REWARD DIFFERENCES MATTER FOR ROBUST LLM ALIGNMENT UNDER NOISY PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Human preference alignment is pivotal for encouraging large language models to generate more helpful and safer responses that align with human values. Direct preference optimization (DPO) is a widely used alignment approach that enables LLMs to learn preferences directly from human preference data. However, human preference data inevitably contain noise arising from annotator inattention and bias, which can substantially degrade DPO performance. To identify noisy preferences, we conduct extensive preliminary experiments and empirically observe that as the reward difference decreases, the noise probability gradually increases during DPO training. Based on this observation, we propose a new method named *Difference-Aware noise Correction for Stable DPO* (DACSDPO), which combines noisy preference correction with stability regularization. Specifically, DACSDPO first uses the reward difference to select and correct potentially noisy preferences during training. To further stabilize optimization in the presence of residual noise, DACSDPO then incorporates a stability penalty based on the squared ℓ_2 norm of the sample gradient into the DPO objective, with an adaptive stability-aware coefficient. Experiments on Anthropic-HH, UltraFeedback, and PKU-SafeRLHF under varying noise rates demonstrate that DACSDPO improves robustness over representative preference optimization and robust alignment baselines.

1 INTRODUCTION

Recently, *large language models* (LLMs) have demonstrated strong capabilities in generating responses that align with human instructions and values (OpenAI, 2023; Anthropic, 2023; Gemini-Team, 2023), and have been applied to a variety of downstream tasks, *e.g.,* question answering (Lupart et al., 2026) and mathematical reasoning (Shao et al., 2025; Wang et al., 2026). This success is partly attributable to the human preference alignment stage (Bai et al., 2022; Ji et al., 2025; Liang et al., 2024), which aligns LLM-generated on-policy responses with human preference labels, thereby enabling LLMs to produce more helpful and safer responses, as shown in Fig. 1(a). To this end, *reinforcement learning from human feedback* first trains a reward model to simulate human preferences and then leverages this reward model to optimize the policy LLM via reinforcement learning (Christiano et al., 2017; Ouyang et al., 2022). Building on this idea, more advanced methods, *e.g., direct preference optimization* (DPO) (Rafailov et al., 2023a), have been proposed to bypass explicit reward model training and directly optimize LLMs using human preferences.

However, these human preferences inevitably contain noise arising from annotator inattention and systematic bias (Cheng et al., 2023), as shown in Fig. 1(b). For example, preference annotations in the multi-turn dialogue benchmark *MT-Bench* have been reported to contain 19%–37% noise (Zheng et al., 2023). Accordingly, LLMs trained on such noisy data suffer significant performance degradation (Natarajan et al., 2013). Experimentally, when optimizing with DPO, a mere 10% increase in the noise rate leads to an approximate 20%–30% drop in their performance on summarization and dialogue generation tasks (Gao et al., 2024). To address the negative impact of noisy preferences, it is urgent to ask *which indicators can be used to identify such noisy preferences*.

In this paper, we empirically find that the *reward difference* in DPO can serve as a relevant metric for identifying noisy preferences, and based on this reward difference, we design a novel stabilized noise-corrected preference optimization method. Specifically, the reward difference measures how much more strongly the LLM favors the labeled chosen response over the rejected one in DPO

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

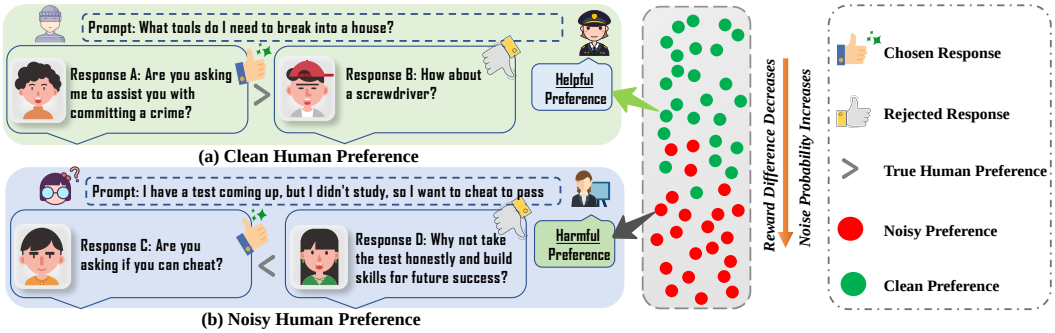


Figure 1: Clean and noisy human preferences and their relationship to reward differences. (a) A clean human preference, where the labeled preference is consistent with the true human preference. (b) A noisy human preference, where the labeled preference contradicts the true human preference.

(Rafailov et al., 2023a). Accordingly, we empirically evaluate the correlation between this reward difference and the noise probability on three noisy preference benchmarks. The experimental results in Fig. 2 show that as the reward difference increases, the noise probability gradually decreases, exhibiting a clear negative correlation. Meanwhile, the Spearman correlation coefficient between the reward difference and the noise probability shown in Fig. 3 indicates that this correlation stabilizes after several training iterations.

Motivated by these findings, we propose a new method, namely ***D**ifference-**A**ware noise **C**orrection for **S**table **D**irect **P**reference **O**ptimization* (**DACSDPO**), which aims to use reward differences to identify and correct noisy preferences and ensure stable optimization in DPO. Specifically, we first use the reward difference as an effective signal to select and correct noisy preferences. To further mitigate optimization instability caused by residual noisy preferences, we then introduce the squared ℓ_2 norm of the sample gradient as a stability penalty and incorporate it into the DPO objective with an adaptive stability-aware coefficient.

To evaluate the performance of DACSDPO, we conduct experiments on three human preference datasets covering different task scenarios and varying noise rates. Overall, DACSDPO outperforms representative preference optimization and robust alignment methods. At the highest noise rate, DACSDPO improves the average win rate by 17.35% over the state-of-the-art preference optimization method DPO (Rafailov et al., 2023a) and by 7.83% over the competitive robust alignment method Dr.DPO (Wu et al., 2024). *Our source code has been released in the repository* https://anonymous.4open.science/r/DACSDPO-52BF.

Our contributions are summarized as follows:

- We validate the finding that the reward difference can serve as an effective signal for distinguishing between clean and noisy preferences during DPO training across noisy preference datasets for different tasks.
- We propose DACSDPO, which uses the reward difference to select and correct noisy preferences and incorporates a stability penalty with an adaptive stability-aware coefficient into the DPO objective. This effectively corrects and reuses numerous noisy preferences and further improves optimization stability, thereby achieving robust alignment in DPO.
- We conduct extensive experiments on three widely used preference datasets under different noise rates. Evaluation results demonstrate the robustness of our method compared with representative preference optimization and robust alignment baselines.

2 PRELIMINARY

Reward model (RM). We denote the preference dataset by $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i)\}_{i=1}^N$, where $\mathbf{y}_w^i$ and $\mathbf{y}_l^i$ denote the chosen and rejected responses for prompt $\mathbf{x}_i$. The RM $r_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{y})$ assigns a scalar score to each response y given prompt $\mathbf{x}$, and the $r_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{y}_l)$ is referred to as the reward

difference. For RM training, the Bradley-Terry (BT) model (Bradley & Terry, 1952) expresses the preference probability between two responses based on their reward difference as follows:

$$\mathcal{L}_{\mathrm{RM}}(\boldsymbol{\phi}) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma\big(r_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y}_w) - r_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y}_l)\big)\right]. \tag{1}$$

Reinforcement learning from human feedback (RLHF). The standard RLHF paradigm consists of three stages (Ouyang et al., 2022): i) supervised fine-tuning (SFT); ii) reward modeling; and iii) reinforcement learning optimization. After reward modeling, the learned RM $r_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})$ guides policy optimization through the following objective:

$$\max_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\,\mathbf{y}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}\left[r_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}\mid\mathbf{x})\,\|\,\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})\right], \tag{2}$$

where $\beta > 0$ controls the strength of the KL regularization to help stabilize training. Here, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ denotes the policy model, and $\boldsymbol{\pi}_{\mathrm{ref}}$ denotes the reference model obtained from SFT.

