PLATEAU-DPO: GRADIENT-ENVELOPE ALLOCATION FOR ROBUST PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust preference optimization must limit the influence of incorrect labels while preserving learning from valid preferences that the model has yet to learn. Both types of comparisons can have negative reference-relative margins, so down-weighting comparisons based on their margins can also weaken useful supervision. For robust losses whose gradient weights concentrate around a single peak, valid comparisons away from that peak receive smaller weights even when they remain useful for learning. To address this issue, we introduce Plateau-DPO, which assigns a constant gradient weight to all comparisons within a prescribed margin interval and smoothly attenuates weights outside it. Comparisons within the interval are therefore not downweighted merely because their margins deviate from a peak, while extreme margins still receive reduced weights. We join the plateau with Gaussian tails and integrate the resulting weighting function to obtain a bounded symmetric preference loss. For a fixed interval, loss range, and integrated tail mass, we prove that the constant plateau maximizes the minimum gradient weight within the interval. Further analysis provides sufficient conditions for one-step descent of a target objective. Experiments under controlled label flips on Anthropic HH and UltraFeedback show improvements in preference accuracy and generation win rate. Gradient diagnostics and shape ablations show that the plateau retains larger normalized gradient weights on hard, unflipped comparisons, supporting the design of preserving learning signals across a finite margin interval. Code is available at `https://anonymous.4open.science/r/Plateau-DPO-5B2F`.

1 INTRODUCTION

Preference optimization is widely used to align language-model behavior with comparative feedback (Kaufmann et al., 2023; Lambert et al., 2023; Shen et al., 2023). RLHF established a practical pipeline based on learned rewards and policy optimization (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), while direct preference optimization (DPO) replaces explicit reward-model training with a pairwise classification loss (Rafailov et al., 2023). In practice, preference data are collected across annotators, prompts, response pairs, and evaluation criteria, increasingly through synthetic or AI feedback (Kim et al., 2023; Lee et al., 2023; Chen et al., 2024b; Cui et al., 2024). Empirical studies document sensitivity to injected preference noise and systematic variation in model-based judgments (Gao et al., 2024; Zheng et al., 2023b; Chen et al., 2024a; Chiang et al., 2024; Li et al., 2023; Dubois et al., 2024; Li et al., 2024). Annotation errors and disagreement arising from ambiguity, underspecified criteria, and heterogeneous preferences are therefore unavoidable features of the data. Since every recorded comparison directs a policy update, the loss must determine how to learn from feedback that can conflict with other judgments or with the evolving policy.

Existing robust objectives address this setting through distinct mechanisms. Unbiased rDPO corrects the risk under a specified label-flip model, gamma-PO adapts target margins, and Hölder-DPO uses a bounded robust loss (Chowdhury et al., 2024; Sun et al., 2025; Fujisawa et al., 2026). These constructions focus on different statistical objectives, but they share an optimization question: how much weight should an update receive when the policy strongly disagrees with an observed label, and how much weight should remain on margins that still carry useful learning signals? Robustness

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.

to extreme disagreement alone does not determine how a method should preserve learning from comparisons that remain unresolved.

Gradient envelopes provide a common language for this question because they are the scalar coefficients that multiply pairwise margin gradients after differentiating a preference loss. They separate scalar weighting from pairwise gradient directions and the statistical assumptions behind each loss. A negative-tail bound caps the coefficient at extremely negative reference-relative margins, while minimum interval coverage records the weakest coefficient over a prescribed active region. DPO, rDPO, and gamma-PO retain nonzero negative-tail coefficients; Hölder-DPO attenuates both tails but concentrates its finite-margin coefficient near one peak. Figure 1 places these choices on a common margin axis and contrasts this peaked allocation with a plateau that spreads coefficient mass across the active interval. The figure concerns update geometry rather than learning performance, and motivates our question: can broader interval coverage preserve useful directional updates while retaining tail attenuation?

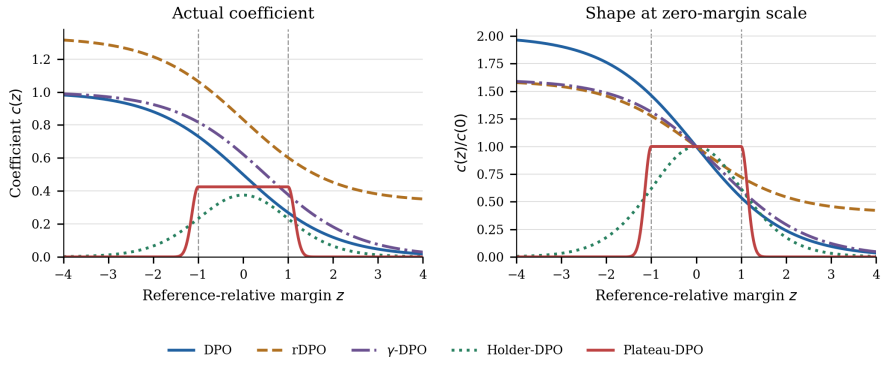


Figure 1: Gradient envelopes on a reference-relative margin axis: coefficients (left) and zero-margin-normalized shapes (right); dashed lines mark $\pm B$. The plot compares geometry rather than performance, and nonintegrable DPO, rDPO, and frozen-target γ-PO are not unit-mass normalized.

To answer this question, we introduce Plateau-DPO. At fixed loss range and tail mass, a constant coefficient maximizes the smallest coefficient in the active interval. We extend the plateau with Gaussian tails and integrate it to obtain a bounded loss. Each pair's coefficient depends on its current reference-relative margin; useful pairs outside the selected interval are attenuated as well, so the interval choice matters. We decompose an entire minibatch into active, transition, and extreme-tail pairs. If the average active-region gradient points sufficiently toward a target objective and the remaining adverse contributions are small, the resulting direction decreases that objective for a sufficiently short gradient-descent step. These conditions do not identify mislabeled pairs from their margins. Our contributions are as follows:

- We characterize four preference objectives through their update coefficients and derive a full-batch conditional descent bound that includes average active-region alignment and transition pairs.
- We formulate fixed-mass interval coverage and construct Plateau-DPO, whose central plateau attains its maximum possible minimum coefficient; Gaussian tails provide one smooth continuation with explicit pointwise bounds.
- Experiments on two model–dataset settings show that Plateau-DPO improves preference accuracy and generation win rate. Training curves on Anthropic HH further show that it avoids the late-training accuracy decline observed in the baselines.

2 Related Work

Preference alignment objectives. Later objectives modify the supervision or update geometry: RRHF and SLiC-HF use ranking or contrastive signals (Yuan et al., 2023; Zhao et al., 2023); ΨPO derives IPO as an alternative to logistic DPO (Azar et al., 2024); KTO accepts binary feedback (Ethayarajh et al., 2024); SimPO and ORPO remove the explicit reference term (Meng et al., 2024; Hong

et al., 2024); and β-DPO adapts the preference temperature (Wu et al., 2024). Reward-ranked fine-tuning and distilled alignment instead train from selected or model-generated responses (Dong et al., 2023; Tunstall et al., 2023). Preference ranking, contrastive objectives, generalized divergence constraints, and game-theoretic formulations further broaden offline alignment (Song et al., 2024; Xu et al., 2024; Wang et al., 2024; Munos et al., 2023; Swamy et al., 2024; Ramesh et al., 2024). Rejection sampling and pair weighting alter the effective training distribution (Liu et al., 2024; Zhou et al., 2024). These choices change the scalar coefficient on a pairwise gradient. We use its envelope as a diagnostic and design object without assuming a statistical interpretation across objectives.

Noisy preferences and correction. Existing responses follow the broader noisy-label taxonomy of selection, reweighting, correction, and robust loss design (Natarajan et al., 2013; Patrini et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019; Liu & Guo, 2020; Song et al., 2022). Curriculum and meta-reweighting methods learn which examples to emphasize, while early-learning and training-dynamics diagnostics seek to prevent or identify memorization (Jiang et al., 2018; Ren et al., 2018; Shu et al., 2019; Liu et al., 2020; Swayamdipta et al., 2020; Pleiss et al., 2020). In preference optimization, reward margins can select comparisons before training (Huang et al., 2025); rDPO unbiasedly estimates clean DPO risk under a known instance-independent symmetric flip rate (Chowdhury et al., 2024); robust RLHF treats corrupted feedback beyond DPO (Zheng et al., 2023a); and flipping-aware DPO estimates instance-dependent errors (Xu et al., 2026). PerpCorrect edits pairs using perplexity differences (Kong et al., 2024), ROPO couples a noise-aware loss with filtering and replacement (Liang et al., 2025), and RE-PO and PLC-DPO infer latent correctness, annotator reliability, or tie-like cases (Cao et al., 2026; Cho et al., 2026). Self-improving robust preference optimization is complementary in using model-generated comparisons (Choi et al., 2025). Plateau-DPO does not infer labels, reconstruct data, or treat a negative margin as proof of corruption, consistent with broader concerns about feedback misspecification (Casper et al., 2023). Its uniform-flip identity follows from loss symmetry and applies only to the unregularized surrogate.

