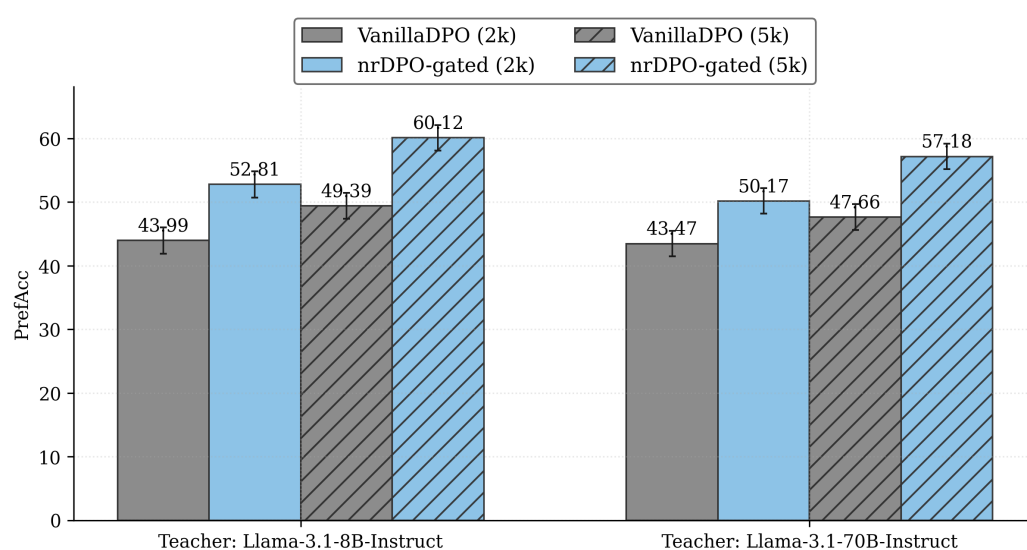


Figure 4. RLAIF on a harmless-base dataset (2k and 5k relabeled): PrefAcc for vanilla DPO vs. nrDPO-gated under two teachers (Llama-3.1-8B and Llama-3.1-70B).

4.4.5. RLAIIF vs. RLHF

Table 8 reports the models’ results for human-labeled RLHF baselines (“H-2k/5k”) with RLAIIF. With human labels (RLHF), ungated DPO variants reach 59–66% PrefAcc. If we replace human labels with teacher preferences (RLAIIF), the same ungated models drop to approx. 44–46% at 2k and approx. 48–50% at 5k—about 10–16 points lower. Using nrDPO-gated largely closes this gap: at 2k, it reaches 52.8% (8B)/50.2% (70B) vs. 59.6% RLHF; at 5k, it reaches 60.1% (8B)/57.2% (70B) vs. 65.5% RLHF. The confidence intervals for RLAIIF-gated often overlap the RLHF-gated baselines (e.g., 60.1% vs. 60.8% at 5k), so the differences are not always statistically clear. Across sizes, the 8B teacher provides stronger RLAIIF results than the 70B. RLAIIF lags RLHF without gating; with gating, it becomes competitive.

Table 8. RLAIIF on a harmless-base dataset (2k and 5k relabeled) vs. RLHF on the same corpus (H-2k and H-5k). We report preference accuracy (PrefAcc, %) with 95% confidence intervals. Test split: 2312 pairs. Bold indicates the highest PrefAcc in each column.

Dataset	Vanilla DPO	nrDPO	nrDPO-Gated
RLAIIF-2k-8B	43.99 (41.91, 46.02)	45.59 (43.51, 47.62)	52.81 (50.74, 54.84)
RLAIIF-2k-70B	43.47 (41.48, 45.50)	42.99 (40.96, 45.03)	50.17 (48.18, 52.21)
H-2k (RLHF)	59.47 (57.48, 61.46)	59.60 (57.61, 61.59)	51.73 (49.74, 53.81)
RLAIIF-5k-8B	49.39 (47.36, 51.43)	50.26 (48.27, 52.29)	60.12 (58.13, 62.11)
RLAIIF-5k-70B	47.66 (45.63, 49.70)	47.02 (45.03, 49.05)	57.18 (55.19, 59.21)
H-5k (RLHF)	64.79 (62.85, 66.74)	65.53 (63.54, 67.43)	60.81 (58.82, 62.80)

