DISENTANGLING OPTIMIZATION SCALE FROM PREFERENCE SCALE IN DPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) uses β both to control the sharpness of the preference likelihood and, through an explicit gradient prefactor, to rescale optimization. Under fixed-learning-rate, finite-budget training, this coupling can make policy movement non-monotone in β: our sweeps exhibit near-zero KL movement at very small β (validation KL below 0.05 nats per token), a peak at an intermediate value, and reduced movement at larger values. A scalar stopping-tolerance model explains this qualitative pattern. We also show that standard DPO losses are difficult to compare across β, because nearly identical loss curves can correspond to substantially different margins and KL divergences. We propose a normalized centered-softplus objective that is a positive affine transformation of DPO for every fixed $\beta > 0$, preserving its ordering and any attained minimizers while removing the explicit β gradient prefactor. This parameterization separates likelihood sharpness from first-order update scale and yields more informative loss trajectories in our experiments. It also has a continuous $\beta \to 0$ limit corresponding to a linear preference-margin objective.

1 INTRODUCTION

Direct Preference Optimization (DPO) is a post-training alignment method for large language models that directly optimizes pairwise human preferences without explicit reward-model training, and it is widely used in instruction tuning and RLHF-style pipelines (Rafailov et al., 2023). The work most closely related to our question studies the sensitivity and dynamics of DPO itself: β-DPO dynamically adapts β to data quality and instance difficulty (Wu et al., 2024), Balanced-DPO reweights preferred- and rejected-response gradients to mitigate gradient imbalance (Ma et al., 2025), and length-focused analyses separate preference quality from verbosity effects in DPO-trained policies (Park et al., 2024).

Across this literature, β is predominantly treated as a control parameter for the preference–regularization trade-off. Under this common KL-control interpretation, smaller β values are often expected to permit larger policy deviation from the reference model (Rafailov et al., 2023; Wu et al., 2024). DPO performance is also empirically known to be highly sensitive to the choice of β (Wu et al., 2024). Prior work has not isolated the direct optimizer-scale role of β: in standard DPO-style objectives, β not only changes the effective inverse preference-noise scale or regularization behavior, but also multiplies the marginal gradient and therefore changes the effective update scale.

Decreasing β proportionally reduces the explicit gradient prefactor, and therefore the initial gradient magnitude at $\Delta = 0$. Later in training the sigmoid factor and the evolving Jacobian also affect the total gradient magnitude. Under SGD, the explicit prefactor is equivalent to lowering the learning rate and complicates hyperparameter tuning. We first demonstrate the effect in standard DPO training: with very small β, gradients stay weak, KL changes little, and the policy remains close to the reference model rather than moving farther toward the preferred responses. The same entanglement also undermines the loss as a monitoring signal: the standard DPO loss depends on β and the preference margin only through their product, so its value can remain nearly flat during training, or coincide across runs whose policies differ several-fold in KL.

Our main contributions are as follows:

- We show that β explicitly multiplies the DPO marginal gradient and therefore changes the effective learning speed. Under a scalar stopping-tolerance model, this coupling produces

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.

a small-β dead zone and a peak margin near $\beta_{\text{peak}} \approx 4.6\,\tau$; our fixed-learning-rate, finite-budget sweeps exhibit the same qualitative pattern (Section 3.3).

- We propose a normalized centered-softplus loss (Section 4) that is a positive affine transformation of DPO for every fixed $\beta > 0$. It preserves the objective ordering and any attained minimizers while removing the explicit β prefactor from the marginal gradient. It also has a well-defined $\beta = 0$ limit corresponding to a linear preference-margin objective.
- We show that standard DPO loss values are difficult to compare across β (Section 3.4): runs with nearly identical loss curves can differ substantially in unscaled margin, KL divergence, validation NLL, and downstream behavior. In the reported sweeps, the normalized loss remains more sensitive to the unscaled margin and separates these runs.

We emphasize that we do not propose a different estimator for fixed $\beta > 0$: learning-rate retuning can compensate for the affine loss rescaling, and adaptive optimizers can partly mask it. Our contribution is a parameterization that makes likelihood sharpness and first-order update scale independently tunable, provides a more informative monitoring signal in the reported sweeps, and admits a continuous $\beta = 0$ endpoint. Finally, the explicit β prefactor is not specific to classical DPO: several DPO-family losses surveyed in Appendix A carry the same multiplicative factor in the marginal gradient. All reported training runs, however, use standard DPO or the corresponding normalized centered-softplus objective.

2 Background: KL-Regularized RLHF and DPO

Consider a reference policy π_{ref} and a reward function $r(x, y)$. The standard KL-regularized RLHF objective is

$$\max_{\pi_\theta} \; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\big[r(x,y)\big] - \lambda\, \mathbb{E}_{x\sim\mathcal{D}}\mathrm{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big), \tag{1}$$

where $\lambda > 0$ is the KL-regularization coefficient controlling the trade-off between reward maximization and staying close to the reference model. A common pipeline first fits a reward model on human preference data under the Bradley–Terry model (Bradley & Terry, 1952) and then optimizes Equation (1) with PPO or a related policy-gradient method. Rafailov et al. (2023) show that this two-stage procedure can be collapsed into a single objective over policy log-probabilities. The optimal policy of Equation (1) induces an implicit reward $r(x, y) = \lambda \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + \lambda \log Z(x)$, where $Z(x)$ is the prompt-dependent partition function. If the Bradley–Terry likelihood has pairwise annotator-noise scale s_a (Sun et al., 2025), substitution cancels $Z(x)$ for a fixed prompt and yields the effective DPO coefficient $\beta = \lambda/s_a$ in Equation (5); the conventional unit-noise formulation sets $s_a = 1$. Thus no separate reward model needs to be trained. The DPO likelihood identifies this ratio rather than λ and s_a separately (Appendix F). Used as a training loss, β therefore sets both the sharpness of $\sigma(\beta\Delta)$ and the scale of the gradient. The closed-form optimal policy and the full derivation are recalled in Appendix F.

The original DPO formulation connects preference optimization to KL-constrained RLHF and establishes the basic log-ratio objective (Rafailov et al., 2023). Subsequent work has refined this framework through alternative optimization principles (Azar et al., 2024; Xu et al., 2024a; Ethayarajh et al., 2024), reference-free or length-normalized margins (Meng et al., 2024), token-level credit (Zeng et al., 2024), robustness to noisy or soft labels (Mitchell, 2023; Furuta et al., 2024; Liang et al., 2025), adaptive margins (Wu et al., 2025), distributional objectives (Melnyk et al., 2024), and the choice of reference policy itself (Liu et al., 2025). Appendix A collects their objectives and marginal gradients; all except IPO retain the explicit β prefactor studied here.

Following the standard DPO setup, let (x, y_w, y_l) denote a prompt, a preferred response, and a rejected response. Define the policy log-probability gap

$$\Delta_\theta = \log \pi_\theta(y_w \mid x) - \log \pi_\theta(y_l \mid x), \tag{2}$$

and analogously for the reference model,

$$\Delta_{\text{ref}} = \log \pi_{\text{ref}}(y_w \mid x) - \log \pi_{\text{ref}}(y_l \mid x). \tag{3}$$

We use the margin variable

$$\Delta = \Delta_\theta - \Delta_{\text{ref}}. \tag{4}$$

The classical DPO objective is

$$\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) = -\log\sigma(\beta\Delta), \tag{5}$$

equivalently $\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) = \log(1+\exp(-\beta\Delta))$, where $\sigma(\cdot)$ denotes the sigmoid function.

Most DPO variants share an explicit multiplicative β factor in the marginal gradient, so reducing β similarly suppresses update magnitude at the level of the loss derivative (Appendix A). We analyze this coupling empirically only for standard DPO (Section 3.2).

3 Gradient-Scale Coupling and Loss Incomparability in Standard DPO

We isolate two related scale effects in the standard DPO parameterization. First, β multiplies the marginal gradient and therefore rescales the effective update. Second, the pairwise loss depends on β and the preference margin only through their product, so similar loss values across β need not represent similar policies.

3.1 When Smaller β Weakens DPO Updates

Figure 1 is the empirical starting point of our analysis. It shows a standard-DPO β sweep on HelpSteer3 with Qwen3-4B-Instruct-2507 (SGD, learning rate 2×10^{-4}, batch size 24), comparing $\beta\in\{5\times10^{-4}, 10^{-3}, 2\times10^{-3}, 10^{-2}\}$. Two observations follow from this sweep. First, when β is made very small, optimization degrades rather than producing a larger policy deviation: the observed gradients are weak, the validation KL remains almost unchanged, and the trained policy stays close to the reference model. Second, the achieved KL is not ordered monotonically in β even away from that regime: final validation KL peaks at the intermediate value $\beta = 2\times10^{-3}$ (0.83 per token) while $\beta = 10^{-2}$ reaches only 0.54, and after epoch 3 the $\beta = 0.002$ run overtakes $\beta = 0.01$ in both logged mean-absolute gradient magnitude and KL. Both are hard to reconcile with the KL-control reading of β, under which a smaller β should permit a larger deviation from the reference model. In KL-regularized RLHF (Equation (1)), larger λ strengthens an explicit KL penalty, and DPO inherits the ratio $\beta = \lambda/s_a$ after eliminating that term (Appendix F); in this finite-horizon sweep, however, the achieved KL is not a monotone function of β in either direction. Section 3.2 identifies the explicit β factor in the marginal gradient, and Section 3.3 uses a stopping-tolerance model to explain the observed pattern. Table 1 already makes the linear suppression visible at the very first optimization step of the same sweep: the measured initial gradient magnitudes G_0 scale linearly with β, with G_0/β constant to within roughly 2% across a 20× range of β. At initialization $\Delta = 0$, so the marginal gradient is exactly $-\beta/2$ (Section 3.2) and G_0/β is predicted to be constant across β. Appendix K relates G_0 to the Euclidean loss gradient and explains why a minibatch log identifies an effective batch-averaged Jacobian scale rather than a typical per-pair norm.

Table 1: Initial gradient magnitudes G_0 (mean absolute per-parameter gradient at the first logged step of epoch 1, averaged over five seeds) from the HelpSteer3 SGD runs in Figure 1.

β	5×10^{-4}	10^{-3}	2×10^{-3}	10^{-2}
G_0	4.12×10^{-6}	8.37×10^{-6}	1.64×10^{-5}	8.28×10^{-5}
G_0/β	8.23×10^{-3}	8.37×10^{-3}	8.22×10^{-3}	8.28×10^{-3}

Neither effect is specific to this dataset or optimizer. Adaptive optimizers such as AdamW can partly obscure the β-dependent scaling of gradient magnitude. They cancel a common multiplicative factor only while the second-moment term dominates the numerical stabilizer ϵ; at sufficiently small β, ϵ dominates the denominator and the update again scales with β, as with SGD (Appendix B). Moment estimates otherwise stabilize step sizes, so the coupling is most apparent when β is already small, exactly where the explicit β prefactor in $\partial\mathcal{L}/\partial\Delta$ suppresses gradients most strongly. Appendix C (Figure C1) documents the same degradation on UltraFeedback with TRL `DPOTrainer` under AdamW at very small β; the same pattern appears on PKU-processed HH-RLHF with Mamba-2 (Figure C2). Compared with those AdamW runs, plain SGD makes the effect easier to see, and it already appears at larger β. An alternative HelpSteer3 SGD configuration at learning rate 5×10^{-4} (Figure C4) and HH-RLHF trajectories are reported in Appendix C.

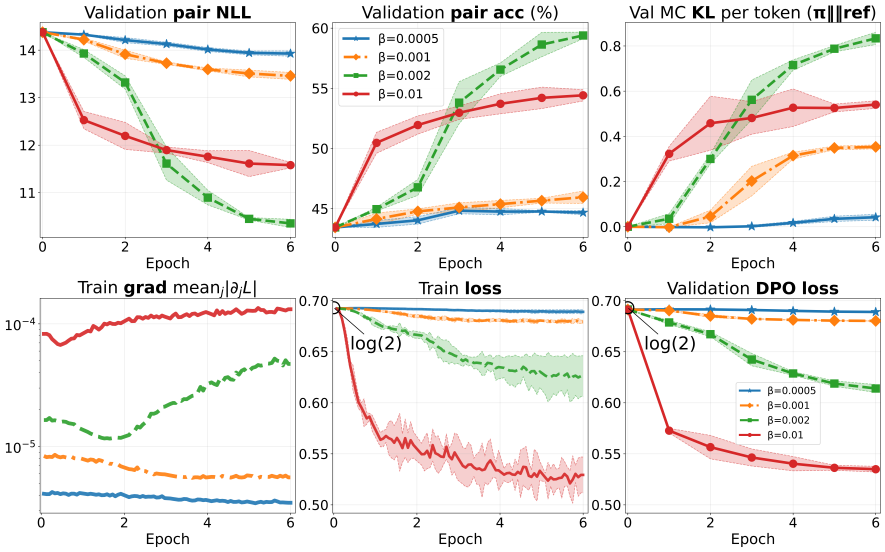

Figure 1: Training and validation dynamics of our DPO implementation on HelpSteer3 with Qwen3-4B-Instruct-2507 (SGD, $\mathrm{lr} = 2 \times 10^{-4}$, batch size 24). Curves show the mean over seeds 43–47; one seed stopped after epoch 4, so epochs 5–6 average four seeds. Shaded bands denote ± 2 standard deviations (no band is drawn for the gradient panel). Final KL is non-monotone in β; see Section 3.3 and Appendix G.

3.2 Mechanism of Gradient-Scale Coupling

Using the margin Δ and classical DPO loss from Section 2 (Equation (5)), the gradient with respect to Δ is

$$\frac{\partial \mathcal{L}_{\mathrm{DPO}}}{\partial \Delta} = -\beta \frac{\exp(-\beta\Delta)}{1 + \exp(-\beta\Delta)} = -\beta\, \sigma(-\beta\Delta). \tag{6}$$

Figures C1 and 1 mirror the linear β prefactor in Equation (6): on HelpSteer3, logged mean-absolute gradient magnitudes scale approximately with β, while the UltraFeedback TRL run in Appendix C uses a very small β under AdamW and maintains comparatively small Euclidean gradient norms throughout training. This linear scaling is exact at initialization: the policy coincides with the reference model, so $\Delta = 0$, $\partial \mathcal{L}_{\mathrm{DPO}}/\partial \Delta = -\beta/2$, and the full parameter gradient $-(\beta/2)\,\nabla_\theta \Delta_\theta$ has norm proportional to β for a fixed model state, i.e., G_0/β should be constant across β. Table 1 quantifies this prediction; measurement of the logged gradient magnitudes is discussed in Appendix K. The classical form therefore carries an explicit multiplicative β factor in the gradient magnitude, which linearly suppresses updates as β decreases. This coupling motivates a normalized reformulation: if β is intended to control the preference–regularization trade-off, it should not also directly rescale the effective update size. As summarized in Table A1 (Appendix A), most DPO-family objectives share this explicit β prefactor in the marginal gradient. That table is an analytical survey: we do not train SimPO, KTO, or the other listed variants.

The same linear β-dependent gradient-scale coupling is not specific to logistic Bradley–Terry (Bradley & Terry, 1952): it holds for any pairwise loss $-\log F(\beta\Delta)$ with a full-line noise CDF F (Appendix D, Eq. (D.3)), including the Thurstone case $F = \Phi$ (Sun et al., 2025); the corresponding marginal gradient always contains an explicit multiplicative β prefactor (Eq. (D.4)).

3.3 Dead Zone, Peak, and Saturation

The explicit β prefactor also enters a simple model of when training effectively stops. We model effective convergence with a scalar residual tolerance $\tau > 0$: updates on a preference pair become negligible once the magnitude of the marginal gradient drops below a noise- and optimizer-scale threshold. Operationally, one useful reading of τ is the noise-dominated regime in which the directed windowed trend of the margin is no larger than its seed-to-seed scatter (Appendix M; Figure M1 gives a five-seed Orca-DPO example). This approximation describes a pairwise margin, not a universal

relationship between that margin and policy-level KL. Operationally, this marginal-gradient tolerance may be related to an effective noise floor ε in per-step margin increments through

$$\tau(\eta_{\mathrm{ref}}) = \frac{\varepsilon}{\eta_{\mathrm{ref}} g_0^2}, \qquad \beta_{\mathrm{left}}(\eta_{\mathrm{ref}}) = 2\tau(\eta_{\mathrm{ref}}) \approx \frac{2\varepsilon}{\eta_{\mathrm{ref}} g_0^2}, \tag{7}$$

where g_0 is the initial margin Jacobian norm. Thus the dead zone is a region in the joint (η, β) plane rather than a fixed interval of β. If ε and g_0 remain approximately fixed across a learning-rate sweep, its left boundary scales inversely with the reference learning rate; evolution of the Jacobian or of the effective noise floor can bend this boundary. The deterministic effect of a finite learning-rate budget is treated separately in Appendix L. By Equation (6), $|\partial \mathcal{L}_{\mathrm{DPO}}/\partial \Delta| = \beta\, \sigma(-\beta\Delta)$, and the stopping condition $\beta\, \sigma(-\beta\Delta) = \tau$ yields

$$\Delta^*_{\mathrm{DPO}} = \frac{1}{\beta} \log\left(\frac{\beta}{\tau} - 1\right), \qquad \beta > 2\tau. \tag{8}$$

Three regimes follow (Appendix G gives the full derivation and the normalized-objective counterpart).

- *Dead zone* ($\beta \le 2\tau$): the marginal gradient is largest at initialization, where $\Delta = 0$ and $|\partial \mathcal{L}_{\mathrm{DPO}}/\partial \Delta| = \beta/2$. If $\beta/2 \le \tau$, the gradient is below the model's effective tolerance from the start.
- *Non-monotone peak* ($\beta > 2\tau$): maximizing Equation (8) over β gives

  $$\beta_{\mathrm{peak}} \approx 4.6\,\tau, \tag{9}$$

  after which Δ^*_{DPO} decays as $\log(\beta/\tau)/\beta$. The equilibrium product is $\beta\, \Delta^*_{\mathrm{DPO}} = \log(\beta/\tau - 1)$. It grows only logarithmically in β, so raising β barely increases the logit-margin seen by the sigmoid, while the log-ratio gap $\Delta^*_{\mathrm{DPO}} = \beta^{-1} \log(\beta/\tau - 1)$ itself shrinks.
- *Right-hand saturation*: the modeled margin decays as $\log(\beta/\tau)/\beta$. In our sweeps, mean margin and KL decline correspondingly, although the scalar model does not imply a universal margin–KL relationship. Appendix L derives an order-of-magnitude finite-horizon scale.

Figure 2 compares the observed dense β-sweep with Equation (8) under the same HelpSteer3 SGD protocol as Figure 1. Validation pair NLL (with $\beta = 1$; Appendix G.1), per-token KL, and mean margin $\overline{\Delta}$ are qualitatively consistent with the predicted shape: a left dead zone, an intermediate peak, and a decaying right slope, with NLL an inverted copy of the saturation curve. Appendix G.1 obtains two descriptive estimates of τ: from Equation (9), $\tau = \beta_{\mathrm{peak}}/4.6 \approx 5.99 \times 10^{-4}$, and from a shape+scale fit of $\overline{\Delta}(\beta)$ to Equation (8) on $\beta > 2\tau$, $\tau \approx 6.37 \times 10^{-4}$. The two overlays in Figure 2 are almost indistinguishable. The $\beta = 0.002$ vs. 0.01 ordering in Figure 1 is consistent with the threshold model and need not indicate an instability.

3.4 The Standard DPO Loss Is Not Comparable Across β

The standard DPO loss depends on β and the preference margin only through $u = \beta\Delta$. Similar loss values can therefore correspond to different unscaled margins and substantially different policies. Moreover, as $\beta \to 0$, the loss approaches $\log 2$ over an increasingly broad range of Δ, making within-run progress harder to discern. Appendix I gives the loss geometry and cross-β ordering in detail.

When β and the learning rate are traded off, the problem becomes especially clear. Under a scalar frozen-Jacobian approximation, matching $\eta\beta$ predicts similar trajectories in u; several sweeps exhibit this behavior approximately, although it is not guaranteed for general minibatch training (Appendix I). The resulting standard-loss curves can nearly coincide while the unscaled margins, KL divergences, and downstream behavior differ substantially. Figure 4 shows the effect on UltraFeedback: the standard-loss pair $(\beta, \mathrm{lr}) = (0.01,\, 10^{-3})$ and $(0.02,\, 5 \times 10^{-4})$ has nearly identical validation DPO losses but a KL ratio of about 3, with validation NLL and AlpacaEval 2 length-controlled win rates (about 14 percentage points) also separated (Tables I2 and J1). Appendix I reports the same construction on HelpSteer3 (KL ratio ≈ 10), Ministral-3B, and HH-RLHF. Consequently, raw standard-DPO loss alone is not a reliable criterion for comparing or selecting runs across β. The normalized loss of Section 4 remains more sensitive to the unscaled margin and separates the examined runs, without constituting a universal cross-β model-selection metric.

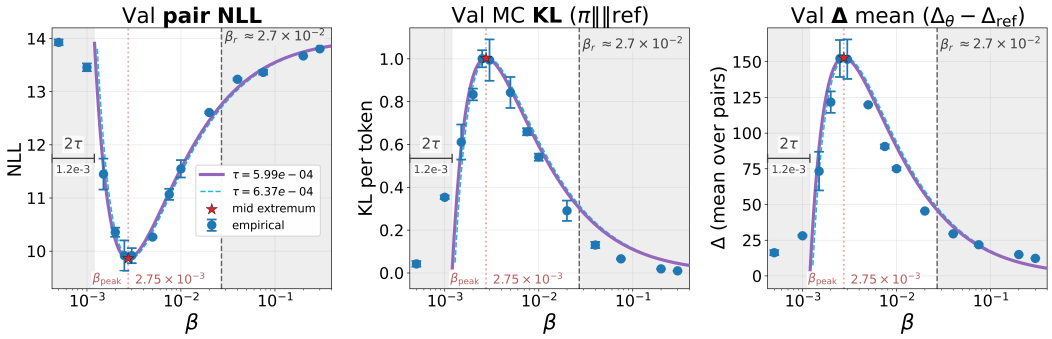

Figure 2: Standard DPO on HelpSteer3 with Qwen3-4B-Instruct-2507 (SGD, $\mathrm{lr} = 2 \times 10^{-4}$, batch size 24): minimum validation pair NLL (evaluated with $\beta = 1$), with per-token Monte Carlo KL and mean validation margin $\overline{\Delta} = \Delta_\theta - \Delta_{\mathrm{ref}}$ at the same epoch, plotted against β. Points are means $\pm$ one standard deviation (Table G1); far-right $\beta \in \{0.2, 0.3\}$ are single-seed. Solid and dashed overlays follow the shape of Equation (8) with τ from β_{peak} ($\tau = 5.99 \times 10^{-4}$) and from a shape+scale fit of $\overline{\Delta}(\beta)$ ($\tau = 6.37 \times 10^{-4}$; Appendix G.1). The left shaded band marks $\beta \le 2\tau$; the broad right-hand band is only an order-of-magnitude finite-horizon estimate (Appendix L).

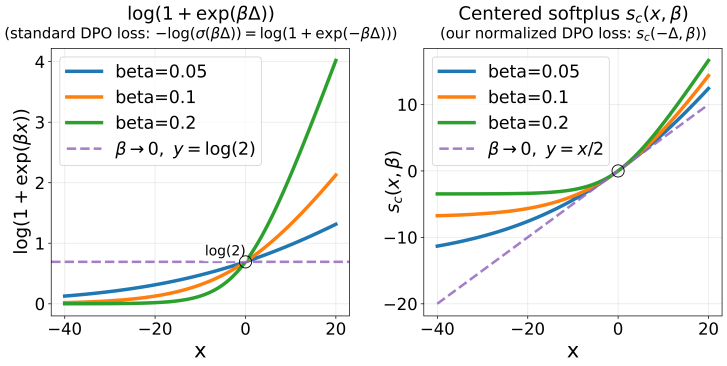

Figure 3: Standard DPO loss $-\log \sigma(\beta\Delta)$ and normalized centered-softplus loss $s_c(-\Delta; \beta)$. As $\beta \to 0$, the standard loss approaches $\log 2$, whereas the normalized loss approaches $-\Delta/2$.

4 Normalized Objective Formulation

4.1 Normalized DPO Loss

A natural normalization is to divide Equation (5) by $\beta > 0$. For every fixed β, this positive rescaling preserves the objective ordering and any attained minimizers. For the standard softplus term, define

$$s(x; \beta) = \frac{1}{\beta} \log\left(1 + e^{\beta x}\right). \tag{10}$$

Using this notation, the normalized DPO objective can be written as

$$\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}(\Delta; \beta) = s(-\Delta; \beta). \tag{11}$$

For $\beta > 0$, its marginal gradient is

$$\frac{\partial \mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}}{\partial \Delta} = -\sigma(-\beta\Delta). \tag{12}$$

Compared with Equation (6), the explicit multiplicative β prefactor has been removed. Thus β still controls the shape and saturation of the sigmoid response as an effective inverse preference-noise-scale coefficient, but no longer appears as a multiplicative gradient prefactor. Quantitatively,

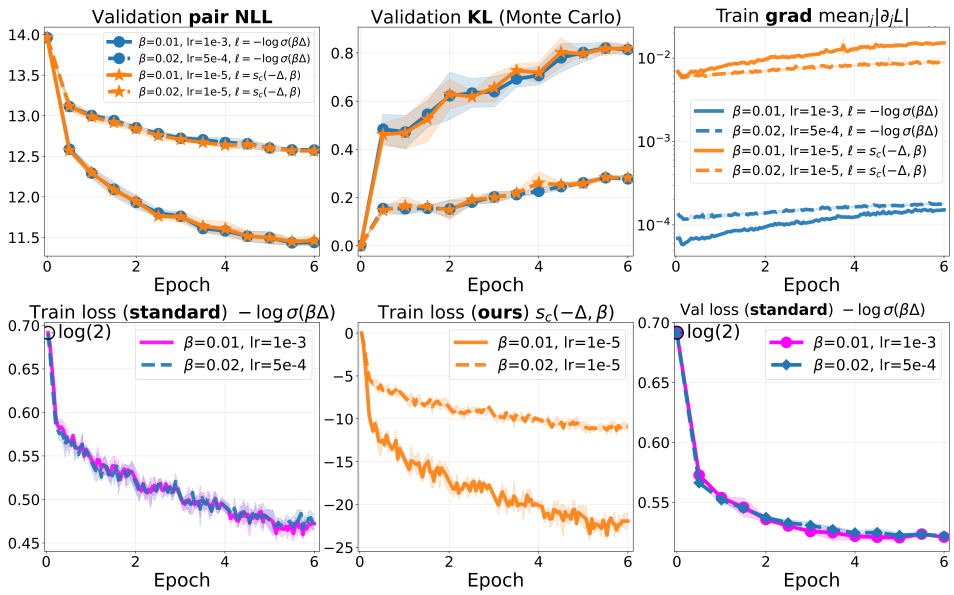

Figure 4: Comparison of training with the normalized centered-softplus DPO objective $s_c(-\Delta;\beta)$ and standard DPO with pairwise loss $-\log\sigma(\beta\Delta)$ on UltraFeedback with Qwen3-4B-Instruct-2507 (batch size 26). Learning rates satisfy $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}}$. With SGD, validation metrics are nearly equivalent. Even when validation metrics differ substantially between $\beta \in \{0.01, 0.02\}$, the standard training-loss curves are almost overlapping, whereas for $s_c(-\Delta;\beta)$ the training losses separate noticeably. Curves show the mean over four seeds; shaded bands denote one standard deviation. The comparison tests the predicted scale equivalence; it is not a claim that the normalized objective outperforms a retuned DPO baseline.

$|\partial\mathcal{L}^{\mathrm{norm}}_{\mathrm{DPO}}/\partial\Delta| = \sigma(-\beta\Delta) \le 1$ while $\partial^2\mathcal{L}^{\mathrm{norm}}_{\mathrm{DPO}}/\partial\Delta^2 = \beta\,\sigma'(\beta\Delta) \le \beta/4$: β bounds the curvature, while the initial marginal-gradient magnitude remains $1/2$ for every β, in line with the low-curvature motivation for centered softplus in Srinivas et al. (2022). For $\beta\Delta \approx 0$, $\log(1+e^{-\beta\Delta})/\beta \approx (\log 2)/\beta - \Delta/2$. The $\beta \to 0$ endpoint is handled rigorously in Appendix E; in particular, the marginal gradient has the well-defined limit $-1/2$.

4.2 Centered Softplus

To eliminate the $\log(2)/\beta$ offset, we adopt the centered-softplus transformation introduced by Srinivas et al. (2022):

$$s_c(x;\beta) = s(x;\beta) - \frac{\log 2}{\beta} = \frac{1}{\beta}\log\left(\frac{1+e^{\beta x}}{2}\right). \tag{13}$$

For every fixed $\beta > 0$, $s_c(-\Delta;\beta)$ is a positive affine transformation of the original DPO loss. The two objectives induce the same ordering and have the same minimizers whenever these are attained; their infima correspond under the same affine map (Appendix E). The centering constant $\log 2$ has a statistical meaning. In the Gumbel-noise view of Bradley–Terry (Sun et al., 2025), $P(y_w \succ y_l) = \sigma((r_w - r_l)/s_a)$. Appendix F shows that substituting $r_w - r_l = \lambda\Delta$ gives $P(y_w \succ y_l) = \sigma(\beta\Delta)$ with $\beta = \lambda/s_a$, and that $\mathcal{L}_{\mathrm{DPO}} - \log 2$ equals the negative log-likelihood ratio of the current preference model against the random-choice baseline $P = 1/2$. Dividing by β removes the effective inverse preference-noise-scale prefactor from the marginal gradient, yielding the normalized centered-softplus loss; it should not be interpreted as measuring literal annotator noise unless the reward/KL normalization is fixed. The same construction carries over to the Gaussian (Thurstone) model (Appendix F). The practical consequence for cross-β comparability of loss values was established in Section 3.4. In the limit $\beta \to 0$, the centered softplus converges to a linear map $s_c(x;0) = x/2$.