Robust objectives and update geometry. Dr. DPO uses distributionally robust weights that couple each pair to the other margins in its batch (Wu et al., 2025); γ-PO adapts target margins to comparison ambiguity (Sun et al., 2025); and Hölder-DPO combines a bounded density-power objective with a redescending influence profile (Fujisawa et al., 2026). Two-sided tail attenuation therefore predates our work. We instead study allocation *within* a finite active interval: for a fixed loss range and tail mass, a constant plateau maximizes the interval's minimum coefficient. This scalar result does not rank methods; experiments test whether the allocation helps across their different objectives and training procedures.

3 Background: revisiting preference losses

3.1 Reference-relative margins and gradient envelopes

For a prompt x, responses (a, b), and observed label $y \in \{-1, +1\}$, define

$$s_\theta = \log \frac{\pi_\theta(a \mid x)}{\pi_{\mathrm{ref}}(a \mid x)} - \log \frac{\pi_\theta(b \mid x)}{\pi_{\mathrm{ref}}(b \mid x)}, \qquad z_\theta = y s_\theta. \tag{1}$$

Here $y = +1$ denotes a preference for a over b, and $y = -1$ denotes the reverse. Response log probabilities are sums over response tokens. The fixed reference policy makes s_θ a difference in relative log probabilities; $z_\theta < 0$ means that this reference-relative difference disagrees with the observed label.

For a differentiable per-pair loss ℓ, the descent direction factors into $c(z_\theta)\nabla_\theta z_\theta$, where $c(z) = -\ell'(z)$. The margin gradient supplies a direction in parameter space; c determines how strongly the loss weights that direction. We call a continuous nonnegative coefficient $w = c$ a *gradient envelope*. Nonnegativity preserves the direction associated with the observed preference. For objectives with auxiliary state ξ, the corresponding coefficient is $c(z; \xi) = -\partial_z \ell(z; \xi)$, with ξ held fixed during the policy update.

3.2 TAIL SUPPRESSION AND MINIMUM INTERVAL COVERAGE

Let w be a continuous nonnegative gradient envelope. For a state-dependent envelope $w(z;\xi)$, fix the auxiliary state ξ during one policy update. For $M > B > 0$, define

$$T_-(M) = \sup_{z \le -M} w(z), \qquad C(B) = \min_{|z| \le B} w(z). \tag{2}$$

The threshold M isolates extreme negative margins, whereas B specifies the active interval.

Definition 3.1 (Negative-tail decay)**.** The envelope has negative-tail decay if

$$\lim_{M \to \infty} T_-(M) = 0.$$

Equivalently, for every $\delta > 0$ there exists M_δ such that $w(z) \le \delta$ for all $z \le -M_\delta$.

Definition 3.2 (Strict active plateau)**.** The envelope has a strict active plateau on $[-B, B]$ if there exists $h > 0$ such that

$$w(z) = h \qquad \text{for every } z \in [-B, B].$$

The two properties constrain different parts of the envelope. $T_-(M)$ bounds the coefficient assigned to sufficiently extreme negative margins, while $C(B)$ measures the weakest coefficient in the active interval. A strict active plateau implies $C(B) = h > 0$ and fixes the coefficient throughout the interval; the weaker condition $C(B) > 0$ permits variation with z. Continuous positive baselines already have $C(B) > 0$ on every finite interval, so this condition alone does not establish a learning advantage.

3.3 GRADIENT ENVELOPES OF EXISTING PREFERENCE OBJECTIVES

Write $r = \beta z$, $p = \sigma(r)$, and $L(r) = -\log \sigma(r)$, with $\beta > 0$. Below $g(r) = -\partial_r \ell(r)$; the coefficient in the coordinate of Eq. 1 is $c(z) = \beta g(\beta z)$.

Likelihood fitting, noise correction, target adaptation, and bounded robust estimation induce different envelopes. We examine these choices through DPO (Rafailov et al., 2023), unbiased rDPO (Chowdhury et al., 2024), the DPO variant of γ-PO (Sun et al., 2025), and the empirical density-power loss of Hölder-DPO (Fujisawa et al., 2026, Eq. (12)). Differentiating their training objectives gives

$$\begin{aligned} g_{\mathrm{DPO}}(r) &= 1 - p, \\ g_{\mathrm{rDPO}}(r) &= 1 - p + \frac{\rho}{1 - 2\rho}, \\ g_{\gamma\text{-DPO}}(r_i; \gamma_i) &= 1 - \sigma(r_i - \gamma_i), \\ g_{\mathrm{H}}(r) &= \kappa(1+\kappa) p^{\kappa} (1-p)^2. \end{aligned} \tag{3}$$

Here $0 \le \rho < 1/2$ is the assumed uniform flip rate and $\kappa > 0$ is a shape parameter. The target $\gamma_i \ge 0$ is computed from current and historical margins and held fixed during each policy update. Loss expressions and coefficient derivations are given in Appendix C.4.

Negative-tail behavior. As $r \to -\infty$, the DPO and rDPO coefficients approach 1 and $(1-\rho)/(1-2\rho)$, respectively. The γ-PO coefficient also approaches 1, uniformly over its nonnegative targets. These three objectives continue to assign nonzero coefficients to extreme disagreement. Their statistical roles differ: in particular, rDPO's correction recovers an unbiased clean DPO risk under its assumed noise model. Hölder-DPO instead has coefficients that vanish in both tails. Risk correction and extreme-margin attenuation therefore express different choices about how noisy feedback should affect optimization.

Finite-interval gradient allocation. All four coefficients are positive on finite intervals. DPO, rDPO, and fixed-target γ-PO allocate progressively smaller coefficients as the margin increases. Hölder-DPO allocates its largest coefficient near a single peak, with smaller coefficients on either side. These shapes specify a preference among margins even within a region where all pairs remain active. The alternative studied here is to allocate a common coefficient throughout that region and let attenuation begin at its boundary.

Appendix Table 5 summarizes these structural properties without ranking learning quality. We next optimize minimum coverage within bounded monotone losses, whose finite endpoint difference constrains margin-space coefficient mass.

4 PLATEAU-DPO: ALLOCATING GRADIENTS OVER AN ACTIVE INTERVAL

We balance extreme-margin suppression with learning on unresolved pairs by maximizing the smallest coefficient over an active interval and extending the resulting plateau into a smooth, bounded loss.

4.1 MAX-MIN GRADIENT ALLOCATION

A negative margin can indicate either mislabeling or a valid, difficult preference. We therefore protect the least-weighted margin throughout $[-B, B]$, $B > 0$, without assigning reliability from its position.

For a bounded decreasing loss, the envelope integral equals its endpoint range. Endpoints 1 and 0 thus give unit margin-space mass, without fixing the batch update norm. Reserving tail mass $\varepsilon \in (0, 1)$ yields

$$\begin{aligned} \underset{w \geq 0}{\text{maximize}} \quad & C(B) = \min_{|z| \leq B} w(z) \\ \text{subject to} \quad & \int_{\mathbb{R}} w(z)\, dz = 1, \qquad \int_{|z|>B} w(z)\, dz = \varepsilon. \end{aligned} \tag{4}$$

The interval contains mass $1 - \varepsilon$, so uniform allocation attains the largest minimum, $(1 - \varepsilon)/(2B)$ (Theorem 5.3), with $w(z) = h_B$ on $[-B, B]$. The scalar coefficient remains constant across the reference-relative boundary, although the parameter update still depends on the margin-gradient geometry.

4.2 SMOOTH PLATEAU LOSS

Outside the active interval, symmetric Gaussian tails $h_B \exp[-(d/t_B)^2]$, $d = |z| - B > 0$, attenuate both extreme disagreement and already-satisfied preferences while joining the plateau with zero derivative. The mass constraints give

$$2Bh_B = 1 - \varepsilon, \qquad 2h_B \int_B^\infty \exp\left[-\left(\frac{z - B}{t_B}\right)^2\right] dz = h_B t_B \sqrt{\pi} = \varepsilon. \tag{5}$$

so that the plateau height and tail scale are determined jointly by B and ε:

$$h_B = \frac{1 - \varepsilon}{2B}, \qquad t_B = \frac{2B\varepsilon}{(1 - \varepsilon)\sqrt{\pi}}. \tag{6}$$

Integrating this envelope yields the Plateau-DPO loss:

$$w_{B,\varepsilon}(z) = h_B \exp\left[-\left(\frac{(|z| - B)_+}{t_B}\right)^2\right], \qquad \ell_{B,\varepsilon}(z) = \int_z^\infty w_{B,\varepsilon}(u)\, du, \tag{7}$$

where $(u)_+ = \max(u, 0)$. The loss is linear on the interval, smoothly saturates to 1 and 0 in the two tails, and satisfies $\ell(z) + \ell(-z) = 1$. Appendix C.5 gives its closed form.