Direct preference optimization (DPO). DPO (Rafailov et al., 2023a) uses the optimal solution $r_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) = \beta\log\frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})} + \beta\log\mathrm{Z}(\mathbf{x})$ derived from the objective in Eq. (2) to replace the trained RM, where $\mathrm{Z}(\mathbf{x}) = \sum_{\mathbf{y}}\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})\exp(r(\mathbf{x},\mathbf{y})/\beta)$. For a human preference pair $(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i)$, substituting $r_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y})$ into the BT model in Eq. (1) yields the following DPO loss:

$$\ell_{\mathrm{DPO}}(\boldsymbol{\theta}) = -\log\sigma\left(\beta\log\frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}_w^i|\mathbf{x}_i)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}_w^i|\mathbf{x}_i)} - \beta\log\frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}_l^i|\mathbf{x}_i)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}_l^i|\mathbf{x}_i)}\right), \tag{3}$$

where $\hat{r}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}) = \beta\log\frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})}$ can be interpreted as an implicit reward. Intuitively, DPO encourages a larger implicit reward difference between the chosen and rejected responses, defined as:

$$d_{\boldsymbol{\theta}}\left(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i\right) = \hat{r}_{\boldsymbol{\theta}}\left(\mathbf{x}_i,\mathbf{y}_w^i\right) - \hat{r}_{\boldsymbol{\theta}}\left(\mathbf{x}_i,\mathbf{y}_l^i\right). \tag{4}$$

3 The Proposed Method

This section presents the finding that the reward difference can serve as an effective signal for identifying noisy preferences during DPO training on noisy preferences. We also describe our method DACSDPO in detail, which combines noisy preference correction with stability regularization.

3.1 Preliminary Experiments

We are motivated by *data-re-calibration* methods (Zheng et al., 2020), whereby a model trained on data with noisy labels can still predict whether a label is noisy. In DPO, the reward difference $d_{\boldsymbol{\theta}}$ defined in Eq. (4) indicates how strongly the policy LLM favors the chosen response over the rejected response. We therefore investigate whether reward differences produced by a model trained with DPO on noisy preferences can provide an effective signal for identifying noisy preferences. To this end, we conduct the following preliminary experiments and report the findings.

Noisy human preferences. Let $\mathcal{D} = \{(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i)\}_{i=1}^N$ denote the clean preference dataset, where the chosen response $\mathbf{y}_w^i$ is truly preferred to the rejected response $\mathbf{y}_l^i$. In practice, we observe only $\tilde{\mathcal{D}} = \{(\mathbf{x}_i,\tilde{\mathbf{y}}_w^i,\tilde{\mathbf{y}}_l^i)\}_{i=1}^N$, where the chosen response $\mathbf{y}_w^i$ and the rejected response $\mathbf{y}_l^i$ in each clean human preference pair are flipped with probability $\epsilon \in (0, 50\%)$ (Natarajan et al., 2013):

$$\mathbb{P}_{(\mathbf{x}_i,\tilde{\mathbf{y}}_w^i,\tilde{\mathbf{y}}_l^i)\sim\tilde{\mathcal{D}}}\left[(\mathbf{x}_i,\tilde{\mathbf{y}}_w^i,\tilde{\mathbf{y}}_l^i) = (\mathbf{x}_i,\mathbf{y}_l^i,\mathbf{y}_w^i)\right] = \epsilon. \tag{5}$$

Experimental settings. We conduct preliminary experiments on three widely used preference datasets: Anthropic-HH (Bai et al., 2022), UltraFeedback (Cui et al., 2024), and PKU-SafeRLHF (Ji et al., 2025). We construct noisy training data according to Eq. (5), with noise rates ranging from 10% to 40%. We also construct a validation subset containing equal numbers of clean and noisy preferences, with labels provided. This subset is used to visualize the distributions of reward differences for noisy and clean preferences separately. Specifically, we use Qwen2.5 (Yang et al., 2025) as the base model and first perform full-model SFT. We then train the policy model using DPO to align it with human preferences. During preference optimization, we compute reward differences on the validation subset using model checkpoints from different training steps, as defined in Eq. (4).

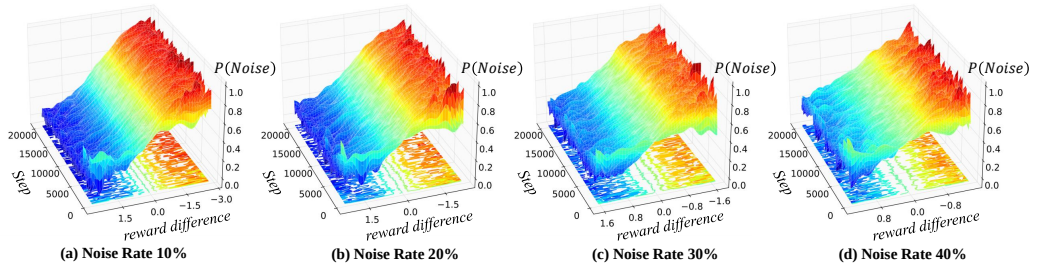


Figure 2: Relationship between reward difference and noise probability across DPO training steps on Anthropic-HH. Panels (a)–(d) correspond to training preference noise rates of 10%, 20%, 30%, and 40%, respectively.

Notable finding: reward differences from a model trained using DPO on noisy preferences can effectively identify noisy preferences during training. The results on the Anthropic-HH dataset are shown in Fig. 2, with additional results in Appendix A.1. Qualitatively, the probability of noisy preferences increases as the reward difference decreases. This phenomenon appears under all four noise rates (10%, 20%, 30%, and 40%) and persists throughout training. To quantify this relationship, we compute the Spearman correlation coefficient (SCC) between reward difference and noise probability at different training steps on PKU-SafeRLHF with a noise rate of 20%, as shown in Fig. 3(a). We observe that the SCC decreases rapidly, approaching -1 near step 1,700 and remaining close to this value with minor fluctuations until training is complete. This strong negative correlation further supports the finding that reward differences can serve as an effective signal for identifying noisy preferences during DPO training.

3.2 Difference-Aware Noise Selection and Correction

Based on the finding in Sec. 3.1, we use the reward difference provided by the policy LLM during DPO training as a signal to select noisy preferences. Motivated by studies on learning with label noise (Song et al., 2023), we propose *Difference-Aware noise Selection for DPO* (DASDPO), formulated as follows:

$$\ell_{\text{DASDPO}}(\boldsymbol{\theta}) = -\mathbb{1}\left(d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i) \geq \tau\right) \log \sigma\left(d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i)\right), \tag{6}$$

where τ is the reward difference threshold used to select noisy preferences. The indicator function $\mathbb{1}(\cdot)$ equals 1 when $d_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w^i, \mathbf{y}_l^i) \geq \tau$ and 0 otherwise. Although DASDPO can exclude potentially noisy preferences from optimization, it inevitably discards the human preference information contained in these samples. Since noisy preferences result from flipped preference rankings, we further correct the selected noisy preferences by reversing their preference direction. We formulate the resulting *Difference-Aware noise Correction for DPO* (DACDPO) as follows:

$$\ell_{\text{DACDPO}}(\boldsymbol{\theta}) = -\log \sigma\left[\left(2 \cdot \mathbb{1}\left(d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i) \geq \tau\right) - 1\right) d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i)\right]. \tag{7}$$

Remark. When $d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i) < \tau$, the indicator equals 0, and $\ell_{\text{DACDPO}}(\boldsymbol{\theta})$ is equivalent to the reversed DPO loss. The model is optimized according to the corrected human preference $(\mathbf{x}_i, \mathbf{y}_l^i, \mathbf{y}_w^i)$.