Accordingly, we define the updated normalized DPO objective as

$$\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}(\Delta;\beta) = \begin{cases} s_c(-\Delta;\beta), & \beta > 0, \\ -\frac{\Delta}{2}, & \beta = 0. \end{cases} \tag{14}$$

This parameterization removes the divergent offset while preserving any attained minimizers of the original objective for $\beta > 0$. At $\beta = 0$, minimizing $\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}$ is equivalent (up to a constant factor and the fixed reference term) to maximizing the policy log-probability margin $\log \pi_\theta(y_w \mid x) - \log \pi_\theta(y_l \mid x)$. The linear branch does not saturate—its marginal gradient remains $-\frac{1}{2}$—and should be read as a limiting endpoint rather than a default: it does not by itself impose a finite preferred margin, so the realized update is still set by the optimizer and the training horizon.

The normalized objective also addresses the Section 3.1 issue that, for small β, the standard DPO loss can remain nearly flat even when optimization is relatively aggressive. Unlike $-\log \sigma(\beta\Delta)$, which flattens toward $\log 2$ and saturates quickly as β decreases, the centered-softplus profile does not collapse on the negative half-axis (Figure 3): its lower plateau $-\log 2/\beta$ deepens as β shrinks, so the marginal gradient $-\sigma(-\beta\Delta)$ stays near $-\frac{1}{2}$ over a $1/\beta$-wide range of margins. Any attained minimizer is unchanged for $\beta > 0$, while the loss retains more variation along a trajectory and separates the examined runs more clearly (Figure 4).

For the centered-softplus objective, $|\partial \mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}/\partial\Delta| = \sigma(-\beta\Delta)$, so the stopping condition $\sigma(-\beta\Delta) = \tau$ gives the saturation margin

$$\Delta^* = \frac{1}{\beta} \log \frac{1-\tau}{\tau}. \tag{15}$$

Within this scalar stopping model, the margin scales as $1/\beta$ with no small-β dead zone or intermediate peak (Appendix G).

4.3 Experiments

To verify that the normalized objective behaves as predicted by the gradient-scale analysis, we compare it against standard DPO under a learning-rate rescaling that equalizes the marginal gradient scale. Because $\nabla_\theta \mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}} = \nabla_\theta \mathcal{L}_{\mathrm{DPO}}/\beta$ for $\beta > 0$, equal SGD updates require

$$\mathrm{lr}_{\mathrm{norm}} = \beta\, \mathrm{lr}_{\mathrm{standard}}.$$

This identification is an SGD statement. Under AdamW in the cancellation regime of Appendix B, the optimizer already removes the explicit β multiplier, so the existing AdamW learning rate should be kept (or retuned independently) when switching to the normalized objective; applying the SGD rescaling would undertrain. The SGD rule is appropriate only in the ϵ-dominated regime, where AdamW behaves like scaled SGD. SGD trajectories should therefore match closely under this learning-rate relation, up to stochasticity and numerical effects. Figure 4 illustrates this prediction in a realistic LLM fine-tuning setup: each centered-softplus run uses $\mathrm{lr}_{\mathrm{norm}} = \beta\, \mathrm{lr}_{\mathrm{standard}}$, and the two objectives produce closely aligned validation trajectories, consistent with their objective-level affine relation. The same figure shows that the normalized training loss remains distinguishable for $\beta \in \{0.01, 0.02\}$, whereas the standard training-loss curves coincide (Section 3.4). Appendix J reports downstream AlpacaEval 2 and IFEval for these four curves (Table J1): the overlapping standard-loss pair differs substantially, while each standard-DPO run agrees with its learning-rate-rescaled centered-softplus counterpart; a HelpSteer3 comparison over eight paired seeds shows no significant seed-wise Wilcoxon difference at $n = 8$ (Table J2, Figure J2), which does not establish practical equivalence. The two *standard-loss* UltraFeedback runs in Figure 4 are the loss-invariance pair of Section 3.4 and Appendix I.

The same identification of the two losses under $\mathrm{lr}_{\mathrm{norm}} = \beta\, \mathrm{lr}_{\mathrm{standard}}$ defines a common (β, η) window. Figure 5 shows the mean margin on that window for centered-softplus training on Orca-DPO with Phi-4-mini 3.8B (SGD, batch size 22, six epochs). Each cell uses one seed; the Matérn interpolant of Appendix H pools neighboring cells and partially smooths seed-level noise rather than repeating seeds at a fixed (β, η). The left panel plots the rectangle in the (β, η) coordinates of standard DPO; the right panel is the same runs in $(\beta, \eta_{\mathrm{norm}})$ with $\eta_{\mathrm{norm}} = \beta\eta$, i.e. the coordinates in which the SGD step of centered softplus matches standard DPO under the rescaling above. Appendix H records the sampling grid (112 completed cells), the interpolant, the companion NLL and KL surfaces, and the gradient field of $\overline{\Delta}$.

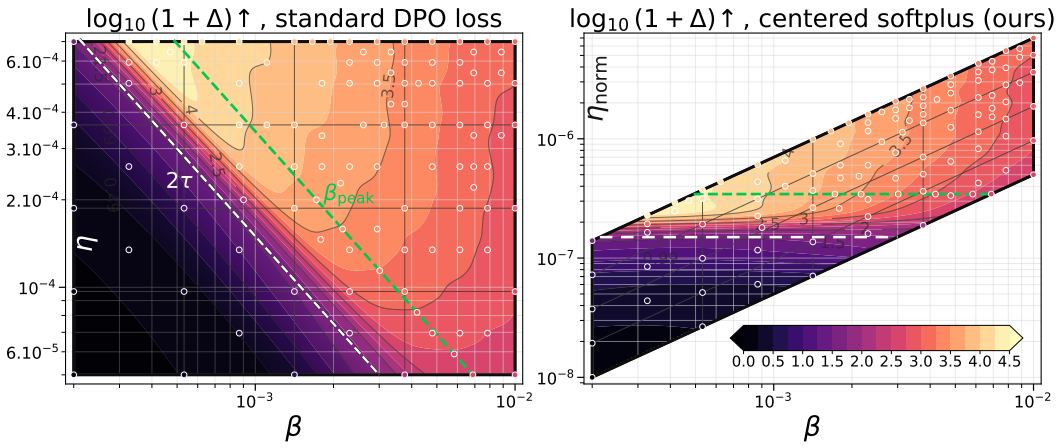

Figure 5: Mean margin $\log_{10}(1+\overline{\Delta})$ under centered-softplus DPO on Orca-DPO / Phi-4-mini 3.8B (SGD, batch size 22, six epochs; one seed). Left: (β, η). Right: same runs in $(\beta, \eta_{\text{norm}})$, $\eta_{\text{norm}} = \beta\eta$. White dashed: 2τ. Colored dashed: $\beta_{\text{peak}} = 4.59\tau$.

For centered softplus the initial marginal gradient is $1/2$ independently of β, so the dead zone is a floor $\eta_{\text{norm}} \lesssim 2\varepsilon/g_0^2$ (white dashed in Figure 5), not a left wall in β. Isolines of $\overline{\Delta}$ run roughly parallel to that floor. Substituting Equation (7) into Equation (15) gives Equation (H.2) (Appendix H): β enters only as $1/\beta$ and the learning rate only inside the logarithm. Under standard DPO the dead zone instead depends on the learning rate, $\beta \leq 2\tau(\eta) = 2\varepsilon/(\eta g_0^2)$ (Equation (7)), and at fixed η the margin often *increases* with β up to $\beta_{\text{peak}} \approx 4.6\,\tau(\eta)$. If β is read as a KL-control parameter, this empirical trend toward larger β is counterintuitive: stronger intended regularization then coincides with a larger, rather than smaller, mean margin. Validation pair NLL and KL co-move with $\overline{\Delta}$ on the same interpolant (Figure H1, Appendix H); fixed-η slices of standard DPO are in Figure 2 and Figure G2. This is the part of the comparison that a learning-rate retuning argument does not absorb. Under standard DPO the dead-zone boundary is the diagonal $\beta\eta = 2\varepsilon/g_0^2$, so the usable ranges of β and η cannot be chosen independently; under centered softplus it is horizontal in $(\beta, \eta_{\text{norm}})$, with η_{norm} alone deciding whether the run moves and β then setting the margin scale as $1/\beta$. In the reported window the two-dimensional search is thus approximately separable rather than merely shifted.

The same pattern holds under AdamW, including the $\beta = 0$ linear branch of Equation (14) (Figures J3 and J4, Appendix J): the endpoint is numerically viable in these experiments, but is less stable than any $\beta > 0$ run because the linear objective is unbounded below.

5 Conclusion

We recommend considering the normalized centered-softplus objective when the explicit β gradient prefactor or weak sensitivity of the standard loss across β is a practical concern. This substitution:

- removes the explicit β prefactor responsible for initial gradient attenuation at small β;
- yields a saturation margin monotone in $1/\beta$ and replaces the β-dependent dead-zone diagonal by a floor on η_{norm} alone, making the two hyperparameters approximately separable;
- extends continuously to $\beta = 0$, a linear margin loss unreachable in the standard form; and
- produces loss trajectories that better distinguish the examined cross-β runs, although their numerical values remain objective- and β-dependent.

The change is a local rewrite of the loss evaluation and drops into existing DPO codebases (e.g., TRL) without restructuring the trainer or the data pipeline.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing (translation, grammar and readability editing, LaTeX formatting), to retrieve and discover related work, to draft sections of the manuscript, to assist with research execution including writing and editing experiment and analysis code, and to draft and check mathematical derivations and proofs in the appendices. We have not used generative AI tools to generate synthetic datasets. All AI-assisted text, literature suggestions, derivations, and code were reviewed and verified by the authors; we take responsibility for the final content of this work, including every mathematical claim and proof.

ETHICS STATEMENT

Our empirical evidence covers four preference datasets and five backbones (Qwen3-4B-Instruct-2507, Qwen2.5-3B-Instruct, Ministral-3B-Instruct, Mamba-2 2.7B, and Phi-4-mini-instruct), all fine-tuned with LoRA at moderate scale, so transfer to larger models, full fine-tuning, and other data regimes remains to be verified. Training runs use standard DPO or its normalized centered-softplus form; Appendix A extends the β prefactor analytically to other DPO-family losses but does not evaluate them empirically. Minimizer equivalence for $\beta > 0$ is an objective-level statement: finite-time trajectories can still differ across optimizers and implementations. HelpSteer3, UltraFeedback, and HH-RLHF sweeps use four to five seeds; the Orca-DPO (η, β) grid (Figure 5) is reported for a single seed. Downstream checks cover a HelpSteer3 comparison (AlpacaEval 2 and IFEval in Table J2) and AlpacaEval 2 and IFEval on the UltraFeedback curves of Figure 4 (Table J1), and broader robustness evaluations (e.g., noisy or adversarial labels, out-of-domain prompts, long-horizon iterative preference updates) remain for future work.

This work reformulates an existing preference-optimization objective and does not introduce a new capability class; its risks are those of language-model alignment more broadly, including misuse of better-aligned systems and biases inherited from preference data and annotator judgments. Downstream AlpacaEval 2 scores use a Qwen2.5-14B-Instruct judge on Qwen candidates, so intra-family preference may affect absolute win rates. We recommend standard deployment safeguards: safety evaluation, bias and robustness assessment, and human oversight in high-stakes settings.

REPRODUCIBILITY STATEMENT

Code for this work is publicly available at `https://anonymous.4open.science/r/bayesian_dpo-86DF/`. We train Qwen3-4B-Instruct-2507, Qwen2.5-3B-Instruct, Ministral-3B-Instruct, Mamba-2 2.7B, and Phi-4-mini-instruct (all with LoRA) on four public pairwise preference datasets from Hugging Face: HelpSteer3-Preference (NVIDIA, 2025; Wang et al., 2025), UltraFeedback Binarized (Hugging Face H4, 2023; Cui et al., 2024), PKU-processed HH-RLHF (PKU-Alignment, 2023; Anthropic, 2022; Bai et al., 2022), and Intel Orca-DPO pairs (Intel, 2024; Lian et al., 2023; Mukherjee et al., 2023). For downstream instruction-following checks (Tables J2 and J1), we additionally use Qwen2.5-14B-Instruct as an LLM judge for AlpacaEval 2 (Dubois et al., 2024) and the official verifiable IFEval benchmark (Zhou et al., 2023) without a judge model. The judge shares a model family with the Qwen candidates, so absolute AlpacaEval win rates may be affected by intra-family preference. All experiments ran on NVIDIA A100-SXM4-80GB GPUs. Dataset processing and splits, per-run hyperparameters, the Monte Carlo KL protocol, training times, and software versions are detailed in Appendix N; full run configurations are included in the released code.

REFERENCES

Anthropic. Anthropic/hh-rlhf. https://huggingface.co/datasets/Anthropic/hh-rlhf, 2022. Hugging Face dataset.

AntonV. Antonv/mamba2-2.7b-hf. https://huggingface.co/AntonV/mamba2-2.7b-hf, 2024. Hugging Face model.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Deep Ganguli, Tom Henighan, Nicholas Joseph, Ben Mann, Chris Olah, Nicholas Schiefer, and et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2204.05862.

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. doi: 10.2307/2334029.

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 (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

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 (ICML)*, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Hiroki Furuta, Kuang-Huei Lee, Shixiang Shane Gu, Yutaka Matsuo, Aleksandra Faust, Heiga Zen, and Izzeddin Gur. Geometric-averaged preference optimization for soft preference labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-1819.

Xiangkun Hu, Lemin Kong, Tong He, and David Wipf. Explicit preference optimization: No need for an implicit reward model. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/hu25l.html.

Hugging Face H4. Huggingfaceh4/ultrafeedback_binarized. https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized, 2023. Hugging Face dataset.

Intel. Intel/orca_dpo_pairs. https://huggingface.co/datasets/Intel/orca_dpo_pairs, 2024. Hugging Face dataset.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and Teknium. Openorca: An open dataset of GPT augmented FLAN reasoning traces. https://huggingface.co/datasets/Open-Orca/OpenOrca, 2023. Hugging Face dataset.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37131–37161, 2025. URL https://proceedings.mlr.press/v267/liang25d.html.

Yixin Liu, Pengfei Liu, and Arman Cohan. Understanding reference policies in direct preference optimization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. doi: 10.18653/v1/2025.findings-naacl.447.

Qinwei Ma, Jingzhe Shi, Can Jin, Jenq-Neng Hwang, Serge Belongie, and Lei Li. Gradient imbalance in direct preference optimization. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2502.20847.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of llms via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-3317.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-3946.

Microsoft. microsoft/phi-4-mini-instruct. https://huggingface.co/microsoft/Phi-4-mini-instruct, 2025a. Hugging Face model.

Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint*, 2025b. URL https://arxiv.org/abs/2503.01743.

ministral. ministral/ministral-3b-instruct. https://huggingface.co/ministral/Ministral-3b-instruct, 2024. Community Hugging Face checkpoint (not an official Mistral AI release).

Eric Mitchell. A note on dpo with noisy preferences & relationship to ipo, 2023. URL https://ericmitchell.ai/cdpo.pdf. Technical note.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2306.02707.

NVIDIA. nvidia/helpsteer3. https://huggingface.co/datasets/nvidia/HelpSteer3, 2025. Hugging Face dataset.

Motoki Omura, Yasuhiro Fujita, and Toshiki Kataoka. Entropy controllable direct preference optimization. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2411.07595. Also presented at ICML 2025 Workshop on MoFA.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Marc Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

PKU-Alignment. Pku-alignment/processed-hh-rlhf. https://huggingface.co/datasets/PKU-Alignment/processed-hh-rlhf, 2023. Hugging Face dataset.

Qwen Team. Qwen/qwen2.5-14b-instruct. https://huggingface.co/Qwen/Qwen2.5-14B-Instruct, 2024. Hugging Face model; Qwen License (Apache 2.0 based).

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2338.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006. ISBN 0-262-18253-X. URL https://gaussianprocess.org/gpml/.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/ray-chowdhury24a.html.

Judy Hanwen Shen, Archit Sharma, and Jun Qin. Towards data-centric RLHF: Simple metrics for preference dataset comparison. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2409.09603.

Suraj Srinivas, Kyle Matoba, Himabindu Lakkaraju, and François Fleuret. Efficient training of low-curvature neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.52202/068431-1882.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking bradley–terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2411.04991. ICLR 2025 version titled "Rethinking Reward Modeling in Preference-based Large Language Model Alignment".

Luca Viano, Ruida Zhou, Yifan Sun, Mahdi Namazifar, Volkan Cevher, Shoham Sabach, and Mohammad Ghavamzadeh. Direct preference optimization with rating information: Practical algorithms and provable gains. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2602.00603.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Secrets of RLHF in large language models part II: Reward modeling. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.06080.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-1448.

Zixian Wang. Adpo: Anchored direct preference optimization: A unified and robust framework from pairwise to listwise preferences. *arXiv preprint*, 2025. URL https://arxiv.org/html/2510.18913v1.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-dpo: Direct preference optimization with dynamic β. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-4128.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Alphadpo: Adaptive reward margin for direct preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67793–67809, 2025. URL https://proceedings.mlr.press/v267/wu25af.html.

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. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a. URL https://proceedings.mlr.press/v235/xu24t.html.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? a comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b. URL https://proceedings.mlr.press/v235/xu24h.html.

Zhiqin Yang, Yonggang Zhang, Wei Xue, Dong Fang, Bo Han, and Yike Guo. Conditional equivalence of DPO and RLHF: Implicit assumption, failure modes, and provable alignment. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2605.20834.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/zeng24c.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2311.07911.

A MARGINAL GRADIENTS IN SELECTED DPO-FAMILY VARIANTS

Table A1: Comparison of objectives and marginal gradients across DPO-family methods. The last column is $\partial L/\partial(\cdot)$ for the margin of that row (Δ, m, M, or v), not always $\partial L/\partial\Delta$. Analytical survey only; we train standard DPO.

Method	Hyperparameters / labels	Loss	Marginal gradient
DPO (Rafailov et al., 2023)	β	$-\log\sigma(\beta\Delta)$	$-\beta\,\sigma(-\beta\Delta)$
IPO (Azar et al., 2024)	β	$\left(\Delta-\frac{1}{2\beta}\right)^2$	$2\left(\Delta-\frac{1}{2\beta}\right)$
cDPO (Mitchell, 2023)	β,ε	$-(1-\varepsilon)\log\sigma(\beta\Delta)-\varepsilon\log\sigma(-\beta\Delta)$	$\beta\big(\sigma(\beta\Delta)-(1-\varepsilon)\big)$
α-DPO (Wu et al., 2025)	β,α,γ	$-\log\sigma(u-c_\alpha)$	$\partial L/\partial m_\theta=-\beta\,\sigma(c_\alpha-u)$
GAPO (Furuta et al., 2024)	$\beta,\hat{p}\in(0,1)$	$-\log\sigma(\beta(2\hat{p}-1)\Delta)$	$-\beta(2\hat{p}-1)\sigma(-\beta(2\hat{p}-1)\Delta)$
β-DPO (Wu et al., 2024)	$\beta_{\mathrm{dyn}}(\cdot)$	$-\log\sigma(\beta_{\mathrm{dyn}}\Delta)$	$-\beta_{\mathrm{dyn}}\,\sigma(-\beta_{\mathrm{dyn}}\Delta)$
SimPO (Meng et al., 2024)	β,γ	$-\log\sigma(\beta m-\gamma)$	$\partial L/\partial m=-\beta\,\sigma(\gamma-\beta m)$
TDPO (Zeng et al., 2024)	β,α	$-\log\sigma(\beta M)$, $M=\Delta-\delta_{\mathrm{SeqKL}}$	$\partial L/\partial M=-\beta\,\sigma(-\beta M)$
ROPO (Liang et al., 2025)	β,α	robust loss inducing gradient weight $w_\alpha(\Delta)$	$\propto-\beta\,w_\alpha(\Delta)$
AOT (Melnyk et al., 2024)	$\beta,\{\Delta_i\}_{i=1}^B$	distributional loss $D_{\mathrm{dist}}\left(\{\beta\Delta_i\}_{i=1}^B\right)$	$\partial L/\partial\Delta_i=\beta\,w_i^{\mathrm{dist}}$
KTO Ethayarajh et al. (2024)	$\beta,\lambda_D,\lambda_U,$ z_0,b	b=1: $-\lambda_D\log\sigma(\beta(v-z_0))$ b=0: $-\lambda_U\log\sigma(-\beta(v-z_0))$	$\partial L/\partial v=-\lambda_D\beta\big(1-\sigma(\beta(v-z_0))\big)$ $\partial L/\partial v=+\lambda_U\beta\big(1-\sigma(-\beta(v-z_0))\big)$

Table A1 summarizes pairwise objectives and their marginal gradients with respect to the listed margin. These identities are analytical. Except for standard DPO, we do not report training runs for the listed methods. Related extensions handle noisy labels, soft preferences, adaptive margins, entropy control, and richer comparison structures (Ray Chowdhury et al., 2024; Furuta et al., 2024; Wu et al., 2025; Omura et al., 2024; Wang, 2025; Viano et al., 2026), as well as explicit-reward formulations such as EXPO (Hu et al., 2025). The unpaired KTO objective (Ethayarajh et al., 2024) instead uses binary labels $b\in\{0,1\}$ with $v=\log\pi_\theta(y\mid x)-\log\pi_{\mathrm{ref}}(y\mid x)$ and reference offset z_0, and its marginal gradients $\partial\mathcal{L}_{\mathrm{KTO}}/\partial v$ likewise carry an explicit β prefactor alongside the class weights λ_D and λ_U. SimPO (Meng et al., 2024) is reference-free: it replaces Δ with the length-normalized policy-only margin $m=\frac{1}{|y_w|}\log\pi_\theta(y_w\mid x)-\frac{1}{|y_l|}\log\pi_\theta(y_l\mid x)$ (the length-normalized counterpart of Δ_θ, to which it reduces when $|y_w|=|y_l|$) and subtracts a target margin γ, yet its marginal gradient $\partial\mathcal{L}_{\mathrm{SimPO}}/\partial m=-\beta\,\sigma(\gamma-\beta m)$ retains the same explicit β prefactor.

Adaptive margin in α-DPO. α-DPO (Wu et al., 2025, Eq. 13) uses the length-normalized policy-only margin

$$m_\theta:=\frac{1}{|y_w|}\log\pi_\theta(y_w\mid x)-\frac{1}{|y_l|}\log\pi_\theta(y_l\mid x),\qquad u:=\beta m_\theta,$$

and the instance-adaptive offset

$$c_\alpha:=\mathrm{sg}[\gamma+\alpha M^*(x,y_w,y_l)],\qquad M^*:=\frac{M-\mu_M}{\sigma_M},$$

where

$$M:=\beta\log\frac{\pi_\theta(y_w\mid x)\pi_{\mathrm{ref}}(y_l\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)\pi_\theta(y_l\mid x)}$$

and sg denotes stop-gradient. Its per-example objective and marginal gradient are therefore

$$\mathcal{L}_{\alpha\text{-DPO}}=-\log\sigma(u-c_\alpha),\qquad\frac{\partial\mathcal{L}_{\alpha\text{-DPO}}}{\partial m_\theta}=-\beta\,\sigma(c_\alpha-u).$$

Thus α-DPO retains the explicit β gradient prefactor, but its adaptive offset is not a fixed margin and the objective is not solely a function of the standard unnormalized DPO margin Δ. TDPO (Zeng et al.,

2024) moves to the token level and augments the margin with a sequential forward-KL correction, optimizing $-\log\sigma(\beta M)$ with $M = \Delta - \delta_{\mathrm{SeqKL}}$, where $\delta_{\mathrm{SeqKL}} = D_{\mathrm{SeqKL}}(x, y_l; \pi_{\mathrm{ref}} \,\|\, \pi_\theta) - D_{\mathrm{SeqKL}}(x, y_w; \pi_{\mathrm{ref}} \,\|\, \pi_\theta)$ (in TDPO$_2$ the correction is weighted by α under a stop-gradient); its marginal gradient $\partial\mathcal{L}_{\mathrm{TDPO}}/\partial M = -\beta\,\sigma(-\beta M)$ likewise carries the prefactor and equals $-\beta/2$ at initialization, exactly as in DPO. The main exception is IPO, where β has a different semantic role (as a target/margin scale in the objective) rather than acting as a direct multiplicative coefficient on the DPO-style gradient term; note, however, that the coupling is inverted rather than removed: at $\Delta = 0$ the IPO marginal gradient equals $-1/\beta$, so decreasing β inflates rather than suppresses the update magnitude. For ROPO and AOT, the β dependence is present but enters indirectly through the robust or distributional weighting terms $w_\alpha(\Delta)$ and w_i^{dist}; the coupling may therefore be partially mediated by these weights in practice, and the effective suppression of update magnitude need not be strictly linear in β for all operating points.

B Adam Compensation Breakdown: Closed-Form Threshold

We derive a coordinate-wise condition under which Adam's adaptive scaling ceases to compensate for the explicit β factor in DPO gradients, and then construct an isotropic instantaneous scale proxy from the available global-gradient logs.

Setup. Write the per-parameter DPO gradient as $g_t = \beta h_t$, where h_t is the full non-β factor of the DPO gradient and $\beta > 0$ is the effective DPO coefficient. By the chain rule, the per-parameter gradient is $\partial\mathcal{L}/\partial\theta_j = -\beta\,\sigma(-\beta\Delta_t)\cdot\partial\Delta_t/\partial\theta_j$, so $h_{t,j} = -\sigma(-\beta\Delta_t)\cdot\partial\Delta_t/\partial\theta_j$ combines the sigmoid factor and the log-probability-gap Jacobian. The quantity $\sqrt{\hat{v}_t^h}$ is therefore the root-mean-square of this product over the moment history, and it evolves during training as both Δ_t and the policy Jacobian change. Under Adam with first and second moment decay rates $\rho_1, \rho_2 \in (0,1)$, optimizer learning rate η, and numerical stabilizer $\epsilon > 0$, the moment updates are:

$$m_t = \rho_1 m_{t-1} + (1-\rho_1)\,\beta h_t, \tag{B.1}$$

$$v_t = \rho_2 v_{t-1} + (1-\rho_2)\,\beta^2 h_t^2. \tag{B.2}$$

Let $\hat{m}_t^h$ and $\hat{v}_t^h$ denote the bias-corrected moment estimates of h_t and h_t^2, respectively, defined in the usual way ($\hat{m}_t^h = m_t^h/(1-\rho_1^t)$ and $\hat{v}_t^h = v_t^h/(1-\rho_2^t)$, where m_t^h, v_t^h accumulate h_t and h_t^2 without the β prefactor). Then $\hat{m}_t = \beta\hat{m}_t^h$ and $\hat{v}_t = \beta^2\hat{v}_t^h$, so the Adam parameter update is:

$$\theta_{t+1} - \theta_t = -\eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} = -\eta\frac{\beta\hat{m}_t^h}{\sqrt{\beta^2\hat{v}_t^h}+\epsilon} = -\eta\frac{\hat{m}_t^h}{\sqrt{\hat{v}_t^h}+\epsilon/\beta}. \tag{B.3}$$

Proposition B1 (Two-regime characterization). For coordinate j, define the critical threshold

$$\beta_{\mathrm{crit},t,j} \;:=\; \frac{\epsilon}{\sqrt{\hat{v}_{t,j}^h}}. \tag{B.4}$$

The coordinate-wise Adam update in Eq. (B.3) satisfies:

- *Cancellation regime* ($\beta \gg \beta_{\mathrm{crit},t,j}$): $\epsilon/\beta \ll \sqrt{\hat{v}_{t,j}^h}$, so the denominator $\approx \sqrt{\hat{v}_{t,j}^h}$ and
$$\theta_{t+1,j} - \theta_{t,j} \;\approx\; -\eta\frac{\hat{m}_{t,j}^h}{\sqrt{\hat{v}_{t,j}^h}},$$
so, for a fixed sequence $h_{t,j}$, Adam cancels the explicit common multiplicative β prefactor. This does not remove the remaining β-dependence of $h_{t,j}$, its moment history, or the optimization trajectory.
- *Non-cancellation regime* ($\beta \ll \beta_{\mathrm{crit},t,j}$): $\epsilon/\beta \gg \sqrt{\hat{v}_{t,j}^h}$, so the denominator $\approx \epsilon/\beta$ and
$$\theta_{t+1,j} - \theta_{t,j} \;\approx\; -\frac{\eta\beta}{\epsilon}\hat{m}_{t,j}^h,$$
which is proportional to β, the same linear suppression as plain SGD.