Gaussian tails are one smooth continuation, not a consequence of the coverage theorem. For $M > B$,

$$C(B) = h_B, \qquad T_-(M) = h_B \exp\left[-\left(\frac{M - B}{t_B}\right)^2\right]. \tag{8}$$

Hence relative attenuation depends on distance beyond the interval in units of t_B. At fixed B, decreasing ε raises coverage and accelerates decay but sharpens the transition, approaching a ramp as $\varepsilon \to 0$ (Appendix C.7). At fixed ε, increasing B widens the plateau while lowering its height.

4.3 PREFERENCE OPTIMIZATION WITH PLATEAU GRADIENTS

Given a preference dataset $\mathcal{D} = \{(x_i, a_i, b_i, y_i)\}_{i=1}^n$, we apply the loss to the signed margins z_i in Eq. 1, keeping the reference policy fixed. The empirical risk and its descent direction are

$$\widehat{R}(\theta) = \frac{1}{n} \sum_i \ell_{B,\varepsilon}(z_i), \qquad -\nabla_\theta \widehat{R} = \frac{1}{n} \sum_i w_{B,\varepsilon}(z_i) \nabla_\theta z_i. \tag{9}$$

For an individual comparison, the resulting policy update takes the form

$$-\nabla_\theta \ell_{B,\varepsilon}(z_i) = w_{B,\varepsilon}(z_i) y_i \big[\nabla_\theta \log \pi_\theta(a_i \mid x_i) - \nabla_\theta \log \pi_\theta(b_i \mid x_i)\big]. \tag{10}$$

The coefficient remains h_B as a pair crosses from disagreement to agreement inside the plateau, decays outside it, and recovers upon re-entry. A valid difficult pair can still remain in the negative tail and receive little update; the loss does not estimate label correctness.

The two shape parameters are B and tail mass ε, which allocates $\varepsilon/2$ per tail and is not an assumed noise rate. A separate margin temperature is redundant because, for $\beta > 0$,

$$\ell_{B,\varepsilon}(\beta z) = \ell_{B/\beta,\varepsilon}(z), \qquad \beta w_{B,\varepsilon}(\beta z) = w_{B/\beta,\varepsilon}(z). \tag{11}$$

Thus B controls both margin scale and plateau extent. We fix (B, ε) and optimize minibatch loss averages; coefficients remain pairwise, without cross-pair normalization. Reference initialization places all pairs at zero margin with coefficient h_B, and weighting differs only after they leave the interval.

5 Theory: gradient allocation and objective descent

An envelope controls scalar coefficients, whereas learning depends on the vector sum of pairwise updates. We connect the two by bounding extreme-pair contributions and projecting active-pair contributions onto a target descent direction. Throughout this section, envelopes are continuous and nonnegative, auxiliary states are fixed during an update, and the update rule is ordinary gradient descent. Proofs appear in Appendix C.

5.1 Negative-tail update bound

Theorem 5.1 (Negative-tail contribution)**.** *At a parameter value θ, let $\mathcal{O}$ be a subset of a batch of size n, with $|\mathcal{O}|/n = \eta$. Suppose $z_i \le -M$ and $\|\nabla_\theta z_i\| \le G$ for $i \in \mathcal{O}$. Its contribution $u_\mathcal{O}$ to the batch descent direction satisfies*

$$u_\mathcal{O} = \frac{1}{n} \sum_{i \in \mathcal{O}} w(z_i) \nabla_\theta z_i, \qquad \|u_\mathcal{O}\| \le \eta G T_-(M). \tag{12}$$

For a step size $\alpha > 0$, the corresponding displacement is at most $\alpha \eta G T_-(M)$.

The bound holds regardless of the directions of the extreme pairs. If these directions oppose a learning objective, $T_-(M)$ limits their potential interference. A common finite gradient bound G makes their aggregate contribution vanish with negative-tail decay.

5.2 Conditional descent under gradient alignment

Let F be a differentiable target objective, for example a population preference risk, with $\nabla F(\theta) \neq 0$. Its unit descent direction is $v = -\nabla F(\theta)/\|\nabla F(\theta)\|$. The analysis uses F to evaluate the update; computing F or v is not part of the training rule. Alignment with v supplies the link between an observed margin update and progress on this target.

Theorem 5.2 (Full-batch conditional descent for Plateau-DPO)**.** *Fix $M > B$ and partition the entire batch into $\mathcal{A} = \{i : |z_i| \le B\}$, $\mathcal{R} = \{i : B < |z_i| < M\}$, $\mathcal{O}_- = \{i : z_i \le -M\}$, and $\mathcal{O}_+ = \{i : z_i \ge M\}$, with fractions $\eta_A, \eta_R, \eta_-, \eta_+$ summing to one and $\eta_A > 0$. Let $\bar{a}_A = |\mathcal{A}|^{-1} \sum_{i \in \mathcal{A}} \langle \nabla z_i, v \rangle$ and suppose $\bar{a}_A \ge a_A > 0$. Assume $\|\nabla z_i\| \le G$ for every pair and that F has an L-Lipschitz gradient along the update segment, with $L > 0$. Set $e_M = \exp[-((M-B)/t_B)^2]$ and define*

$$u = \frac{1}{n} \sum_i w_{B,\varepsilon}(z_i) \nabla z_i, \tag{13}$$

$$q = h_B \big[\eta_A a_A - G\eta_R - G e_M(\eta_- + \eta_+)\big], \qquad H = G h_B \big[\eta_A + \eta_R + e_M(\eta_- + \eta_+)\big]. \tag{14}$$

Then $\langle u, v \rangle \ge q$, $\|u\| \le H$, and

$$F(\theta + \alpha u) \le F(\theta) - \alpha \|\nabla F(\theta)\| q + \frac{L \alpha^2 H^2}{2}. \tag{15}$$

If $q > 0$, every $0 < \alpha < 2\|\nabla F(\theta)\| q/(LH^2)$ gives strict descent.

On the plateau, the active-region projection equals $\eta_A h_B \bar{a}_A$ exactly, so individual active pairs may have negative projections. The residual term accounts for all transition pairs, which may oppose the target direction regardless of label correctness. The tail terms bound either sign of extreme margin. A positive q certifies a descent direction under these local assumptions; H controls the allowable step size. This result concerns ordinary gradient descent, not an adaptive optimizer or general convergence.

If the transition and positive-tail sets are empty, the first-order condition reduces to $\eta_- < a_A/(a_A + Ge_M)$. In the general batch, the transition fraction also enters q, so tail decay alone cannot certify progress. These region fractions are not annotation-error rates. The bound on G is an independent condition on parameter-space margin gradients; bounded scalar coefficients or batch-level clipping do not supply it automatically.

5.3 Optimal interval coverage at fixed tail mass

Theorem 5.3 (Optimal coverage at fixed tail mass)**.** *Fix $B > 0$ and $\varepsilon \in (0, 1)$. Among absolutely continuous nonincreasing losses with endpoints $\ell(-\infty) = 1$, $\ell(+\infty) = 0$ and $\int_{|z|>B} -\ell'(z)\, dz = \varepsilon$, let $C(B) = \operatorname{ess\,inf}_{|z|\le B} -\ell'(z)$. Then*

$$C(B) \le \frac{1-\varepsilon}{2B}. \tag{16}$$

Equality holds exactly when the envelope is constant almost everywhere on $[-B, B]$. Equation 7 attains equality.

Uniform allocation removes the below-average coefficients induced by any nonuniform envelope with the same interval mass. It does not by itself maximize q or downstream performance, which also depend on tail shape, region proportions, and gradient geometry. Symmetry additionally gives $R_\rho = (1 - 2\rho)R_0 + \rho$ under independent uniform flips $\rho < 1/2$, so the unregularized clean and corrupted risks share minimizers (Ghosh et al., 2017). This property is not unique to the plateau and does not imply probability calibration or extend unchanged to a fixed KL/SFT penalty; Appendix C.7 gives the precise statements.

6 Empirical Evaluation

We evaluate aggregate quality under controlled label corruption and diagnose coefficient allocation on clean, hard-clean, and flipped comparisons.

6.1 Experimental setup

Following Dr. DPO, we train Pythia-2.8B from a common SFT initialization on Anthropic HH with a fixed reference. We compare Plateau-DPO with DPO, unbiased rDPO, the DPO variant of γ-PO, and Hölder-DPO using equal training and tuning budgets. Preference accuracy uses the HH test set; win rate uses its single-turn subset against the chosen response.

In the HH experiment, every method receives the same seeded uniform-flip mask at each $\nu \in \{0, 0.1, 0.2, 0.3, 0.4\}$; the Mistral-7B study instead uses instance-dependent flips. Evaluation data remain unchanged, and $\nu = 0$ means no additional flips. Unless stated otherwise, Plateau-DPO uses the default envelope parameters $B = 1$ and $\varepsilon = 0.4$. Validation selects hyperparameters and checkpoints, with test data used only for final evaluation. Results use three policy-training seeds (mean and standard deviation where shown), and generation win rates use paired bootstrap intervals. Appendix A gives both protocols, source configurations, and diagnostic definitions.