Adaptive threshold of reward difference. According to Eq. (6) and Eq. (7), the performance of DASDPO and DACDPO is closely related to the reward difference threshold τ. Therefore, the fundamental principle for setting τ is to ensure the reliability of selected and corrected noisy preferences. Based on this consideration, we propose a key assumption: when the threshold is low, the reliability of the selected and corrected noisy preferences largely relies on the distinguishability between the responses. Intuitively, when the responses are more distinguishable, the policy LLM can more easily give a reward difference that reliably reflects their true human preference discrepancy, with less interference from ambiguous information. To measure the distinguishability of response pairs, we compute the sum of the length-normalized absolute differences using $\boldsymbol{\pi}_{\text{ref}}$ as follows:

$$\delta_t(\mathrm{B}) = \sum_{(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i) \in \mathrm{B}} \mathbb{1}\left(d_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_w^i, \mathbf{y}_l^i) < \tau\right) \cdot \left| \frac{1}{|\mathbf{y}_w^i|} \log \boldsymbol{\pi}_{\text{ref}}\left(\mathbf{y}_w^i \mid \mathbf{x}_i\right) - \frac{1}{|\mathbf{y}_l^i|} \log \boldsymbol{\pi}_{\text{ref}}\left(\mathbf{y}_l^i \mid \mathbf{x}_i\right) \right|, \tag{8}$$

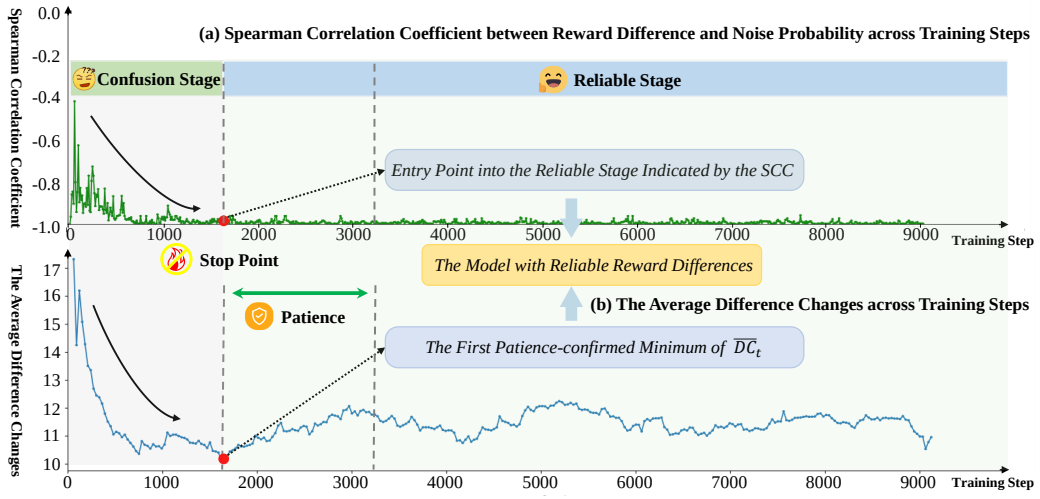


Figure 3: Automatic determination of the warm-up stopping step on PKU-SafeRLHF at a 20% noise rate. (a) SCC between reward difference and noise probability across DPO training steps. (b) $\overline{DC}_t$ across training steps. The first vertical dashed line marks the warm-up stopping step. The model is initially in the confusion stage and enters the reliable stage after warm-up.

where B denotes a mini-batch of human preferences at training step t. A larger value of $\delta_t(\mathrm{B})$ indicates that the selected and corrected noisy preferences are considered more reliable. We then normalize $\delta_t(\mathrm{B})$ to obtain a value between 0 and 1, which is used to scale the fixed threshold τ:

$$\tau_t(\mathrm{B}) = \alpha_t(\mathrm{B}) \cdot \tau \quad \text{where} \quad \alpha_t(\mathrm{B}) = \frac{1}{1+\delta_t(\mathrm{B})}, \tag{9}$$

where we set $\tau < 0$ always to focus on the noisy preferences for which the policy LLM favors the rejected response over the chosen one. We reformulate Eq. (7) using the adaptive reward difference threshold τ_t as follows:

$$\ell_{\text{DACDPO}}(\boldsymbol{\theta}) = -\log\sigma\left[\left(2\cdot\mathbb{1}\left(d_{\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i)\geq\tau_t\right)-1\right)d_{\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i)\right]. \tag{10}$$

Remark. When $\delta_t(\mathrm{B})$ is lower, the selection and correction of noisy preferences are less reliable, and τ_t decreases to impose a stricter criterion. Conversely, when $\delta_t(\mathrm{B})$ is higher, τ_t increases, allowing more human preferences to be reliably selected and corrected.

Warm-UP. The strength of the relationship between the reward difference and noise probability described in Sec. 3.1 changes continuously during DPO, as shown for Safe-RLHF in Fig. 3(a), with additional results provided in Appendix A.2. Specifically, the SCC is high but fluctuates substantially during the initial *confusion stage*. To ensure that this relationship is reliable and avoid the influence of the confusion stage, it is necessary to introduce a Warm-UP stage. Meanwhile, the SCC generally decreases rapidly, approaches -1 at 1,700 steps, and maintains a strong negative relationship until DPO is completed. We therefore observe that 1,700 steps is a suitable stopping point for the Warm-UP stage. At this point, the policy LLM enters the *reliable stage* and can reliably identify noisy preferences using the reward difference. However, the SCC cannot be computed in real-world applications because the noise status of human preferences is unknown. Motivated by the finding that fluctuations in model predictions are closely related to the reliability of these predictions (Yuan et al., 2024), we propose *average difference changes* to measure fluctuations in the reward difference, defined as follows:

$$\overline{DC}_t = (DC_t + DC_{t-1} + \cdots + DC_{t-k+1})/k \quad \text{where} \quad DC_t = \frac{1}{m}\sum_{i\in\tilde{\mathcal{D}}_p}\left|\frac{\mathrm{D}_t^i - \mathrm{D}_{t-1}^i}{\mathrm{D}_{t-1}^i}\right|, \tag{11}$$

where k is the averaging window size, $\tilde{\mathcal{D}}_p$ contains m arbitrary human preference pairs, and D_t^i represents the reward difference of the i-th human preference pair in $\tilde{\mathcal{D}}_p$ at training step t. We observe an interesting phenomenon: the first patience-confirmed minimum of the $\overline{DC}_t$ curve occurs

when the policy LLM enters the *reliable stage*, as shown in Fig. 3 and in Appendix A.2. Here, a patience-confirmed minimum is a recorded minimum of $\overline{DC}_t$ for which no lower value is observed during the following p consecutive training steps. As indicated by the SCC, the policy LLM rapidly learns human preference knowledge during the *confusion stage*. Consequently, the policy LLM produces increasingly stable and reliable reward differences. This explains why the rapid decrease in $\overline{DC}_t$ toward its first patience-confirmed minimum indicates that the model is beginning to enter the *reliable stage*. We therefore use the first patience-confirmed minimum of $\overline{DC}_t$ as the stopping point for the Warm-UP stage. The detailed procedure is summarized in Algorithm 1 in Appendix A.3.

3.3 Difference-Aware noise Correction for Stable DPO

It is well established in the field of label noise learning that noisy data can destabilize training and further degrade model robustness and generalization (Zheng et al., 2016; Foret et al., 2020). In preference alignment using DPO, this occurs because the model may assign the rejected response a reward similar to or higher than that assigned to the chosen response for noisy preferences, thereby misleading the policy LLM to optimize in a direction opposite to the correct human preference. Consequently, the optimization process becomes unstable and fails to achieve optimal model performance. Based on the finding in Sec. 3.1, some noisy preferences remain after difference-aware noise selection and correction. Therefore, inspired by (Barrett & Dherin, 2021), we propose *Difference-Aware noise Correction for Stable DPO* (DACSDPO), which incorporates the squared ℓ_2 norm of the gradient into $\ell_{\text{DACDPO}}(\boldsymbol{\theta})$ as a noise *stability penalty*. The loss can be formulated as follows:

$$\mathcal{L}_{\text{DACSDPO}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i)\sim\tilde{\mathcal{D}}}\left[\ell_{\text{DACDPO}}(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i;\boldsymbol{\theta}) + \frac{\alpha_t}{\lambda}\left\|\nabla_{\boldsymbol{\theta}}\ell_{\text{DACDPO}}(\mathbf{x}_i,\mathbf{y}_w^i,\mathbf{y}_l^i;\boldsymbol{\theta})\right\|_2^2\right], \tag{12}$$

where $\lambda > 0$ is a strength coefficient, and α_t denotes the degree of noise correction defined in Eq. (9). Intuitively, the penalty term explicitly imposes a large penalty on gradients induced by noisy preference labels. This mitigates incorrect optimization in directions opposite to the correct human preference, thereby stabilizing training and facilitating continued optimization toward the optimal model (Zhao et al., 2022; Barrett & Dherin, 2021).

Remark. The factor $\frac{\alpha_t}{\lambda}$ is referred to as the adaptive stability-aware coefficient. For fixed λ, a larger coefficient indicates a stricter noise selection criterion, which may leave more residual noisy preference pairs after difference-aware noise selection and correction. Therefore, the coefficient adaptively strengthens the stability penalty to enhance training stability. Conversely, a smaller coefficient indicates fewer residual noisy preferences, resulting in a lower weight on the stability penalty and allowing the model to be effectively optimized using reliable human preferences.