4.4.6. Influence of SFT

Table 9 shows that swapping the base reference for a light SFT reference yields consistent, modest gains at 10–20% noise; however, little-to-no gain at 30%—unless gating is used. For vanilla DPO, SFT-ref improves from 62.98 to 64.23 at n10 (+1.25) and from 60.99 to 63.06 at n20 (+2.07), while the n30 change is negligible (58.52→58.69, +0.17). nrDPO exhibits similar small lifts (e.g., n20: 61.25→62.20, +0.95). Confidence intervals largely overlap at n10–n20, indicating stabilization rather than a regime change, and they converge at n30. Crucially, with gating, the SFT-ref becomes decisive at high noise: nrDPO-gated reaches 62.28 at n30, outperforming the best ungated settings by approximately +3.6–3.9 points. In short, SFT-ref regularizes DPO under moderate noise, and SFT-ref + gating is needed to maintain accuracy under severe noise.

Table 9. Comparison of ref = SFT vs. ref = base with noise injection dataset; nrDPO-gated uses ref = SFT by default. Test split: 2312 pairs.

Models	H-5k-n10		H-5k-n20		H-5k-n30	
	PrefAcc	95% CI	PrefAcc	95% CI	PrefAcc	95% CI
SFT	56.36	(54.33, 58.39)	56.62	(54.58, 58.65)	56.23	(54.20, 58.26)
Vanilla DPO (Base)	62.98	(61.03, 64.92)	60.99	(59.56, 62.37)	58.52	(56.49, 60.51)
Vanilla DPO (ref = SFT)	64.23	(62.28, 66.13)	63.06	(61.12, 65.01)	58.69	(56.66, 60.73)
nrDPO (Base)	64.14	(62.15, 66.13)	61.25	(59.26, 63.24)	58.43	(56.44, 60.42)
nrDPO (ref = SFT)	64.66	(62.72, 66.61)	62.20	(60.25, 64.14)	58.48	(56.49, 60.51)
nrDPO-gated	63.24	(61.29, 65.27)	63.15	(61.25, 65.10)	62.28	(60.29, 64.23)

5. Discussion

We observe a consistent trend across all tested conditions:

- (i) On clean small data (2k/5k/10k), nrDPO matches or slightly improves over vanilla DPO, while nrDPO-gated can trail on the largest clean split because gating removes some useful (but low-margin) signals;
- (ii) Under synthetic noise (10/20/30% label flips), nrDPO degrades more slowly than vanilla, and nrDPO-gated is the only variant that remains competitive at the highest noise rate;
- (iii) In RLAIIF, where teacher relabeling flips roughly half of the pairs ($\approx 52\text{--}54\%$ at 2k and 5k), nrDPO-gated is decisively better than ungated methods and narrows the performance gap with RLHF.

These patterns follow from how the objective treats reliability signals in the data. We next explain why weighting and gating lead to the observed behavior and how the reference policy moderates these effects.

5.1. Why Weighting Helps

nrDPO treats the reference model's log-likelihood margin as an efficient proxy for label reliability. Pairing the references allows us to be confident about large $|\Delta_{\text{ref}}|$ receiving more weight; pairs whose margin collapses when context is truncated are down-weighted as unstable; extreme response-length imbalances are subject to a mild correction. This rebalances gradients toward pairs where the model family itself is internally consistent, which is especially valuable when labels are noisy or partially anti-aligned with the base policy.

5.2. Why Gating Helps (and When It Can Hurt)

The gate applies a simple sign-consistency check: if a pair's label disagrees with the reference margin (beyond a threshold), we drop it.