6.2 Robustness to controlled label corruption

Table 1 follows the Dr. DPO paired accuracy/win-rate layout (Wu et al., 2025) and compares DPO, rDPO, γ-PO (Sun et al., 2025), Hölder-DPO (Fujisawa et al., 2026), and Plateau-DPO under one protocol.

Model	Method	Preference accuracy (%) ↑ $\nu = 0$	$\nu = 0.1$	$\nu = 0.2$	$\nu = 0.3$	$\nu = 0.4$	Win rate (%) ↑ $\nu = 0$	$\nu = 0.4$
Pythia-2.8B	DPO	63.59	62.27	61.28	58.54	55.23	53.38	49.00
Pythia-2.8B	rDPO	64.40	62.72	62.53	60.56	57.11	51.27	53.54
Pythia-2.8B	γ-PO	64.88	63.20	62.91	61.12	57.64	53.86	54.50
Pythia-2.8B	Hölder-DPO	65.34	63.95	63.50	61.80	58.22	54.49	55.50
Pythia-2.8B	Plateau-DPO	**66.17**	**64.71**	**64.24**	**62.52**	**60.91**	**55.31**	**56.23**

Table 1: Anthropic HH preference acc and win rate under controlled flips. Bold marks best result.

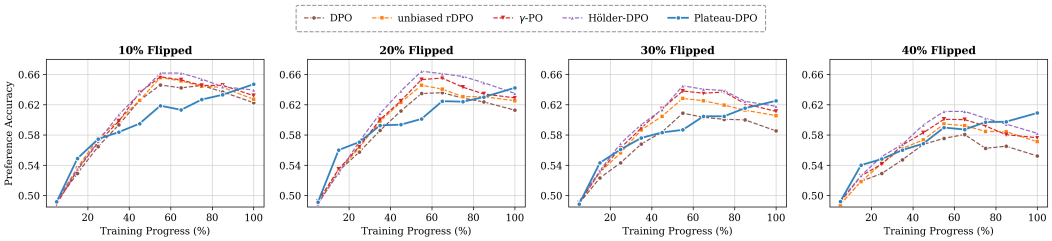


Figure 2: Pythia-2.8B preference-accuracy trajectories under 10–40% flips (three-seed means); final checkpoints correspond to Table 1.

Plateau-DPO has the highest final accuracy at every corrupted setting. Figure 2 shows that baselines often decline after a mid-training peak, whereas Plateau-DPO remains stable or improves, with the clearest separation at 40% flips.

On UltraFeedback/Mistral-7B (Table 2), Plateau-DPO leads the strongest baseline by 2.05/4.06 accuracy/win-rate points at 20% flips and 5.20/6.75 points at 40%. These values are not directly comparable with the HH experiment because the model and data differ; the following diagnostics examine the proposed gradient mechanism.

Method	Mistral-7B 20% flips Acc	WR	40% flips Acc	WR
DPO	71.18	45.89	61.93	35.48
γ-PO	73.12	47.61	63.38	34.45
rDPO	71.81	44.92	63.22	36.51
Hölder-DPO	74.64	56.42	70.21	50.22
Plateau-DPO	**76.69**	**60.48**	**75.41**	**56.97**

Table 2: UltraFeedback/Mistral-7B accuracy and win rate.

6.3 Learning from initially misclassified clean comparisons

The controlled study provides latent labels $y_i^\star$. A clean pair retains that label after corruption, and a hard-clean pair is additionally misclassified by a common frozen predictor before preference optimization. Training and test subsets are fixed before method-specific training; the predictor is independent of the policy/reference initialization, and no such partition is inferred on real preference data. Appendix A gives the formal sets.

Figure 3 connects the accuracy trends to the proposed update geometry. Plateau-DPO retains larger normalized coefficients on hard-clean pairs than Hölder-DPO, while the adverse projection from flipped pairs decreases over training and the net contribution remains positive. The result supports the intended balance between preserving unresolved clean comparisons and suppressing extreme disagreement. These are group-level diagnostics constructed from known latent labels; Plateau-DPO itself neither observes those groups nor infers whether an individual pair is flipped.

6.4 Disentangling interval allocation from tail shape

We compare a centered Gaussian, a flatter generalized Gaussian ($p = 4$), and a plateau with super-Gaussian tails while matching envelope integral, B, ε, data, training budget, tuning budget, and three seeds. Appendix B gives the exact envelopes, their matched scales, and sensitivity to the default parameters.

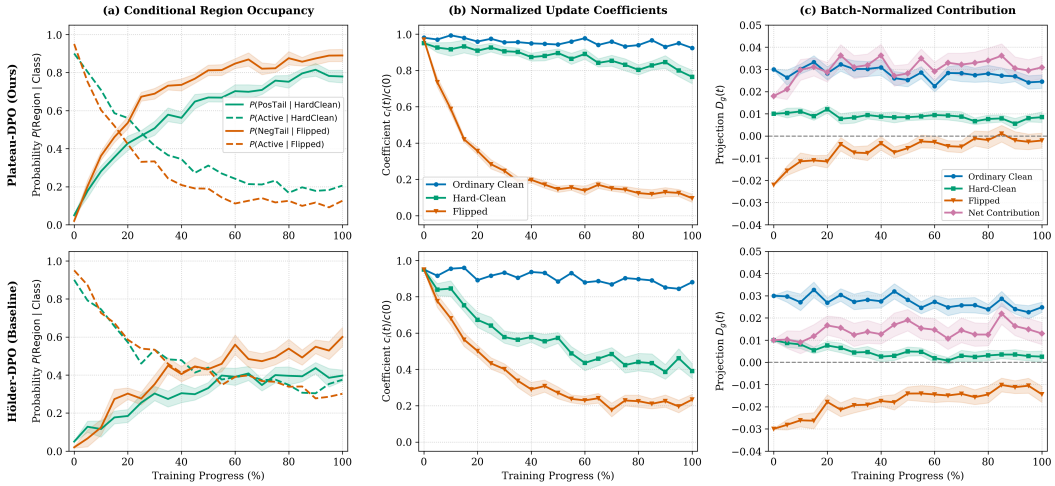


Figure 3: Gradient diagnostics at $\nu = 0.3$ for Plateau-DPO (top) and Hölder-DPO (bottom): region occupancy, zero-margin-normalized coefficients, and signed projections along the clean target direction. Shading is ± 1 standard deviation over 3 seeds; the net curve sums the three disjoint groups.

	Accuracy (%)		Mechanism at $\nu = 0.4$		
Envelope variant	$\nu = .1$	$\nu = .4$	$\bar{c}_{\text{hard}} \uparrow$	$\bar{c}_{\text{flip}} \downarrow$	$D_{\text{net}} \uparrow$
Centered Gaussian	64.05±.41	59.45±.51	0.45±.04	**0.12±.02**	0.33±.03
Generalized Gaussian ($p = 4$)	64.38±.38	60.15±.42	0.68±.03	0.18±.02	0.50±.04
Plateau + super-Gaussian tail	64.65±.32	60.82±.48	0.81±.02	0.16±.03	0.65±.03
Plateau-DPO	**64.71±.35**	**60.91±.45**	**0.82±.02**	0.15±.02	**0.67±.02**

Table 3: Matched shape ablation. Mechanism coefficients are zero-margin-normalized, and D_{net} is the batch-normalized clean-direction projection. Values are mean ± standard deviation over three seeds; bold marks the best mean.

Table 3 shows that Plateau-DPO's advantage over the centered Gaussian grows from 0.66 to 1.46 accuracy points as noise increases. Although the centered Gaussian most strongly suppresses flipped pairs, it also reduces the hard-clean coefficient to 0.45 and yields the lowest net projection. The generalized Gaussian lies between the peaked and plateau envelopes, while the two plateau variants differ by at most 0.09 accuracy points and retain the largest net projections. This ordering associates the gain with broad central coverage rather than the particular choice of Gaussian tails.

7 Conclusion

We studied robust preference optimization through its gradient envelope, separating extreme-margin attenuation from learning-signal allocation over a prescribed finite interval. This view yields Plateau-DPO, a bounded loss with a constant active-region coefficient and smoothly decaying tails. At fixed loss range and tail mass, the plateau maximizes the smallest coefficient within the interval. A batch decomposition further bounds adverse transition and tail contributions and gives conditional one-step descent under explicit assumptions on alignment, gradient norms, and smoothness. These guarantees remain conditional and do not establish unconditional convergence or identify corrupted labels. Across both experimental settings, Plateau-DPO attains the highest preference accuracy and generation win rate at the evaluated corruption levels. On Anthropic HH, its training trajectories also avoid the late-stage accuracy decline observed in the baselines. Diagnostics show stronger weighting of difficult clean pairs and smaller adverse flipped-pair contributions. Matched-shape ablations associate the gains with broad central coverage, while the tested tail variants perform similarly when the plateau is retained. The parameter-sensitivity study further indicates that strong performance persists across a broad neighborhood of the default interval width and tail mass. Interval selection nevertheless matters: valid comparisons outside the chosen interval can still be attenuated.