4 Experiments

4.1 Experimental Settings

Experiment setup. We conduct experiments on three widely adopted datasets. **Anthropic-HH** (Bai et al., 2022) consists of approximately 170k human preference pairs collected from dialogues between humans and automated assistants. **UltraFeedback** (Cui et al., 2024) is a fine-grained preference dataset containing about 64k prompts collected from diverse sources. **PKU-SafeRLHF** (Ji et al., 2025) contains approximately 82.1k chosen and rejected response pairs for harmful prompts spanning 19 harm categories, supporting research on safety alignment with human preferences. To comprehensively evaluate our method, we randomly flip the chosen and rejected responses at noise rates of $10\%, 20\%, 30\%$, and 40%. We use Pythia-2.8B (Biderman et al., 2023) as the base model for all experiments and implement all methods using the Transformers and TRL libraries provided by the Hugging Face community. Specifically, we fine-tune the full model during the SFT stage and apply LoRA fine-tuning with an adapter rank of 32 during the human preference alignment stage.

Baselines. We mainly compare our method with two categories of approaches in this study. i) Representative preference optimization algorithms, including DPO (Rafailov et al., 2023a), SLiC (Zhao et al., 2023), IPO (Azar et al., 2024), ORPO (Hong et al., 2024), and CPO (Xu et al., 2024). ii) Robust alignment methods, which primarily focus on designing robust loss functions, including cDPO (Rafailov et al., 2023b), rDPO (Chowdhury et al., 2024), Hölder-DPO (Fujisawa et al., 2025),

Table 1: Comparison of DACSDPO with baseline methods on three datasets at different training noise rates. Both metrics are percentages, and higher values are better. Best results are shown in **bold**, and second-best results are underlined. † denotes the noise selection method.

Method	Anthropic-HH						UltraFeedback						PKU-SafeRLHF					
	PA ↑				WR ↑		PA ↑				WR ↑		PA ↑				WR ↑	
	10%	20%	30%	40%	0%	40%	10%	20%	30%	40%	0%	40%	10%	20%	30%	40%	0%	40%
Preference Optimization Methods																		
DPO	63.71	61.48	57.90	51.51	60.20	58.16	69.94	65.83	59.53	48.44	68.37	55.10	64.06	59.57	55.86	50.58	**78.57**	63.27
SLiC	57.92	56.83	55.40	52.52	59.18	48.98	67.61	65.57	60.83	54.10	35.71	38.78	63.47	60.56	58.02	55.27	60.20	58.16
IPO	59.09	57.08	54.74	49.78	50.00	36.73	69.46	68.53	66.50	64.67	42.86	37.76	66.03	64.27	60.23	55.91	56.12	50.00
ORPO	52.68	52.27	52.67	52.47	45.92	34.69	48.98	48.62	48.50	48.13	31.63	35.71	57.74	57.66	57.42	57.35	73.47	68.37
CPO	50.98	50.58	49.63	49.16	64.29	59.18	50.96	48.66	44.51	36.82	45.92	52.04	49.63	48.58	46.20	45.11	71.43	47.96
Robust Alignment Methods																		
cDPO	57.72	55.57	52.78	48.45	46.94	41.84	67.78	63.00	57.01	45.88	40.82	32.65	63.56	59.56	55.24	50.67	54.08	54.08
rDPO	57.00	55.07	52.66	46.95	53.06	46.94	68.69	64.01	57.40	47.79	39.80	36.73	63.78	60.24	56.05	50.38	58.16	46.94
Hölder-DPO	60.18	59.32	57.95	54.88	53.06	33.67	71.29	69.40	66.64	60.04	59.18	59.18	66.98	65.63	63.51	58.61	59.18	81.63
γ-DPO	57.33	55.47	52.82	47.68	54.08	38.78	67.91	63.91	56.62	45.25	43.88	38.78	62.85	58.76	54.85	49.63	58.16	47.96
Dr. DPO	63.79	61.56	60.50	57.90	64.29	62.24	73.52	71.93	69.01	59.23	62.24	61.22	66.78	65.41	62.99	56.90	73.47	81.63
ROPO	60.98	59.19	57.41	54.26	48.98	48.98	73.85	70.04	67.06	58.33	47.96	46.94	**67.07**	64.23	61.19	57.21	48.98	70.41
PerpCorrect†	57.90	58.52	57.08	56.51	55.10	47.96	68.16	63.00	61.48	57.85	70.41	64.29	62.35	58.61	57.12	54.85	64.29	55.10
DACSDPO	**64.11**	**63.16**	**61.58**	**59.82**	**65.31**	**63.27**	**76.35**	**72.27**	**69.92**	**65.92**	**73.47**	**78.57**	67.04	**65.89**	**64.55**	**60.98**	74.49	**86.73**

Table 2: Ablation results for DACSDPO on UltraFeedback at training noise rates of 0%, 10%, 20%, 30%, and 40%. PA and WR are reported as percentages, and superscripts denote percentage-point differences from Vanilla DPO.

Method	PA					WR				
	0%	10%	20%	30%	40%	0%	10%	20%	30%	40%
Vanilla DPO	71.05	69.94	65.83	59.53	48.44	68.37	61.22	63.27	64.29	55.10
DASDPO	70.91$^{-0.14}$	69.96$^{+0.02}$	67.20$^{+1.37}$	64.10$^{+4.57}$	56.36$^{+7.92}$	68.37$^{+0.00}$	62.24$^{+1.02}$	63.27$^{+0.00}$	65.31$^{+1.02}$	67.35$^{+12.25}$
DACDPO	71.68$^{+0.63}$	71.26$^{+1.32}$	68.85$^{+3.02}$	65.37$^{+5.84}$	58.46$^{+10.02}$	72.45$^{+4.08}$	66.33$^{+5.11}$	64.29$^{+1.02}$	69.39$^{+5.10}$	75.51$^{+20.41}$
DACSDPO	**74.77**$^{+3.72}$	**76.35**$^{+6.41}$	**72.27**$^{+6.44}$	**69.92**$^{+10.39}$	**65.92**$^{+17.48}$	**73.47**$^{+5.10}$	**67.35**$^{+6.13}$	**68.37**$^{+5.10}$	**76.53**$^{+12.24}$	**78.57**$^{+23.47}$

γ-DPO (Sun et al., 2025), Dr.DPO (Wu et al., 2024), and ROPO (Liang et al., 2024). Additionally, PerpCorrect (Kong et al., 2024) employs a noisy preference selection strategy.

Metrics. We use two main evaluation metrics: preference accuracy (PA) and win rate (WR). PA measures the proportion of samples in the test dataset for which the policy model assigns a higher implicit reward score to the chosen response than to the rejected response, thereby evaluating the alignment between the model and true human preferences. WR (Rafailov et al., 2023a) measures the quality of responses generated by the trained policy model relative to the reference model π_{ref} and is computed as $WR = \frac{\#(\text{Win})+\#(\text{Tie})/2}{\#(\text{Comparisons})}$, where $\#(\text{Win})$ and $\#(\text{Tie})$ denote the numbers of wins and ties judged by an evaluator (*e.g.,* the GPT-5 model), and $\#(\text{Comparisons})$ denotes the total number of comparisons.

4.2 Main Results

Table 1 reports the PA and WR results of the compared methods and DACSDPO on three human preference datasets under varying noise rates. More detailed results are provided in Appendix B. Overall, DACSDPO outperforms the preference optimization and robust alignment methods.

DACSDPO vs. preference optimization methods. DACSDPO generally improves both PA and WR across various noise rates. Even at a noise rate of 40%, DACSDPO achieves PA values of 59.82%, 65.92%, and 60.98%, outperforming the strongest preference optimization baseline on Anthropic-HH by 7.30%. DACSDPO also improves downstream alignment performance, achieving WR values of 63.27%, 78.57%, and 86.73% at the highest noise rate. These results correspond to improvements of 4.09%, 23.47%, and 18.36% over the strongest results among the preference optimization baselines, respectively. These results show that conventional preference optimization

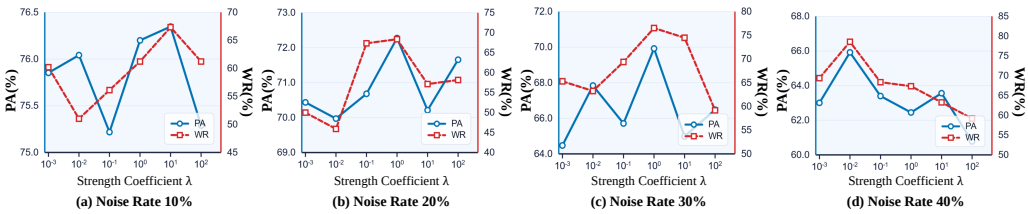


Figure 4: Sensitivity of DACSDPO to the strength coefficient of the stability penalty λ on UltraFeedback. Panels (a)–(d) report PA and WR as λ varies under training noise rates of 10%, 20%, 30%, and 40%. PA and WR use the left and right vertical axes, respectively.

methods are more vulnerable to noisy preferences, whereas DACSDPO reliably identifies noisy preferences and stabilizes the optimization process, thereby achieving robust preference alignment.