Under severe noise or RLAIIF (where the teacher LLM disagrees with human preferred labels about half the time), this prevents systematic anti-learning from flipped examples and yields the largest gains. On clean data, however, some low-margin, borderline pairs still carry useful supervision; removing too many of them can slightly reduce the achievable performance, which is exactly what we observed on H-10k.

5.3. The Role of the Reference Policy

Using an SFT reference generally stabilizes DPO under moderate noise (10–20%) and makes nrDPO's weights more informative; confidence intervals shrink without fundamentally changing the ranking. Under 30% noise, SFT-ref alone is not enough; gating is required to keep prevent the accuracy from decreasing. Practically: with mild label noise, SFT-ref is a good default; with heavy or systematic noise (including RLAIIF), enable gating.

5.4. RLAIIF vs. RLHF

With human labels (RLHF), vanilla and nrDPO reach the high 50s to mid-60s on HH test sets. Pure RLAIIF drops into the mid-40s to ~ 50 for ungated training, which is consistent with the observed $\sim 50\%$ flips from teachers. Adding nrDPO-gated recovers much of the gap (e.g., $>50\%$ at 2k and $\sim 60\%$ at 5k with the 8B teacher), and its confidence intervals often overlap RLHF with gating. Without gating, RLAIIF lags RLHF; with gating, it becomes competitive in small-data regimes. This study focuses on introducing reference confidence and sign gating within the DPO framework to suppress bias and noise; therefore, we do not provide a systematic comparison with alternative training paradigms, such as RRHF, PRO, or ORPO, which operate without a reference model or explicit rewards. Preliminary observations indicate that these methods are not incompatible with our weighting and

gating approach from an engineering standpoint. In future work, we plan to combine them and evaluate generalization and robustness on broader datasets.

6. Limitations

While nrDPO and nrDPO-gated demonstrate promising robustness in noisy and small-scale alignment scenarios, several limitations warrant consideration. First, our evaluations are confined to English-language subsets of the HH-RLHF dataset, focusing on harmlessness and helpfulness splits. This restricts generalizability to multilingual contexts, diverse domains (e.g., code generation or creative writing), or other preference datasets, like UltraFeedback, where noise patterns may differ. Extending to non-English or broader benchmarks could reveal unaddressed biases.

Second, we primarily rely on preference accuracy (PrefAcc) as the evaluation metric, which, while standard for DPO variants, may not capture nuanced aspects of alignment, such as factual correctness, coherence, or long-term safety. Richer assessments, including human evaluations (e.g., via Likert scales or side-by-side comparisons) or automated metrics, like harmlessness scores from reward models, would provide a more holistic view but were beyond our small-scale scope of this article.

Third, the gated variant risks over-filtering in clean datasets, as evidenced by its underperformance (e.g., 3–8% drops) compared to non-gated nrDPO. This suggests that aggressive sign-consistency checks may discard informative pairs with subtle margins, particularly in high-quality human-labeled data.

Additionally, experiments used compact models (e.g., Llama-3.2-1B-Instruct) and limited training sizes (up to 10k pairs), potentially underestimating scalability issues in larger setups. RLAIF evaluations were restricted to specific teachers (8B/70B), overlooking ensemble effects or weaker/stronger alternatives. Noise injection simulated random flips but not other real-world corruptions, like distributional shifts or adversarial labels.

Future work could address these by incorporating adaptive thresholds, multilingual testing, and hybrid human–AI evaluations to enhance nrDPO’s applicability. Despite these constraints, the proposed approach offers a valuable step toward efficient, noise-tolerant alignment.

7. Conclusions

In this work, we explored preference alignment in scenarios with limited and noisy supervision, comparing traditional RLHF with RLAIF, where a teacher LLM relabels preference pairs to reduce human annotation costs. We presented nrDPO, a weighted extension of DPO that prioritizes pairs with high margins, contextual stability, and balanced lengths, along with nrDPO-gated, which additionally filters pairs where the reference model’s preference sign conflicts with the label.