REFERENCES

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. PMLR, 2024.

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

Xiaoyang Cao, Zelai Xu, Mo Guang, Kaiwen Long, Michiel Bakker, Yu Wang, and Chao Yu. Re-po: Robust enhanced policy optimization as a general framework for llm alignment. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 109970–109991, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b292ef48c211e2d8288146ff045dd61d-Paper-Conference.pdf.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. pp. 8301–8327, 2024a.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024b.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Boryeong Cho, Sumyeong Ahn, and Se-Young Yun. Plc-dpo: Posterior label correction in noisy and ambiguous preference optimization. *arXiv preprint arXiv:2608.30597*, 2026.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Olivier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. 2025:4047–4068, 2025.

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. volume 30, 2017.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Qiang Yue, 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.

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

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Masahiro Fujisawa, Masaki Adachi, and Michael A Osborne. Scalable valuation of human feedback through provably robust model alignment. volume 38, pp. 134187–134235, 2026.

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.

Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. 31(1), 2017.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Kexin Huang, Junkang Wu, Ziqian Chen, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. Larger or smaller reward margins to select preferences for llm alignment? In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.

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

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. pp. 13677–13700, 2023.

Keyi Kong, Xilie Xu, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. Perplexity-aware correction for robust alignment with noisy preferences. volume 37, pp. 28296–28321, 2024.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Hanzhu Chen, Feng Wu, and Jieping Ye. ROPO: Robust preference optimization for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. volume 33, pp. 20331–20342, 2020.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *International conference on learning representations*, volume 2024, pp. 54605–54624, 2024.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pp. 6226–6236. PMLR, 2020.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 124198–124235. Curran Associates, Inc., 2024. doi: 10.52202/079017-3946. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e099c1c9699814af0be873a175361713-Paper-Conference.pdf.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. volume 26, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. volume 33, pp. 17044–17056, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. volume 36, pp. 53728–53741, 2023.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. volume 37, pp. 37100–37137, 2024.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. volume 32, 2019.

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*, volume 38, 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, 2022.

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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

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, 2025. URL https://aclanthology.org/2025.findings-acl.282/.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. 2024:10450–10480, 2024.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 322–330. IEEE, 2019.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. \beta-dpo: Direct preference optimization with dynamic \beta. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 129944–129966. Curran Associates, Inc., 2024. doi: 10.52202/079017-4128. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ea888178abdb6fc233226d12321d754f-Paper-Conference.pdf.

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. In *International Conference on Learning Representations*, volume 2025, pp. 16123–16150, 2025.

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.

Yifan Xu, Xichen Ye, Yifan Chen, and Qiaosheng Zhang. When human preferences flip: An instance-dependent robust loss for rlhf. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38057–38065, 2026.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10935–10950. Curran Associates, Inc., 2023. doi: 10.52202/075280-0482. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/23e6f78bdec844a9f7b6c957de2aae91-Paper-Conference.pdf.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. volume 31, 2018.

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 Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023a. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Lianmin Zheng et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023b.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8328–8340, 2024.

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

A Additional experimental setup

A.1 Source protocols and shared training configuration

The Pythia-2.8B/Anthropic HH experiment follows Appendix D.1 of Dr. DPO (Wu et al., 2025). Anthropic HH contains approximately 170k dialogues. We train the supervised fine-tuned model for one epoch with global batch size 64 and learning rate 5×10^{-7}, then run one additional preference-optimization epoch. The SFT checkpoint initializes both the trainable policy and its frozen reference. The margin temperature is $\beta = 0.1$, and all Pythia experiments use four 80GB A100 GPUs.

The Mistral-7B/UltraFeedback experiment follows the large-model configuration of Xu et al. (2026). Preference pairs with score margin at most 0.5 are removed before artificial corruption, leaving 42.4k training and 1.4k clean test pairs from the respective full sets of 61.1k and 2k. The SFT stage runs for three epochs and the preference stage for one epoch. Both use per-device batch size 16, two gradient-accumulation steps, gradient checkpointing, maximum token length 512, and LoRA with rank 128 and scaling parameter 16. The SFT and preference-stage learning rates are 5×10^{-5} and 10^{-5}, respectively.

Setting	Pythia-2.8B / HH	Mistral-7B / UltraFeedback
SFT epochs / preference epochs	$1/1$	$3/1$
Batch size	64 (global)	16 per device
SFT / preference learning rate	5×10^{-7} / 5×10^{-7}	5×10^{-5} / 10^{-5}
Margin temperature β	0.1	method configuration

Table 4: Training configurations inherited from the two source protocols.

A.2 Controlled corruption protocols

For the HH experiment, we independently reverse each training pair with probability $\nu \in \{0, 0.1, 0.2, 0.3, 0.4\}$, following Dr. DPO. The seed and resulting mask are fixed before optimization and shared by all methods at a given ν. The setting $\nu = 0$ means no additional artificial flips and does not imply that the original data are noise-free. Evaluation pairs and generation prompts are unchanged. On this uniform-flip benchmark, rDPO receives the injected flip rate.

For UltraFeedback, we follow the instance-dependent construction of Xu et al. (2026). A randomly initialized logistic flipping model maps length-based pair features to a flip probability. A pair is flipped when that probability exceeds $\tau = 0.8$; the logistic parameters are fitted so that the fraction above the threshold matches the target rate (20% or 40%). Only length-based features are used for simulated corruption, as in the source protocol. Each generated corruption mask is then frozen and shared by all objectives, so the comparison changes the preference loss rather than the corrupted data.

A.3 Evaluation and reporting

For Pythia-2.8B, preference accuracy is the fraction of HH test pairs for which the policy's implicit reward ranks the chosen response above the rejected response. Win rate uses the single-turn HH test subset: GPT-4 compares a policy response with the dataset's chosen response. Following the reported generation controls, responses are sampled at temperature 0.7 with at most 1024 new tokens.

For Mistral-7B, accuracy is averaged over five evaluations, each drawing 1k pairs from the clean filtered test set. Win rate uses 1k test prompts. GPT-4o compares the trained policy response with the corresponding SFT response using helpfulness, harmlessness, and honesty criteria and returns win, loss, or tie. We compute

$$\mathrm{WR} = \frac{N_{\mathrm{win}} + \frac{1}{2} N_{\mathrm{tie}}}{N_{\mathrm{comparisons}}}. \tag{17}$$

A.4 TUNING, SELECTION, AND UNCERTAINTY

The common search covers learning rate and margin temperature where applicable, the rDPO correction rate, γ-PO target parameters, the Hölder-DPO shape parameter, and (B, ε) for Plateau-DPO. For Plateau-DPO, temperature is represented by the effective width B/β (Eq. 11) rather than an additional search dimension. Hyperparameters and checkpoints are selected on a validation split shared across methods; test sets are reserved for final evaluation.

Training trajectories report checkpoint means over three policy-training seeds. Mechanism plots use the same number of seeds, with shading denoting one standard deviation. Generation win rates additionally use paired bootstrap intervals over evaluation prompts. For the matched-shape ablation, all envelope variants share the initialization, data, corruption mask, training budget, tuning budget, and three seeds; learning rates are tuned separately under that common budget.

A.5 DIAGNOSTIC SUBSETS

The controlled study provides a latent preference label $y_i^\star$ for each pair. Artificial corruption produces the training label $\widetilde{y}_i$; a pair is clean when $\widetilde{y}_i = y_i^\star$. To define difficulty independently of method-specific preference optimization, let $f_0(x_i, a_i, b_i) = \log \pi_0(a_i \mid x_i) - \log \pi_0(b_i \mid x_i)$ be the raw score difference of a common frozen predictor. The predictor and the resulting subsets are fixed before method-specific training. Excluding zero-score ties, define

$$\begin{aligned} \mathcal{H}_{\text{train}} &= \{i \in \mathcal{D}_{\text{train}} : \widetilde{y}_i = y_i^\star,\ y_i^\star f_0(x_i, a_i, b_i) < 0\}, \\ \mathcal{H}_{\text{test}} &= \{i \in \mathcal{D}_{\text{test}} : y_i^\star f_0(x_i, a_i, b_i) < 0\}. \end{aligned} \tag{18}$$

The first set supports training diagnostics and the second defines held-out hard-clean accuracy. Ordinary-clean, hard-clean, and flipped training pairs are disjoint in the mechanism decomposition. The policy's initial reference-relative margin cannot define these sets because it is identically zero when the policy and reference initially coincide. No corresponding clean/hard partition is inferred for naturally collected preference data without known latent labels.