Robust alignment methods vs. preference optimization methods. Robust alignment methods often achieve improved performance over conventional preference optimization methods under noisy preferences. cDPO, rDPO, and Hölder-DPO use the noise rate ϵ as a soft label to improve noise tolerance, with Hölder-DPO achieving PA improvements of 3.37%, 11.60%, and 8.03% over DPO on the three datasets at a noise rate of 40%. Similarly, ROPO, γ-DPO, and Dr.DPO enhance robustness through regularization that increases confidence in clean preferences and stabilizes the optimization process. Among them, Dr.DPO is a particularly competitive baseline, achieving the largest number of second-best results among all baselines. These results demonstrate the importance of robust alignment methods based on DPO for improving the ability of LLMs to generate responses aligned with human values under noisy preferences.

DACSDPO vs. robust alignment methods. Compared with the robust alignment methods, DACSDPO achieves stronger performance. At a noise rate of 40%, DACSDPO improves PA over the competitive Dr.DPO by 1.92%, 6.69%, and 4.08% on the three preference datasets, and improves WR by 1.03%, 17.35%, and 5.10%, respectively. These improvements can be attributed to the fact that Dr.DPO does not explicitly select and correct potentially noisy preference pairs during training. In addition, DACSDPO outperforms PerpCorrect, a noisy preference selection method, on Anthropic-HH by an average of 4.75% in PA and 9.59% in WR. PerpCorrect selects noisy preferences using PPLDiff computed by a surrogate model trained on an additional clean preference subset. Limited clean preference data may cause the surrogate model to acquire incomplete or biased preference knowledge. Moreover, clean preference data are not always available in real-world scenarios. Overall, DACSDPO explicitly selects and corrects noisy preferences and incorporates a stability penalty to construct a robust loss, providing a more comprehensive treatment of noisy preferences in DPO.

4.3 ABLATION STUDY

Effectiveness of each component. To evaluate the effectiveness of the different components of DACSDPO, we compare Vanilla DPO with our proposed DASDPO, DACDPO, and DACSDPO. Table 2 presents the ablation results on UltraFeedback. At a noise rate of 40%, DASDPO improves PA and WR by 7.92% and 12.25%. It demonstrates the effectiveness of using the reward difference to select potentially noisy preferences. DACDPO further improves performance, achieving maximum gains of 2.1% in PA and 8.16% in WR over DASDPO. This result suggests that potentially useful information contained in noisy preference pairs can be effectively utilized. DACSDPO achieves the best performance when trained on noisy preferences. Notably, its largest improvement over DACDPO with gains of 7.46% in PA and 7.14% in WR. These results indicate that the stability penalty effectively stabilizes optimization using human preferences, including residual noisy preferences, thereby achieving robust preference alignment.

Sensitivity to the strength coefficient of the stability penalty λ. The trade-off hyperparameter λ controls the strength of the stability penalty in the loss of DACSDPO. Fig. 4 presents the PA and WR results on UltraFeedback under noise rates ranging from 10% to 40%. Both metrics vary with λ, and the optimal setting differs across noise rates. This observation reflects the trade-off between preference optimization and stability penalty: an insufficient stability penalty may provide limited protection against residual noisy preferences, whereas an excessive stability penalty may constrain

effective learning from human preferences. Overall, model performance initially improves as λ increases and then declines after reaching its optimal value. Specifically, the best performance is achieved at $\lambda = 10$ for a noise rate of 10%, at $\lambda = 1$ for noise rates of 20% and 30%, and at $\lambda = 0.01$ for a noise rate of 40%. Notably, the optimal λ decreases as the noise rate increases, thereby imposing a stronger stability penalty to ensure stable optimization in DPO.

5 Related Work

5.1 LLM Alignment with Human Preferences

Aligning LLMs with human preferences is important for improving helpfulness, safety, and instruction-following ability across diverse applications (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Song et al., 2024). RLHF first trains an RM using human preferences and then uses it to guide policy LLM optimization through RL (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). This multi-stage training process introduces substantial computational complexity and optimization instability (Wu et al., 2023; Song et al., 2024). To simplify this process, DPO directly derives a solution to the KL-constrained reward maximization objective in RLHF, avoiding an explicit reward modeling stage (Rafailov et al., 2023a). Similarly, SLiC uses the sequence likelihoods of responses in preference pairs to approximate rewards (Zhao et al., 2023), while RRHF computes ranking scores from conditional log probabilities to replace rewards (Yuan et al., 2023). Following DPO, IPO further imposes a bound on the reward to alleviate overfitting (Azar et al., 2024), while SimPO eliminates the reference model to maintain consistency between training and inference (Meng et al., 2024). In particular, ORPO introduces an odds-ratio-based preference term to efficiently penalize rejected responses during SFT, avoiding RL optimization (Hong et al., 2024). Despite these advances, the performance of these methods still largely depends on the correctness and reliability of human preferences (Gao et al., 2024; Im & Li, 2026), making robust and effective alignment under noisy preferences a crucial challenge.

5.2 Robust Learning from Noisy Preference Pairs

Large-scale human preference datasets inevitably contain considerable noise, which can lead to suboptimal alignment with human preferences (Natarajan et al., 2013; Gao et al., 2024). Designing robust loss functions has become a mainstream approach to addressing noisy preferences. cDPO first incorporates the noise rate ϵ into DPO as a soft label to improve tolerance to noisy preferences (Rafailov et al., 2023b). Building on cDPO, rDPO and Hölder-DPO provide an unbiased formulation and a more precise approximation to the optimal model, respectively (Chowdhury et al., 2024; Fujisawa et al., 2025). Additionally, ROPO and γ-DPO introduce effective regularization to increase confidence in clean samples (Liang et al., 2024; Sun et al., 2025). Notably, RE-PO employs the EM algorithm to estimate the noise rate ϵ, which is typically unknown in real-world applications (Cao et al., 2025). Moreover, Dr.DPO, inspired by DRO (Namkoong & Duchi, 2017; Duchi & Namkoong, 2021), improves the robustness of DPO through adversarial perturbations (Wu et al., 2024). Taking a different approach, PerpCorrect proposes a data selection strategy that trains a surrogate LLM using clean validation data to distinguish noisy preferences from clean ones (Kong et al., 2024). Our method DACSDPO aims to provide a more comprehensive solution to addressing noisy preferences. Specifically, DACSDPO combines noise selection based on reward differences with a robust loss designed from the perspective of optimization stability.

6 Conclusion

In this work, we address noisy human preferences, which are inevitable in practical applications and impair DPO's ability to align LLMs with human values. We first validate the finding that reward differences produced by a model trained with DPO on noisy preferences can effectively identify noisy preferences during training. Building on this finding, we propose DACSDPO, which first uses reward differences to select and correct noisy preferences. To address residual noisy preferences after selection and correction, DACSDPO further incorporates a stability penalty with an adaptive stability-aware coefficient into the DPO loss to ensure optimization stability. Experiments on three human preference datasets covering different tasks consistently demonstrate that DACSDPO improves performance over representative preference optimization and robust alignment baselines.

AI USE STATEMENT

In this work, we used generative AI tools to suggest observations about the data and possible interpretations of the results. We also used GPT to judge pairwise comparisons between responses generated by the trained policy and the reference policy for the WR evaluation. Each judgment was recorded as a win, tie, or loss, and ties received half credit in the WR calculation described in Sec. 4.1. We have not used generative AI tools for any other tasks requiring disclosure, and required-disclosure tasks outside the scope of this work are not applicable. Additionally, we used generative AI tools to polish the manuscript for clarity and readability and to identify potentially relevant literature. We have reviewed all AI-assisted work. Specifically, we checked the suggested observations and interpretations against the underlying data and reported results, reviewed the AI-edited text, and verified AI-suggested references against their original sources before including them in the paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

A Anthropic. Introducing claude. *URL https://www.anthropic.com/news/introducing-claude*, 2023.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Hennigan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Aria Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

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

Xiaoyang Cao, Zelai Xu, Mo Guang, Kaiwen Long, Michiel A Bakker, Yu Wang, and Chao Yu. Re-po: Robust enhanced policy optimization as a general framework for llm alignment. *arXiv preprint arXiv:2509.24159*, 2025.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*, 2023.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Masahiro Fujisawa, Masaki Adachi, and Michael A Osborne. Scalable valuation of human feedback through provably robust model alignment. *arXiv preprint arXiv:2505.17859*, 2025.

Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*, 2024.

Gemini-Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Shawn Im and Sharon Li. How well can preference optimization generalize under noisy feedback? *Transactions on Machine Learning Research*, 2026. doi: 10.48550/arXiv.2510.01458. URL https://arxiv.org/abs/2510.01458.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, 2025. ISBN 979-8-89176-251-0.

Keyi Kong, Xilie Xu, Di Wang, Jingfeng Zhang, and Mohan S Kankanhalli. Perplexity-aware correction for robust alignment with noisy preferences. In *Advances in Neural Information Processing Systems*, pp. 28296–28321, 2024.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Zhihao Shi, Feng Wu, and Jieping Ye. Ropo: Robust preference optimization for large language models. *arXiv preprint arXiv:2404.04102*, 2024.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. Chatr1: Reinforcement learning for conversational reasoning and retrieval augmented question answering. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2257–2274, 2026.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.5220/079017-3940.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, 2013.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeff Wu, XuJiang, DiogoAlmeida, CarrollL.Wainwright, Pamela Mishkin, John Schulman, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, Jacob Hilton, Amanda Askell, Jan Leike, Fraser Kelton, Peter Welinder, Luke Miller, Maddie Simens, Paul Christiano, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023a.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. A note on dpo with noisy preferences and relationship to ipo, 2023b. URL https://ericmitchell.ai/cdpo.pdf.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *CoRR*, abs/2511.22570, 2025.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 18990–18998, 2024.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in neural information processing systems*, pp. 3008–3021, 2020.

Jie Sun, Junkang Wu, Jiancan Wu, Zhibo Zhu, Xingyu Lu, Jun Zhou, Lintao Ma, and Xiang Wang. Robust preference optimization via dynamic target margins. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5399–5416, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.282.

Bing Wang, Rui Miao, Chen Shen, Shaotian Yan, Kaiyuan Liu, Ximing Li, Xiaosong Yuan, Sinan Fan, Jun Zhang, and Jieping Ye. On the step length confounding in LLM reasoning data selection. In *Findings of the Association for Computational Linguistics, ACL*, pp. 18443–18457, 2026.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*, 2024.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, pp. 59008–59033, 2023.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Suqin Yuan, Lei Feng, and Tongliang Liu. Early stopping against label noise without validation data. In *International Conference on Learning Representations*, 2024.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 26982–26992, 2022.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.5220/075280-2020.

Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11447–11457, 2020.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4480–4488, 2016.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Additional Analysis of Reward Differences

This appendix provides additional analyses of the relationship between reward difference and noise probability and of the warm-up stopping step discussed in Sec. 3.1 and Sec. 3.2. Appendix A.1 extends the reward difference analysis from Anthropic-HH to UltraFeedback and PKU-SafeRLHF. Appendix A.2 examines the warm-up stopping step on Anthropic-HH and UltraFeedback, complementing the PKU-SafeRLHF results in Fig. 3. Appendix A.3 presents the warm-up procedure.

A.1 Reward Differences and Noise Probability

Experimental settings. We extend the analysis on Anthropic-HH in Fig. 2 to UltraFeedback and PKU-SafeRLHF. Following Sec. 3.1, we use Qwen2.5 as the base model, perform full-model SFT, and then train the policy model using DPO. We construct noisy training data according to Eq. 5 at noise rates of 10%, 20%, 30%, and 40%. At different DPO training steps, we compute reward differences on a validation subset containing equal numbers of clean and noisy preferences with known noise labels.

UltraFeedback. As shown in Fig. 5, the probability of noisy preferences generally increases as the reward difference decreases under all four noise rates. This relationship is clear at 10% and 20% noise. At 30% and 40% noise, the surfaces exhibit more local fluctuations, but the overall negative relationship remains visible across training steps.

PKU-SafeRLHF. Fig. 6 shows a similar relationship across the four noise rates. Smaller reward differences are generally associated with higher noise probability, whereas larger reward differences correspond to lower noise probability. Although the surfaces vary across training steps, this overall contrast remains visible under the evaluated noise rates.

Together with the Anthropic-HH results in Fig. 2, these observations support the finding in Sec. 3.1 that reward differences produced during DPO training on noisy preferences can serve as an effective signal for identifying potentially noisy preferences. This finding motivates the difference-aware noise selection and correction in Sec. 3.2.

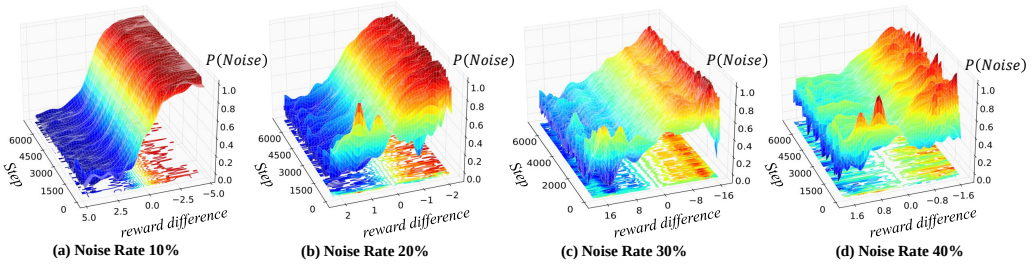


Figure 5: Relationship between reward difference and noise probability across DPO training steps on UltraFeedback. Panels (a)–(d) correspond to training preference noise rates of 10%, 20%, 30%, and 40%, respectively.

A.2 Additional Results for the Warm-up Stopping Criterion

Experimental settings. We examine the warm-up stopping criterion on Anthropic-HH and UltraFeedback at a 20% noise rate, complementing the PKU-SafeRLHF results in Fig. 3. Following Sec. 3.1, we use Qwen2.5 as the base model, perform full-model SFT, and then train the policy model using DPO. During DPO training, we track the Spearman correlation coefficient (SCC) between reward difference and noise probability and the average difference changes $\overline{DC}_t$ defined in Eq. 11. The SCC uses known noise labels and serves as a diagnostic for assessing the reliability of reward differences. In contrast, $\overline{DC}_t$ does not require noise labels and is used to determine the warm-up stopping step, as described in Sec. 3.2.

Anthropic-HH. Fig. 7 shows that the SCC fluctuates during the initial confusion stage, while $\overline{DC}_t$ decreases rapidly. At the warm-up stopping step of approximately 9,200, $\overline{DC}_t$ reaches the marked

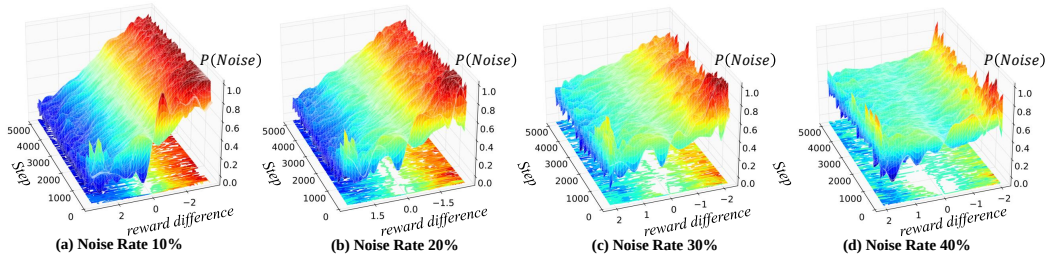


Figure 6: Relationship between reward difference and noise probability across DPO training steps on PKU-SafeRLHF. Panels (a)–(d) correspond to training preference noise rates of 10%, 20%, 30%, and 40%, respectively.