Proof. Equation (B.3) follows directly from the linearity of the moment recurrences in $g_t = \beta h_t$. The two limiting cases follow from comparing ϵ/β to $\sqrt{\hat{v}_t^h}$. □

Numerical estimates for DPO. The PyTorch AdamW default is $\epsilon = 10^{-8}$. At initialization, $\Delta \approx 0$ so $\sigma(-\beta\Delta) \approx 1/2$, giving

$$\sqrt{\hat{v}^h}\big|_{\text{init}} \;\approx\; \tfrac{1}{2}\,\mathrm{RMS}_j\left(\frac{\partial\Delta}{\partial\theta_j}\right),$$

where $\mathrm{RMS}_j(\cdot)$ is the per-element root-mean-square of the log-probability-gap Jacobian. The sigmoid factor $1/2$ alone does *not* determine $\sqrt{\hat{v}^h}$; the actual magnitude is set by the per-element Jacobian $|\partial\Delta/\partial\theta_j|$, which is architecture- and dataset-dependent. Consequently, the distribution of $\beta_{\text{crit},t,j} = \epsilon/\sqrt{\hat{v}^h_{t,j}}$ is best estimated from Adam's stored second moments and can vary substantially across coordinates, models, and training stages.

Estimating an isotropic scale proxy from gradient logs. For the TRL runs used in this appendix, the logged gradient metric is the global pre-clipping norm $G_t^{\text{norm}} = \|g_t\|_2$. With $g_t = \beta_{\text{ref}} h_t$, this gives

$$G_t^{\text{norm}} \;=\; \beta_{\text{ref}}\|h_t\|_2.$$

The logged global norm does not determine Adam's coordinate-wise, history-dependent second moments. In the absence of saved optimizer states, we use the isotropic instantaneous approximation $\sqrt{\hat{v}^h_t} \approx \|h_t\|_2/\sqrt{d}$, where d is the number of *trainable* parameters that actually receive gradients and enter Adam's per-parameter moment estimates. In our LoRA runs, d is exactly the number of LoRA adapter parameters (the frozen base-model weights contribute no gradient and are excluded from both G_t^{norm} and the moment estimates), and we take d directly from the `adapter_config.json` / PEFT `print_trainable_parameters()` log for each run rather than treating it as a free parameter. Then

$$\widetilde{\beta}_{\text{crit},t} \;:=\; \frac{\epsilon\,\beta_{\text{ref}}\sqrt{d}}{G_t^{\text{norm}}}. \tag{B.5}$$

This is an isotropic instantaneous proxy for the 50%-attenuation scale, not a measurement of Adam's coordinate-wise attenuation point.

For a practical "noticeable" proxy, define $\widetilde{\beta}_{\text{notice}}(\delta)$ by requiring a relative attenuation δ of the compensated Adam step. From Eq. (B.3), the attenuation factor is $A(\beta) = 1/(1+\epsilon/(\beta\sqrt{\hat{v}^h_t}))$, hence

$$\widetilde{\beta}_{\text{notice}}(\delta) \;=\; \frac{1-\delta}{\delta}\,\widetilde{\beta}_{\text{crit}}. \tag{B.6}$$

At $\delta = 10\%$, this gives $\widetilde{\beta}_{\text{notice}}(10\%) = 9\,\widetilde{\beta}_{\text{crit}}$.

Application to the reported experiments. We estimate the run-specific isotropic proxy from the directly logged TRL `grad_norm` values, together with the exact LoRA trainable-parameter count d for each run. Both runs use LoRA with rank $r = 16$ and $\alpha = 32$, but with different target modules and therefore different d: for Mamba-2 2.7B (target modules `in_proj`, `x_proj`, `dt_proj`), $d = 13{,}451{,}264$; for Qwen3-4B (target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj` across 36 layers), $d = 33{,}030{,}144$. Using PyTorch default $\epsilon = 10^{-8}$ and Eq. (B.5) with the corresponding d for each run, we report $\widetilde{\beta}_{\text{crit}}$ and $\widetilde{\beta}_{\text{notice}}(10\%)$. No parameter is fitted, but both quantities rely on the isotropic instantaneous approximation rather than the stored Adam moments. Table B1 lists two rows per run (first and last epoch): for Mamba-2 on HH-RLHF, step 100 and step 47,910; for Qwen on UltraFeedback, step 10 and step 22,890.

Three conclusions follow directly from the table.

(i) Early in training, both runs sit close to their isotropic proxy scales. For Mamba-2, $\beta_{\text{ref}} = 10^{-5}$ is within a factor of 1.1 of $\widetilde{\beta}_{\text{crit}} \approx 1.1\times10^{-5}$, consistent with the weak epoch-1 gradients ($\sim 3\times10^{-5}$) and the DPO loss remaining pinned at $\log 2$ (Appendix C). For UltraFeedback, $\beta_{\text{ref}} = 10^{-6}$ is within a factor of 1.5 of $\widetilde{\beta}_{\text{crit}} \approx 6.5 \times 10^{-7}$, and $\widetilde{\beta}_{\text{notice}}(10\%) \approx 5.8 \times 10^{-6}$ is of the same order of magnitude as the empirically reported onset scale $\beta \sim 10^{-5}$ for Adam-visible coupling. These agreements are descriptive because the actual coordinate-wise moment distribution was not logged.

(ii) The isotropic proxies are strongly stage-dependent. As training progresses and G_t^{norm} increases by three to four orders of magnitude, both proxies drop by a similar amount (e.g., $\widetilde{\beta}_{\text{crit}}$ falls from

Table B1: Isotropic instantaneous proxies $\widetilde{\beta}_{\mathrm{crit}}$ (50% attenuation scale) and $\widetilde{\beta}_{\mathrm{notice}}(10\%)$ (10% attenuation scale) from training logs. Rows alternate first/last epoch within each run (see text above). $G_t = \texttt{grad_norm}$ from TRL `DPOTrainer` (global gradient norm, pre-clipping, standard DPO loss). d is the exact number of trainable LoRA parameters for that run (from `adapter_config.json` / PEFT `print_trainable_parameters()`). Numerics use $\epsilon = 10^{-8}$.

Dataset	Model	Epoch	β_{ref}	d	G_t	$\widetilde{\beta}_{\mathrm{crit}}$	$\widetilde{\beta}_{\mathrm{notice}}$
HH-RLHF	Mamba-2 2.7B	first	10^{-5}	1.35×10^{7}	3.4×10^{-5}	1.1×10^{-5}	9.8×10^{-5}
HH-RLHF	Mamba-2 2.7B	last	10^{-5}	1.35×10^{7}	2.0×10^{-1}	1.8×10^{-9}	1.7×10^{-8}
UltraFB	Qwen3-4B	first	10^{-6}	3.30×10^{7}	8.9×10^{-5}	6.5×10^{-7}	5.8×10^{-6}
UltraFB	Qwen3-4B	last	10^{-6}	3.30×10^{7}	1.3×10^{-3}	4.4×10^{-8}	4.0×10^{-7}

1.1×10^{-5} to 1.8×10^{-9} for Mamba-2). This pattern is consistent with Adam's ϵ-dominated regime being most consequential early in training, but the global norms do not recover its coordinate-wise history.

(iii) Mamba-2 and UltraFeedback differ in both d and early gradient scale. Writing $\widetilde{\beta}_{\mathrm{crit}} = \epsilon\sqrt{d}/\|h_t\|_2$ with $\|h_t\|_2 = G_t^{\mathrm{norm}}/\beta_{\mathrm{ref}}$ separates the two effects within the isotropic proxy. Mamba-2's initial non-β gradient scale is much smaller than Qwen's ($\|h\|_2 \approx 3.4$ vs. 89), which by itself raises Mamba-2's proxy by $\approx 26\times$; its $2.5\times$ fewer trainable LoRA parameters (1.35×10^7 vs. 3.30×10^7, from the different target-module sets) partially offset this by lowering the proxy by $\sqrt{2.5} \approx 1.6\times$. The net effect, $26/1.6 \approx 17\times$, matches the tabulated proxy ratio early in training, showing how the gradient-norm scale and trainable-parameter count enter this approximation in opposite directions.

The instantaneous gradient-scale proxy is β-asymmetric in the experimental figures; the actual Adam moment distribution was not logged. For *small* β, the sigmoid stays near $1/2$ and $\sqrt{\hat{v}^h}$ tracks the Jacobian: as the policy moves, the Jacobian can grow and gradients increase. Across all reported runs, in Figure C3 (AdamW, UltraFeedback, $\beta = 10^{-6}$) gradient norms trend upward throughout training; in Figure C5 (HH-RLHF, SGD) small-β trajectories show a delayed surge in gradient magnitude; and in Figure 1 (HelpSteer3, SGD), larger β generally produces larger gradient magnitudes, while each curve rises over the run, with the $\beta = 2 \times 10^{-3}$ curve showing the most pronounced increase. The growing global gradient norm lowers $\widetilde{\beta}_{\mathrm{crit}}$ over time, consistent with partial self-correction of the non-compensation condition. However, this correction is delayed: learning is suppressed during the early phase before the proxy has fallen sufficiently. For *large* β, sigmoid saturation ($\sigma(-\beta\Delta) \to 0$ as Δ grows) suppresses h_t even as the Jacobian increases, keeping $\sqrt{\hat{v}^h}$ bounded and β_{crit} relatively stable; in this regime $\beta \gg \beta_{\mathrm{crit}}$ from the outset.

For $\beta \in [0.01, 0.2]$ (a common DPO range), the reported isotropic proxies lie many orders of magnitude below β, consistent with effective Adam compensation of the explicit prefactor.

Effect of gradient clipping and ϵ choices. If gradients entering Adam's moments are clipped, clipping reduces their effective per-element magnitude and can raise coordinate-wise cancellation thresholds; the TRL metric used above is pre-clipping and does not quantify this effect. Increasing ϵ (sometimes used for numerical stability in low-gradient regimes) directly raises β_{crit} and can push moderate β values into the partial non-cancellation regime. Both effects can exacerbate the gradient-scale coupling in practice, particularly in combination with very small β.

C Additional Training Dynamics

UltraFeedback TRL baseline (Figure C1). To make the small-β slowdown visible in a realistic setting, we train Qwen3-4B-Instruct-2507 on UltraFeedback for six epochs with the TRL `DPOTrainer`, comparing $\beta \in \{10^{-8}, 7 \times 10^{-7}, 10^{-6}, 10^{-3}\}$ (AdamW, learning rate 5×10^{-7}, effective batch size 16, five seeds per setting). We report both training dynamics, including gradient norm, reward accuracy, and log-probability gap, and validation dynamics, including pair NLL, pair accuracy, and policy-to-reference KL $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ per token. The training-time metrics are taken from the TRL trainer logs. Gradient norms scale approximately with β: for $\beta \leq 10^{-6}$ they remain low, the training loss stays pinned near $\log 2$, and validation KL stays close to zero, so these policies remain near the reference model, whereas the $\beta = 10^{-3}$ run learns clearly. Under the KL-control interpretation, the smaller-β runs would instead be expected to deviate at least as much. The pattern does not depend on the model or dataset: Figure C2 below shows the same dynamics on PKU-processed HH-RLHF with Mamba-2.

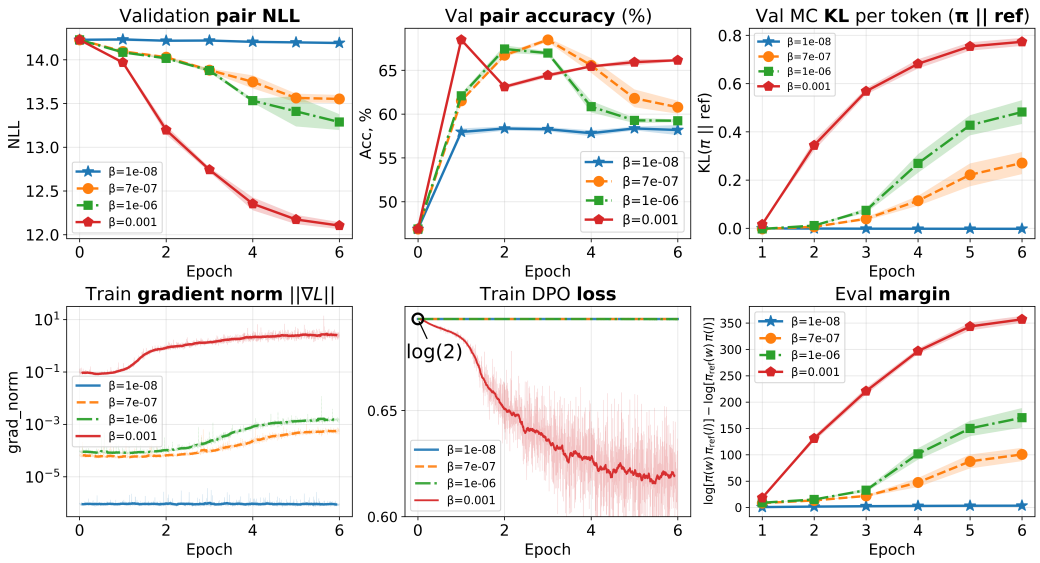

Figure C1: Training and validation dynamics of standard DPO on UltraFeedback with Qwen3-4B-Instruct-2507 using TRL `DPOTrainer` (AdamW, learning rate 5×10^{-7}, effective batch size 16), comparing $\beta \in \{10^{-8}, 7 \times 10^{-7}, 10^{-6}, 10^{-3}\}$. Curves show the mean over five seeds; shaded bands denote one standard deviation; see the discussion above.

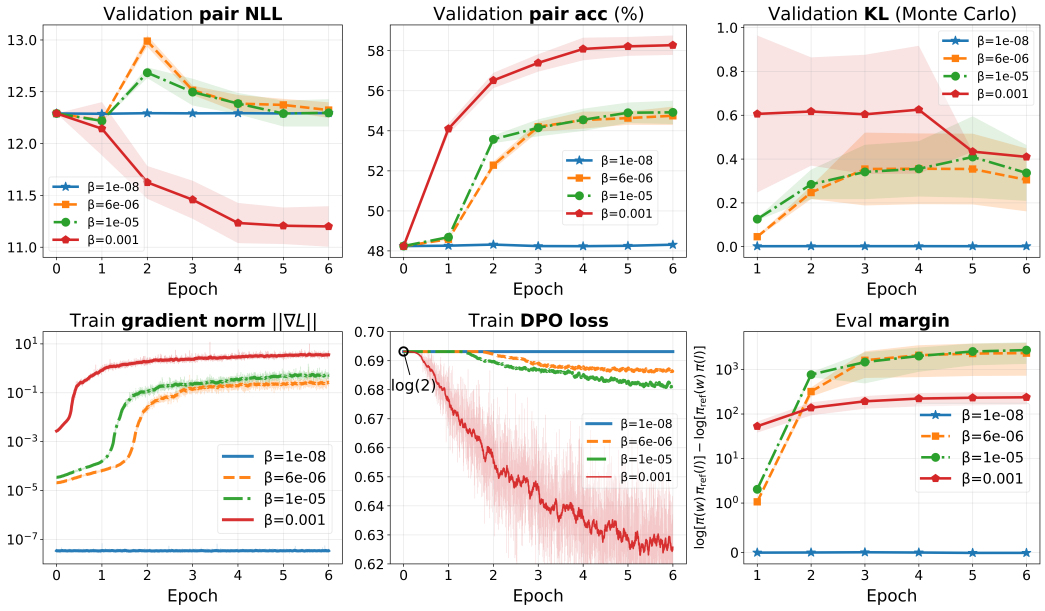

Figure C2: Training and validation dynamics of standard DPO on PKU-processed HH-RLHF with Mamba-2 using TRL `DPOTrainer` (AdamW, learning rate 5×10^{-6}, effective batch size 20). Curves show the mean over five seeds; shaded bands denote one standard deviation.

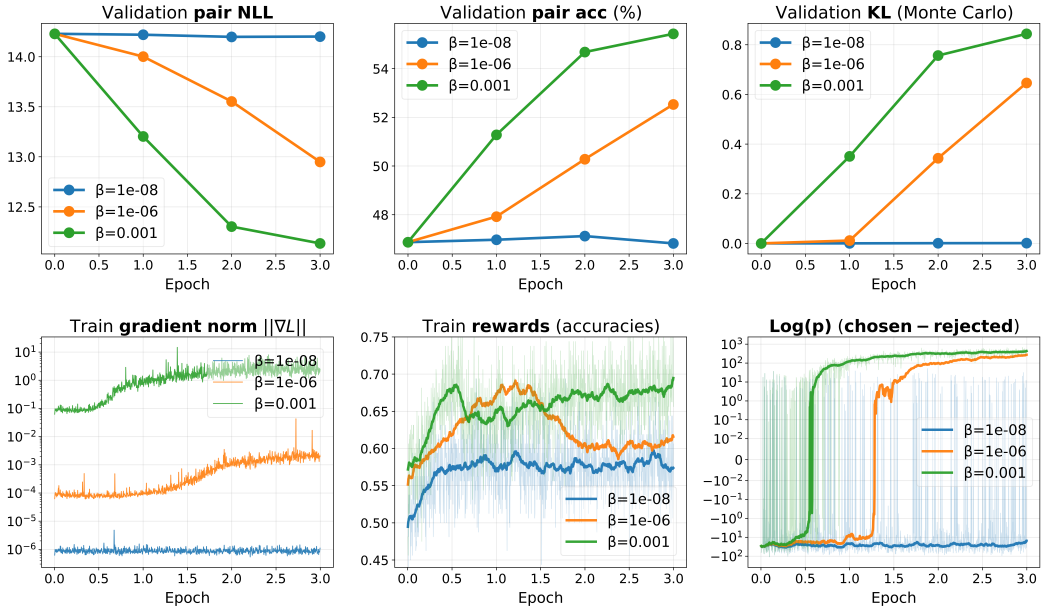

Figure C3: Training and validation dynamics of standard DPO on UltraFeedback with Qwen3-4B-Instruct-2507 using TRL `DPOTrainer` (AdamW, learning rate 1×10^{-6}, effective batch size 16), comparing $\beta \in \{10^{-8}, 10^{-6}, 10^{-3}\}$. For the two smaller β values, gradient norms remain low and validation KL (per token) stays close to zero, so the policy remains near the reference model; only the $\beta = 10^{-3}$ run moves appreciably.

Figure C4 shows an alternative HelpSteer3 SGD configuration at learning rate 5×10^{-4}, comparing $\beta \in \{5 \times 10^{-5}, 10^{-3}, 10^{-1}\}$.

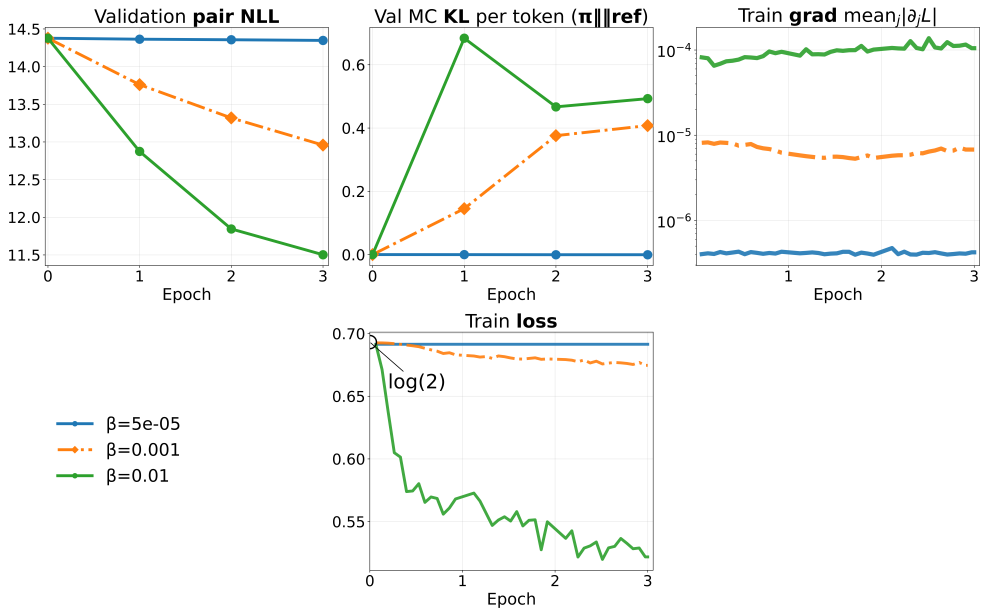

Figure C4: Training and validation dynamics of our DPO implementation on HelpSteer3 with Qwen3-4B-Instruct-2507 (SGD, $\mathrm{lr} = 5 \times 10^{-4}$, batch size 24).

Figure C5 shows a more intricate regime: across the HH-RLHF SGD trajectories ($\beta \in \{0.001, 0.002, 0.005, 0.01, 0.02\}$), a more aggressive policy trajectory can induce a delayed increase in gradient magnitude along the smaller-β runs, partially compensating for the nominal linear β prefactor in the DPO objective.

This delayed surge reflects stage-dependent growth of G_t^{norm}. Within the isotropic instantaneous approximation of Appendix B, such growth lowers $\widetilde{\beta}_{\mathrm{crit}}$; the actual coordinate-wise Adam moments were not logged. Under AdamW, growth of the corresponding second moments can move coordinates from the ϵ-dominated regime, in which updates retain the explicit β suppression, toward the cancellation regime. Thus the slowdown can be strongest early in training and weaken once gradients have grown. The HH-RLHF curves in Figure C5 use SGD rather than Adam, so there is no moment-based β_{crit} transition; nevertheless, the same rising Jacobian enlarges raw gradient norms and partially offsets the smaller β. Under plain SGD with learning rate η, using the decomposition $g_t = \beta h_t$ from Appendix B,

$$\theta_{t+1} - \theta_t = -\eta\, g_t, \qquad g_t = \beta\, h_t, \qquad \|g_t\|_2 = \beta\, \|h_t\|_2, \tag{C.1}$$

where h_t collects the sigmoid factor and the log-probability-gap Jacobian. A smaller β therefore directly shrinks $\|g_t\|_2$, but stage-dependent growth of $\|h_t\|_2$ as the policy moves can partially offset this reduction, as in the delayed surge visible in Figure C5.

For sufficiently small β, however, this late-stage compensation remains weaker than the β-induced attenuation during the early phase, yielding dynamics that are noticeably non-monotone across practically used tuning ranges. Such interaction effects between the inverse preference-noise scale and the optimization trajectory help explain why the phenomenon has often remained underappreciated in routine hyperparameter searches. Consistent with this view, the $\beta = 0.001$ trajectory attains substantially lower pair NLL and higher pair accuracy than the older $\beta = 0.01$ run, while its temporal profile remains unusual and departs from the behavior one would expect from a simple β rescaling. The sharp late-stage surge in gradient norm and the qualitatively different curve shapes relative to Figures C3 and 1 are, among other factors, compounded by the comparatively large learning rate ($\mathrm{lr} = 10^{-3}$, five times the HelpSteer3 SGD setting in Figure 1) acting on HH-RLHF preference pairs that are comparatively noisy and less separable. Data-centric comparisons of public RLHF corpora quantify effective sample size, noise invariance, and information content and report lower informativeness for HH-RLHF relative to several alternatives (Shen et al., 2024), while reward-modeling analyses of the same corpus emphasize annotator disagreement, ambiguous or incorrect preferences, and weak pairwise margins (Wang et al., 2024).

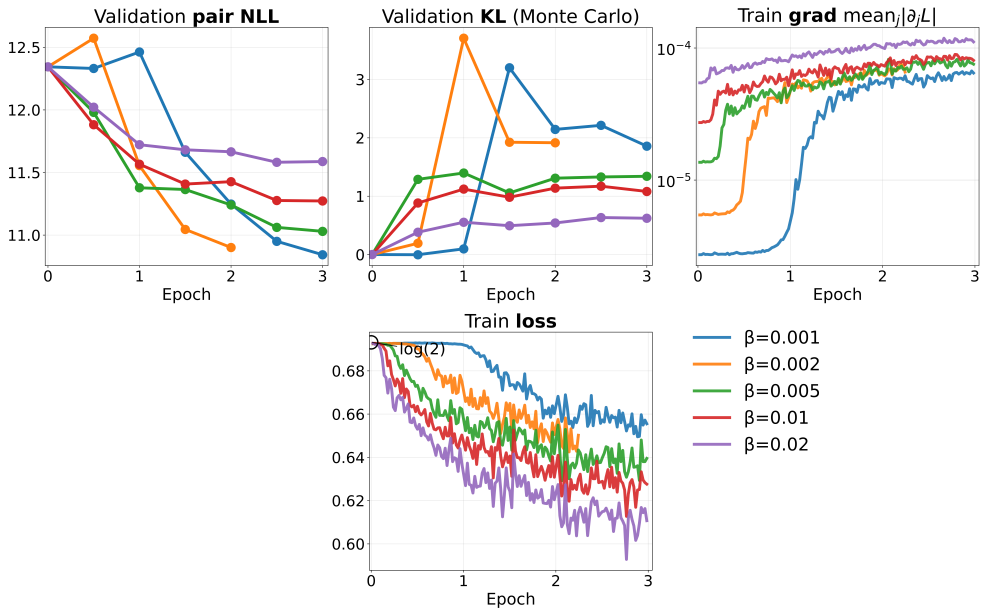

Figure C5: Training and validation dynamics of our DPO implementation on PKU-processed HH-RLHF with Qwen2.5-3B-Instruct (SGD, $\mathrm{lr} = 10^{-3}$, batch size 32).

D Thurstone Gradient and Comparison with DPO

General link functions. Following the annotator-noise viewpoint of Sun et al. (2025), let ε denote a pairwise comparison noise variable with CDF F and PDF $f = F'$ on $\mathbb{R}$. We assume the standard regularity conditions

$$F(-\infty) = 0, \quad F(+\infty) = 1, \quad f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\, dx = 1, \tag{D.1}$$

together with strict monotonicity ($f(x) > 0$ for all x, hence $F(x) \in (0, 1)$) and a finite first moment $\mathbb{E}[|\varepsilon|] < \infty$. An annotator prefers y_w over y_l when the latent margin exceeds the noise, $\Delta > \varepsilon$, so

$$P(y_w \succ y_l \mid x) = P(\varepsilon < \Delta) = F(\Delta). \tag{D.2}$$

Writing the effective inverse preference-noise scale in policy-margin coordinates as $\beta > 0$ gives the scaled link $P(y_w \succ y_l \mid x) = F(\beta\Delta)$. As shown in Appendix F, if the underlying reward model has pairwise noise scale s_a and the RLHF objective has KL coefficient λ, then $\beta = \lambda/s_a$. The corresponding negative log-likelihood is

$$\mathcal{L}_F(\Delta; \beta) = -\log F(\beta\Delta), \tag{D.3}$$

and, by the chain rule,

$$\frac{\partial \mathcal{L}_F}{\partial \Delta} = -\beta\, \frac{f(\beta\Delta)}{F(\beta\Delta)}. \tag{D.4}$$

Thus the explicit multiplicative β prefactor is not specific to the logistic or Gaussian branches: it follows for *any* link CDF in this class. Bradley–Terry/DPO and Thurstone correspond to $F = \sigma$ and $F = \Phi$; other examples on $\mathbb{R}$ with finite first moments include Laplace noise and symmetric Student-t_ν noise with $\nu > 1$. Distributions without finite first moments (e.g., Cauchy) are outside this interpretable class even though $P(\varepsilon < \Delta)$ is formally defined, because the latent-utility/noise-scale story no longer has finite expected bias. Symmetry or membership in an exponential family is not required. For plain SGD, decreasing β therefore still linearly suppresses update magnitude across this class unless the learning rate is retuned. Dividing by $\beta > 0$ generalizes as well:

$$\frac{\partial}{\partial \Delta}\left[-\frac{1}{\beta}\log F(\beta\Delta)\right] = -\frac{f(\beta\Delta)}{F(\beta\Delta)}, \tag{D.5}$$

with the same $\arg\min_\Delta \mathcal{L}_F$ because $1/\beta$ is a positive constant.

Thurstone case ($F = \Phi$). Beyond the logistic Bradley–Terry branch (Bradley & Terry, 1952), the Gaussian annotator-noise model of Sun et al. (2025) induces preferences $P(y_w \succ y_l) = \Phi(\beta\Delta)$, yielding the Thurstone pairwise loss

$$\mathcal{L}_{\mathrm{Th}}(\Delta;\beta) = -\log \Phi(\beta\Delta), \tag{D.6}$$

the special case $F = \Phi$ of Equation (D.3), where Φ is the standard normal CDF and $\phi = \Phi'$ is the corresponding PDF. Writing $u = \beta\Delta$, so that $\mathcal{L}_{\mathrm{Th}}(\Delta;\beta) = -\log \Phi(u)$ with $\partial u/\partial\Delta = \beta$, and using $\Phi'(u) = \phi(u)$, the chain rule gives

$$\frac{\partial \mathcal{L}_{\mathrm{Th}}}{\partial \Delta} = -\frac{1}{\Phi(u)}\,\phi(u)\,\frac{\partial u}{\partial \Delta} = -\beta\,\frac{\phi(\beta\Delta)}{\Phi(\beta\Delta)}, \tag{D.7}$$

which instantiates Equation (D.4). As in the general case, dividing $\mathcal{L}_{\mathrm{Th}}$ by $\beta > 0$ removes the explicit β prefactor while preserving any attained minimizers, yielding the normalized Thurstone marginal gradient $-\phi(\beta\Delta)/\Phi(\beta\Delta)$.