The findings can be summarized as follows:

1. On clean small data (2k/5k/10k), nrDPO matches or slightly outperforms vanilla DPO, so the weighting does not hurt when labels are clean.
2. On noisy data (10–30% flips), as noise grows, vanilla DPO degrades faster. nrDPO is more stable, and nrDPO-gated is the only method that remains effective at the highest noise.
3. RLAIF (teacher relabeling). Teachers flipped about half of the pairs, which is a challenging setting. nrDPO-gated closes much of the gap with RLHF: about 53% at 2k and about 60% at 5k with an 8B teacher versus about 60% and 66% for RLHF on the same splits. In several cases, confidence intervals overlap with the RLHF baselines.
4. Reference choice. An SFT reference helps at moderate noise; under severe noise or RLAIF, adding the gate is important to retain accuracy.

5. The ablations further reveal that hyper-parameter choices, such as moderate margin scaling γ and limited context retention k , provide reliable defaults for practical use.

Overall, these results suggest that noise-robust preference optimization is a promising direction for scalable alignment, particularly in settings where reliable human feedback is scarce. The presented approach narrows the performance gap between RLAIIF and RLHF, showing that careful weighting and gating can mitigate noisy supervision. Future work could extend these techniques to larger-scale preference datasets, multi-turn dialogue alignment, test multilingual and domain-shifted distributions, and hybrid pipelines that combine human and AI feedback to further reduce annotation costs while maintaining reliability.

Practical Takeaways

We summarize actionable guidance for small-scale alignment under noisy supervision:

- Clean labels (2k/5k/10k). Choose vanilla DPO or nrDPO; avoid gating, as it may discard informative borderline pairs.
- Random label noise (10–30%). Use nrDPO at low noise (10–20%); switch to nrDPO-gated at higher noise ($\approx 30\%$).
- RLAIIF (AI-judge relabeling). With flip rates near one half, nrDPO-gated is recommended; performance gains are most pronounced at 5k.
- Reference policy. A light SFT reference stabilizes training under mild noise (10–20%); under severe/systematic noise (RLAIIF or $\geq 30\%$ flips), combine an SFT reference with gating.
- Hyper-parameters (robust defaults). Margin weight $\gamma \in [1.0, 1.5]$ (increase under noise); context window $K \approx 32$; length regularization $\lambda \in [10^{-3}, 2 \cdot 10^{-3}]$; DPO temperature $\beta \approx 0.1$.
- Training recipe (small models/data). One epoch typically suffices to reveal trends; AdamW, policy learning rate 5×10^{-6} , SFT reference learning rate 5×10^{-5} , batch size 1; mask prompt tokens when computing answer likelihoods.
- When not to gate. In clean-data regimes or on the largest clean split, gating can remove useful supervision and is not advised.

Author Contributions: Conceptualization, A.T. and G.T.; Methodology, A.T. and G.T.; Validation, A.T. and G.T.; Resources, G.T.; Data Curation, G.T., A.P. and A.J.; Writing—Original Draft Preparation, A.T. and G.T.; Writing—Review and Editing, A.T. and G.T.; Visualization, G.T.; Project Administration, A.P.; Funding Acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under grant number AP23489782.

Data Availability Statement: The data presented in this study are publicly available. The corresponding URLs to the used datasets are provided at the following link: <https://huggingface.co/datasets/HuggingFaceH4/hh-rlhf> (accessed on 2 December 2024).

Acknowledgments: We utilized GenAI—Grammarly <https://www.grammarly.com/> (access on 2 July 2025) to ensure accuracy in grammar and sentence structure.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1 shows that across all RLAIIF settings, nrDPO-gated attains the highest PrefAcc, with 90%, 95%, and 99% CIs being clearly greater than vanilla's (e.g., RLAIIF-5k-8B: 0.6012; 90% 0.5843–0.6181, 95% 0.5813–0.6211, 99% 0.5757–0.6272 vs. vanilla 0.4939; 90% 0.4771–0.5108,

95% 0.4736–0.5147, 99% 0.4676–0.5212). nrDPO generally matches or slightly exceeds vanilla (especially with 8B relabels). Scaling from 2k to 5k increases the performance of all methods, but mostly for nrDPO-gated, making gating the key driver of robustness under AI-judge relabeling.