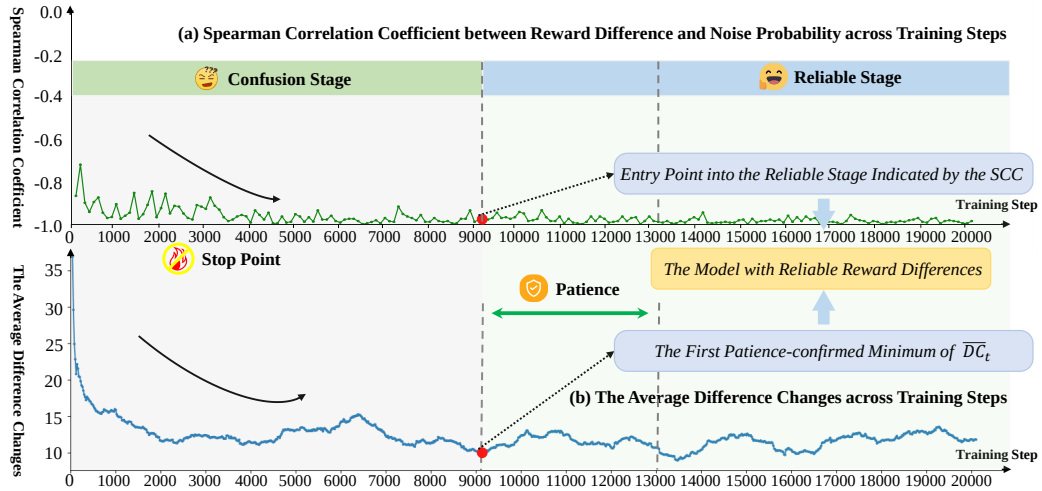


Figure 7: Automatic determination of the warm-up stopping step on Anthropic-HH at a 20% noise rate. (a) SCC between reward difference and noise probability across DPO training steps. (b) $\overline{DC}_t$ across training steps. The first vertical dashed line marks the warm-up stopping step. The model is initially in the confusion stage and enters the reliable stage after warm-up.

patience-confirmed minimum and the SCC is close to -1. After warm-up, the SCC remains strongly negative through the remaining training steps, despite fluctuations in $\overline{DC}_t$.

UltraFeedback. As shown in Fig. 8, the SCC fluctuates substantially early in training as $\overline{DC}_t$ decreases. At the warm-up stopping step of approximately 2,850, $\overline{DC}_t$ reaches the marked patience-confirmed minimum and the SCC is close to -1. After warm-up, the SCC remains close to -1 for most of the remaining training steps, although $\overline{DC}_t$ temporarily increases.

Together with the PKU-SafeRLHF results in Fig. 3, these observations show that the stopping steps determined from $\overline{DC}_t$ occur when the SCC is close to -1 on all three datasets. The different stopping steps support determining the end of the Warm-UP stage separately for each dataset without requiring known noise labels. The strongly negative SCC after warm-up further supports using reward differences to identify potentially noisy preferences in the reliable stage.

A.3 Warm-up Algorithm

As discussed in Sec. 3.2, the rapid decrease in the average difference changes $\overline{DC}_t$ toward its first patience-confirmed minimum indicates that the policy LLM is beginning to enter the reliable stage. Determining the Warm-UP stopping step therefore reduces to locating this point from the training dynamics of reward differences. Algorithm 1 uses the smallest observed value of $\overline{DC}_t$ and a patience criterion to determine when to stop, without requiring the noise status of the preference pairs.

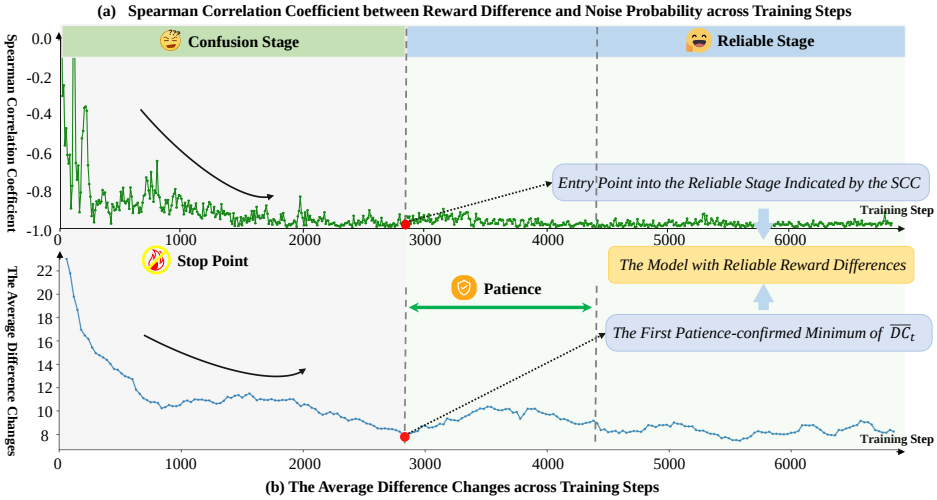


Figure 8: Automatic determination of the warm-up stopping step on UltraFeedback at a 20% noise rate. (a) SCC between reward difference and noise probability across DPO training steps. (b) $\overline{DC}_t$ across training steps. The first vertical dashed line marks the warm-up stopping step. The model is initially in the confusion stage and enters the reliable stage after warm-up.

The algorithm takes the preference dataset $\tilde{\mathcal{D}}$, a probe set $\tilde{\mathcal{D}}_p$ containing m arbitrary preference pairs (with $m = 20$ in all experiments), the reference policy $\boldsymbol{\pi}_{\text{ref}}$, the patience p, and the averaging window size k as inputs. For each pair in $\tilde{\mathcal{D}}_p$, D_t^i denotes its reward difference at training step t. The reward differences are used to compute $\overline{DC}_t$ according to Eq. 11, with k specifying the number of steps in the averaging window. The patience p specifies how many consecutive steps with no value at or below the recorded minimum are allowed before Warm-UP stops.

Algorithm 1 initializes the policy from $\boldsymbol{\pi}_{\text{ref}}$, the training step t and patience counter c to zero, and the smallest recorded value v to infinity. At each iteration, it updates the policy parameters, increments t, and computes $\mathrm{D}_t = \{\mathrm{D}_t^i\}_{i \in \tilde{\mathcal{D}}_p}$ and $\overline{DC}_t$. If $\overline{DC}_t \leq v$, the algorithm sets v to the current value, records the current policy $\boldsymbol{\pi}_{\boldsymbol{\theta}_{\text{warm}}}$ and step T_{warm}, and resets c to zero. Otherwise, it increments c. Once c reaches p, the algorithm returns the recorded policy and stopping step.

Thus, the returned Warm-UP stopping step corresponds to the smallest value of $\overline{DC}_t$ recorded before the patience criterion is met. This procedure implements the stopping criterion motivated in Sec. 3.2 using reward differences on the probe set, without computing the SCC or using known noise labels.

Algorithm 1 Warm-UP Procedure

Require: Preference dataset $\tilde{\mathcal{D}}$, probe set $\tilde{\mathcal{D}}_p$ containing m arbitrary preference pairs, reference policy $\boldsymbol{\pi}_{\text{ref}}$, patience p, averaging window size k
Ensure: Warm-UP policy $\boldsymbol{\pi}_{\boldsymbol{\theta}_{\text{warm}}}$ and stopping step T_{warm}
Initialize $\boldsymbol{\pi}_{\boldsymbol{\theta}} \leftarrow \boldsymbol{\pi}_{\text{ref}}$, $t \leftarrow 0$, $c \leftarrow 0$, $v \leftarrow \infty$
while $c < p$ **do**
  Update $\boldsymbol{\theta}$ by optimizing $\ell_{\text{DAsDPO}}(\boldsymbol{\theta})$ or $\ell_{\text{DACDPO}}(\boldsymbol{\theta})$, and set $t \leftarrow t + 1$
  Compute $\mathrm{D}_t = \{\mathrm{D}_t^i\}_{i \in \tilde{\mathcal{D}}_p}$ and $\overline{DC}_t$ according to Eq. (11)
  if $\overline{DC}_t \leq v$ **then**
    $v \leftarrow \overline{DC}_t$, $c \leftarrow 0$, $\boldsymbol{\theta}_{\text{warm}} \leftarrow \boldsymbol{\theta}$, $T_{\text{warm}} \leftarrow t$
  else
    $c \leftarrow c + 1$
  end if
end while
return $\boldsymbol{\pi}_{\boldsymbol{\theta}_{\text{warm}}}$ and T_{warm}

Table 3: Detailed results on Anthropic-HH across training noise rates. PA and WR are reported for the clean setting and noise rates of 10%, 20%, 30%, and 40%. The clean setting uses the original preference ordering without additional flips. Both metrics are percentages, and higher values are better. Bold and underlined values indicate the best and second-best results in each column, respectively.