B ENVELOPE-ABLATION SPECIFICATION AND ADDITIONAL ROBUSTNESS

B.1 EXACT MATCHED ENVELOPE FUNCTIONS

All four ablation variants are symmetric probability densities over the margin, so their total coefficient mass is one. They also share an active half-width $B > 0$ and tail mass $\varepsilon \in (0, 1)$:

$$\int_{\mathbb{R}} w(z)\, dz = 1, \qquad \int_{|z|>B} w(z)\, dz = \varepsilon. \tag{19}$$

Let $Q(a, x) = \Gamma(a, x)/\Gamma(a)$ denote the regularized upper incomplete gamma function. Define the centered generalized-Gaussian family

$$g_p(z; s_p) = \frac{p}{2 s_p \Gamma(1/p)} \exp\left[-\left(\frac{|z|}{s_p}\right)^p\right], \qquad s_p = \frac{B}{[Q^{-1}(1/p, \varepsilon)]^{1/p}}. \tag{20}$$

The centered Gaussian uses $w_{\text{CG}} = g_2$; equivalently, $s_2 = B/\operatorname{erfc}^{-1}(\varepsilon)$. The flatter generalized Gaussian uses $w_{\text{GG}} = g_4$. The scale in Eq. 20 makes the mass outside $[-B, B]$ exactly ε, rather than fitting the plotted curves post hoc.

For the two plateau variants, set $h_B = (1 - \varepsilon)/(2B)$ and

$$p_p(z) = h_B \exp\left[-\left(\frac{(|z| - B)_+}{t_p}\right)^p\right], \qquad t_p = \frac{\varepsilon p}{2 h_B \Gamma(1/p)}. \tag{21}$$

The plateau with super-Gaussian tails uses $w_{\text{SG}} = p_4$, whereas Plateau-DPO uses the Gaussian continuation $w_{\text{P}} = p_2$. Thus

$$t_2 = \frac{2B\varepsilon}{(1 - \varepsilon)\sqrt{\pi}}, \qquad t_4 = \frac{4B\varepsilon}{(1 - \varepsilon)\Gamma(1/4)}. \tag{22}$$

For every variant j, the optimized scalar loss is defined without an additional normalization as $\ell_j(z) = \int_z^\infty w_j(u)\,du$. Equations 20–22 imply Eq. 19; hence the comparison changes coefficient shape while holding endpoint range, boundary, and integrated tail mass fixed.

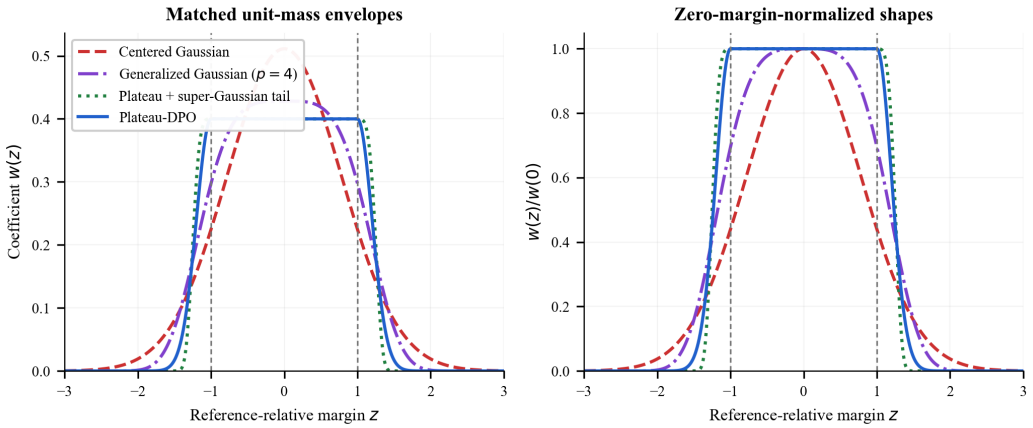


Figure 4: Exact matched ablation envelopes for the illustrative choice $B = 1$ and $\varepsilon = 0.2$. Left: unit-mass coefficients satisfying Eq. 19. Right: the same functions normalized by their zero-margin coefficient. Dashed vertical lines mark $\pm B$; training uses the selected B and ε, not necessarily the illustrative values.

B.2 Supplementary robustness summary

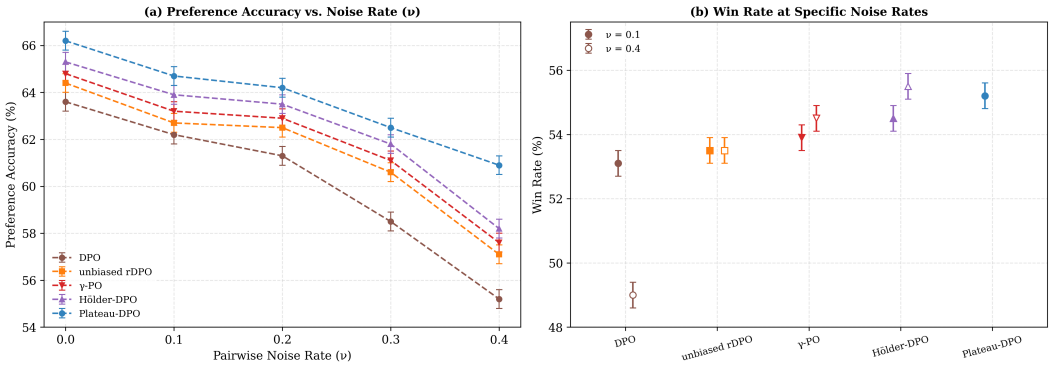


Figure 5: Supplementary Pythia-2.8B/Anthropic HH robustness summary. Left: preference accuracy across pairwise noise rates. Right: win rate at $\nu = 0.1$ and $\nu = 0.4$ for the five compared objectives.

B.3 Sensitivity to the default envelope parameters

Figure 6 varies one envelope parameter at a time on Pythia-2.8B/Anthropic HH while keeping the remaining training configuration fixed. The width sweep fixes $\varepsilon = 0.4$ and varies B relative to the default $B^\star = 1$; the tail-mass sweep fixes $B = 1$ and varies ε around the default $\varepsilon^\star = 0.4$. Each point is the mean over three policy-training seeds, and the shaded region denotes one standard deviation.

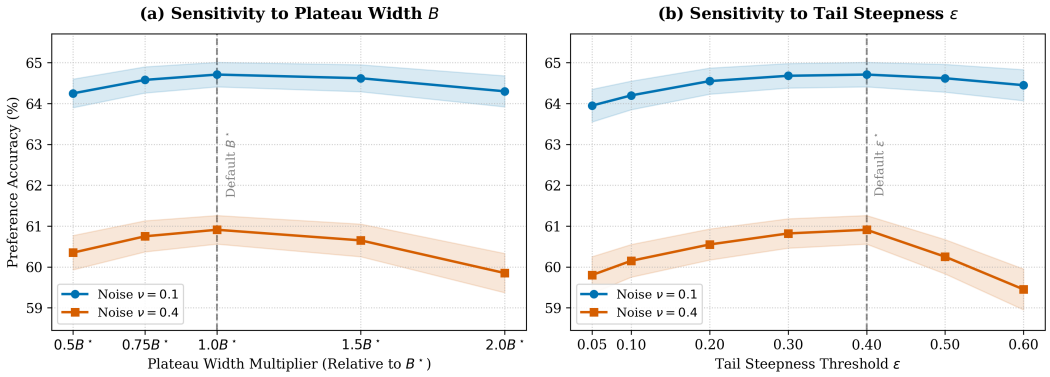


Figure 6: Preference-accuracy sensitivity at low ($\nu = 0.1$) and high ($\nu = 0.4$) label-noise rates. Left: active half-width B relative to $B^\star = 1$, with $\varepsilon = 0.4$. Right: tail mass ε, with $B = 1$. Dashed lines mark the default configuration.

The default gives 64.71% and 60.91% accuracy at $\nu = 0.1$ and $\nu = 0.4$, respectively, matching the main ablation. Widths from $0.75B^\star$ to $1.5B^\star$ remain close to this point, whereas the high-noise result declines more clearly at $2B^\star$. The ε sweep is likewise smooth near the default: moderate tail masses preserve accuracy, while either extreme is less favorable, particularly at $\nu = 0.4$. This pattern is consistent with the intended tradeoff: a narrow plateau or small tail mass can attenuate unresolved comparisons too early, whereas a wide plateau or large tail mass retains more influence from corrupted extremes. The broad neighborhood around the default indicates that the result is not tied to a sharply isolated parameter choice.

C Mathematical proofs and coefficient derivations

This appendix supplies derivations for the negative-tail bound, full-batch conditional descent, optimal interval coverage, baseline coefficients, and the closed form of Plateau-DPO. A restricted two-tail corollary allows unequal tail cutoffs when no transition pairs are present.

C.1 Proof of Theorem 5.1

The tail bound is a batch-level consequence of a pointwise envelope bound. It is stated separately because it does not require active examples to align with a target direction: it only quantifies how much an extreme subset can move the update when pairwise margin gradients are bounded.