Comparison with the logistic branch. The Thurstone objective in Equations (D.6)–(D.7) shares the same explicit β prefactor as standard DPO (Eq. (6)): only the shape factor changes from $\sigma(-\beta\Delta)$ to the Mills ratio $\phi(\beta\Delta)/\Phi(\beta\Delta)$. At $\beta\Delta \approx 0$,

$$\sigma(0) = \tfrac{1}{2}, \qquad \frac{\phi(0)}{\Phi(0)} = \sqrt{\frac{2}{\pi}}, \tag{D.8}$$

so the Gaussian branch is locally steeper by the constant factor

$$\frac{\phi(0)/\Phi(0)}{\sigma(0)} = \frac{\sqrt{2/\pi}}{1/2} = \sqrt{\frac{8}{\pi}} \approx 1.60. \tag{D.9}$$

For large negative margins, $\phi(x)/\Phi(x) \sim |x|$ grows rather than saturating, so the Thurstone loss increases quadratically, $-\log \Phi(\beta\Delta) \sim (\beta\Delta)^2/2$ as $\beta\Delta \to -\infty$, and penalizes strongly misranked pairs more aggressively than the logistic branch, whose loss grows only linearly, $-\log \sigma(\beta\Delta) \sim -\beta\Delta$, with marginal factor $\sigma(-\beta\Delta)$ bounded above by 1.

E Formal Proof of Proposition 1 and Relation to Prior Work

Novelty statement. The centered-softplus function s_c was introduced by Srinivas et al. (2022) as an activation for lightweight convex neural networks. We distinguish our contribution: we *diagnose* the β-entanglement problem in DPO (Section 3.1–3.2), show that applying s_c as a *loss normalization* removes the explicit gradient prefactor while preserving the DPO objective ordering (Proposition 1 below), extend the resulting objective to $\beta = 0$ via a limiting argument (Proposition 2 below), and provide a theoretical analysis of the Adam compensation regime (Appendix B). These results are not reported in Srinivas et al. (2022) or in prior DPO work.

Proposition 1 (Positive affine equivalence for $\beta > 0$). For $\beta > 0$,

$$\mathcal{L}^{\mathrm{norm}}_{\mathrm{DPO}}(\Delta;\beta) = \frac{1}{\beta}\,\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) - \frac{\log 2}{\beta}, \tag{E.1}$$

where $\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) = -\log \sigma(\beta\Delta)$ and $\mathcal{L}^{\mathrm{norm}}_{\mathrm{DPO}}(\Delta;\beta) = s_c(-\Delta;\beta)$. Hence,

$$\arg\min_{\Delta} \mathcal{L}^{\mathrm{norm}}_{\mathrm{DPO}}(\Delta;\beta) = \arg\min_{\Delta} \mathcal{L}_{\mathrm{DPO}}(\Delta;\beta). \tag{E.2}$$

This equality includes the case in which neither objective attains a minimum; whenever minimizers exist, they coincide. More generally, the two losses have the same ordering, and their infima correspond under the affine map above. Equivalently, subtracting the additive constant $\log 2/\beta$ from the normalized softplus loss does not change any attained minimizers:

$$\arg\min_{\Delta} s(-\Delta;\beta) = \arg\min_{\Delta} \big[s(-\Delta;\beta) - \log 2/\beta\big]. \tag{E.3}$$

Proof of Proposition 1 (algebraic identity). We verify the identity $\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}(\Delta;\beta) = \frac{1}{\beta}\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) - \frac{\log 2}{\beta}$ directly. Recall $\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) = -\log\sigma(\beta\Delta) = \log(1+e^{-\beta\Delta})$ and $\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}(\Delta;\beta) = s_c(-\Delta;\beta)$. By the definition of s_c,

$$\begin{aligned} s_c(-\Delta;\beta) &= \frac{1}{\beta}\log\frac{1+e^{-\beta\Delta}}{2} \\ &= \frac{1}{\beta}\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) - \frac{\log 2}{\beta}, \end{aligned} \tag{E.4}$$

which is the required identity. □

Proof of Proposition 1 (argmin equivalence over θ). Let Θ be the parameter space and write

$$F(\theta) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\mathcal{L}_{\mathrm{DPO}}(\Delta_\theta;\beta)\big], \quad G(\theta) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\mathcal{L}_{\mathrm{DPO}}^{\mathrm{norm}}(\Delta_\theta;\beta)\big], \tag{E.5}$$

where $\Delta_\theta(x,y_w,y_l)$ is the log-ratio gap induced by θ. From the algebraic identity in Equation (E.4),

$$G(\theta) = \frac{1}{\beta}F(\theta) - \frac{\log 2}{\beta} \qquad \forall\,\theta\in\Theta. \tag{E.6}$$

For any $\theta_1,\theta_2\in\Theta$,

$$G(\theta_1)\le G(\theta_2) \iff \frac{1}{\beta}F(\theta_1) - \frac{\log 2}{\beta} \le \frac{1}{\beta}F(\theta_2) - \frac{\log 2}{\beta} \iff F(\theta_1)\le F(\theta_2), \tag{E.7}$$

where the last equivalence uses $1/\beta > 0$. Hence the two objectives induce the same ordering (equivalently, the same total preorder) on Θ, and in particular $\arg\min_\theta G(\theta) = \arg\min_\theta F(\theta)$; if the infimum is not attained, both argmin sets are empty. □

Continuity at $\beta = 0$ (Proposition 2). For fixed $x\in\mathbb{R}$, define

$$h_x(\beta) := \log\left(\frac{1+e^{\beta x}}{2}\right), \qquad \beta\ge 0. \tag{E.8}$$

Then for $\beta > 0$,

$$s_c(x;\beta) = \frac{h_x(\beta)}{\beta}. \tag{E.9}$$

Applying this identity at $x = -\Delta$ gives

$$s_c(-\Delta;\beta) = \frac{h_{-\Delta}(\beta)}{\beta}. \tag{E.10}$$

Since $h_{-\Delta}(0) = 0$, this is the right difference quotient of $h_{-\Delta}$ at 0. Because $h_{-\Delta}$ is differentiable at $\beta = 0$,

$$\lim_{\beta\to 0^+} s_c(-\Delta;\beta) = h'_{-\Delta}(0). \tag{E.11}$$

Moreover,

$$h'_{-\Delta}(\beta) = \frac{(-\Delta)e^{-\beta\Delta}}{1+e^{-\beta\Delta}} = -\frac{\Delta}{1+e^{\beta\Delta}}, \tag{E.12}$$

so $h'_{-\Delta}(0) = -\Delta/2$. Therefore

$$\lim_{\beta\to 0^+} s_c(-\Delta;\beta) = -\frac{\Delta}{2}, \tag{E.13}$$

which matches the piecewise definition at $\beta = 0$ in Equation (14). □

Gradient at $\beta = 0$. For $\beta > 0$, $\partial s_c(-\Delta;\beta)/\partial\Delta = -\sigma(-\beta\Delta)$. At $\beta = 0$, from $s_c(-\Delta;0) = -\Delta/2$, we have $\partial s_c(-\Delta;0)/\partial\Delta = -1/2$. To verify continuity at the boundary, use that σ is globally Lipschitz with constant $1/4$:

$$|\sigma(u) - \sigma(v)| \le \tfrac{1}{4}|u - v|. \tag{E.14}$$

By direct substitution,

$$\left|\frac{\partial s_c(-\Delta;\beta)}{\partial\Delta} - \frac{\partial s_c(-\Delta;0)}{\partial\Delta}\right| = \left|-\sigma(-\beta\Delta) + \tfrac{1}{2}\right| = |\sigma(-\beta\Delta) - \sigma(0)|. \tag{E.15}$$

Applying Equation (E.14) to Equation (E.15) with $u = -\beta\Delta$ and $v = 0$ yields

$$|\sigma(-\beta\Delta) - \sigma(0)| \leq \tfrac{1}{4}|\beta\Delta|. \tag{E.16}$$

As $(\Delta, \beta) \to (\Delta_0, 0^+)$, Δ remains bounded in a neighborhood of Δ_0, so the right-hand side tends to 0. Therefore $\partial s_c(-\Delta;\beta)/\partial\Delta$ is continuous at $\beta = 0$ (from the admissible side $\beta \geq 0$), confirming that no special optimizer treatment is required at the boundary.

Hard-margin limit of normalized softplus. At the opposite scale, the same $1/\beta$ normalization has the standard hard-margin limit of softplus, which matches the soft-margin-ranking interpretation of DPO (Yang et al., 2026, Prop. 4.1):

$$\lim_{\beta\to\infty} \log\left(1 + e^{-\beta\Delta}\right)/\beta = \max\{-\Delta, 0\}. \tag{E.17}$$

Thus β controls the sharpness of a soft hinge: finite β gives a smooth transition, while $\beta \to \infty$ recovers the hard hinge boundary, consistent with the margin-ranking interpretation of DPO (Yang et al., 2026). Centering subtracts the constant $\log 2/\beta$ and therefore changes only the vertical offset: $s_c(0;\beta) = 0$ exactly, so the decision boundary remains at $\Delta = 0$. As $\beta \to \infty$, $s_c(-\Delta;\beta)$ likewise converges to $\max\{-\Delta, 0\}$.

F Bradley–Terry Scale and Log-Likelihood-Ratio View

Annotator-noise foundations. Following Sun et al. (2025), let $\Delta_r = r_{x,y_1} - r_{x,y_2}$ denote the latent utility gap and $\Delta_b = b(x, y_2, A) - b(x, y_1, A)$ the pairwise annotator-bias difference. Then

$$P(y_1 \succ y_2 \mid x) = P(\Delta_r > \Delta_b). \tag{F.1}$$

Logistic difference (Assumption 3). If the two annotator-bias terms are independently Gumbel-distributed with common scale s_a, $b(x, y_i, A) \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0, s_a)$, then their difference is $\Delta_b = b(x, y_2, A) - b(x, y_1, A) \sim \mathrm{Logistic}(0, s_a)$:

$$P(y_1 \succ y_2 \mid x) = \sigma\left(\frac{\Delta_r}{s_a}\right). \tag{F.2}$$

Gaussian difference (Assumption 4). If the two annotator-bias terms are independently Gaussian with common variance $s_a^2/2$, $b(x, y_i, A) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, s_a^2/2)$, then their difference is $\Delta_b = b(x, y_2, A) - b(x, y_1, A) \sim \mathcal{N}(0, s_a^2)$:

$$P(y_1 \succ y_2 \mid x) = \Phi\left(\frac{\Delta_r}{s_a}\right), \tag{F.3}$$

where Φ is the standard normal CDF. This parameterization absorbs the $\sqrt{2}$ factor from subtracting two independent Gaussian biases into the per-item variance, so s_a denotes the standard deviation of the pairwise bias difference, matching the role of s_a in the logistic branch. This matches the Thurstone case in Remark 4 of Sun et al. (2025) after this rescaling of the per-item variance. At $s_a = 1$, Eqs. (F.2)–(F.3) reproduce the unit-scale logistic and Gaussian curves compared in Sun et al. (2025); there Δ_b is modeled directly as $\mathrm{Logistic}(0, 1)$ or $\mathcal{N}(0, 1)$.

For an observed preference $y_w \succ y_l$ with $\Delta_r = r_w - r_l$, the corresponding pairwise reward-modeling losses are

$$\mathcal{L}_{\mathrm{BT}} = -\log\sigma\left(\frac{\Delta_r}{s_a}\right), \qquad \mathcal{L}_{\mathrm{Th}} = -\log\Phi\left(\frac{\Delta_r}{s_a}\right). \tag{F.4}$$

The shared denominator in Eq. (F.4) reflects the same pairwise-difference construction: s_a is the scale of the induced pairwise bias difference in both the logistic and Gaussian cases. It is distinct from the KL coefficient λ in Equation (1).

Connection to DPO. The optimal policy of the KL-regularized RLHF objective (1) has the closed form

$$\pi^*(y \mid x) = \frac{1}{Z(x)}\,\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{r^*(x, y)}{\lambda}\right), \tag{F.5}$$

which can be inverted to express the reward through the policy, $r^*(x, y) = \lambda \log(\pi^*(y \mid x)/\pi_{\text{ref}}(y \mid x)) + \lambda \log Z(x)$. Substituting this implicit reward into the Bradley–Terry likelihood (F.2) cancels the partition function $Z(x)$ and gives

$$P(y_w \succ y_l \mid x) = \sigma\left(\frac{\lambda}{s_a}\Delta\right) = \sigma(\beta\Delta), \qquad \Delta = \log \frac{\pi(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}. \tag{F.6}$$

Here

$$\beta := \frac{\lambda}{s_a} \tag{F.7}$$

is the effective inverse preference-noise scale in policy-log-ratio coordinates. The DPO likelihood identifies this ratio, not λ and s_a separately; interpreting either factor physically requires a fixed reward normalization. This matches comparative studies: Xu et al. (2024b) contrast PPO's explicit KL penalty with DPO, which has no explicit KL term and controls deviation only through the preference likelihood and β, and β-DPO (Wu et al., 2024) adapts it as a loss-sharpness knob. The DPO loss is the negative log-likelihood of this model over observed pairs: $\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta\Delta)$. At fixed λ, smaller s_a (more decisive annotations) increases β and sharpens the sigmoid, whereas larger s_a decreases β and yields softer preferences. At fixed s_a, changing λ changes the same effective coefficient through the reward–policy conversion.

Connection to a log-likelihood-ratio baseline. Each training example presents an observation O: "annotator chose y_w over y_l". We compare two models of the world:

- H_0 (null): both completions equally good, $r(y_w) = r(y_l)$, so $P(O \mid H_0) = \sigma(0) = \frac{1}{2}$.
- H_1 (current policy): the policy assigns log-ratio gap Δ, so $P(O \mid H_1) = \sigma(\beta\Delta)$.

The per-example likelihood ratio against this null baseline is

$$\Lambda = \frac{P(O \mid H_1)}{P(O \mid H_0)} = \frac{\sigma(\beta\Delta)}{1/2} = 2\,\sigma(\beta\Delta). \tag{F.8}$$

Taking logarithms and recalling $\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta\Delta)$,

$$\log \Lambda = \log 2 - \mathcal{L}_{\text{DPO}}. \tag{F.9}$$

Hence **minimizing the DPO loss is equivalent to maximizing** $\log \Lambda$: the policy is trained to make each observed preference as much more likely than the "equal-quality" baseline as possible. Subtracting $\log 2$ removes the random-choice baseline, and $\mathcal{L}_{\text{DPO}} - \log 2 = -\log \Lambda$ is the negative log-likelihood ratio relative to chance. Dividing by the effective coefficient gives $(\mathcal{L}_{\text{DPO}} - \log 2)/\beta = (s_a/\lambda)(\mathcal{L}_{\text{DPO}} - \log 2)$. This normalization removes the direct β prefactor from the marginal gradient. It equals a conversion to annotator-noise units only under an explicitly fixed reward/KL normalization.

Note also that Δ itself is a log-likelihood ratio comparing the current policy to the reference:

$$\Delta = \underbrace{\log \frac{\pi(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}}_{\text{log-LR for } y_w} - \underbrace{\log \frac{\pi(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}}_{\text{log-LR for } y_l}. \tag{F.10}$$

Thus $\beta\Delta = (\lambda/s_a)\Delta$ is a scale-controlled evidence score: how strongly the current policy departs from the reference in the direction of the observed preference. In ML terms, $\sigma(\beta\Delta)$ is a soft preference score. At fixed λ, varying the annotator-noise scale gives, for $\Delta \neq 0$ in the first limit,

$$\begin{aligned} &s_a \to 0\ (\beta \to \infty): \quad \sigma(\beta\Delta) \to \mathbf{1}[\Delta > 0] \quad \text{(hard preference boundary, near-deterministic labels)}, \\ &s_a \to \infty\ (\beta \to 0): \quad \sigma(\beta\Delta) \to \tfrac{1}{2} \quad \text{(maximally noisy labels, no discrimination)}. \end{aligned} \tag{F.11}$$

At $\Delta = 0$, the first limit is $1/2$. Because only $\beta = \lambda/s_a$ appears in the DPO likelihood, a sweep over β varies an effective inverse scale; it cannot by itself distinguish a change in annotator noise from a change in reward/KL normalization.

Implication for the disentanglement. The two roles of the effective coefficient $\beta = \lambda/s_a$ identified in the main paper are:

- *Statistical inverse scale* $\beta = \lambda/s_a$: the ratio of the KL/reward scale to the annotator-noise scale sets the width of the soft decision boundary in policy-log-ratio coordinates.
- *Optimization scale* β (gradient multiplier): an unintended coupling to effective step size that arises from the standard DPO parameterization.

The centered-softplus reformulation removes the second role while preserving the first. It lets the practitioner tune the effective statistical coefficient β and the optimizer independently. Separating λ from s_a additionally clarifies the modeling interpretation, but the two quantities require an external normalization or additional information to be identified individually.

G Saturation Thresholds: Where Training Stops as a Function of β

Setup. We model effective convergence with a residual stopping tolerance $\tau > 0$: updates on a preference pair become negligible once the magnitude of the marginal gradient drops below this noise- and optimizer-scale threshold. The same τ can be estimated from the shape of a β-sweep, from $\beta_{\text{peak}}/4.59$, or from a drift–noise crossover. Under the scalar closure these procedures target the same residual stopping scale; Jacobian evolution, stochastic fluctuations, pair heterogeneity, and minibatch Gram terms may cause their empirical estimates to differ. Finite-horizon limitation is treated separately through the cumulative budget I in Appendix L. The invariant quantity behind this tolerance is the per-step drift of the margin. A parameter step $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ with $\nabla_\theta \mathcal{L} = \ell'(\Delta)\, \nabla_\theta \Delta$ changes the scalar margin by $\Delta \leftarrow \Delta + \eta\, g^2\, |\ell'(\Delta)|$ to first order, where $g = \|\nabla_\theta \Delta\|_2$: the factor g^2 is the squared Jacobian of the margin, not a square of the marginal loss derivative (Appendix L, Equation (L.4)). A run stalls once this drift falls below a noise-floor scale $\varepsilon > 0$ of the stochastic margin increments (Appendix M),

$$\eta\, g^2\, \beta\, \sigma(-\beta\Delta) \;\lesssim\; \varepsilon.$$

Under the frozen initial-Jacobian closure, let $g_0 := g(\theta_0)$, choose a reference learning-rate scale η_{ref}, and define

$$\tau := \frac{\varepsilon}{\eta_{\text{ref}}\, g_0^2}. \tag{G.1}$$

For a constant learning rate, $\eta_{\text{ref}} = \eta$; for the linear-decay schedules considered below, we use $\eta_{\text{ref}} = \eta_{\max}$. Across a learning-rate sweep, approximate constancy of ε for a fixed model, dataset, batch construction, optimizer, and schedule family is a modeling assumption. Under this assumption, $\tau(\eta_{\text{ref}}) \propto \eta_{\text{ref}}^{-1}$. Equation (G.1) provides an alternative operational estimate of the same tolerance τ, rather than defining a second noise-specific threshold. Under the learning-rate rescaling $\eta \propto 1/\beta$ of Appendix I, the same drift criterion instead reduces to a condition on $\sigma(-\beta\Delta)$ alone (used there). This is a coarse description of where training stops, not an exact optimization statement; it turns the claim that β controls saturation into scaling predictions that can be checked against the reported runs.

Normalized objective: a clean $1/\beta$ law. For the centered-softplus objective, $|\partial \mathcal{L}_{\text{DPO}}^{\text{norm}}/\partial\Delta| = \sigma(-\beta\Delta)$, so the stopping condition $\sigma(-\beta\Delta) = \tau$ gives the saturation margin in Equation (15),

$$\Delta^* = \frac{1}{\beta} \log \frac{1-\tau}{\tau}. \tag{G.2}$$

The tolerance enters only through the numerator, so for any fixed $\tau < 1/2$ the product $\beta\, \Delta^*$ is a constant independent of β. This is a statement about the saturation margin, not about $\beta \cdot \text{KL}$: per-token KL is only an empirical co-moving readout, and we do not claim that $\beta \cdot \text{KL}$ is constant (Appendix H). Unlike Equation (8), there is no dead zone and no intermediate peak: Δ^* is strictly monotone in $1/\beta$.

Standard DPO: dead zone, drifting product, and a peak. Section 3.3 states the main result: for the classical loss, $|\partial \mathcal{L}_{\text{DPO}}/\partial\Delta| = \beta\, \sigma(-\beta\Delta)$, and the stopping condition $\beta\, \sigma(-\beta\Delta) = \tau$ yields Equation (8). We reproduce the derivation here. Setting $\beta\, \sigma(-\beta\Delta) = \tau$ and solving for Δ gives

$$\Delta^*_{\text{DPO}} = \frac{1}{\beta} \log\!\left(\frac{\beta}{\tau} - 1\right), \qquad \beta > 2\tau. \tag{G.3}$$

This scalar is the threshold-limited margin of the idealized one-pair stopping model—not a prediction of the pair-averaged validation margin, not a finite-time solution of the training dynamics, and not a quantity recovered by re-evaluating frozen weights on the training pairs. Aggregated empirical margin statistics need not coincide with this one-pair scale and may lie on either side of it because pairs have different Jacobians, alignments, and effective stopping thresholds. We therefore compare primarily the β-shape and allow a free vertical scale (Section G.1). Three regimes follow (as summarized in Section 3.3).

- *Dead zone* ($\beta \le 2\tau$): in the one-pair monotone closure initialized at $\Delta = 0$, the marginal-gradient magnitude starts at $\beta/2$ and decreases as the modeled margin moves in the positive direction. If $\beta/2 \le \tau$, the positive drift is below the effective tolerance from the start and is predicted to remain suppressed. This is not an absolute statement for minibatch training, where cross-pair updates and stochastic fluctuations can still produce movement. The normalized objective has no analogous small-β dead zone in this model: its initial marginal gradient is $1/2$ for every β.
- *No clean scaling* ($\beta > 2\tau$): the product $\beta\,\Delta^*_{\mathrm{DPO}} = \log(\beta/\tau - 1)$ grows logarithmically in β, in contrast to the constant product in Equation (15).
- *Non-monotone peak and right-hand saturation*: maximizing Equation (8) over β gives Equation (9); the detailed derivation is given below. The threshold model therefore predicts this non-monotone shape for the saturation margin. In the HelpSteer3 SGD sweep of Figure 2, per-token KL co-moves with the mean margin over a wide β range (a left dead zone, an intermediate peak, and a decaying right slope), which we treat as an empirical observation rather than as a general mapping from Δ to KL.

Derivation of the peak location. Differentiating Equation (8) with respect to β,

$$\frac{\mathrm{d}\Delta^*_{\mathrm{DPO}}}{\mathrm{d}\beta} = \frac{1}{\beta^2}\left[\frac{\beta/\tau}{\beta/\tau - 1} - \log\left(\frac{\beta}{\tau} - 1\right)\right], \tag{G.4}$$

so the stationarity condition is

$$\frac{\beta/\tau}{\beta/\tau - 1} = \log\left(\frac{\beta}{\tau} - 1\right). \tag{G.5}$$

Substituting $v = \beta/\tau - 1$ turns the left-hand side into $(v+1)/v = 1 + 1/v$, giving the transcendental equation

$$1 + \frac{1}{v} = \log v. \tag{G.6}$$

The left-hand side decreases from $+\infty$ to 1 and the right-hand side increases from $-\infty$, so Equation (G.6) has a unique root; numerically $v^* \approx 3.591$ (at $v = 3.591$: $\log v \approx 1.2784$ and $1 + 1/v \approx 1.2785$). Hence

$$\beta_{\mathrm{peak}} = \tau\,(1 + v^*) \approx 4.59\,\tau \approx 4.6\,\tau, \tag{G.7}$$

which is Equation (9). The second derivative is negative at v^* (the bracket above changes sign from positive to negative), so this stationary point is the unique maximum. The peak value itself has a compact form: using $\log v^* = 1 + 1/v^*$,

$$\Delta^*_{\mathrm{DPO}}(\beta_{\mathrm{peak}}) = \frac{\log v^*}{\tau(1 + v^*)} = \frac{1 + 1/v^*}{\tau(1 + v^*)} = \frac{1}{\tau\, v^*} \approx \frac{0.278}{\tau}, \tag{G.8}$$

so the maximal threshold-limited margin in this one-pair model is set entirely by the tolerance τ. For $\beta \gg \beta_{\mathrm{peak}}$, $\Delta^*_{\mathrm{DPO}} \approx \log(\beta/\tau)/\beta$ decays toward zero. These two identities are linked by $\beta_{\mathrm{peak}}\,\Delta^*_{\mathrm{DPO}}(\beta_{\mathrm{peak}}) = \log v^* \approx 1.278$: they are two writings of the same one-dimensional threshold model, not independent estimators.

Empirical illustration. Figure 2 overlays Equation (8) on a dense HelpSteer3 SGD sweep of the same protocol as Figure 1. The four-point subset in Section 3.1 is already non-monotone: seed-averaged final per-token KL $\{0.04,\ 0.35,\ 0.83,\ 0.54\}$ peaks at the nearest sampled cell $\beta = 2 \times 10^{-3}$. On the dense grid of Figure 2 and Table G1, the minimum NLL, maximum KL, and maximum mean $\overline{\Delta}$ occur between $\beta = 2.5 \times 10^{-3}$ and 3×10^{-3}. The ordering anomaly noted in Section 3.1 (the $\beta = 0.002$ trajectory overtaking $\beta = 0.01$) is thus a predictable consequence of the threshold model rather than an instability.

G.1 Estimating τ from β_{peak} and $\Delta(\beta)$

Two readings of τ. Equations (G.7)–(G.8) suggest two practical ways to recover τ from a β-sweep under standard DPO at fixed learning rate. Figure 2 overlays both resulting curves:

1. *From the peak location* (Eq. (G.7)): $\tau = \beta_{\text{peak}}/(1+v^*)$. This is our primary estimate: β_{peak} is read jointly from validation NLL, KL, and preference-margin statistics.
2. *From the $\Delta(\beta)$ curve*: either plug a height into Eq. (G.8), or fit the shape of the equilibrium margin $\Delta^*_{\text{cl}}(\beta;\tau) = \frac{1}{\beta}\log(\beta/\tau - 1)$ ($\beta > 2\tau$) to an empirical margin statistic across β.

The second route is a consistency check, not a replacement for the first: the one-dimensional saturating scalar Δ^*_{cl} is not directly observed.

Why Δ^*_{cl} is not observed. The threshold derivation treats a single preference pair with scalar margin Δ. On validation we instead see a heavy-tailed distribution of margins over ~ 1920 pairs. Empirical summaries—mean $\overline{\Delta}$, median, and upper quantile Δ_{p95}—are therefore not interchangeable with Δ^*_{cl}. In particular, $\overline{\Delta}$ is typically several times smaller than the one-pair peak height implied by Eq. (G.8), so a naive G.7-style inversion of the mean is not a valid τ reading. The upper quantile Δ_{p95} is closer in magnitude and can be used cautiously for height-based checks; the preferred Δ-based estimator below uses the *shape* of $\overline{\Delta}(\beta)$ with a free vertical scale.

Uncertainty in locating β_{peak}. On a discrete β grid the exact peak is often ambiguous: neighboring cells can agree within seed noise for mean Δ and min NLL, while the median may peak one cell away. It is therefore useful to keep several β points near the apparent maximum and to report a mid-cell anchor with a grid half-width. On HelpSteer3 / Qwen3-4B (SGD, $\text{lr} = 2\cdot 10^{-4}$, batch 24, 6 epochs; Figure 2), mean Δ and min NLL at $\beta = 0.0025$ and $\beta = 0.003$ are tied within seed error, while the median is slightly higher at 0.003. We take

$$\beta^{\text{mid}}_{\text{peak}} = 0.00275, \qquad \tau^{\text{mid}}_{\text{G.6}} = \frac{0.00275}{1+v^*} \approx 5.99\cdot 10^{-4},$$

with grid uncertainty given by the neighboring cells 0.0025 and 0.003:

$$\tau_{\text{G.6}} = (5.99 \pm 0.54)\cdot 10^{-4} \quad (\pm \sim 9\% \text{ relative to mid}),$$

since $\delta\tau/\tau = \delta\beta_{\text{peak}}/\beta_{\text{peak}}$. With this τ, the left shaded region of Figure 2 is the dead zone $\beta \le 2\tau \approx 1.2\times 10^{-3}$: it contains $\beta = 5\times 10^{-4}$ (near-zero KL) and the edge point $\beta = 10^{-3}$. On the right, the rough finite-horizon scale $\beta_{\text{r}} \sim 10^{-2}$ (the scalar proxy gives 2.7×10^{-2}; Appendix L) marks the broad region where the $\log(\beta/\tau)/\beta$ ceiling is already decreasing; by $\beta = 0.2$–0.3 the measured KL has collapsed.