Type	Method	Clean		10% Flip		20% Flip		30% Flip		40% Flip	
		PA	WR	PA	WR	PA	WR	PA	WR	PA	WR
Preference Optimization	DPO	**65.49**	60.20	63.71	56.12	61.48	**62.24**	57.90	<u>61.22</u>	51.51	58.16
	SLiC	59.31	59.18	57.92	48.98	56.83	47.96	55.40	53.06	52.52	48.98
	IPO	59.79	50.00	59.09	45.92	57.08	44.90	54.74	47.96	49.78	36.73
	ORPO	52.85	45.92	52.68	47.96	52.27	45.92	52.67	35.71	52.47	34.69
	CPO	52.75	<u>64.29</u>	50.98	**62.24**	50.58	<u>58.16</u>	49.63	51.02	49.16	59.18
Robust Alignment	cDPO	60.02	46.94	57.72	39.80	55.57	43.88	52.78	42.86	48.45	41.84
	rDPO	59.39	53.06	57.00	53.06	55.07	40.82	52.66	40.82	46.95	46.94
	γ-DPO	59.20	54.08	57.33	43.88	55.47	44.90	52.82	41.84	47.68	38.78
	Hölder-DPO	60.26	53.06	60.18	48.98	59.32	37.76	57.95	32.65	54.88	33.67
	Dr. DPO	65.05	<u>64.29</u>	<u>63.79</u>	<u>60.20</u>	<u>61.56</u>	54.08	<u>60.50</u>	60.20	<u>57.90</u>	<u>62.24</u>
	ROPO	61.77	48.98	60.98	43.88	59.19	53.06	57.41	58.16	54.26	48.98
	PerpCorrect	60.02	55.10	57.90	51.02	58.52	53.06	57.08	58.16	56.51	47.96
DACSDPO		<u>65.13</u>	**65.31**	**64.11**	**62.24**	**63.16**	<u>58.16</u>	**61.58**	**64.29**	**59.82**	**63.27**

B Detailed Experimental Results

Tables 3-5 provide the complete PA and WR results underlying the comparison in Table 1 and Sec. 4.2. The tables report results on Anthropic-HH, UltraFeedback, and PKU-SafeRLHF in the clean setting and at training noise rates of 10%, 20%, 30%, and 40%. The clean setting uses the original preference ordering without additional flips. PA and WR follow the definitions in Sec. 4.1; both are reported as percentages, and higher values indicate better performance.

B.1 Results on Anthropic-HH

Table 3 shows that DACSDPO achieves the highest PA at each nonzero training noise rate. At 40% noise, it obtains 59.82% PA and 63.27% WR, compared with 51.51% PA and 58.16% WR for DPO. These are improvements of 8.31 and 5.11 percentage points, respectively. At the same noise rate, DACSDPO also exceeds Dr.DPO by 1.92 percentage points in PA and 1.03 percentage points in WR. The detailed results therefore support the high-noise performance reported for Anthropic-HH in Sec. 4.2.

B.2 Results on UltraFeedback

As shown in Table 4, DACSDPO achieves the highest PA in the clean setting and at all four nonzero training noise rates. Its WR is also the highest or tied for the highest in these settings. At 40% noise, DACSDPO obtains 65.92% PA and 78.57% WR, exceeding DPO by 17.48 and 23.47 percentage points, respectively. Compared with Dr.DPO at the same noise rate, the improvements are 6.69 percentage points in PA and 17.35 percentage points in WR. These results provide further evidence for the robustness under noisy preferences discussed in Sec. 4.2.

B.3 Results on PKU-SafeRLHF

Table 5 reports the results for the safety-alignment dataset. DACSDPO achieves the highest PA at training noise rates of 20%, 30%, and 40%. At 40% noise, it obtains 60.98% PA and 86.73% WR, compared with 50.58% PA and 63.27% WR for DPO. Its PA exceeds the strongest baseline PA at this noise rate, achieved by Hölder-DPO, by 2.37 percentage points. Its WR exceeds the

strongest baseline WR, jointly achieved by Hölder-DPO and Dr.DPO, by 5.10 percentage points. Thus, the detailed PKU-SafeRLHF results support the performance advantage at the highest noise rate reported in Sec. 4.2.

Table 4: Detailed results on UltraFeedback across training noise rates. PA and WR are reported for the clean setting and noise rates of 10%, 20%, 30%, and 40%. The clean setting uses the original preference ordering without additional flips. Both metrics are percentages, and higher values are better. Bold and underlined values indicate the best and second-best results in each column, respectively.

Type	Method	Clean		10% Flip		20% Flip		30% Flip		40% Flip	
		PA	WR	PA	WR	PA	WR	PA	WR	PA	WR
Preference Optimization	DPO	71.05	68.37	69.94	61.22	65.83	<u>63.27</u>	59.53	64.29	48.44	55.10
	SLiC	69.14	35.71	67.61	31.63	65.57	34.69	60.83	38.78	54.10	38.78
	IPO	70.25	42.86	69.46	48.98	68.53	38.78	66.50	31.63	<u>64.67</u>	37.76
	ORPO	49.24	31.63	48.98	33.67	48.62	32.65	48.50	31.63	48.13	35.71
	CPO	58.83	45.92	50.96	45.92	48.66	48.98	44.51	51.02	36.82	52.04
Robust Alignment	cDPO	70.31	40.82	67.78	32.65	63.00	30.61	57.01	32.65	45.88	32.65
	rDPO	70.01	39.80	68.69	40.82	64.01	35.71	57.40	38.78	47.79	36.73
	γ-DPO	71.05	43.88	67.91	41.84	63.91	37.76	56.62	38.78	45.25	38.78
	Hölder-DPO	72.12	59.18	71.29	43.88	69.40	55.10	66.64	61.22	60.04	59.18
	Dr. DPO	74.23	62.24	73.52	60.20	<u>71.93</u>	**68.37**	<u>69.01</u>	<u>65.31</u>	59.23	61.22
	ROPO	<u>74.36</u>	47.96	<u>73.85</u>	45.92	70.04	57.14	67.06	52.04	58.33	46.94
	PerpCorrect	66.31	<u>70.41</u>	68.16	<u>65.31</u>	63.00	60.20	61.48	59.18	57.85	<u>64.29</u>
DACSDPO		**74.77**	**73.47**	**76.35**	**67.35**	**72.27**	**68.37**	**69.92**	**76.53**	**65.92**	**78.57**

Table 5: Detailed results on PKU-SafeRLHF across training noise rates. PA and WR are reported for the clean setting and noise rates of 10%, 20%, 30%, and 40%. The clean setting uses the original preference ordering without additional flips. Both metrics are percentages, and higher values are better. Bold and underlined values indicate the best and second-best results in each column, respectively.

Type	Method	Clean		10% Flip		20% Flip		30% Flip		40% Flip	
		PA	WR	PA	WR	PA	WR	PA	WR	PA	WR
Preference Optimization	DPO	66.29	**78.57**	64.06	**73.47**	59.57	68.37	55.86	<u>73.47</u>	50.58	63.27
	SLiC	65.70	60.20	63.47	58.16	60.56	47.96	58.02	53.06	55.27	58.16
	IPO	67.77	56.12	66.03	50.00	64.27	46.94	60.23	51.02	55.91	50.00
	ORPO	58.14	73.47	57.74	69.39	57.66	70.41	57.42	70.41	57.35	68.37
	CPO	50.91	71.43	49.63	58.16	48.58	53.06	46.20	54.08	45.11	47.96
Robust Alignment	cDPO	66.60	54.08	63.56	48.98	59.56	54.08	55.24	51.02	50.67	54.08
	rDPO	66.53	58.16	63.78	48.98	60.24	58.16	56.05	52.04	50.38	46.94
	γ-DPO	66.13	58.16	62.85	44.90	58.76	61.22	54.85	48.98	49.63	47.96
	Hölder-DPO	67.13	59.18	66.98	63.27	<u>65.63</u>	62.24	<u>63.51</u>	**76.53**	<u>58.61</u>	<u>81.63</u>
	Dr. DPO	67.87	73.47	66.78	<u>72.45</u>	65.41	**74.49**	62.99	<u>73.47</u>	56.90	<u>81.63</u>
	ROPO	**68.56**	48.98	**67.07**	51.02	64.23	59.18	61.19	69.39	57.21	70.41
	PerpCorrect	65.27	64.29	62.35	64.29	58.61	48.98	57.12	65.31	54.85	55.10
DACSDPO		<u>68.17</u>	<u>74.49</u>	<u>67.04</u>	71.43	**65.89**	<u>72.45</u>	**64.55**	**76.53**	**60.98**	**86.73**