Proof. The triangle inequality yields

$$\|u_{\mathcal{O}}\| \le \frac{1}{n} \sum_{i \in \mathcal{O}} w(z_i) \|\nabla z_i\| \le \frac{|\mathcal{O}|}{n} G T_-(M). \tag{23}$$

Multiplication by α gives the displacement bound. Taking $M \to \infty$ is valid with a common finite G. □

Thus, the negative-tail quantity controls the magnitude of the extreme subset's contribution independently of the signs of its individual gradient directions. The bound becomes useful in the descent result below once the remaining active comparisons provide a positive directional contribution.

C.2 Proof of Theorem 5.2

The proof bounds every part of the batch. The plateau makes the active contribution exact after averaging; the transition region and both extreme tails are bounded pessimistically.

Objective	Negative-tail decay	Strict active plateau	Fixed per-pair envelope
DPO	×	×	✓
rDPO	×	×	✓
γ-PO (DPO variant)†	×	×	×
Hölder-DPO	✓	×	✓
Plateau-DPO	✓	✓	✓

Table 5: Update geometry at finite, nondegenerate parameters. A strict active plateau has positive height and width; a fixed per-pair envelope depends only on that pair's margin after fixing hyperparameters and the reference. †The γ-PO target adapts across updates.

Proof. Since $w_{B,\varepsilon} = h_B$ on $\mathcal{A}$, its projected contribution is exactly $\eta_A h_B \bar{a}_A \geq \eta_A h_B a_A$. For $i \in \mathcal{R}$, $w_{B,\varepsilon}(z_i) \leq h_B$; for $i \in \mathcal{O}_- \cup \mathcal{O}_+$, $w_{B,\varepsilon}(z_i) \leq h_B e_M$. Cauchy–Schwarz therefore bounds each projected contribution in these regions below by $-Gh_B$ and $-Gh_B e_M$, respectively. Averaging gives $\langle u, v\rangle \geq q$. The same coefficient bounds and the triangle inequality give $\|u\| \leq H$. Smoothness gives

$$F(\theta + \alpha u) \leq F(\theta) + \alpha\langle \nabla F(\theta), u\rangle + \frac{L\alpha^2}{2}\|u\|^2 \tag{24}$$

$$\leq F(\theta) - \alpha\|\nabla F(\theta)\| q + \frac{L\alpha^2 H^2}{2}. \tag{25}$$

If $q > 0$, then $H \geq \|u\| \geq q > 0$; the stated strict step-size inequality makes the change negative. □

Consequently, $q > 0$ requires the average active contribution to dominate the worst-case transition and two-tail opposition. The step-size restriction controls only the quadratic smoothness term; the claim remains local and conditional.

C.3 Proof of Theorem 5.3

The plateau theorem is a finite-interval allocation result. Once the total mass outside $[-B, B]$ is fixed, the interval contains a fixed amount of coefficient mass, so its smallest attainable value cannot exceed the interval average. The equality case identifies the unique allocation principle needed by the construction.

Proof. Let $w = -\ell' \geq 0$ almost everywhere. Absolute continuity on finite intervals and the endpoint limits give $\int_{\mathbb{R}} w = 1$, hence $\int_{-B}^{B} w = 1 - \varepsilon$. Its essential infimum cannot exceed its interval average. Equality holds if and only if the nonnegative difference $w - C(B)$ integrates to zero, equivalently $w = C(B)$ almost everywhere on the interval. For Eq. 7, the center mass is $2Bh_B = 1 - \varepsilon$ and the two tails have mass $h_B t_B \sqrt{\pi} = \varepsilon$, so equality is attained.

For continuous envelopes and an otherwise unconstrained local pair geometry, the infimum of $w(z)\langle d, v\rangle$ over $|z| \leq B$ and $\langle d, v\rangle \geq a$ is $aC(B)$. It is attained by a minimizing margin and $d = av$. Maximizing this quantity over the continuous comparison class therefore gives $a(1-\varepsilon)/(2B)$. Real datasets need not realize these worst-case pairs. □

The equality condition is the source of the constant central coefficient in Plateau-DPO. It is an allocation statement rather than a claim that every observed comparison has the same reliability; the tail profile can still be changed without altering the interval optimum, provided its integrated mass is preserved.

C.4 Gradient coefficients of baseline losses

The coefficients in Section 3 follow by differentiating the actual training losses. DPO uses $\ell_{\mathrm{DPO}}(r) = L(r)$; the other three objectives are

$$\ell_{\mathrm{rDPO}}(r) = \frac{(1-\rho)L(r) - \rho L(-r)}{1-2\rho}. \tag{26}$$

$$\ell_{\gamma\text{-DPO}}(r_i;\gamma_i) = L(r_i - \gamma_i). \tag{27}$$

$$\ell_{\mathrm{H}}(r) = -(1+\kappa)p^{\kappa} + \kappa p^{\kappa+1}. \tag{28}$$

Equation 28 is the final empirical training objective of Fujisawa et al. (2026, Eq. (12)). Its derivative differs from the contamination weight $p^{\kappa}(1-p)$ in that paper's influence-function analysis, which concerns an estimator and requires additional assumptions.

With $p = \sigma(r)$, $p' = p(1-p)$ and $L'(r) = -(1-p)$ yield the DPO coefficient. For rDPO, differentiating Eq. 26 gives

$$-\ell'_{\mathrm{rDPO}}(r) = \frac{(1-\rho)(1-p) + \rho p}{1-2\rho} = 1 - p + \frac{\rho}{1-2\rho}. \tag{29}$$

The DPO and rDPO coefficients both have derivative $-p(1-p) < 0$ with respect to r, so neither has a constant interval. Taking $p \to 0$ gives the negative-tail limits in Section 3. The subtraction in Eq. 26 is the unbiased correction of Chowdhury et al. (2024). It differs from the positive label-smoothing mixture also called rDPO in parts of Sun et al. (2025).

For a frozen target γ_i, differentiating Eq. 27 gives the corresponding coefficient in Eq. 3. For $r \le -R$ and $\gamma_i \ge 0$,

$$0 \le 1 - g_{\gamma\text{-DPO}}(r;\gamma_i) = \sigma(r-\gamma_i) \le \sigma(-R) \longrightarrow 0. \tag{30}$$

The limit is uniform over nonnegative targets. At a fixed target, $\partial_r g = -\sigma(r-\gamma_i)[1-\sigma(r-\gamma_i)] < 0$. Writing the simplex weights as ω_i to distinguish them from sigmoid probabilities, the targets in Appendix C of Sun et al. (2025) take the form $\gamma_i = m_n \omega_i \gamma_0$, with $\gamma_0 > 0$ and m_n the number of margins. Current and historical margins determine these weights. The conditional derivative holds the resulting target fixed; differentiating through target construction would add chain-rule terms.

For Hölder-DPO's empirical loss,

$$-\ell'_{\mathrm{H}}(r) = \big[(1+\kappa)\kappa p^{\kappa-1} - \kappa(1+\kappa)p^{\kappa}\big]p(1-p) = \kappa(1+\kappa)p^{\kappa}(1-p)^2. \tag{31}$$

Its log derivative is $\partial_r \log g_{\mathrm{H}} = \kappa - (\kappa+2)p$. It changes sign once, at $p = \kappa/(\kappa+2)$, establishing a single peak and excluding any constant interval of positive length. The factors p^{κ} and $(1-p)^2$ give decay in the negative and positive tails, respectively. Adding 1 to the loss sets its endpoints to 1 and 0 without changing the coefficient; its total gradient mass is therefore one. For each fixed temperature and shape parameter, matching the interval and integrated tail mass makes its minimum coverage strictly smaller than the plateau optimum, by the equality condition of Theorem 5.3. The size of this difference depends on those parameters.

C.5 Closed-form Plateau-DPO loss

The envelope in Eq. 7 specifies the update coefficient, whereas optimization requires the corresponding scalar loss. Integrating from the current margin to $+\infty$ gives the following closed form:

$$\ell_{B,\varepsilon}(z) = \begin{cases} 1 - \dfrac{\varepsilon}{2}\operatorname{erfc}\left(\dfrac{-z-B}{t_B}\right), & z < -B, \\ \dfrac{1}{2} - h_B z, & |z| \le B, \\ \dfrac{\varepsilon}{2}\operatorname{erfc}\left(\dfrac{z-B}{t_B}\right), & z > B. \end{cases} \tag{32}$$

The three branches correspond directly to the intended learning regimes. For $|z| \le B$, the loss is affine with derivative $-h_B$, so every comparison in the active interval receives the same scalar coefficient even as its margin crosses zero. For $z < -B$, the loss approaches its upper bound 1 and its derivative vanishes as the disagreement becomes more extreme. For $z > B$, the loss approaches

0 and its derivative likewise vanishes as the observed preference becomes strongly satisfied. These tail statements describe attenuation by the current margin and do not identify whether an individual label is correct.