Why validation NLL uses $\beta = 1$. The training DPO loss is $-\log\sigma(\beta\Delta)$ with the run's own β. That quantity is not comparable across a β-sweep: it depends on the product $\beta\Delta$, so a smaller training β pushes the loss toward $\log 2$ even when the policy ranks chosen vs. rejected more strongly (Section 3.4). The left panel of Figure 2 therefore reports a different diagnostic, *validation pair NLL with $\beta = 1$*:

$$\mathbb{E}\big[-\log\sigma(z_w - z_l)\big] = \mathbb{E}\big[-\log\sigma(\Delta_\theta)\big],$$

where $z_w - z_l = \Delta_\theta = \log\pi_\theta(y_w \mid x) - \log\pi_\theta(y_l \mid x)$ on held-out pairs, and the training β of that run is not inserted into the sigmoid. The pairwise NLL is decreasing in each individual policy gap $\Delta_{\theta,i} = \Delta_i + \Delta_{\text{ref},i}$, with $\Delta_{\text{ref},i}$ fixed by the reference model and data. Thus a pointwise increase, or a suitable stochastic ordering, of the margin distribution decreases aggregate NLL; an increase in the mean margin alone is insufficient. Empirically, the mean-margin peak and minimum validation NLL occur at nearly the same location in Figure 2. We therefore use the following relation only as a descriptive fit, not as a consequence of the scalar saturation model:

$$\text{NLL}(\beta) \approx a - b\,\Delta^*_{\text{cl}}(\beta;\tau), \qquad b > 0, \tag{G.9}$$

where the intercept a and slope b are obtained by least squares on the active region $\beta > 2\tau$ (or with an empirical margin statistic in place of Δ^*_{cl}). These two constants absorb the unknown map from the one-pair ceiling to the pair-averaged NLL; they are not estimators of τ. The tolerance continues to be read from β_{peak} or from the shape of $\overline{\Delta}(\beta)$ as below.

Table G1: HelpSteer3 / Qwen3-4B standard DPO (SGD, $\mathrm{lr} = 2 \cdot 10^{-4}$): min validation pair NLL (with $\beta = 1$); per-token MC KL, $\beta \cdot \mathrm{KL}$, and validation margin quantiles at the same epoch (rounded). Same sweep as Figure 2; uncertainties are mean±std over seeds.

β	n	min NLL	KL	$\beta \cdot \mathrm{KL}$	mean Δ	med. Δ	$p95$ Δ
0.0005	4	13.93 ± 0.03	0.042 ± 0.007	$2.1 \cdot 10^{-5}$	16.2 ± 0.9	10.2 ± 0.5	112 ± 7
0.001	4	13.46 ± 0.04	0.353 ± 0.004	$3.5 \cdot 10^{-4}$	28.1 ± 0.3	18.0 ± 0.0	177 ± 1
0.0015	3	11.45 ± 0.15	0.611 ± 0.041	$9.2 \cdot 10^{-4}$	73.3 ± 6.8	40.2 ± 2.0	431 ± 44
0.002	4	10.35 ± 0.04	0.833 ± 0.014	$1.7 \cdot 10^{-3}$	122 ± 4	69.5 ± 1.9	607 ± 13
0.0025	**5**	9.92 ± 0.14	$\mathbf{1.00 \pm 0.02}$	$2.5 \cdot 10^{-3}$	$\mathbf{152 \pm 6}$	87.1 ± 4.4	$\mathbf{708 \pm 30}$
0.003	4	$\mathbf{9.91 \pm 0.07}$	0.994 ± 0.048	$3.0 \cdot 10^{-3}$	152 ± 7	$\mathbf{88.4 \pm 2.1}$	707 ± 24
0.005	2	10.26 ± 0.01	0.842 ± 0.036	$4.2 \cdot 10^{-3}$	120 ± 0	72.1 ± 0.2	536 ± 0
0.0075	2	11.07 ± 0.05	0.659 ± 0.008	$4.9 \cdot 10^{-3}$	90.6 ± 0.7	58.0 ± 0.0	399 ± 4
0.01	4	11.55 ± 0.08	0.540 ± 0.008	$5.4 \cdot 10^{-3}$	75.1 ± 0.6	49.2 ± 0.9	321 ± 6
0.02	2	12.61 ± 0.01	0.290 ± 0.024	$\mathbf{5.8 \cdot 10^{-3}}$	45.4 ± 0.0	33.0 ± 1.4	177 ± 1
0.04	2	13.23 ± 0.01	0.130 ± 0.007	$5.2 \cdot 10^{-3}$	29.4 ± 0.0	24.0 ± 0.0	106 ± 0
0.075	2	13.36 ± 0.02	0.066 ± 0.001	$4.9 \cdot 10^{-3}$	21.9 ± 0.3	19.0 ± 1.4	68 ± 0
0.2	1	13.67	0.019	$3.8 \cdot 10^{-3}$	14.9	14.0	40
0.3	1	13.80	0.010	$2.9 \cdot 10^{-3}$	12.3	12.0	32

Validation NLL / KL and margin statistics. Table G1 reports the numbers plotted in Figure 2: for each run we take the epoch of minimum validation pair NLL (with $\beta = 1$), and at that same epoch report per-token MC KL, $\beta \cdot \mathrm{KL}$, and validation margin quantiles $\Delta = \Delta_\theta - \Delta_{\mathrm{ref}}$ (full validation set). Uncertainties are mean±std over seeds when $n > 1$, matching the error bars in Figure 2; far-right points $\beta \in \{0.2, 0.3\}$ are single-seed (seed 43). Bold entries mark selected extrema: maximum KL, mean Δ, and $p95$ Δ at $\beta = 0.0025$; minimum NLL and maximum median Δ at $\beta = 0.003$; and maximum $\beta \cdot \mathrm{KL}$ at $\beta = 0.02$.

Shape+scale fit of $\overline{\Delta}(\beta)$. On the active region $\beta > 2\tau$ we relate the observed mean to the one-pair formula by

$$\overline{\Delta}(\beta) \approx s \cdot \Delta^*_{\mathrm{cl}}(\beta; \tau). \tag{G.10}$$

In practice we form a rough peak estimate $\beta^{\mathrm{mid}}_{\mathrm{peak}} = 0.00275$, set $\tau_{\mathrm{rough}} = \beta^{\mathrm{mid}}_{\mathrm{peak}}/(1 + v^*)$, and fit on all sweep points with $\beta > 2\tau_{\mathrm{rough}}$ (here $\beta \in \{0.0015, \ldots, 0.075, 0.2, 0.3\}$; $n = 12$). Parameters are obtained by log-RMSE in two steps. First, for each candidate τ on a grid, set $\mathrm{pred}(\beta) = \Delta^*_{\mathrm{cl}}(\beta; \tau)$. Second, the optimal scale at that τ is the closed-form log-space least-squares value

$$\log s(\tau) = \frac{1}{n} \sum_i \Big(\log \overline{\Delta}(\beta_i) - \log \mathrm{pred}(\beta_i) \Big), \qquad s(\tau) = e^{\log s(\tau)}. \tag{G.11}$$

We then choose τ to minimize

$$\frac{1}{n} \sum_i \Big(\log \overline{\Delta}(\beta_i) - \log\big(s(\tau)\, \mathrm{pred}(\beta_i)\big) \Big)^2. \tag{G.12}$$

Thus τ is read from the *shape* of the curve and $s(\tau)$ from its height. Once τ is obtained, the implied peak $\beta^{\mathrm{pred}}_{\mathrm{peak}} = \tau(1 + v^*)$ follows from Eq. (G.7) without reusing Δ. Figure G1 shows the objective versus τ; the minimum is at $\tau_{\mathrm{fit}} \approx 6.37 \cdot 10^{-4}$ with $s \approx 0.35$. The two overlays in Figure 2 are this shape+scale curve and the G.6 curve from $\beta^{\mathrm{mid}}_{\mathrm{peak}}$.

Comparison of τ estimates. Table G2 compares G.6 anchors on the peak grid cell, the shape+scale fit, and a height-based check that inverts Eq. (G.8) using Δ_{p95} at $\beta = 0.0025$ (not the mean). Upright entries are observed or fitted inputs; italic entries are quantities derived by the indicated map ($\beta \to \tau$ or $\tau \to \beta^{\mathrm{pred}}_{\mathrm{peak}}$).

Takeaway. The primary estimate remains Eq. (G.7) from β_{peak}, with an explicit grid uncertainty when the peak cell is not unique. Fitting $\overline{\Delta} \approx s\, \Delta^*_{\mathrm{cl}}$ on $\beta > 2\tau_{\mathrm{rough}}$ recovers a compatible τ (here $+6\%$ relative to mid and -2.5% relative to the $\beta = 0.003$ G.6 anchor), providing an internal consistency check that the shape of the mean margin tracks the threshold model once a height scale is freed; the two curves in Figure 2 nearly coincide. Because both estimates use the same β-sweep, their

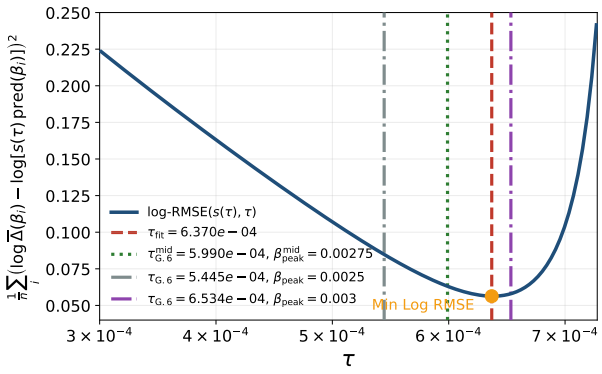

Figure G1: Shape+scale log-RMSE versus τ for mean validation Δ on $\beta > 2\tau_{\mathrm{rough}}$ (HelpSteer3 / Qwen3-4B, including $\beta = 0.2, 0.3$). Vertical lines: τ_{fit} and G.6 anchors at $\beta_{\mathrm{peak}} \in \{0.0025, 0.00275, 0.003\}$.

Table G2: Consistency of τ estimates on HelpSteer3 / Qwen3-4B, corresponding to the two overlays in Figure 2. Reference: $\tau^{\mathrm{mid}}_{\mathrm{G.6}}$ at $\beta^{\mathrm{mid}}_{\mathrm{peak}} = 0.00275$. Shape+scale uses $\beta > 2\tau_{\mathrm{rough}}$ with $\tau_{\mathrm{rough}} = \tau^{\mathrm{mid}}_{\mathrm{G.6}}$.

method	map	τ	β_{peak}	$\tau/\tau^{\mathrm{mid}}_{\mathrm{G.6}}$
G.6, mid $[0.0025, 0.003]$	$\beta \to \tau$	$5.99 \cdot 10^{-4}$	0.00275	1.00
G.6, lower grid cell	$\beta \to \tau$	$5.45 \cdot 10^{-4}$	0.0025	0.91
G.6, upper grid cell	$\beta \to \tau$	$6.53 \cdot 10^{-4}$	0.003	1.09
shape+scale, mean, $\beta > 2\tau_{\mathrm{rough}}$	$\tau \to \beta$	$6.37 \cdot 10^{-4}$	0.00292	1.06
Eq. (G.8) on Δ_{p95} at $\beta = 0.0025$	$\Delta_{p95} \to \tau \to \beta$	$3.94 \cdot 10^{-4}$	0.00181	0.66

agreement is an internal consistency check rather than an out-of-sample validation. A height-only inversion of Eq. (G.8) should use an upper quantile such as Δ_{p95}, not the raw mean.

The same overlays on a second backbone show that Equation (G.3) can miss the vertical scale while still locating the left dead zone relative to the peak. Figure G2 is a dense standard-DPO β sweep on Orca-DPO / Phi-4-mini 3.8B (SGD, batch size 22, six epochs, seed 42) at $\eta \in \{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}\}$. Metrics are taken at the integer epoch of minimum validation pair NLL (with $\beta = 1$), as in Figure 2. A shape+scale fit of $\overline{\Delta}(\beta) \approx s\,\Delta^*_{\mathrm{cl}}(\beta; \tau)$ on $\beta > 2\tau$ gives $\tau = 1.45 \times 10^{-3}, 6.94 \times 10^{-4}$ and 3.96×10^{-4} at the three learning rates; the G.6 readings from the observed $\overline{\Delta}$ peaks ($\beta = 0.006$, 0.0035 and 0.0017) are $\tau = 1.31 \times 10^{-3}$, 7.62×10^{-4} and 3.70×10^{-4}. Each doubling of η from 5×10^{-5} to 10^{-4} to 2×10^{-4} roughly halves both readings, and the 2τ band and the peak shift left by the same factor, as required by Equation (G.1) when ε and g_0 are held fixed ($\tau \propto \eta^{-1}$). The implied peaks $\beta_{\mathrm{peak}} = \tau_{\mathrm{fit}}(1 + v^*)$ sit at 0.0066, 0.0032 and 0.0018, next to the empirical maxima, and the shaded band $\beta \leq 2\tau$ still covers the suppressed small-β cells. The same τ does not fix the height: the scaled one-pair curve understates the Δ peak at $\eta = 2 \times 10^{-4}$ and cannot follow the sharp KL spike near $\beta = 0.0011$. Thus Equation (G.3) is used here as a shape template for the pair $(2\tau, \beta_{\mathrm{peak}})$, not as a calibrated amplitude law. The two-dimensional centered-softplus counterpart of these slices, on Phi-4-mini with the same seed, is in Appendix H: substituting $\tau(\eta_{\mathrm{norm}})$ into Equation (15) gives Equation (H.2).

Scope and caveats. First, the threshold model predicts the β-shape of the saturation margin, and pair NLL is monotone in that gap via Equation (G.9). We do not treat per-token KL as a general monotone function of the pair margin. In the HelpSteer3 SGD sweep the two quantities co-move over a wide range, which is why Figure 2 reports both, but we do not fit a global KL–Δ law or proportionality constants. Second, the threshold model applies only when training reaches saturation before the epoch budget ends. For the very small β values in Figure C1, the implied Δ^*_{DPO} is too large to reach in practice; those runs stop early because updates are weak and Adam partially compensates (Appendix B), not because the loss has saturated. Third, under a decaying learning-rate schedule, τ is only a coarse effective tolerance, so we interpret its magnitude and the functional form, not

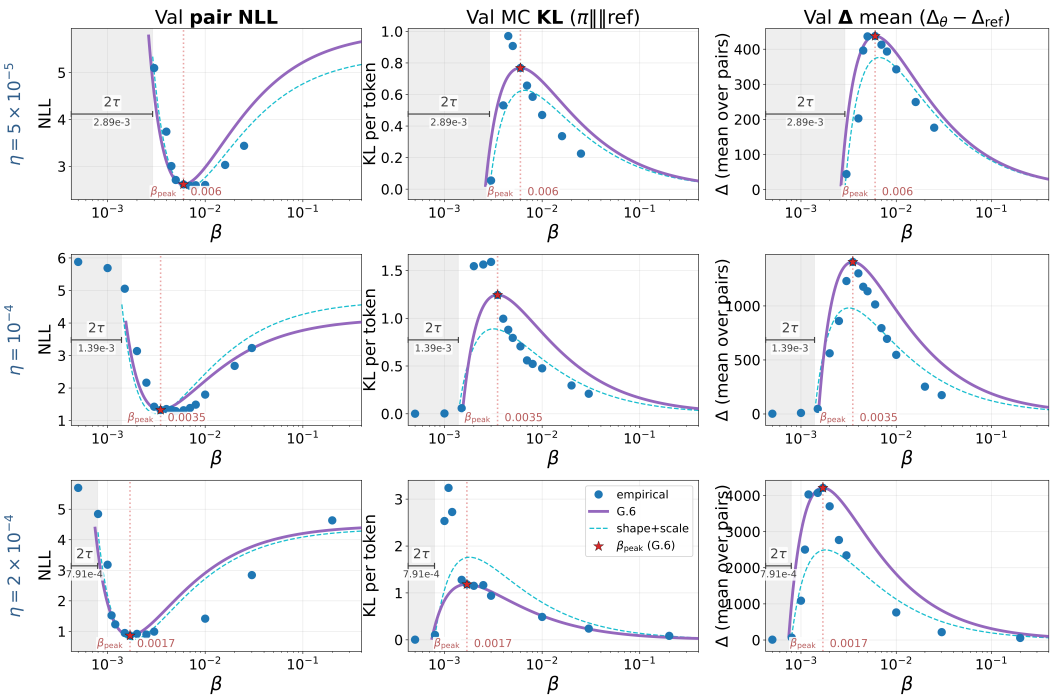

Figure G2: Standard DPO on Orca-DPO / Phi-4-mini 3.8B (SGD, batch size 22, six epochs, seed 42). Rows: learning rate $\eta = 5 \times 10^{-5}$ (top), $\eta = 10^{-4}$ (middle) and $\eta = 2 \times 10^{-4}$ (bottom). Points: validation pair NLL (with $\beta = 1$), per-token Monte Carlo KL, and mean margin $\overline{\Delta} = \Delta_\theta - \Delta_{\text{ref}}$, at the integer epoch of minimum NLL. Solid: G.6 overlay of Equation (G.3) with $\tau = \beta_{\text{peak,obs}}/(1+v^*)$ (1.31×10^{-3}, 7.62×10^{-4} and 3.70×10^{-4}), pinned through the empirical $\overline{\Delta}$ peak. Dashed: the same formula with the shape+scale $\tau_{\text{fit}} = 1.45 \times 10^{-3}$, 6.94×10^{-4} and 3.96×10^{-4}. The shaded band is $\beta \le 2\tau_{\text{fit}}$; the star is the observed β_{peak}. Each doubling of η halves τ, as predicted by Equation (G.1).

a precise fit. Finally, the extra β prefactor in standard DPO (Equation (8)) is a fixed-learning-rate effect: rescaling the learning rate by $1/\beta$ (Section 4) or operating Adam in its cancellation regime (Appendix B) largely removes it and restores the cleaner $1/\beta$ scaling of the normalized objective (Equation (15)). Appendix L complements the left dead zone $\beta \le 2\tau$ with an order-of-magnitude right-hand horizon scale obtained from a scalar margin closure. On the HelpSteer3 SGD sweep, the early weighted-Jacobian proxy gives $\beta_{\text{r}} \approx 0.027$, but time variation and cross-pair Gram terms limit this estimate to a broad $\sim 10^{-2}$ region rather than a calibrated cutoff.

H Orca (β, η) Landscape under Centered Softplus

This appendix records the construction of Figure 5 in Section 4.3. The main-text figure shows only $\log_{10}(1 + \overline{\Delta})$; Figure H1 below is the full interpolant, including validation pair NLL and per-token KL, and Figure H2 is the corresponding gradient field of the margin.

Setup. All cells use the centered-softplus objective on Orca-DPO with Phi-4-mini 3.8B, SGD, batch size 22, and six epochs. Checkpoint, LoRA configuration, and dataset splits are in Appendix N. Each cell is one training trajectory (seed 42); the interpolant below pools neighboring cells rather than repeating seeds at a fixed (β, η). The scanned window is the rectangle $\beta \in [2 \times 10^{-4}, 10^{-2}]$, $\eta \in [5 \times 10^{-5}, 7 \times 10^{-4}]$.

Two views of one window. Because $\nabla_\theta \mathcal{L}^{\text{norm}}_{\text{DPO}} = \nabla_\theta \mathcal{L}_{\text{DPO}}/\beta$ for $\beta > 0$, equal SGD updates require $\text{lr}_{\text{norm}} = \beta\, \text{lr}_{\text{standard}}$, or equivalently $\eta_{\text{norm}} = \beta\eta$. The left column of each figure plots

this rectangle in the (β, η) coordinates of standard DPO; the right column is the same set of runs in $(\beta, \eta_{\mathrm{norm}})$. Panel titles "standard DPO loss" therefore name the *axes*, not a second training objective: every cell is trained with centered softplus.

Grid. We start from a 5×5 log-spaced vertex lattice

$$\beta_k = 2 \times 10^{-4} \cdot 50^{k/4}, \qquad \eta_\ell = 5 \times 10^{-5} \cdot 14^{\ell/4}, \qquad k, \ell \in \{0, 1, 2, 3, 4\}, \tag{H.1}$$

add the 4×4 cell centers $(\beta_{k+1/2}, \eta_{\ell+1/2})$ (15 completed), and then place local probes around ridges and NLL kinks. The interpolant uses 112 completed cells (25 vertices, 15 completed centers, 72 probes).

Metrics. At each cell we take the minimum validation pair NLL over integer epochs 1–6 (initialization excluded) and the mean margin $\overline{\Delta}$ and per-token Monte Carlo KL at that epoch. The plotted Δ field is $\log_{10}(1 + \overline{\Delta})$.

Interpolant. The 112 cells are an irregular design. We interpolate them with three independent Gaussian-process regressors—one each for NLL, KL, and $\log_{10}(1 + \overline{\Delta})$—in $(\log_{10} \beta, \log_{10} \eta_{\mathrm{norm}})$, using scikit-learn's `GaussianProcessRegressor` (Pedregosa et al., 2011): a separate scalar GP per metric, not a multi-output kernel. Each covariance is an anisotropic Matérn kernel of smoothness $\nu = 2.5$ (Rasmussen & Williams, 2006), times a constant amplitude, plus a white-kernel nugget; length-scales, amplitude, and nugget are chosen by maximizing the log marginal likelihood (four restarts). A diagonal observation variance, inflated when the selected epoch is not 6, keeps the posterior mean from interpolating the cells. That mean is plotted on a log–log grid and masked to a slight expansion of the convex hull of the design (Virtanen et al., 2020) and to the scanned window. The left column only changes the plotted vertical axis ($\eta = \eta_{\mathrm{norm}}/\beta$); it is not a second fit. White markers are the 112 cells. The Matérn prior and the nugget let neighboring cells inform one another, so spatially rough, seed-idiosyncratic fluctuations are partially absorbed as observation noise rather than drawn as geometry. With one seed per cell this spatial pooling therefore smooths part of the Monte Carlo error of a raw scatter; it is not a multi-seed interval at a fixed (β, η).

Dead-zone overlays. Equation (15) is the saturation law at fixed τ. Substituting $\tau(\eta_{\mathrm{norm}}) = \varepsilon/(\eta_{\mathrm{norm}} g_0^2)$ from Equation (G.1) eliminates τ:

$$\Delta^* = \frac{1}{\beta} \log\Big(\frac{\eta_{\mathrm{norm}} g_0^2}{\varepsilon} - 1 \Big), \qquad \eta_{\mathrm{norm}} > \frac{2\varepsilon}{g_0^2}. \tag{H.2}$$

Now β appears only in the prefactor and η_{norm} only inside the logarithm. White dashed curves are the floor $\eta_{\mathrm{norm}} \lesssim 2\varepsilon/g_0^2$ (the standard-DPO locus $2\tau(\eta)$ from Equation (7) when plotted in (β, η)), and colored dashed curves are $\beta_{\mathrm{peak}} = 4.59\tau$. For standard DPO the same closure gives $\beta \le 2\tau(\eta) = 2\varepsilon/(\eta g_0^2)$, hence a diagonal in the left panels; β remains inside the logarithm, $\Delta^*_{\mathrm{DPO}} = (1/\beta) \log(\beta \eta g_0^2/\varepsilon - 1)$. These overlays use a shared (ε, g_0) transferred across learning rates; they are predicted loci, not a fit to the 112 cells. Fixed-η slices of *standard* DPO on the same Orca-DPO / Phi-4-mini setting are in Figure G2.

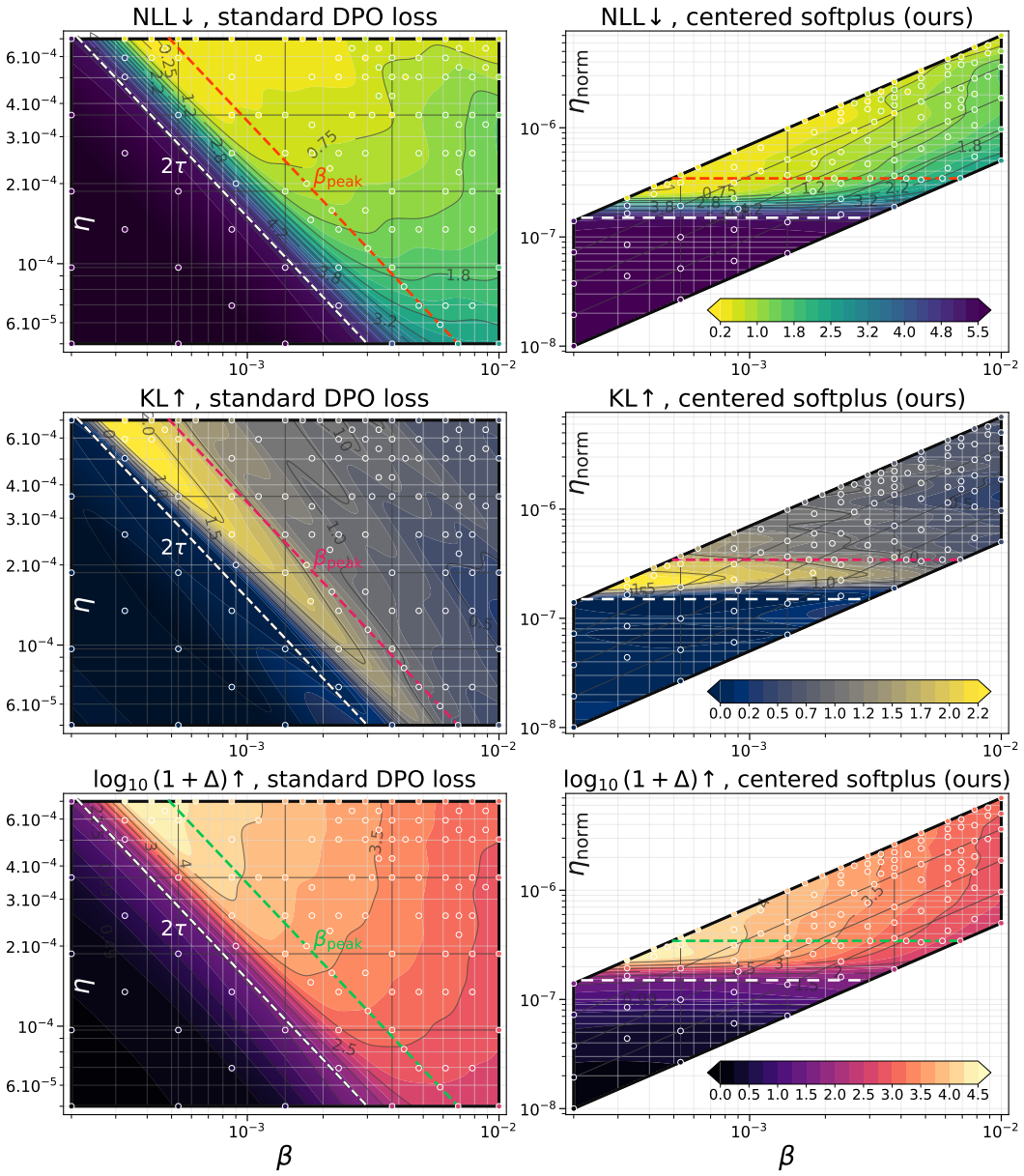

Figure H1: Full interpolant for the Orca centered-softplus grid of Figure 5 (Phi-4-mini 3.8B, SGD, batch size 22, six epochs, seed 42). Rows: validation pair NLL↓ (minimum over epochs 1–6), per-token KL↑, and $\log_{10}(1+\overline{\Delta}) \uparrow$ at that epoch. Left: (β, η). Right: $(\beta, \eta_{\mathrm{norm}})$. Both columns are the same 112 runs. White dashed: 2τ. Colored dashed: $\beta_{\mathrm{peak}} = 4.59\tau$.

On the right column the NLL valley and the KL ridge track the high-margin region: larger $\overline{\Delta}$ is not an artifact of a collapsed likelihood. Isolines of $\overline{\Delta}$ run roughly parallel to the η_{norm} floor, so the roles of β and of the effective step are separated; the remaining dead zone is in η_{norm}, not in β. Equation (H.2) is a law for Δ^*, not for $\beta \cdot \mathrm{KL}$. Above the η_{norm} floor the product $\beta\,\overline{\Delta}$ is essentially uncorrelated with $\log \beta$ (matching the $1/\beta$ margin), whereas $\beta \cdot \mathrm{KL}$ still rises with β (quartile means among cells above the floor: $1.1 \to 4.7 \times 10^{-3}$). Per-token KL co-moves with the margin but is not a constant KL budget.

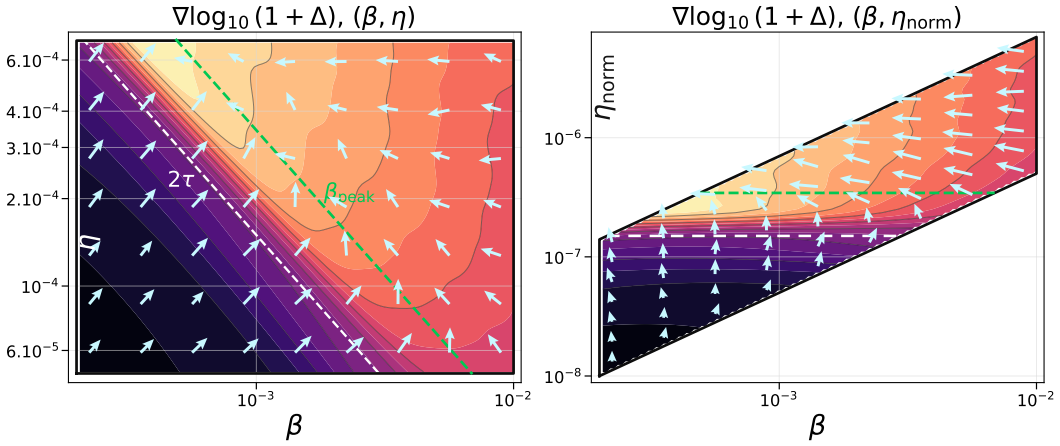

Figure H2: Gradient field of $\log_{10}(1+\overline{\Delta})$ on the interpolant of Figure H1. Arrows are sampled along log-spaced rows and point toward increasing mean margin (smaller β, larger η or η_{norm}). The dead zone remains a floor on η_{norm}, not a cutoff in β.