Table A1. Preference accuracy (PrefAcc) with 90%, 95%, and 99% bootstrap CIs on the harmless-base test data for vanilla DPO, nrDPO, and nrDPO-gated across RLAIIF settings.

Setting	Variant	PrefAcc	90% CI	95% CI	99% CI
RLAIIF-2k-8B	Vanilla DPO	0.4399	(0.4278, 0.4522)	(0.4256, 0.4542)	(0.4213, 0.4589)
RLAIIF-2k-8B	nrDPO	0.4559	(0.4390, 0.4728)	(0.4360, 0.4766)	(0.4304, 0.4827)
RLAIIF-2k-8B	nrDPO-gated	0.5281	(0.5112, 0.5454)	(0.5078, 0.5489)	(0.5017, 0.5554)
RLAIIF-2k-70B	Vanilla DPO	0.4347	(0.4178, 0.4520)	(0.4144, 0.4550)	(0.4083, 0.4615)
RLAIIF-2k-70B	nrDPO	0.4299	(0.4131, 0.4472)	(0.4100, 0.4503)	(0.4031, 0.4567)
RLAIIF-2k-70B	nrDPO-gated	0.5017	(0.4849, 0.5190)	(0.4818, 0.5221)	(0.4758, 0.5281)
RLAIIF-5k-8B	Vanilla DPO	0.4939	(0.4771, 0.5108)	(0.4736, 0.5147)	(0.4676, 0.5212)
RLAIIF-5k-8B	nrDPO	0.5026	(0.4853, 0.5199)	(0.4823, 0.5234)	(0.4758, 0.5298)
RLAIIF-5k-8B	nrDPO-gated	0.6012	(0.5843, 0.6181)	(0.5813, 0.6211)	(0.5757, 0.6272)
RLAIIF-5k-70B	Vanilla DPO	0.4766	(0.4645, 0.4888)	(0.4622, 0.4911)	(0.4585, 0.4959)
RLAIIF-5k-70B	nrDPO	0.4702	(0.4578, 0.4823)	(0.4559, 0.4846)	(0.4518, 0.4892)
RLAIIF-5k-70B	nrDPO-gated	0.5718	(0.5599, 0.5837)	(0.5575, 0.5863)	(0.5530, 0.5906)

References

1. Tolegen, G.; Toleu, A.; Mussabayev, R. Enhancing Low-Resource NER via Knowledge Transfer from LLM. In *Computational Collective Intelligence: 16th International Conference, ICCCI 2024, Leipzig, Germany, 9–11 September 2024*; Nguyen, N.T., Franczyk, B., Ludwig, A., Núñez, M., Treur, J., Vossen, G., Kozierkiewicz, A., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 14810. [\[CrossRef\]](#)
2. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [\[CrossRef\]](#)
3. Toleu, A.; Tolegen, G.; Ualiyeva, I. Fine-Tuning Large Language Models for Kazakh Text Simplification. *Appl. Sci.* **2025**, *15*, 8344. [\[CrossRef\]](#)
4. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
5. Huang, S.; Zhao, J.; Li, Y.; Wang, L. Learning Preference Model for LLMs via Automatic Preference Data Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 9187–9199. [\[CrossRef\]](#)
6. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 10–16 December 2023.
7. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**, arXiv:2204.05862. [\[CrossRef\]](#)
8. Wirth, C.; Akrou, R.; Neumann, G.; Fürnkranz, J. A Survey of Preference-Based Reinforcement Learning Methods. *J. Mach. Learn. Res.* **2017**, *18*, 1–46.
9. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-Tuning Language Models from Human Preferences. *arXiv* **2020**, arXiv:1909.08593. [\[CrossRef\]](#)
10. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. [\[CrossRef\]](#)
11. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P.F. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 3008–3021.

12. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 27730–27744.
13. Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; Huang, S.; Huang, F. RRHF: Rank Responses to Align Language Models with Human Feedback without tears. *arXiv* **2023** arXiv:2304.05302. [[CrossRef](#)]
14. Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; Wang, H. Preference Ranking Optimization for Human Alignment. *arXiv* **2024**, arXiv:2306.17492. [[CrossRef](#)]
15. Hong, J.; Lee, N.; Thorne, J. ORPO: Monolithic Preference Optimization without Reference Model. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 11170–11189. [[CrossRef](#)]
16. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073. [[CrossRef](#)]
17. Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; et al. RLAIIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv* **2024**, arXiv:2309.00267.
18. Guo, S.; Zhang, B.; Liu, T.; Liu, T.; Khalman, M.; Llinares, F.; Ramé, A.; Mesnard, T.; Zhao, Y.; Piot, B.; et al. Direct Language Model Alignment from Online AI Feedback. *arXiv* **2024**, arXiv:2402.04792. [[CrossRef](#)]
19. Yuan, W.; Pang, R.Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; Weston, J. Self-Rewarding Language Models. *arXiv* **2025**, arXiv:2401.10020. [[CrossRef](#)]
20. Liu, A.; Bai, H.; Lu, Z.; Kong, X.; Wang, S.; Shan, J.; Cao, M.; Wen, L. Direct Large Language Model Alignment Through Self-Rewarding Contrastive Prompt Distillation. *arXiv* **2024**, arXiv:2402.11907. [[CrossRef](#)]
21. Yang, K.; Klein, D.; Celikyilmaz, A.; Peng, N.; Tian, Y. RLCD: Reinforcement Learning from Contrastive Distillation for Language Model Alignment. *arXiv* **2024**, arXiv:2307.12950. [[CrossRef](#)]
22. Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
23. Chen, G.H.; Chen, S.; Liu, Z.; Jiang, F.; Wang, B. Humans or LLMs as the Judge? A Study on Judgement Bias. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 8301–8327. [[CrossRef](#)]
24. Dubois, Y.; Galambosi, B.; Liang, P.; Hashimoto, T.B. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv* **2025**, arXiv:2404.04475.
25. Sharma, A.; Keh, S.; Mitchell, E.; Finn, C.; Arora, K.; Kollar, T. A Critical Evaluation of AI Feedback for Aligning Large Language Models. *arXiv* **2024**, arXiv:2402.12366. [[CrossRef](#)]
26. Chowdhury, S.R.; Kini, A.; Natarajan, N. Provably Robust DPO: Aligning Language Models with Noisy Feedback. *arXiv* **2024**, arXiv:2403.00409. [[CrossRef](#)]
27. Liang, X.; Chen, C.; Qiu, S.; Wang, J.; Wu, Y.; Fu, Z.; Shi, Z.; Wu, F.; Ye, J. ROPO: Robust Preference Optimization for Large Language Models. *arXiv* **2024**, arXiv:2404.04102.
28. Bukharin, A.; Hong, I.; Jiang, H.; Li, Z.; Zhang, Q.; Zhang, Z.; Zhao, T. Robust Reinforcement Learning from Corrupted Human Feedback. *arXiv* **2024**, arXiv:2406.15568. [[CrossRef](#)]
29. Yan, Y.; Lou, X.; Li, J.; Zhang, Y.; Xie, J.; Yu, C.; Wang, Y.; Yan, D.; Shen, Y. Reward-Robust RLHF in LLMs. *arXiv* **2024**, arXiv:2409.15360.
30. Zhang, C.; Shen, W.; Zhao, L.; Zhang, X.; Xu, X.; Dou, W.; Bian, J. Policy Filtration for RLHF to Mitigate Noise in Reward Models. *arXiv* **2025**, arXiv:2409.06957.
31. Addula, S.R.; Shukla, A.; Chaurasia, V.R.; Kumar, A.; Akula, A.P.; Harini, R. Dynamic Load Balancing in Cloud Computing using Hybrid Kookaburra–Pelican Optimization Algorithms. In Proceedings of the International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA), Manipal, India, 20–21 December 2024; pp. 1–7. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.