The values at the center and the two joins are

$$\ell(-B) = 1 - \frac{\varepsilon}{2}, \qquad \ell(0) = \frac{1}{2}, \qquad \ell(B) = \frac{\varepsilon}{2}. \tag{33}$$

Hence the active interval accounts for loss variation $1-\varepsilon$, while each tail accounts for the remaining variation $\varepsilon/2$. At both joins, the derivative equals $-h_B$ and the second derivative equals zero, matching the affine center to the Gaussian tails through second order. The closed form therefore preserves the plateau coefficient without introducing a gradient discontinuity at $\pm B$.

The parameters retain the allocation interpretation from Eq. 6. The half-width B determines how far the constant-gradient region extends, and ε determines how much of the unit loss range remains for the two saturating tails. The complementary error function is the cumulative form of the Gaussian continuation, so Eq. 32 can be evaluated directly without numerical integration. Differentiating it recovers $-\ell'(z) = w_{B,\varepsilon}(z)$ and, on the negative tail, the bound in Eq. 8.

C.6 Restricted two-tail case with unequal cutoffs

The main theorem already handles both tails and the transition region with a common cutoff. When no transition pairs are present, a stronger per-pair active alignment assumption permits distinct tail cutoffs. Define

$$T_+(M) = \sup_{z \geq M} w(z), \qquad M > B. \tag{34}$$

The following result treats tail membership only as a margin condition, rather than as an identification of label noise.

Theorem C.1 (Conditional descent with two extreme tails)**.** *Partition a batch into $\mathcal{A}$, $\mathcal{O}_-$, and $\mathcal{O}_+$, with fractions η_A, η_-, and η_+ satisfying $\eta_A + \eta_- + \eta_+ = 1$. Assume $|z_i| \leq B$ and $\langle \nabla z_i, v\rangle \geq a > 0$ for $i \in \mathcal{A}$, $z_i \leq -M_-$ for $i \in \mathcal{O}_-$, and $z_i \geq M_+$ for $i \in \mathcal{O}_+$, where $M_-, M_+ > B$. Suppose $\|\nabla z_i\| \leq G$ for every pair and that F has an L-Lipschitz gradient along the update segment, with $L > 0$. With $u = n^{-1}\sum_i w(z_i)\nabla z_i$ and $W_A = \max_{|z|\leq B} w(z)$, define*

$$q_\pm = \eta_A a C(B) - G\big[\eta_- T_-(M_-) + \eta_+ T_+(M_+)\big], \tag{35}$$

$$H_\pm = G\big[\eta_A W_A + \eta_- T_-(M_-) + \eta_+ T_+(M_+)\big]. \tag{36}$$

Then $\langle u, v\rangle \geq q_\pm$, $\|u\| \leq H_\pm$, and

$$F(\theta + \alpha u) \leq F(\theta) - \alpha \|\nabla F(\theta)\| q_\pm + \frac{L\alpha^2 H_\pm^2}{2}. \tag{37}$$

If $q_\pm > 0$, every $0 < \alpha < 2\|\nabla F(\theta)\| q_\pm / (L H_\pm^2)$ gives strict descent.

Proof. For $i \in \mathcal{A}$, $w(z_i)\langle \nabla z_i, v\rangle \geq aC(B)$. For $i \in \mathcal{O}_-$, Cauchy–Schwarz gives $w(z_i)\langle \nabla z_i, v\rangle \geq -GT_-(M_-)$; the same argument gives $w(z_i)\langle \nabla z_i, v\rangle \geq -GT_+(M_+)$ for $i \in \mathcal{O}_+$. Averaging these three bounds yields $\langle u, v\rangle \geq q_\pm$. The triangle inequality and the definitions of W_A, $T_-(M_-)$, and $T_+(M_+)$ give $\|u\| \leq H_\pm$. Applying L-smoothness to $F(\theta + \alpha u)$ then gives Eq. 37. Since $q_\pm \leq \langle u, v\rangle \leq \|u\| \leq H_\pm$, $q_\pm > 0$ implies $H_\pm > 0$, and the stated step-size condition makes the right-hand-side increment strictly negative. □

For Plateau-DPO, symmetry of the envelope gives $T_+(M) = T_-(M)$. Writing $e_\pm = \exp[-((M_\pm - B)/t_B)^2]$, Eq. 36 becomes

$$q_\pm = h_B\big[\eta_A a - G(\eta_- e_- + \eta_+ e_+)\big], \qquad H_\pm = G h_B\big[\eta_A + \eta_- e_- + \eta_+ e_+\big]. \tag{38}$$

When both cutoffs equal M, let $\eta_T = \eta_- + \eta_+$ and $e_M = \exp[-((M-B)/t_B)^2]$. The first-order condition then reduces to $\eta_T < a/(a + G e_M)$. When the cutoffs coincide, this is the $\eta_R = 0$ case of Theorem 5.2, with the stronger per-pair condition implying its average-alignment condition. The special case $\eta_+ = 0$ gives a one-sided tail bound.

C.7 SMOOTHNESS, CURVATURE, AND SYMMETRIC NOISE

The construction leaves two properties to verify beyond interval coverage and tail decay. First, the Gaussian continuation should join the plateau without introducing a kink that destabilizes gradient-based optimization. Second, the symmetry induced by the even envelope should translate into a statistical statement under the uniform label-flip model. The next lemma establishes both properties, while keeping their roles distinct: curvature concerns the local geometry of the loss, whereas the risk identity concerns its population minimizers.

Lemma C.2 (Smoothness and risk under uniform label flips)**.** *Equation 7 defines a C^2 loss with $\ell(z) + \ell(-z) = 1$ and*

$$\sup_z |\ell''(z)| = \frac{\sqrt{2\pi/e}(1-\varepsilon)^2}{4B^2\varepsilon}. \tag{39}$$

Let $R_0(\theta) = \mathbb{E}[\ell(ys_\theta)]$. If $\widetilde{y}$ is obtained by flipping y independently of the example with probability $0 \le \rho < 1/2$, define $R_\rho(\theta) = \mathbb{E}[\ell(\widetilde{y}s_\theta)]$. Then

$$R_\rho(\theta) = (1-2\rho)R_0(\theta) + \rho. \tag{40}$$

Whenever minimizers exist, their sets coincide.

Proof. The envelope is even, has unit mass, and is continuously differentiable with derivative zero at $\pm B$. Integration gives symmetry and C^2 regularity. The tail derivative magnitude is $2h_B x e^{-x^2}/t_B$, where $x = (|z| - B)/t_B$. Its maximum occurs at $x = 1/\sqrt{2}$, giving $\sqrt{2/e}\, h_B/t_B$ and Eq. 39. Finally,

$$\mathbb{E}[\ell(\widetilde{y}s) \mid y, s] = (1-\rho)\ell(ys) + \rho\ell(-ys) = (1-2\rho)\ell(ys) + \rho. \tag{41}$$

Taking expectations proves the identity. Since $1 - 2\rho > 0$, the affine transformation preserves the ordering of all finite surrogate risks and hence their minimizers. □

The curvature expression quantifies the optimization price of concentrating gradient mass inside the active interval. At fixed B, decreasing ε raises the plateau height and shortens the Gaussian transition, but the maximum curvature grows as $(1-\varepsilon)^2/\varepsilon$. In particular, the limit $\varepsilon \to 0$ approaches the nonsmooth ramp: tail mass vanishes, while the transition curvature diverges. At fixed ε, increasing B reduces the curvature as B^{-2}, together with the reduction in plateau height already described in Eq. 6. This margin-space curvature bound does not by itself imply smoothness of the parameter-space objective, which also depends on the derivatives of s_θ; it instead isolates the contribution introduced by the surrogate loss.

The risk identity supplies a separate statistical interpretation. Because uniform flips transform R_0 by a positive affine map, Plateau-DPO does not need the flip probability ρ to preserve the clean-risk minimizers at the population level. This invariance follows from loss symmetry rather than from the plateau shape: the plateau controls where gradient mass is allocated, while symmetry controls how opposite labels combine in expectation. The identity requires flips to be independent of the example and symmetric between labels, and therefore does not extend directly to instance-dependent or asymmetric corruption. It also concerns population surrogate risks and does not assert equality of finite-sample optimization trajectories.

Symmetry does not calibrate preference probabilities. For a freely chosen score s and preference probability p, the conditional risk is $r_p(s) = (1-p) + (2p-1)\ell(s)$; for $p > 1/2$, its infimum is approached as $s \to +\infty$ regardless of the magnitude of p, with the opposite direction for $p < 1/2$. Moreover, adding a fixed KL or SFT penalty breaks the clean/noisy affine equivalence: $(1-2\rho)R_0 + \rho + \lambda\Omega$ is not a positive affine transform of $R_0 + \lambda\Omega$ at the same λ.

D USE OF LLM

Large language models (LLMs) were used as assistive tools for text editing and improving the clarity of exposition. They were not involved in the design of algorithms, implementation, or experimental analysis. All technical content, theoretical results, and experimental findings were produced and verified by the authors. We take full responsibility for the content of this paper. We only used LLMs for language polishing.