I Loss-Level Scale Invariance: Identical Loss Curves, Different Policies

Loss geometry and cross-β ordering. For the standard DPO loss, $\mathcal{L}_{\text{DPO}}(\Delta;\beta) = -\log\sigma(\beta\Delta)$, the pair (β, Δ) enters only through $u = \beta\Delta$. Its level sets are therefore hyperbolas, so equal loss values need not imply equal unscaled margins. Equivalently, writing $x = -\Delta$, for $\beta_1 < \beta_2$ and $x < 0$,

$$\log\sigma(\beta_1 x) > \log\sigma(\beta_2 x). \tag{I.1}$$

As $\beta \to 0$, the standard loss approaches the constant $\log 2$, whereas the normalized centered-softplus loss approaches $-\Delta/2$ and retains first-order sensitivity to the unscaled margin.

Observation. Throughout this appendix, the *KL ratio* at a given checkpoint is the ratio of Monte Carlo forward per-token validation KL values, $\text{KL ratio} = \mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})_{\beta_1} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})_{\beta_2}$, with the smaller-$\beta$ run in the numerator and the larger-β run in the denominator. In the UltraFeedback comparison of Figure 4, the two standard-DPO runs ($(\beta, \text{lr}) = (0.01,\ 10^{-3})$ and $(0.02,\ 5 \times 10^{-4})$) produce training- and validation-loss curves that are nearly indistinguishable, while their validation NLL and KL differ substantially. Table I2 quantifies this from the six-epoch logs (mean over four seeds; validation every half epoch): from epoch 1 onward the validation DPO losses differ by at most 0.005 nats, while the KL ratio stays near 2.9–4.1. Table I1 reports the same metrics for the three-epoch single-run comparison of Figure J1 (seed 42): from epoch 1 onward the validation DPO losses differ by at most 0.0017 nats, while the KL ratio is stable at 3.3–3.9. The same pattern holds on HelpSteer3 (Figure I1, Table I3): with $(\beta, \text{lr}) = (0.05,\ 2 \times 10^{-4})$ and $(0.1,\ 1 \times 10^{-4})$ (seed 51, effective batch size 16), validation DPO losses again agree to the third decimal, but the KL ratio reaches ≈ 10 by epoch 6 because the larger-β run remains close to the reference model. A third TRL `DPOTrainer` comparison on UltraFeedback Binarized with Ministral-3B-Instruct (Figure I2, Table I4): with $(\beta, \text{lr}) = (0.01,\ 3 \times 10^{-4})$ and $(0.03,\ 1 \times 10^{-4})$ (seed 52, effective batch size 16, $\eta\beta = 3 \times 10^{-6}$), validation DPO losses stay within 0.009 nats while the KL ratio stabilizes near 6.5. By contrast, on PKU-processed HH-RLHF with Qwen2.5-3B-Instruct (Figure I3, Table I5), the same

scale-equivalent setting $(\beta, \mathrm{lr}) = (0.015,\ 10^{-3})$ and $(0.005,\ 3 \times 10^{-3})$ (seed 52, $\eta\beta = 1.5 \times 10^{-5}$) yields separated raw training losses and validation DPO losses that differ by up to 0.03 nats early on, even though the KL ratio remains stable at ≈ 1.8–3.1.

Table I1: Validation DPO loss and per-token Monte Carlo KL for the two standard-loss runs of Figure J1 (UltraFeedback, Qwen3-4B-Instruct-2507, SGD, learning rate scaled as $1/\beta$, seed 42). KL ratio $= \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.01} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.02}$ at each checkpoint. The loss curves coincide to the third decimal from epoch 1 onward, while the KL ratio stays near 3.3–3.9.

Epoch	Val. DPO loss		Val. KL per token		KL ratio
	$\beta = 0.01$	$\beta = 0.02$	$\beta = 0.01$	$\beta = 0.02$	
0.5	0.5728	0.5620	0.436	0.128	3.4
1.0	0.5552	0.5557	0.455	0.140	3.3
1.5	0.5463	0.5462	0.579	0.168	3.5
2.0	0.5387	0.5385	0.616	0.158	3.9
2.5	0.5361	0.5368	0.622	0.189	3.3
3.0	0.5346	0.5363	0.630	0.194	3.3

Table I2: Validation DPO loss and per-token Monte Carlo KL for the two standard-loss runs of Figure 4 (UltraFeedback, Qwen3-4B-Instruct-2507, SGD, learning rate scaled as $1/\beta$; mean over four seeds). KL ratio $= \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.01} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.02}$ at each checkpoint. From epoch 1 onward the validation DPO losses differ by at most 0.005 nats, while the KL ratio stays near 2.9–4.1.

Epoch	Val. DPO loss		Val. KL per token		KL ratio
	$\beta = 0.01$	$\beta = 0.02$	$\beta = 0.01$	$\beta = 0.02$	
0.5	0.5727	0.5662	0.466	0.154	3.0
1.0	0.5539	0.5520	0.484	0.154	3.1
1.5	0.5458	0.5450	0.545	0.155	3.5
2.0	0.5357	0.5373	0.622	0.152	4.1
2.5	0.5303	0.5328	0.633	0.180	3.5
3.0	0.5257	0.5307	0.639	0.203	3.2
3.5	0.5246	0.5272	0.692	0.212	3.3
4.0	0.5216	0.5245	0.706	0.226	3.1
4.5	0.5209	0.5247	0.780	0.245	3.2
5.0	0.5204	0.5223	0.802	0.261	3.1
5.5	0.5235	0.5233	0.820	0.282	2.9
6.0	0.5210	0.5219	0.814	0.278	2.9

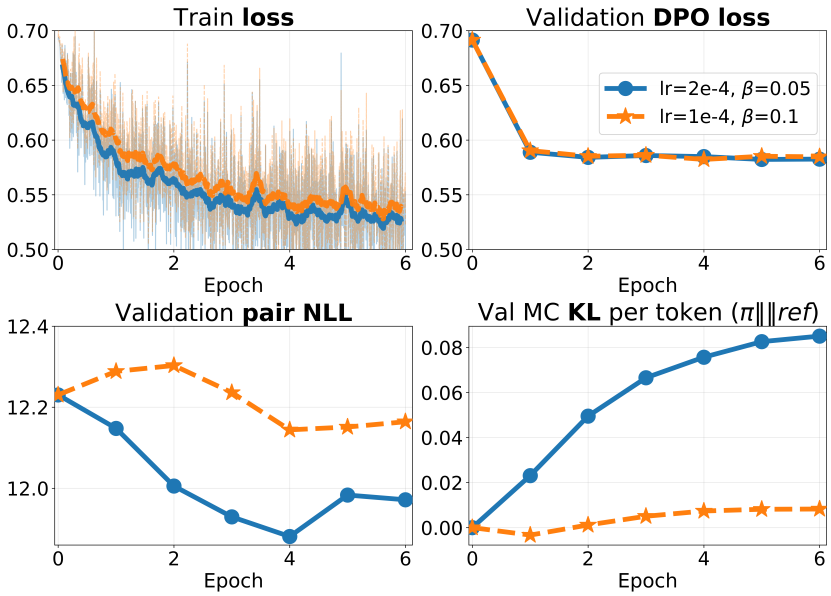

Figure I1: HelpSteer3 standard-DPO comparison with scale-equivalent learning rates using TRL `DPOTrainer` (Qwen3-4B-Instruct-2507, SGD, batch size 4, gradient accumulation 4, seed 51): $(\beta, \mathrm{lr}) = (0.05,\ 2 \times 10^{-4})$ and $(0.1,\ 1 \times 10^{-4})$, so $\eta\beta = 10^{-5}$ in both runs. Table I3 lists the same checkpoints.

Table I3: Validation DPO loss and per-token Monte Carlo KL for the HelpSteer3 TRL `DPOTrainer` runs of Figure I1 (SGD, $\eta\beta = 10^{-5}$, seed 51). KL ratio $= \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.05} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.1}$; entries marked "—" are omitted when the denominator KL is non-positive or too small for a stable ratio (Monte Carlo noise near the reference policy).

Epoch	Val. DPO loss		Val. KL per token		KL ratio
	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.1$	
1	0.5889	0.5904	0.0231	−0.0033	—
2	0.5842	0.5852	0.0495	0.0012	—
3	0.5859	0.5862	0.0666	0.0050	13.2
4	0.5849	0.5823	0.0757	0.0074	10.3
5	0.5823	0.5852	0.0826	0.0081	10.1
6	0.5826	0.5847	0.0851	0.0083	10.3

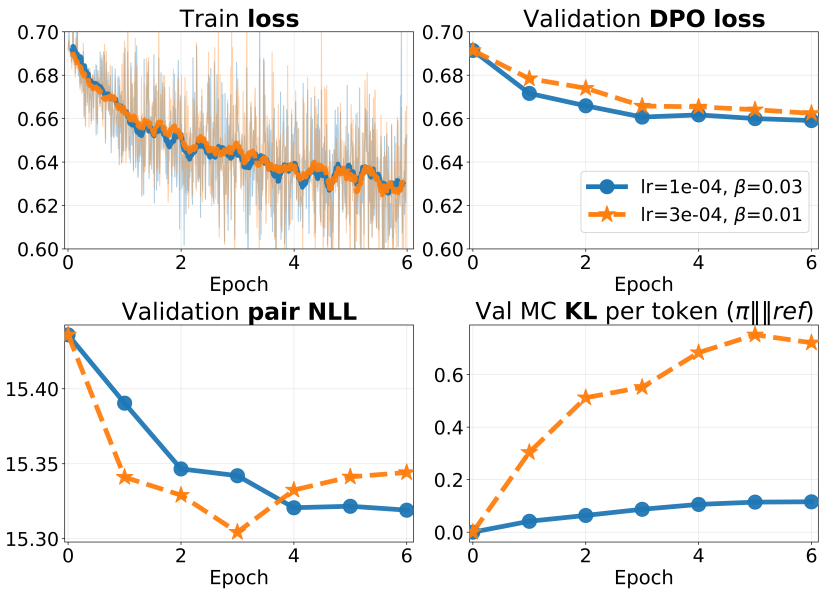

Figure I2: UltraFeedback Binarized standard-DPO comparison with scale-equivalent learning rates using TRL `DPOTrainer` (Ministral-3B-Instruct, SGD, batch size 4, gradient accumulation 4, seed 52): $(\beta, \mathrm{lr}) = (0.01,\ 3 \times 10^{-4})$ and $(0.03,\ 1 \times 10^{-4})$, so $\eta\beta = 3 \times 10^{-6}$ in both runs. Table I4 lists the same checkpoints.

Table I4: Validation DPO loss and per-token Monte Carlo KL for the UltraFeedback Binarized TRL `DPOTrainer` runs of Figure I2 (Ministral-3B-Instruct, SGD, $\eta\beta = 3 \times 10^{-6}$, seed 52). KL ratio $= \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.01} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.03}$ at each checkpoint.

Epoch	Val. DPO loss		Val. KL per token		KL ratio
	$\beta = 0.01$	$\beta = 0.03$	$\beta = 0.01$	$\beta = 0.03$	
1	0.6784	0.6716	0.304	0.041	7.3
2	0.6740	0.6659	0.512	0.064	8.0
3	0.6658	0.6607	0.552	0.087	6.3
4	0.6654	0.6617	0.683	0.106	6.5
5	0.6641	0.6600	0.752	0.114	6.6
6	0.6624	0.6591	0.721	0.116	6.2

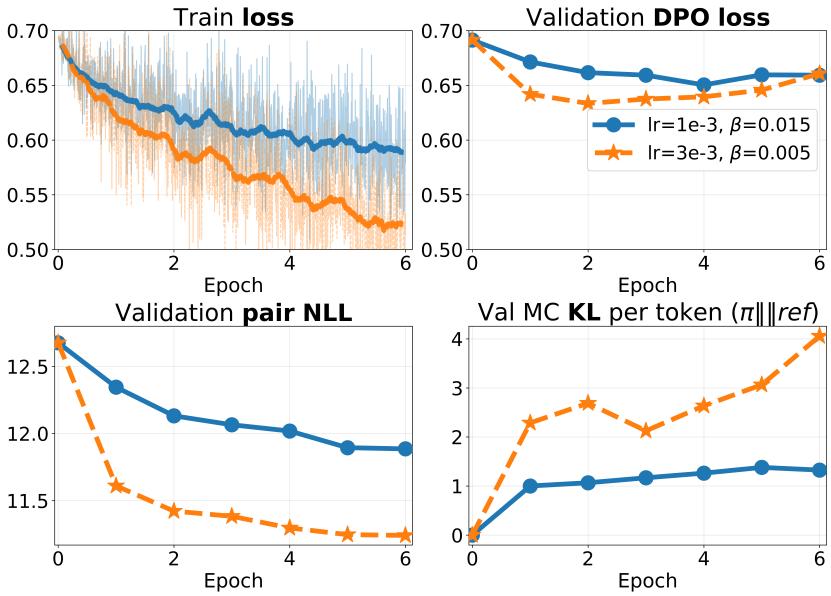

Figure I3: PKU-processed HH-RLHF standard-DPO comparison with scale-equivalent learning rates using TRL `DPOTrainer` (Qwen2.5-3B-Instruct, SGD, batch size 4, gradient accumulation 4, seed 52): $(\beta, \mathrm{lr}) = (0.015,\ 10^{-3})$ and $(0.005,\ 3 \times 10^{-3})$, so $\eta\beta = 1.5 \times 10^{-5}$ in both runs. Table I5 lists the same checkpoints. Raw training losses (top-left) do not coincide, unlike the smoothed HelpSteer3 and Ministral-3B curves above.

Table I5: Validation DPO loss and per-token Monte Carlo KL for the PKU-processed HH-RLHF TRL `DPOTrainer` runs of Figure I3 (Qwen2.5-3B-Instruct, SGD, $\eta\beta = 1.5 \times 10^{-5}$, seed 52). KL ratio $= \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.005} / \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})_{\beta=0.015}$ at each checkpoint.

Epoch	Val. DPO loss		Val. KL per token		KL ratio
	$\beta = 0.015$	$\beta = 0.005$	$\beta = 0.015$	$\beta = 0.005$	
1	0.6714	0.6419	1.001	2.287	2.3
2	0.6616	0.6335	1.067	2.688	2.5
3	0.6593	0.6375	1.168	2.125	1.8
4	0.6504	0.6395	1.263	2.634	2.1
5	0.6596	0.6457	1.380	3.064	2.2
6	0.6593	0.6607	1.325	4.056	3.1

When do the losses coincide? Matching $\eta\beta$ is necessary but not sufficient for overlapping loss curves. The standard loss depends only on $u = \beta\Delta$ (Eq. (I.2)), so training and validation losses coincide only if the two runs follow a common u-trajectory. With $\mathrm{lr} \propto 1/\beta$, the unscaled margin grows at comparable rates ($\mathrm{d}\Delta/\mathrm{d}t \propto \sigma(-u)$), but u advances faster for larger β, so that run initially leads. Saturation, however, is a condition on u alone: the per-step drift criterion $\eta g^2 \beta\, \sigma(-\beta\Delta) \lesssim \varepsilon$ (Appendix G, Setup) depends on (η, β) only through the matched product $\eta\beta$, so both runs share one threshold $\sigma(-u) \lesssim \varepsilon/(\eta\beta g^2)$: the larger-$\beta$ run reaches it first and slows down, while the smaller-β run—still at smaller u for the same Δ—keeps catching up until both settle near a common u^*. Their losses then coincide even though $\Delta = u/\beta$ (and KL) remain different. This catch-up succeeds on UltraFeedback, HelpSteer3, and Ministral-3B (Tables I1–I4): validation DPO losses agree to 10^{-3}–10^{-2} nats and smoothed training losses are nearly overlapping. It does not fully occur on HH-RLHF (Figure I3, Table I5): both runs move far from the reference (KL $\gtrsim 1$ nat/token), the corpus is comparatively noisy, the β ratio is 3, and the early transient does not fully close within six epochs—validation DPO losses differ by up to 0.03 nats and raw training losses remain separated (final values 0.62 vs. 0.53). This breakdown is consistent with the limitations of the scalar frozen-Jacobian closure (Appendices K and L): heterogeneous pair Jacobians and cross-pair Gram cancellations can prevent the two runs from sharing a common u-trajectory, so the catch-up argument need not close on a noisy

corpus. In all cases, however, the KL ratio is a more stable readout of physical displacement: it stays near 2.9–4.1 (UltraFeedback, six epochs) and 3.3–3.9 (three epochs), ≈ 10 (HelpSteer3), 6.2–8.0 (Ministral-3B), and ≈ 1.8–3.1 (HH-RLHF) once the early transient passes, even when the loss curves do not fully coincide. Table I6 collects these scale-equivalent comparisons side by side.

Table I6: Summary of scale-equivalent standard-DPO pairs ($\mathrm{lr} \propto 1/\beta$, matched $\eta\beta$). $|\Delta\mathcal{L}_{\mathrm{val}}|$: absolute difference between the two validation DPO losses (nats); max / final. Train-loss rows: *all* = overall verdict; *sm.* = moving-average curve (2.5% window); *raw* = final per-step values logged at the last checkpoint (HelpSteer3 raw values differ by ≈ 0.025 despite overlapping smoothed curves). KL ratio $= \mathrm{KL}_{\beta_{\mathrm{small}}}/\mathrm{KL}_{\beta_{\mathrm{large}}}$; HelpSteer3 from epoch 3 onward.

Data	Model	$\eta\beta$	β_2/β_1	$\lvert\Delta\mathcal{L}_{\mathrm{val}}\rvert$ max / final	Train loss (all / sm. / raw)	KL ratio
UF 3 ep.	Q3-4B	10^{-5}	2	$\leq$0.0017 / $\leq$0.0017	overlap overlap overlap	3.3–3.9
UF 6 ep.	Q3-4B	10^{-5}	2	0.0051 / 0.0009	overlap overlap overlap	2.9–4.1
HSteer3	Q3-4B	10^{-5}	2	0.0029 / 0.0021	overlap overlap diverge	≈ 10
UF bin.	M3-3B	3×10^{-6}	3	0.0081 / 0.0033	overlap overlap overlap	6.2–8.0
HH-RLHF	Q2.5-3B	1.5×10^{-5}	3	0.0295 / 0.0014	diverge diverge diverge	1.8–3.1

Invariance of the loss. The coincidence is structural, not accidental. The standard pairwise loss depends on (β, Δ) only through the dimensionless product $u = \beta\Delta$:

$$\mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) = \log\bigl(1 + e^{-\beta\Delta}\bigr) = \mathrm{softplus}(-u), \qquad \mathcal{L}_{\mathrm{DPO}}(\Delta/c;\, c\beta) = \mathcal{L}_{\mathrm{DPO}}(\Delta;\beta) \quad \forall\, c > 0. \tag{I.2}$$

Its level sets in the (β, Δ) plane are hyperbolas $\beta\Delta = \mathrm{const}$: the loss value identifies the dimensionless margin u and is blind to the unscaled margin Δ, and can therefore hide large differences in policy displacement, including KL from the reference. For example, $(\beta, \Delta) = (0.01, 100)$ and $(0.02, 50)$ yield the identical loss $\log(1 + e^{-1}) \approx 0.313$, although their unscaled margins differ by a factor of two. The corresponding policy displacement and KL need not scale proportionally. The same invariance holds for every objective of the form $L(\beta\Delta)$ in Table A1.

Why the two runs can approach similar u-trajectories. Under the scalar frozen-Jacobian approximation of Appendix G, $\mathrm{lr} \propto 1/\beta$ gives $\mathrm{d}\Delta/\mathrm{d}t \propto \sigma(-u)$ and the same drift-based saturation threshold $\sigma(-u) \lesssim \varepsilon/(\eta\beta g^2)$ for both runs. This approximation predicts late-time catch-up in u, consistent with several reported sweeps, but it is not a general trajectory-equivalence result for minibatch training because the Jacobians, cross-pair Gram terms, and stochastic gradients evolve during optimization. The KL ratio is not fixed by this identity alone: Table I2 gives $\approx$ 2.9–4.1 on UltraFeedback (six epochs) and Table I1 gives $\approx$ 3.3–3.9 (three epochs), Table I4 gives $\approx$ 6.2–8.0 on UltraFeedback Binarized with Ministral-3B-Instruct, Table I3 gives ≈ 10 once the larger-β run has moved only slightly from the reference model, and Table I5 gives $\approx$ 1.8–3.1 on HH-RLHF.

Early transient. In runs that exhibit the scalar-model catch-up described above, the loss curves differ most during the initial transient; persistent differences remain possible under general minibatch dynamics, as in the HH-RLHF example. While $u \ll 1$, both runs grow the unscaled margin at nearly the same rate because $\sigma(-u) \approx 1/2$, so over the same time interval $u = \beta\Delta$ is initially larger for the larger-β run and its loss drops faster. Table I1 makes this visible at epoch 0.5: the $\beta = 0.02$ validation loss is 0.5620 versus 0.5728 for $\beta = 0.01$, even though the smaller-β run already has higher KL (0.436 vs. 0.128). Once saturation sets in, the larger-β run slows first and the smaller-β run catches up in u; by epoch 1 the validation losses differ by only 5×10^{-4} nats and remain within 10^{-3} thereafter.

The normalized loss reads the unscaled-margin scale. The centered-softplus objective is *not* a function of u alone: $s_c(-\Delta;\beta) = \big(\mathrm{softplus}(-u) - \log 2\big)/\beta$, so at equal u its value scales as $1/\beta$ and the two runs are predicted to separate by a factor of 2. The final logged training losses confirm this: -17.5 for $\beta = 0.01$ versus -9.0 for $\beta = 0.02$, a ratio of 1.95 (bottom-right panel of Figure J1); the four-seed six-epoch comparison in Figure 4 exhibits the same factor-of-two separation (bottom-middle panel). The overlap of the standard losses and the $2\times$ separation of the normalized losses are thus two sides of the same identity (I.2).

J Per-run Trajectories under the Normalized Objective

Figure 4 in the main text reports the six-epoch UltraFeedback comparison of standard DPO and the centered-softplus objective (four-seed mean $\pm$ one standard deviation). Figure J1 shows the corresponding three-epoch single-run trajectories.

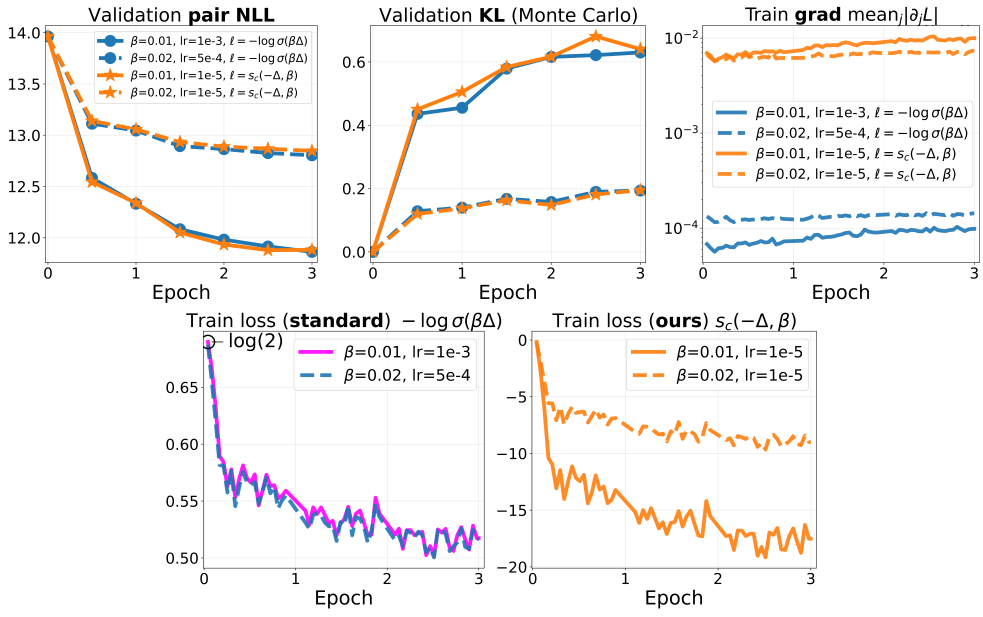

Figure J1: Comparison of training with the normalized centered-softplus DPO objective $s_c(-\Delta;\beta)$ and standard DPO with pairwise loss $-\log\sigma(\beta\Delta)$ on UltraFeedback with Qwen3-4B-Instruct-2507 (batch size 26). Learning rates satisfy $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}}$. With SGD, validation metrics are nearly equivalent. Even when validation metrics differ substantially between $\beta \in \{0.01, 0.02\}$, the standard training-loss curves are almost overlapping, whereas for $s_c(-\Delta;\beta)$ the training losses separate noticeably (Appendix I). Single-run, three epochs. The comparison tests the predicted scale equivalence; it is not a claim that the normalized objective outperforms a retuned DPO baseline.

Downstream evaluation on UltraFeedback. Table J1 reports AlpacaEval 2 and IFEval scores for the four curves of Figure 4 (Qwen3-4B-Instruct-2507, SGD, batch size 26, four paired seeds). We use the same protocol as Table J2: official AlpacaEval 2 instructions and GPT-4-turbo reference outputs,

scored by a fixed `Qwen/Qwen2.5-14B-Instruct` judge, and the official verifiable IFEval benchmark without an LLM judge. The judge shares a model family with the Qwen candidates, so absolute AlpacaEval win rates may be inflated by intra-family preference; relative comparisons between standard DPO and centered-softplus on the same backbone are less exposed to this confound. Entries are mean $\pm$ std. With only four paired seeds we do not report a Wilcoxon test: Table J1 is descriptive, unlike the eight-seed HelpSteer3 analysis in Table J2. The two standard-DPO configurations have nearly identical training and validation DPO losses (Figure 4), yet their length-controlled win rates differ by about 14 percentage points and their IFEval accuracies by about 2 points in the same direction; the centered-softplus pair shows the same downstream split, while each standard-DPO setting agrees with its learning-rate-rescaled centered-softplus counterpart.

Table J1: AlpacaEval 2 and IFEval scores for the four UltraFeedback curves of Figure 4 (Qwen3-4B-Instruct-2507, SGD, batch size 26). Centered softplus uses a shared $\mathrm{lr}_{\mathrm{norm}} = 10^{-5}$, and the paired standard-DPO rates satisfy $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}}$. AlpacaEval 2 scores use `Qwen/Qwen2.5-14B-Instruct` as the pairwise judge; IFEval uses the official verifiable benchmark. Entries are mean$_{\pm\mathrm{std}}$ over 4 seeds (%). With $n = 4$ we do not report a significance test; the comparison is descriptive.

			AlpacaEval 2		IFEval			
Objective	β	lr	WR	LC WR	Strict prompt	Strict instr	Loose prompt	Loose instr
Standard DPO	0.01	10^{-3}	$41.46_{\pm 2.98}$	$48.61_{\pm 2.11}$	$78.19_{\pm 0.40}$	$84.98_{\pm 0.38}$	$81.42_{\pm 0.44}$	$87.35_{\pm 0.32}$
Standard DPO	0.02	5×10^{-4}	$61.43_{\pm 1.88}$	$62.37_{\pm 1.31}$	$79.81_{\pm 0.49}$	$86.06_{\pm 0.44}$	$83.50_{\pm 0.63}$	$88.70_{\pm 0.54}$
Centered softplus	0.01	10^{-5}	$41.74_{\pm 1.15}$	$49.37_{\pm 2.04}$	$77.77_{\pm 0.80}$	$84.68_{\pm 0.45}$	$81.05_{\pm 1.09}$	$87.17_{\pm 0.75}$
Centered softplus	0.02	10^{-5}	$62.48_{\pm 1.91}$	$63.41_{\pm 1.77}$	$79.71_{\pm 0.35}$	$86.09_{\pm 0.17}$	$83.46_{\pm 0.57}$	$88.67_{\pm 0.52}$

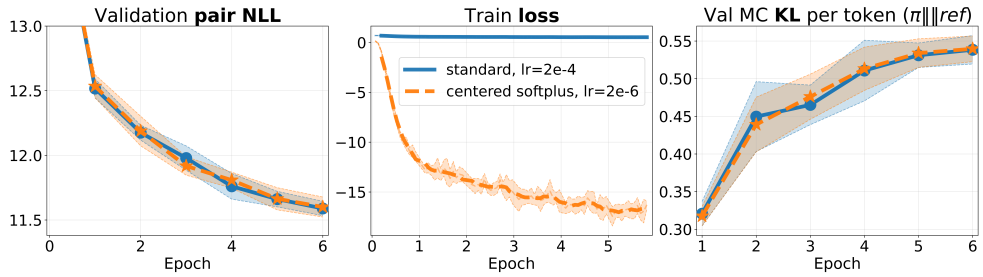

Figure J2: Comparison of training with the normalized centered-softplus DPO objective $s_c(-\Delta;\beta)$ and standard DPO with pairwise loss $-\log\sigma(\beta\Delta)$ on HelpSteer3 with Qwen3-4B-Instruct-2507 (batch size 24). SGD with $\beta = 0.01$. The standard run uses $\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-4}$ and the centered-softplus run uses $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-6}$. Curves show the mean over eight experiments; shaded bands denote one standard deviation.

Downstream evaluation. Table J2 summarizes post-training instruction-following checks for the HelpSteer3 SGD comparison of Figure J2: standard DPO ($\beta = 0.01$, $\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-4}$) versus the centered-softplus objective with $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-6}$, across eight paired seeds (43–50). For *AlpacaEval 2* (Dubois et al., 2024), we use the official instruction set and GPT-4-turbo reference outputs, generate candidate completions from each fine-tuned checkpoint, and score pairwise preferences with a fixed `Qwen/Qwen2.5-14B-Instruct` judge; we report raw win rate, length-controlled (LC) win rate, and length-filtered win rate. Because the judge and the Qwen3-4B candidates share a model lineage, absolute win rates may reflect intra-family preference; the paired standard-versus-normalized comparison on a shared backbone is less sensitive to this bias. For *IFEval* (Zhou et al., 2023), we use the official verifiable instruction-following benchmark (strict and loose prompt- and instruction-level accuracy) without an LLM judge. Entries report mean $\pm$ std; the p column gives two-sided Wilcoxon signed-rank test p-values for seed-matched pairs. No statistically

significant difference is detected at $n = 8$ for any reported metric. This absence of significance does not establish practical equivalence or exclude small systematic effects.

Table J2: Post-training benchmarks for the HelpSteer3 SGD comparison of Figure J2 (Qwen3-4B-Instruct-2507, batch size 24, $\beta = 0.01$, eight paired seeds). Standard DPO uses $\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-4}$; centered softplus uses $\mathrm{lr}_{\mathrm{norm}} = \beta\,\mathrm{lr}_{\mathrm{standard}} = 2 \times 10^{-6}$. AlpacaEval 2 scores use `Qwen/Qwen2.5-14B-Instruct` as the pairwise judge; IFEval uses the official verifiable benchmark. Entries are mean $\pm$ std (%); p is the two-sided Wilcoxon signed-rank test p-value for seed-matched pairs. No statistically significant difference is detected at $n = 8$; this test does not establish practical equivalence.

Metric	Standard DPO	Centered softplus	p
Alpaca win	54.89 ± 2.82	55.47 ± 1.83	0.94
Alpaca LC win	54.72 ± 2.37	55.50 ± 1.44	0.63
Alpaca len-filtered win	43.83 ± 2.67	45.09 ± 1.91	0.63
Tie	0.02 ± 0.04	0.02 ± 0.04	1.00
Loss	45.09 ± 2.82	44.52 ± 1.85	0.94
IFEval strict prompt	80.11 ± 0.98	81.22 ± 0.94	0.11
IFEval strict instr	86.18 ± 0.72	87.02 ± 0.65	0.08
IFEval loose prompt	83.76 ± 0.74	84.61 ± 1.27	0.23
IFEval loose instr	88.74 ± 0.48	89.44 ± 0.81	0.16

Figure J3 shows the seed-averaged HelpSteer3 AdamW trajectories for the centered-softplus objective analyzed in Section 4; Figure J4 shows the corresponding single-run trajectories. These runs include the $\beta = 0$ linear branch of Equation (14), which is unbounded below and comparatively less stable than any $\beta > 0$ run; stronger policy movement still yields the fastest decrease in the normalized training loss. Consistent with the $1/\beta$ readout in Appendix I, the normalized training losses separate across β, tracking the physical aggressiveness of each run rather than collapsing onto a common curve. The validation pair NLL follows the same ranking ($\beta = 0$ and $\beta = 0.002$ lowest, $\beta = 0.015$ highest), whereas the validation *standard* DPO loss inverts it ($\beta = 0.015$ lowest; $\beta = 0$ sits identically at $\log 2$), as predicted by Equation (I.1).

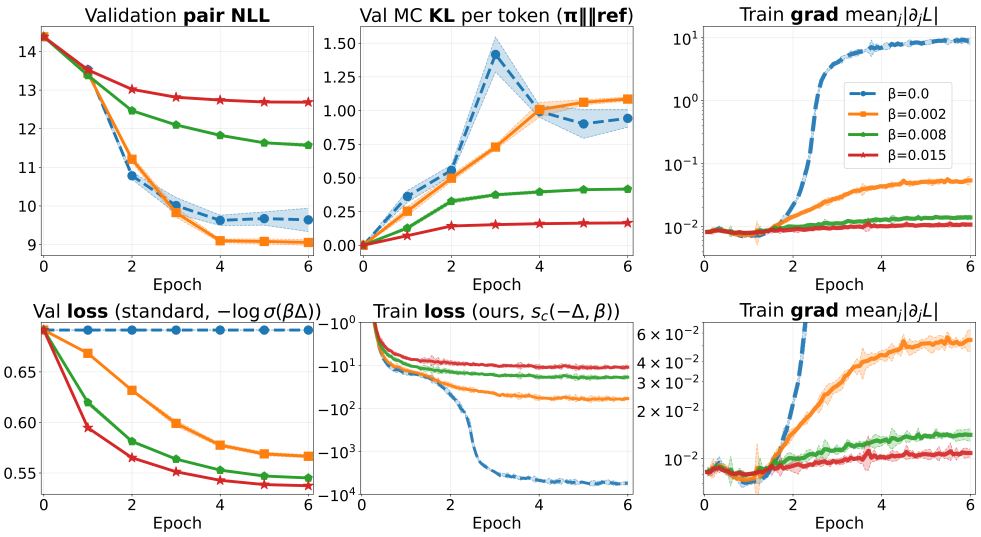

Figure J3: Training and validation dynamics on HelpSteer3 with Qwen3-4B-Instruct-2507 using the centered-softplus DPO objective $s_c(-\Delta;\beta)$ (batch size 24, $\mathrm{lr} = 1 \times 10^{-6}$, Adam). Curves show the mean over four seeds (43–46); shaded bands denote ± 2 standard deviations (no band is drawn for the gradient panel).

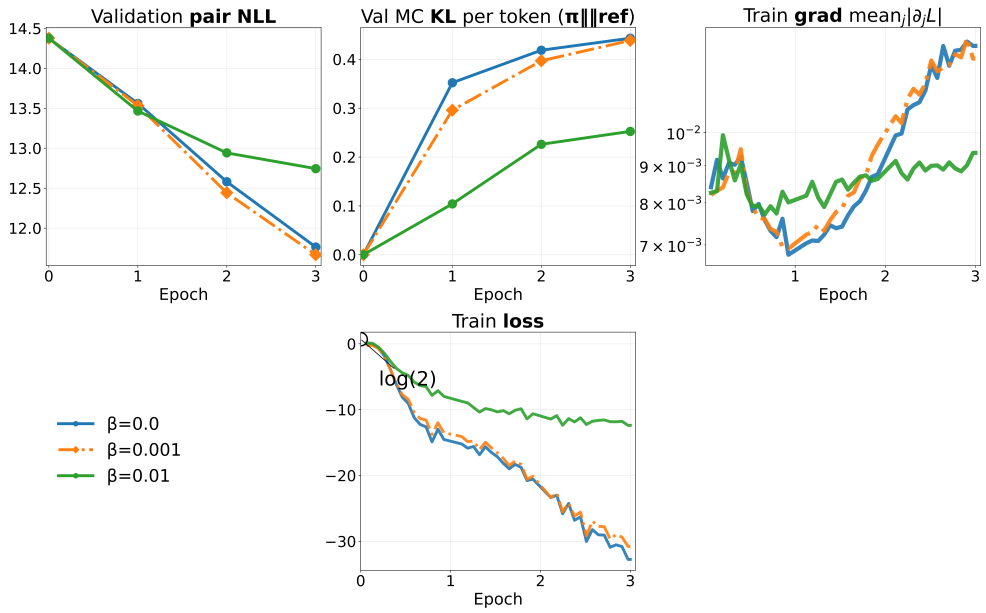

Figure J4: Centered-softplus loss $s_c(-\Delta;\beta)$ on HelpSteer3 with Qwen3-4B-Instruct-2507 (batch size 24, $\mathrm{lr} = 1 \times 10^{-6}$), trained with Adam.

K Logged Gradients, Mean-Abs Magnitudes, and the Jacobian Factor g

This appendix collects the gradient questions that arise in the main text: what the training logs record, how that scalar relates to $\nabla_\theta \mathcal{L}$, and how both relate to the Jacobian scale used in the saturation and horizon formulas. The exact identities below first apply to a single preference pair. For a minibatch, the logged norm identifies a coherence-weighted norm of the batch-averaged margin Jacobian rather than a typical per-pair Jacobian; we make this distinction explicit before using the logged value as an order-of-magnitude proxy in Appendix L.

Chain rule. For a single preference pair the standard DPO loss depends on θ only through the scalar margin $\Delta(\theta)$:

$$\mathcal{L}(\theta) = \ell\big(\Delta(\theta)\big), \qquad \ell(u) = -\log \sigma(\beta u), \qquad \ell'(u) = -\beta\, \sigma(-\beta u). \tag{K.1}$$

The parameter gradient is therefore exactly

$$\nabla_\theta \mathcal{L} = \ell'(\Delta)\, \nabla_\theta \Delta = -\beta\, \sigma(-\beta\Delta)\, \nabla_\theta \Delta. \tag{K.2}$$

Taking Euclidean norms,

$$\|\nabla_\theta \mathcal{L}\|_2 = \beta\, \sigma(-\beta\Delta)\, g, \qquad g := \|\nabla_\theta \Delta\|_2. \tag{K.3}$$

Thus every loss-gradient magnitude factors as

$$\underbrace{\beta}_{\text{explicit scale}} \times \underbrace{\sigma(-\beta\Delta)}_{\text{sigmoid / saturation}} \times \underbrace{g}_{\text{Jacobian of the margin}} .$$

At initialization $\Delta = 0$ and $\sigma(0) = 1/2$, so

$$\|\nabla_\theta \mathcal{L}\|_2 = \frac{\beta}{2}\, g \qquad \Rightarrow \qquad g = \frac{2}{\beta}\, \|\nabla_\theta \mathcal{L}\|_2. \tag{K.4}$$

Equation (K.3) is an identity for one pair. Treating g as constant along a trajectory and identifying it from a minibatch-gradient log are separate approximations used only for the rough horizon estimate.

Minibatch correction. For a minibatch of B pairs, write $j_i = \nabla_\theta \Delta_i$ and $a_i = \sigma(-\beta \Delta_i)$. Then

$$\nabla_\theta \mathcal{L}_B = -\frac{\beta}{B} \sum_{i=1}^{B} a_i j_i. \tag{K.5}$$

Consequently, a global gradient norm does not recover a typical $g_i = \|j_i\|_2$. At initialization, where $a_i = 1/2$,

$$\frac{2}{\beta} \|\nabla_\theta \mathcal{L}_B\|_2 = \left\| \frac{1}{B} \sum_{i=1}^{B} j_i \right\|_2 =: g_B. \tag{K.6}$$

The quantity g_B includes cross-pair alignment and cancellation. It may therefore differ substantially from both $B^{-1} \sum_i \|j_i\|_2$ and a typical per-pair norm. The Gram-matrix form of the resulting margin drift is given in Appendix L; throughout the numerical horizon estimate we use the observed g_B only as an effective scalar proxy.

Logged scalars: mean-abs versus Euclidean. Our DPO implementation logs two scalars every 100 steps over trainable (LoRA) coordinates $q = \nabla_\theta \mathcal{L}$:

$$G := \mathrm{mean}_j |q_j| = \texttt{grad_abs_mean}, \tag{K.7}$$

$$\|\nabla_\theta \mathcal{L}\|_2 := \sqrt{\sum_j q_j^2} = \texttt{grad_norm}. \tag{K.8}$$

TRL `DPOTrainer` TensorBoard curves report the Euclidean quantity as `train/grad_norm`. The two are related by a shape factor

$$\|\nabla_\theta \mathcal{L}\|_2 = C\, G, \qquad C = \sqrt{d}\, r, \qquad r = \frac{\mathrm{RMS}_j |q_j|}{\mathrm{mean}_j |q_j|} \geq 1, \tag{K.9}$$

with d the number of trainable parameters. Equality $r = 1$ holds only if all $|q_j|$ are equal; empirically, for Qwen3-4B, $r \approx 2.8$–3.1 on HelpSteer3 and $r \approx 3.7$–4.1 on UltraFeedback (below). Table 1 in the main text uses G_0 in the sense of (K.7).

The dynamics panels in the main text and appendices use a mixed convention inherited from the original experimental logs: the SGD runs of our implementation recorded G, and their panels are labeled $\mathrm{mean}_j |\partial_j \mathcal{L}|$ accordingly. TRL panels report Euclidean $\|\nabla_\theta \mathcal{L}\|_2$. The two scalars are not interchangeable, and we do not claim that G equals the Euclidean norm. By (K.9) the two differ by the nearly constant factor $C = \sqrt{d}\, r$ (r varies only weakly along a run), so cross-β orderings and time trends are the same for either scalar. Horizon formulas use the Euclidean reading, not G. For actual minibatch logs, Equation (K.6), rather than the one-pair inversion in (K.4), is the appropriate interpretation.

Three-factor decomposition along training. Equation (K.3) splits every loss-gradient magnitude into an explicit β scale, a sigmoid/saturation factor, and the Jacobian g. Figures K1–K3 isolate those three factors along a run:

1. left: the logged loss-gradient magnitude (G for our implementation; $\|\nabla_\theta \mathcal{L}\|_2$ for TRL);
2. center: the same quantity divided by β, which removes the explicit scale (at $t = 0$ the curves collapse);
3. right: division by $\beta\, \sigma(-\beta \Delta_t)$ as well, leaving a scalar proxy for $g_B(t)$ (using a linear interpolation of the empirical final mean validation margin for runs of our implementation, and the reported mean $|\Delta|$ for TRL).

A drop on the left panel need not mean that g shrinks: under (K.3) the sigmoid factor alone drives the loss gradient down as Δ grows at fixed g. The right panel is the diagnostic for whether g itself is moving.

HelpSteer3 numerical check (our implementation, SGD, $\eta_{\max} = 2 \times 10^{-4}$). Two seed-43 runs log both G and $\|\nabla\mathcal{L}\|_2$ with $d = 33\,030\,144$ LoRA parameters. Early steps (100–500):

β	G/β	$r = \|\nabla\mathcal{L}\|_2/(\sqrt{d}\,G)$	$g_B = 2\|\nabla\mathcal{L}\|_2/\beta$
0.0025 (peak)	8.24×10^{-3}	2.76	261
0.075 (right slope)	7.17×10^{-3}	2.94	242

So G/β is essentially the same as in Table 1, while the Euclidean reading gives $r \approx 2.8$–2.9 (within the HelpSteer3 / Qwen3-4B range 2.8–3.1) and an early weighted-Jacobian proxy $g_B \approx 250$, not the isotropic $r = 1$ value $g_B \approx 95$. The isotropic choice $r = 1$ is a lower bound on r (hence an upper bound on the scalar horizon proxy $\beta_{\mathrm{r}} \propto 1/g_B^2 \propto 1/r^2$): it would give $g_B \approx 95$ and $\beta_{\mathrm{r}} \approx 0.19$, whereas the measured L2 norm gives $g_B \approx 250$ and $\beta_{\mathrm{r}} \approx 0.027$ (Appendix L). Neither value estimates a typical per-pair Jacobian. By the end of epoch 1, with validation mean $\Delta \approx 20.9$ ($\beta = 0.0025$) and $\Delta \approx 13.3$ ($\beta = 0.075$),

$$g_B(t) = \frac{\|\nabla\mathcal{L}\|_2}{\beta\,\sigma(-\beta\Delta)}$$

has already risen to ≈ 360 and ≈ 370 respectively under the scalar reconstruction: the effective batch proxy grows, while at $\beta = 0.075$ the raw G falls mainly because $\sigma(-\beta\Delta)$ shrinks. Because these measurements come from steps 100–500, $g_B \approx 250$ is an early weighted-Jacobian proxy rather than a measurement of $g_B(\theta_0)$. The reconstruction additionally substitutes an aggregate validation margin into the sigmoid, so it is diagnostic rather than exact. Freezing the early proxy ignores its subsequent time variation.

Takeaways for the rest of the paper.

- Main-text G_0 and the dynamics figures from our implementation use mean-abs G; their panels are labeled $\mathrm{mean}_j|\partial_j\mathcal{L}|$, whereas TRL panels report the Euclidean norm. Euclidean $\|\nabla_\theta\mathcal{L}\|_2$ differs from G by a near-constant factor $C = \sqrt{d}\,r$ (empirically, for Qwen3-4B, $r \approx 2.8$–3.1 on HelpSteer3 and $r \approx 3.7$–4.1 on UltraFeedback), so the two are interchangeable for cross-β comparisons.
- The scalar horizon formula uses the logged batch quantity g_B as a proxy for the one-pair g in (K.3). An exact minibatch calculation instead requires cross-pair Gram products.
- For the right-hand scale β_{r} introduced in Appendix L and shown as the broad band in Figure 2, the scalar closure $\beta_{\mathrm{r}} \propto 1/g_B^2$ makes the present Jacobian proxy consequential: time variation and cross-pair cancellation in g_B preclude a precise cutoff. The resulting β_{r} is reported only to order of magnitude; factor-of-few shifts occupy less than one decade on the logarithmic β axis of the sweep.

L Time to Saturation and the Right Cutoff β_{r}

Appendix G gives the one-pair saturation margin $\Delta^*_{\mathrm{DPO}}(\beta)$ and the *left* dead zone $\beta \le 2\tau$. On the right slope, $\Delta^*_{\mathrm{DPO}} \sim \log(\beta/\tau)/\beta \to 0$, but that asymptotic alone does not say whether a finite training budget reaches the threshold-limited margin. This appendix seeks only an order-of-magnitude crossover scale β_{r} between a horizon-limited margin and the threshold-limited margin on the right slope, represented by the broad right-hand band of Figure 2, rather than a calibrated cutoff. Two approximations dominate its uncertainty. First, the scalar closure freezes the initial Jacobian proxy even though the reconstructed proxy usually grows during training (Figure L1). Second, the global minibatch gradient identifies the norm of a batch-averaged Jacobian, including cross-pair cancellation, rather than a typical per-pair norm. An exact calculation would retain the full cross-pair Gram matrix. We state that correction below, then deliberately keep the scalar closure because the goal is a rough location on a logarithmic β axis. Notation for the loss gradient, the one-pair Jacobian $g = \|\nabla_\theta\Delta\|_2$, and the effective batch proxy g_B is fixed in Appendix K.

Inputs already fixed elsewhere. We reuse the one-pair saturation ceiling $\Delta^*_{\mathrm{DPO}}(\beta)$ and peak location $\beta_{\mathrm{peak}} \approx 4.59\,\tau$ from Appendix G (Equation (8)) without re-deriving them. For the HelpSteer3

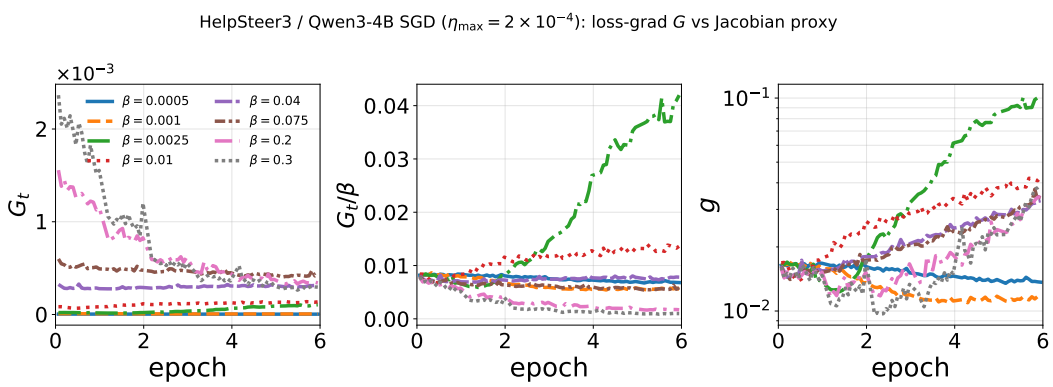

Figure K1: HelpSteer3 / Qwen3-4B, our DPO implementation, SGD, $\eta_{\max} = 2 \times 10^{-4}$. The three panels isolate the factors in (K.3): logged $G = \texttt{grad_abs_mean}$ (left), G/β (center), and a scalar batch-Jacobian proxy $g_B(t)$ after dividing out $\beta\sigma(-\beta\Delta)$ (right). The last quantity is not a typical per-pair Jacobian norm.

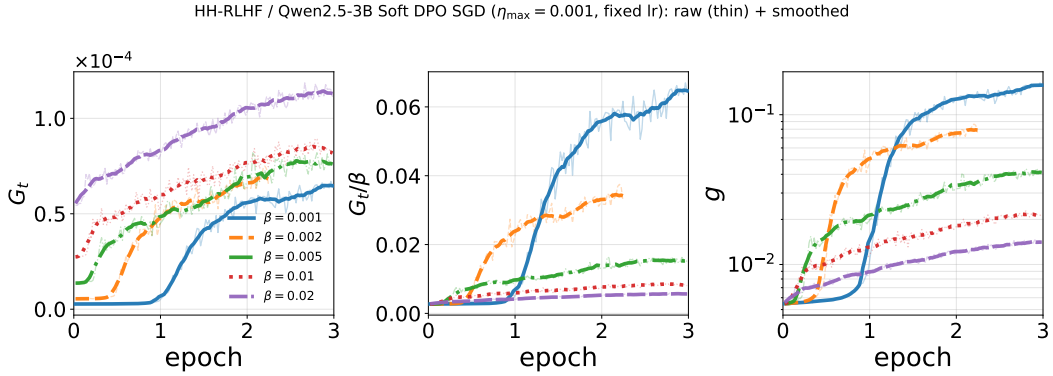

Figure K2: HH-RLHF / Qwen2.5-3B, our DPO implementation, SGD, $\eta_{\max} = 10^{-3}$ (fixed across β). Same three-factor layout as Figure K1. Thin curves: raw; thick: moving average. Growth of the reconstructed batch-Jacobian proxy is stronger than on HelpSteer3 across the sweep.

/ Qwen3-4B standard-DPO SGD sweep we take the shape+scale estimate $\tau = 6.37 \times 10^{-4}$ from Appendix G.1; the rough β_{r} estimate below additionally folds in the learning-rate budget and an early weighted-Jacobian proxy g_B.

Learning-rate budget. Write η_t for the per-step learning rate, T_{budget} for the planned number of SGD steps, and

$$H := \sum_{t=0}^{T_{\mathrm{budget}}-1} \eta_t, \qquad \eta_{\mathrm{eff}} := H/T_{\mathrm{budget}}. \tag{L.1}$$

For a linear warmup+decay (triangular) schedule that rises from 0 to $\eta_{\max}$ and returns to 0, $H \approx \eta_{\max} T_{\mathrm{budget}}/2$.

L.1 From a parameter step to an equation on the margin

The DPO loss on a single pair depends on θ only through $\Delta(\theta)$. Appendix K records the chain rule

$$\nabla_\theta \mathcal{L} = \ell'(\Delta)\, \nabla_\theta \Delta = -\beta\, \sigma(-\beta\Delta)\, \nabla_\theta \Delta, \qquad g(\theta) := \|\nabla_\theta \Delta(\theta)\|_2, \tag{L.2}$$

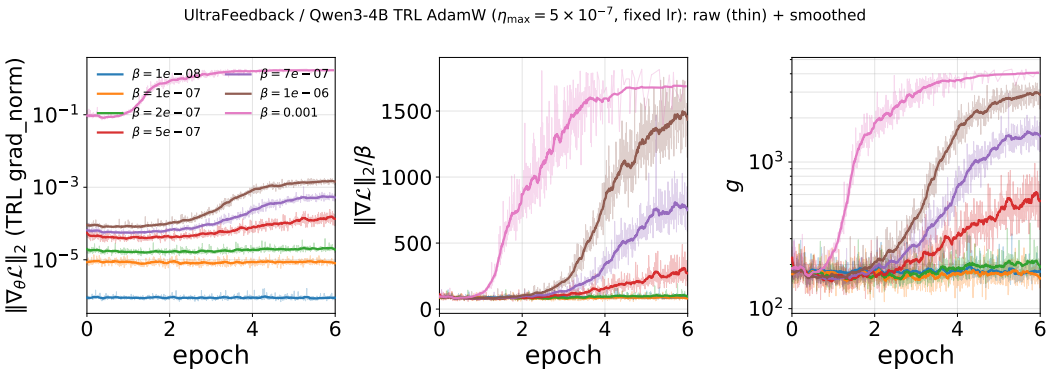

Figure K3: UltraFeedback / Qwen3-4B, TRL `DPOTrainer`, AdamW, $\eta_{\max} = 5 \times 10^{-7}$. Left: Euclidean $\|\nabla_\theta \mathcal{L}\|_2$ from `train/grad_norm`; center and right apply the same decomposition as Figure K1. Absolute scale is not comparable to the SGD figures above (different optimizer and η).

so $\|\nabla_\theta \mathcal{L}\|_2 = \beta\, \sigma(-\beta\Delta)\, g$. An SGD step $\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}(\theta_t)$ therefore changes the scalar margin by a first-order Taylor expansion

$$\Delta(\theta_{t+1}) - \Delta(\theta_t) = \nabla_\theta \Delta(\theta_t)^\top (\theta_{t+1} - \theta_t) + O(\|\theta_{t+1} - \theta_t\|^2). \tag{L.3}$$

Substituting the update and (L.2) gives the discrete drift

$$\Delta_{t+1} - \Delta_t = \eta_t\, g(\theta_t)^2\, \beta\, \sigma(-\beta\Delta_t) + O(\eta_t^2), \tag{L.4}$$

where $\Delta_t := \Delta(\theta_t)$. The factor g^2 appears because the SGD step is parallel to $\nabla_\theta \Delta$, so the induced change in the scalar margin is $\nabla_\theta \Delta^\top (\eta |\ell'|\, \nabla_\theta \Delta) = \eta |\ell'|\, \|\nabla_\theta \Delta\|_2^2$; it is the squared Jacobian of Δ, not $(\partial \mathcal{L}/\partial \Delta)^2$. Identifying one step with $dt = 1$ yields the mean-field ODE

$$\dot{\Delta}(t) = \eta(t)\, g\big(\theta(t)\big)^2\, \beta\, \sigma\big(-\beta \Delta(t)\big). \tag{L.5}$$

Here β enters explicitly through the sigmoid; g enters through $\theta(t)$, and the path $\theta(t)$ itself depends on β, so logged $g(t)$ need not be the same across a β-sweep (Appendix K, Figures K1–K3).

Exact minibatch correction. For a training batch, let $j_i = \nabla_\theta \Delta_i$, $a_i = \sigma(-\beta\Delta_i)$, and $K_{ki} = j_k^\top j_i$. The first-order drift of pair k is

$$\Delta_{k,t+1} - \Delta_{k,t} = \frac{\eta_t \beta}{B} \sum_{i=1}^{B} a_i K_{ki} + O(\eta_t^2). \tag{L.6}$$

Thus the exact dynamics depend on cross-pair alignment, not only on $\|j_k\|_2^2$. For a mean over M probe pairs, the corresponding drift is

$$\overline{\Delta}_{t+1} - \overline{\Delta}_t = \frac{\eta_t \beta}{MB}\, \mathbf{1}^\top K \mathbf{a} + O(\eta_t^2),$$

where K may be the cross-Gram matrix between probe and training pairs. A precise horizon estimate would evaluate or approximate these Gram contractions along the trajectory and integrate them under the actual learning-rate schedule.

Frozen-Jacobian closure. For a deliberately simpler order-of-magnitude estimate, we return to the one-dimensional closure and freeze $g(\theta(t)) \equiv g_0 := g(\theta_0)$. Numerically, g_0 is replaced by the early weighted-Jacobian proxy reconstructed from steps 100–500 in Appendix K; this does not make it an initialization measurement or a typical per-pair Jacobian. Then (L.5) separates:

$$\big(1 + e^{\beta\Delta}\big)\, d\Delta = \eta(t)\, g_0^2\, \beta\, dt. \tag{L.7}$$

With a variable schedule the right-hand side integrates to $g_0^2 \beta\, H$. If the proxy is not frozen, let $I(t)$ denote the cumulative impulse, with $I(0) = 0$ and $dI = \eta(t) g_B(t)^2\, dt$. Its terminal value and the corresponding learning-rate-weighted RMS proxy are

$$I_H := I(T_{\text{budget}}) = \sum_{t=0}^{T_{\text{budget}}-1} \eta_t g_B^2(t) = g_H^2 H, \qquad g_H^2 := \frac{\sum_t \eta_t g_B^2(t)}{\sum_t \eta_t}. \tag{L.8}$$

Thus g_H is the learning-rate-weighted RMS proxy along the trajectory, not its terminal value. Treating the logged proxy as a prescribed function of time, Equation (L.5) is the same ODE in the cumulative impulse $I(t)$:

$$\left(1 + e^{\beta\Delta}\right) d\Delta = \beta\, dI \qquad \text{i.e.} \qquad \frac{d\Delta}{dI} = \beta\, \sigma(-\beta\Delta). \tag{L.9}$$

The frozen closure (L.7) is the special case $dI = \eta(t)\, g_0^2\, dt$. Integrating (L.9) from $\Delta(0) = 0$ to the terminal budget I_H yields the finite-horizon margin Δ_H:

$$\Delta_H + \frac{e^{\beta\Delta_H} - 1}{\beta} = \beta I_H. \tag{L.10}$$

With $u_H = \beta\Delta_H$ this is $u_H + e^{u_H} - 1 = I_H \beta^2$, hence $\Delta_H = u_H/\beta$. This is the *horizon-only* solution: without a stopping rule it grows without bound, although only logarithmically,

$$\Delta_H(I_H, \beta) \sim \frac{\log(I_H \beta^2)}{\beta} \qquad (I_H \beta^2 \to \infty). \tag{L.11}$$

When the residual stopping threshold and finite horizon are modeled as separate mechanisms, the achieved scalar margin is

$$\Delta_{\text{ach}}(I_H, \beta; \tau) = \min\{\Delta_H(I_H, \beta), \Delta^*_{\text{DPO}}(\beta; \tau)\}, \qquad \beta > 2\tau. \tag{L.12}$$

Thus $\Delta_{\text{ach}} \to \Delta^*_{\text{DPO}}$ as $I_H \to \infty$; Δ_H itself does not converge to a finite ceiling.

Logged versus frozen impulse. Figure L1 is a diagnostic of how (L.9)–(L.10) sit against logged SGD, not a calibration of Δ_H or of β_{r}. The columns are standard DPO on Orca-DPO / Phi-4-mini at $\text{lr} = 10^{-4}$, with β left of the empirical peak, near the peak, and on the right slope. Blue and black are *empirical*: the mean and median margin of 256 preference pairs drawn at random from the validation split and then held fixed for every logged checkpoint and all three runs (the same probe, not resampled, and not a single pair). Orange and dotted purple are the *scalar ODE* (L.10): $\Delta_H\big(I(t), \beta\big)$ with the logged impulse $I(t) = \sum_s \eta_s g_B(s)^2$, versus the frozen special case $I(t) = g_0^2 H(t)$. The bottom row is likewise empirical $g_B(t)$ reconstructed from the logged loss gradient, with a horizontal line at the frozen g_0 that enters the ODE. On this backbone the reconstructed proxy $g_B(t)$ grows by one to two orders of magnitude relative to $g_0 \approx 98$, and the learning-rate-weighted RMS satisfies $g_H/g_0 \in [14, 60]$ on the three runs, so the frozen ODE remains $O(1)$–$O(10)$ while the empirical probe mean reaches $O(10^2)$–$O(10^3)$. Substituting the logged $I(t)$ into the ODE restores the correct order of magnitude and approaches the empirical mean on the right slope at the horizon; it does not track the path at every epoch. The HelpSteer3 / Qwen3-4B estimate $\beta_{\text{r}}^{(0)}$ below freezes the early proxy only because its subsequent change is much milder: $g_B \approx 250$ at steps 100–500 rises to $\sim$ 360–370, i.e. a factor of about 1.5, whose effect is quoted after (L.18). Where that ratio is 10–60, the frozen special case $I_H = g_0^2 H$ does not locate β_{r} even to an order of magnitude; we therefore do not report a frozen β_{r} for Orca-DPO / Phi-4-mini.

Stochasticity. Equation (L.5) is a scalar mean-field closure of the minibatch recursion (L.6); random batch sampling adds stochastic variation around the corresponding Gram-weighted drift. Appendix M gives an optional stochastic interpretation of this variation and describes how an effective drift–noise crossover could be estimated from windowed margin trajectories. It does not assume convergence to an equilibrium process.

L.2 Time to saturation and definition of β_{r}

The cumulative impulse required for the horizon-only dynamics to reach the threshold-limited margin is obtained by setting $\Delta_H = \Delta^*_{\text{DPO}}$ in (L.10):

$$I_*(\beta, \tau) = \frac{1}{\beta^2}\left[\log\left(\frac{\beta}{\tau} - 1\right) + \frac{\beta}{\tau} - 2\right], \qquad \beta > 2\tau, \tag{L.13}$$

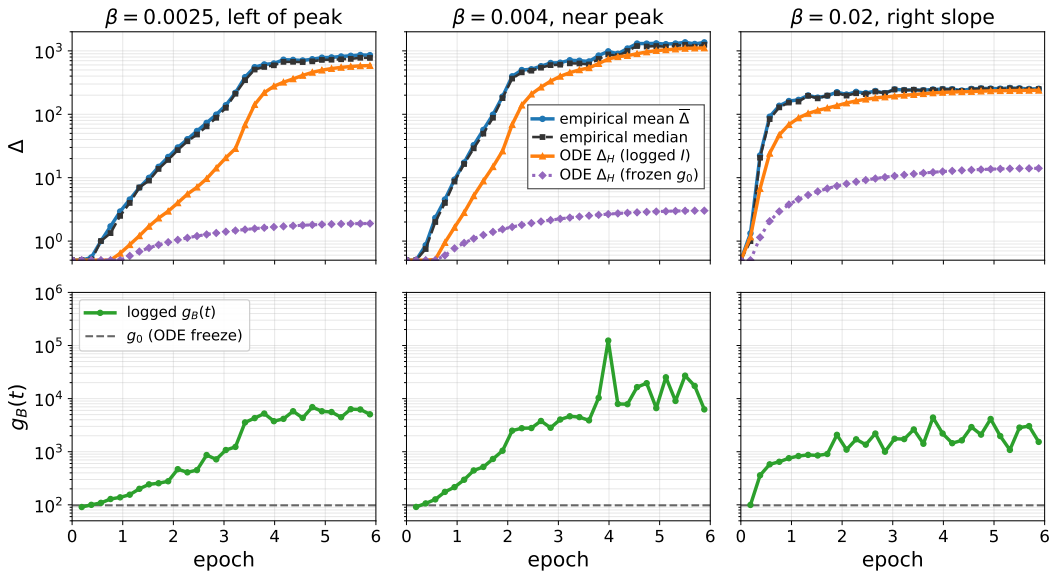

Figure L1: Empirical SGD versus the scalar ODE (L.9)–(L.10) on Orca-DPO / Phi-4-mini 3.8B (standard DPO, $\mathrm{lr} = 10^{-4}$, batch size 22, six epochs). Columns: β left of the empirical peak, near the peak, and on the right slope. Top: empirical mean and median margin on a frozen 256-pair probe (pairs drawn at random from the validation set and held fixed across all logs), versus the ODE solution Δ_H with logged impulse $I(t)$ and with frozen g_0. Bottom: empirically reconstructed $g_B(t)$ versus the frozen g_0 used in the ODE. Shared log Δ scale. The ODE with logged I matches the order of the empirical mean and approaches it on the right slope at the horizon; the frozen-g_0 ODE does not.

where we used $e^{\beta\Delta^*} = \beta/\tau - 1$ from Equation (8). Under the frozen closure, $I_H = \eta g_0^2 T$ for constant η, so the corresponding time to reach the threshold is

$$T_{\mathrm{DPO}}(\beta) = \frac{1}{\eta g_0^2 \beta^2}\left[\log\left(\frac{\beta}{\tau} - 1\right) + \frac{\beta}{\tau} - 2\right] \xrightarrow[\beta \gg \tau]{} \frac{1}{\eta g_0^2 \beta \tau}. \tag{L.14}$$

Thus $T_{\mathrm{DPO}} \to 0$ at both ends of the admissible interval: Δ^* collapses as $\beta \to (2\tau)^+$ and as $\beta \to \infty$. For a trajectory budget $I_H(\beta)$, the crossover is defined exactly and generally only implicitly by

$$I_H(\beta_{\mathrm{r}}) = I_*(\beta_{\mathrm{r}}, \tau). \tag{L.15}$$

On the right branch, $I_*(\beta, \tau) \sim 1/(\beta\tau)$. Under the frozen-proxy closure, $I_H = g_0^2 H$, and the leading right-branch approximation to (L.15) is therefore

$$\beta_{\mathrm{r}}^{(0)} \simeq \frac{1}{g_0^2\, \tau\, H} = \frac{\eta_{\mathrm{ref}}}{\varepsilon H}, \tag{L.16}$$

where the second equality uses Equation (G.1). For a linear-decay schedule, $H \simeq \eta_{\max} T_{\mathrm{budget}}/2$ and $\eta_{\mathrm{ref}} = \eta_{\max}$, hence

$$\beta_{\mathrm{r}}^{(0)} \simeq \frac{2}{\varepsilon T_{\mathrm{budget}}}. \tag{L.17}$$

Replacing $g_0^2 H$ by a trajectory-dependent budget $I_H(\beta)$ gives the corresponding leading right-branch approximation

$$\beta_{\mathrm{r}} \simeq \frac{1}{I_H(\beta_{\mathrm{r}})\tau} \simeq \beta_{\mathrm{r}}^{(0)} \left(\frac{g_0}{g_H}\right)^2. \tag{L.18}$$

If g_H/g_0 is independent of the learning-rate scale, this correction only shifts the boundary toward smaller β. If the ratio grows with the learning rate, the boundary bends toward smaller β at larger learning rates, but it does not generically become a $1/\eta$ hyperbola. For $\beta \gtrsim \beta_{\mathrm{r}}$ the scalar budget is sufficient to reach the threshold-limited margin $\Delta^*(\beta)$, but that margin is already small, so policy movement is practically negligible. Equations (L.15)–(L.18) therefore provide a characteristic crossover scale, not a sharp threshold on the β-axis. Under minibatch training they are scalar proxies for the Gram dynamics in Equation (L.6), so we retain only their order of magnitude.

Table L1: HelpSteer3 / Qwen3-4B standard DPO (SGD): right slope of the sweep in Figure 2 and Table G1. Min validation pair NLL (with $\beta = 1$), per-token KL, and mean validation margin; last column is the one-pair ceiling $\Delta^*_{\mathrm{DPO}}(\beta)$ from Equation (8) at $\tau = 6.37 \times 10^{-4}$. Highlighted row: the rough frozen-proxy estimate $\beta_{\mathrm{r}}^{(0)}$ from (L.19); NLL, $\overline{\Delta}$, and KL are interpolated log-linearly in β between the neighboring measured points 0.02 and 0.04, while $\Delta^*_{\mathrm{DPO}}(0.027)$ is the closed-form ceiling. The highlighted value is an order-of-magnitude guide, not a measured or calibrated cutoff.

	β	min NLL	KL / token	mean val. $\overline{\Delta}$	$\Delta^*_{\mathrm{DPO}}(\beta)$
β_{peak}	0.0025	9.92 ± 0.14	1.00 ± 0.02	152 ± 6	429
	0.01	11.55 ± 0.08	0.540 ± 0.008	75.1 ± 0.6	269
	0.02	12.61 ± 0.01	0.290 ± 0.024	45.4 ± 0.0	171
$\beta_{\mathrm{r}}^{(0)}$	0.027	≈ 12.88	≈ 0.221	≈ 38.5	138
	0.04	13.23 ± 0.01	0.130 ± 0.007	29.4 ± 0.0	103
	0.075	13.36 ± 0.02	0.066 ± 0.001	21.9 ± 0.3	63
	0.2	13.67	0.019	14.9	29
	0.3	13.80	0.010	12.3	21

L.3 Jacobian input

At exact initialization, $g_B(\theta_0) = 2\|\nabla_\theta \mathcal{L}_B(\theta_0)\|_2/\beta$, with the Euclidean minibatch-loss gradient (not mean-abs G). This identity gives the norm of the batch-averaged Jacobian, not a typical $\|\nabla_\theta \Delta_i\|_2$. Because the initialization gradient was not logged, the numerical estimate below substitutes the early weighted-Jacobian proxy g_B^{early} reconstructed in Appendix K.

L.4 Numerical estimate on HelpSteer3

Protocol and inputs. Standard DPO on HelpSteer3 / Qwen3-4B-Instruct-2507 with SGD, batch size 24, six epochs (same sweep as Figure 2 and Table G1): $T_{\mathrm{budget}} = 9078$, $\eta_{\max} = 2 \times 10^{-4}$, hence $H = \eta_{\max} T_{\mathrm{budget}}/2 = 0.9078$, and $\tau = 6.37 \times 10^{-4}$ as above. Appendix K supplies $g_B^{\mathrm{early}} \approx 250$ from the measured Euclidean loss gradient at steps 100–500. We substitute it for g_0 in the scalar closure. Substituting into (L.16) yields

$$\beta_{\mathrm{r}}^{(0)} \approx 0.027. \tag{L.19}$$

The reconstructed proxy reaches roughly 360–370 later in training. Holding the other inputs fixed, Equation (L.18) would move β_{r} downward by about a factor of two, from the frozen estimate 0.027 to roughly 0.012–0.013. Cross-pair cancellation introduces an additional, unquantified difference between g_B and a typical per-pair Jacobian. These factor-of-few shifts occupy less than one decade on the logarithmic sweep and are why we report a broad order-of-magnitude region rather than a precise boundary.

Log diagnostic of the scalar ODE. Windowed SGD logs give a diagnostic of (L.5)–(L.7), not a recalibration of (L.16). We compare the empirical window drift of the training-mean margin to the frozen-proxy right-hand side, using g_B reconstructed from the logged Euclidean loss gradient and, where recorded, the batch factor $\beta \,\mathrm{mean}_i\, \sigma(-\beta\Delta_i)$ in place of the scalar $\beta\sigma(-\beta\overline{\Delta})$. When the mean margin remains moderate and $\mathrm{mean}_i\, \sigma(-\beta\Delta_i) \approx \sigma(-\beta\overline{\Delta})$, the predicted $d\Delta/dH$ tracks the empirical drift up to an $O(1)$ factor. At large $\overline{\Delta}$, the scalar closure can understate the remaining drift for two reasons. First, the reconstructed Jacobian proxy grows relative to the early value used in the frozen estimate. Second, where the per-example sigmoid factor is recorded, we empirically observe that $\mathrm{mean}_i\, \sigma(-\beta\Delta_i)$ can exceed $\sigma(-\beta\overline{\Delta})$. This second effect reflects margin heterogeneity; it is not a general consequence of Jensen's inequality because the sigmoid changes convexity at zero. The growing Jacobian, and any empirically verified increase in the batch-averaged sigmoid factor, imply that the frozen estimate $\beta_{\mathrm{r}}^{(0)} = 0.027$ is shifted to the right relative to an instantaneous estimate. Substituting the later proxy $g \approx 360$–370 gives $\beta_{\mathrm{r}} \approx 0.012$–$0.013$. This diagnostic does not replace the Gram dynamics (L.6); we keep β_{r} as a broad order-of-magnitude band.

Comparison with the empirical right slope. Table L1 places (L.19) on the same β-axis as Figure 2 (highlighted row). The empirical entries at $\beta_{\mathrm{r}}^{(0)}$ (NLL ≈ 12.88, KL ≈ 0.221, $\overline{\Delta} \approx 38.5$) sit smoothly

between $\beta = 0.02$ and $\beta = 0.04$; the 138 in the last column is $\Delta^*_{\mathrm{DPO}}(0.027)$ from Equation (8), between the neighboring ceilings 171 and 103. As throughout Appendix G.1, the pair-averaged $\overline{\Delta}$ lies well below the one-pair ceiling. Training remains clearly active at $\beta = 0.075$; by $\beta = 0.2$–0.3, pair NLL has risen most of the way back toward the dead-zone floor and per-token KL has collapsed, placing these runs on the empirical right-hand saturation slope. Equations (L.14)–(L.16) provide only a rough horizon scale under the scalar frozen-proxy closure; they do not predict the vertical scale of $\overline{\Delta}(\beta)$.

Takeaway. The left dead zone $\beta \le 2\tau$ is a property of the marginal gradient at $\Delta = 0$. The exact scalar crossover satisfies $I_H(\beta_{\mathrm{r}}) = I_*(\beta_{\mathrm{r}}, \tau)$; on the right branch this gives $\beta_{\mathrm{r}} \simeq 1/[I_H(\beta_{\mathrm{r}})\tau]$. The reported frozen value $\beta_{\mathrm{r}}^{(0)} \simeq 1/((g_B^{\mathrm{early}})^2 \tau H)$ uses the special case $I_H = g_0^2 H$ with the early proxy substituted for the unobserved initialization value. On this HelpSteer3 SGD sweep the frozen estimate is $\beta_{\mathrm{r}}^{(0)} \approx 0.027$, placing the right-hand crossover within the observed logarithmic range; Figure L1 is the contrasting case where g_H/g_0 is 10–60 and that special case is not even the correct decade. The windowed-log diagnostic above indicates that the frozen scalar closure is biased toward a larger β_{r} once g grows and $\overline{\Delta}$ is large, so the reported value should be read as sitting on the right of that scale. The precise minibatch value would require the trajectory-dependent Gram contractions in Equation (L.6); therefore neither 0.027 nor the shaded band should be read as a sharp prediction.

M Stochastic Dynamics of the Preference Margin

This appendix is not required for the main results of the paper. Its purpose is to provide additional intuition for the effective stopping tolerance τ, the nonzero movement observed inside the nominal dead zone, and the transition from directed motion to noise-dominated training. The deterministic threshold formulas in Sections 3.3 and Appendix G remain phenomenological finite-training descriptions; the stochastic view below suggests how their effective tolerance could be estimated from training dynamics without fitting the cross-β curve of Figure 2.

A coarse-grained stochastic observable. Let m_k denote a scalar margin observable at logging window k, ideally the mean margin on a fixed probe set (Figure M1) and otherwise the mean training margin aggregated over a window of optimizer steps. At a fixed model state, minibatch sampling makes its update random. We write the coarse-grained recursion as

$$m_{k+1} - m_k = D_k + E_k, \qquad \mathbb{E}[E_k \mid \mathcal{F}_k] = 0, \qquad \mathrm{Var}(E_k \mid \mathcal{F}_k) = Q_k^2, \tag{M.1}$$

where D_k is the directed window drift and Q_k is the innovation scale, both in units of margin per window. Let $H_k := \sum_{t \in k} \eta_t$ be the learning-rate impulse in window k. For standard DPO, the scalar mean-field closure of Appendices G–L writes

$$D_k \approx H_k g_{\mathrm{eff}}^2(k)\, a_k, \qquad a_k := \beta\, \mathbb{E}_i[\sigma(-\beta \Delta_i(k))]\,.$$

The exact minibatch drift contains the Gram contractions in Equation (L.6); g_{eff} is only their scalar proxy.

The corresponding continuous notation is

$$dm = b(m,t)\, dt + q(t)\, dW. \tag{M.2}$$

This process has no finite equilibrium under the positive DPO drift alone. What appears as a plateau in finite training is instead produced by a combination of sigmoid attenuation, stochastic fluctuations, learning-rate decay, and the finite horizon. Local perturbations around the moving mean path can contract, but this local loss of memory should not be confused with convergence of the mean to a fixed point.

Operational noise scale and stopping tolerance. We define a noise-dominated crossover at a chosen observation scale by

$$|D_k| \lesssim c\, Q_k,$$

where $c = O(1)$ specifies the desired signal-to-noise criterion. Because the window-scale effective mobility satisfies $H_k g_{\mathrm{eff}}^2(k) \approx D_k/a_k$, it can be eliminated to define the window-scale effective tolerance

$$\widehat{\tau}_k = c\, a_k\, \frac{Q_k}{|D_k|}.$$

At the crossover window k_*, where $|D_{k_*}| \approx cQ_{k_*}$,

$$\widehat{\tau} \approx a_{k_*} = \beta\, \mathbb{E}_i[\sigma(-\beta\Delta_i(k_*))]\,.$$

Thus a single well-resolved trajectory can provide an exploratory estimate: estimate its local drift and innovation scale, locate the noise-dominated crossover, and read out the marginal-gradient factor there. Repeated seeds or repeated fixed-state minibatch probes are needed for a reliable separation of drift and stochastic variation.

Practical estimation from windowed logs. Let W be the logging window in optimizer steps; in the present experiments $W = 100$. For seed s and window k, let $X_{s,k}$ be the mean training margin aggregated over that window. The following procedure gives a window-scale estimate:

1. Fit a smooth trajectory $\widehat{m}_\beta(k)$ to the aligned windows, preferably jointly across seeds.
2. Estimate the directed window drift by $D_{\beta,k} = \widehat{m}_\beta(k+1) - \widehat{m}_\beta(k)$.
3. Form residuals $r_{s,k} = X_{s,k} - \widehat{m}_\beta(k)$ and fit locally $r_{s,k+1} = \phi_{\beta,k} r_{s,k} + u_{s,k}$. Use $Q_{\beta,k} = \mathrm{Std}(u_{s,k})$ as the innovation scale.
4. Locate the first sustained region, for example three consecutive windows, in which $|D_{\beta,k}| \lesssim cQ_{\beta,k}$.
5. Compute $a_{\beta,k}$ from the per-example margins before averaging: $a_{\beta,k} = \beta\ \mathrm{mean}_i\, \sigma(-\beta\Delta_i)$. If this quantity is unavailable, $\beta\sigma(-\beta X_{s,k})$ is only a scalar approximation because sigmoid and averaging do not commute.
6. Report $\widehat{\tau}_\beta = a_{\beta,k_*}$ and combine the independently obtained estimates across β and seeds, for example by a median and a bootstrap interval.

The within-window standard deviation of the training margin is not the innovation scale $Q_{\beta,k}$ above: it mixes stochastic policy motion, deterministic trend, and changing minibatch composition. It can still be used to weight the window means. A cleaner estimate follows if one also records the window-endpoint margins, the mean and variance of per-step increments, their lag-one covariance, and the per-example factor $\beta\ \mathrm{mean}_i\, \sigma(-\beta\Delta_i)$. Evaluating a small fixed probe set on the same grid would additionally remove minibatch-composition noise. Figure M1 and Table M1 carry this out on Orca-DPO: a frozen 256-pair validation probe, five seeds, and D, Q estimated from the seed-wise increments of the probe mean rather than from AR residuals of training-window means.

Probe and a_t. The observable is the mean margin on a *fixed* set of $M = 256$ preference pairs, drawn once from the Orca-DPO validation split ($n_{\mathrm{val}} = 1286$) with `probe_seed` $= 0$ and held identical across all β and seeds. Every $W = 100$ optimizer steps we store the M pair margins $\Delta_i(t)$ and the scalar

$$a_t = \beta\, \frac{1}{M} \sum_{i=1}^{M} \sigma\big(-\beta\Delta_i(t)\big). \tag{M.3}$$

The sigmoid is applied *per pair* and then averaged. The commuting approximation $\beta\sigma(-\beta\overline{\Delta})$ is not used: Jensen's inequality would bias a_t once the probe margins spread. At initialization $\Delta_i \approx 0$, so $a_0 = \beta/2 = 4 \times 10^{-3}$ for $\beta = 0.008$, as in panel (a). This a_t is the DPO marginal-gradient factor in the scalar closure $\dot{\Delta} \approx \eta g^2 a_t$.

Trend, noise, and $\widehat{\tau}$. Write $X_{s,k}$ for the probe mean $\overline{\Delta}$ of seed s at snapshot k (step kW). The directed window increment and the seed-scale noise are

$$D_k = \frac{1}{S} \sum_{s=1}^{S} \big(X_{s,k+1} - X_{s,k}\big), \qquad Q_k = \mathrm{Std}_s\big(X_{s,k+1} - X_{s,k}\big), \tag{M.4}$$

with $S = 5$ and Std the sample standard deviation ($S - 1$ in the denominator). Thus D is the shared trend of $\overline{\Delta}$ over 100 steps and Q is the seed-to-seed scatter of that same move, not the within-probe standard deviation of Δ_i. A window is noise-dominated when $|D_k| \lesssim Q_k$ (we take $c = 1$). We take the first *sustained* such region—three consecutive windows—while the learning rate is still live, and read

$$\widehat{\tau} = \frac{1}{S} \sum_{s=1}^{S} a_{t_*}^{(s)}, \tag{M.5}$$

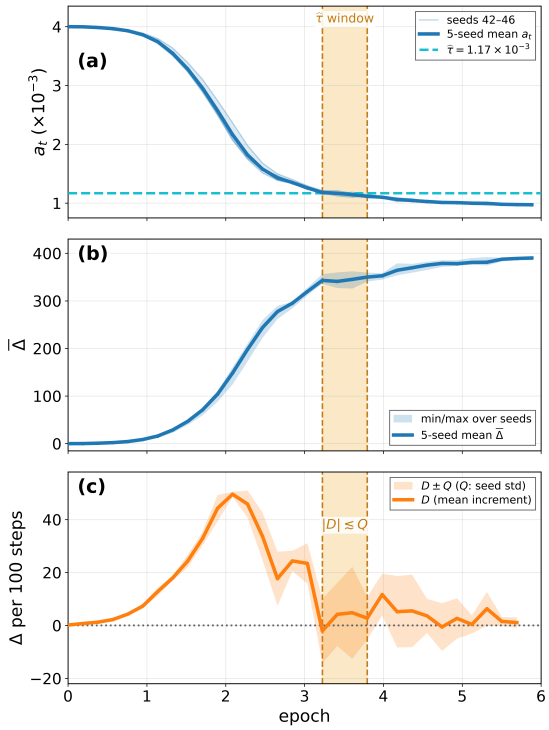

Figure M1: Drift–noise readout of τ for standard DPO on Orca-DPO / Phi-4-mini 3.8B (SGD, batch size 22, $\eta_{\max} = 5 \times 10^{-5}$, $\beta = 0.008$, linear warmup+decay, six epochs, $T = 3162$ steps). Five independent training seeds $\{42, 43, 44, 45, 46\}$ share a frozen 256-pair validation probe, evaluated every 100 steps. (a) Per-seed $a_t = \beta \operatorname{mean}_i \sigma(-\beta \Delta_i)$ and the five-seed mean. (b) Five-seed mean probe margin $\overline{\Delta}$ with the min/max envelope. (c) Window increment D (mean over seeds) and the band $D \pm Q$, where Q is the sample standard deviation of the five per-seed increments. The shaded band is the first sustained run of windows with $|D| \lesssim Q$, while η is still about half of $\eta_{\max}$. The cyan dashed line in (a) is $\widehat{\tau} = 1.17 \times 10^{-3}$, read from the five-seed mean a_t inside that band. The terminal $\eta \approx 0$ tail of a_t is not used.

where t_* is the readout snapshot inside that region. On this run the first three windows with $|D| < Q$ are $1700 \to 1800$, $1800 \to 1900$, and $1900 \to 2000$ (epochs 3.23–3.80), while η remains of order one half of $\eta_{\max}$. The five-seed mean a_t at step 1700 is 1.18×10^{-3}; at step 1800 it is 1.17×10^{-3}, which is the plotted readout. We report $\widehat{\tau} = 1.17 \times 10^{-3}$ with $t_* = 1800$. After the crossover a_t still creeps downward as η decays (1.12×10^{-3} at step 2000, 0.97×10^{-3} at $\eta \approx 0$); those later values are not τ.

The same protocol on $\beta = 0.0045$ never produces a sustained $|D| \lesssim Q$ before η collapses, so that cell is not used for the readout.

Table M1: Five-seed window increments of the probe mean $\overline{\Delta}$ at $\beta = 0.008$, $\eta_{\max} = 5 \times 10^{-5}$. D is the mean of the five increments and Q their sample standard deviation. $\overline{a}$ is the five-seed mean of a_t at the *start* of the window. The highlighted row is the readout window inside the first sustained run with $|D| < Q$, from which $\widehat{\tau}$ is read.

window	s42	s43	s44	s45	s46	D	Q	$\lvert D\rvert/Q$	$\overline{a}$
$1000 \to 1100$	$+45.2$	$+48.9$	$+47.4$	$+43.8$	$+36.1$	$+44.3$	5.0	8.9	2.57×10^{-3}
$1600 \to 1700$	$+35.7$	$+20.8$	$+14.9$	$+21.4$	$+24.4$	$+23.5$	7.7	3.1	1.26×10^{-3}
$1700 \to 1800$	-14.4	-4.7	$+17.2$	$+0.7$	-9.0	-2.1	12.1	0.17	1.18×10^{-3}
$1800 \to 1900$	-15.9	$+10.9$	$+2.3$	$+12.1$	$+11.7$	$+4.2$	11.9	0.35	1.17×10^{-3}
$1900 \to 2000$	$+28.3$	$+13.7$	-17.9	-2.1	$+2.0$	$+4.8$	17.4	0.28	1.14×10^{-3}

For reference, per-seed a_t at the onset step 1700: 1.149, 1.180, 1.176, 1.187, 1.213 (all $\times 10^{-3}$); mean 1.181×10^{-3}, standard error 0.010×10^{-3}. On the same $\eta = 5 \times 10^{-5}$ row of Figure G2, the shape+scale fit of $\overline{\Delta}(\beta)$ gives $\tau_{\text{fit}} = 1.45 \times 10^{-3}$ and the G.6 reading from the observed $\overline{\Delta}$ peak ($\beta = 0.006$) is $\tau = 1.31 \times 10^{-3}$. The probe run uses $\beta = 0.008$, slightly to the right of that observed peak, so a_t crosses the fitted tolerance scale while the learning rate is still live. The drift–noise readout $\widehat{\tau} = 1.17 \times 10^{-3}$ is the same order of magnitude and slightly below both of those estimates. The frozen closure $\tau = \varepsilon/(\eta g_0^2)$ on this learning-rate row, with the heatmap pair $(\varepsilon, g_0) = (7.27 \times 10^{-4}, 98.5)$ of Appendix H, gives 1.50×10^{-3}. This comparison should be interpreted as an order-of-magnitude diagnostic rather than as an identity between two estimators. The cross-β sweep estimates the phenomenological threshold $\tau(\eta_{\max})$ that controls the dead-zone and peak locations under a frozen reference scale. The trajectory readout instead gives a window-scale crossover $\widehat{\tau}_W = a_{k_*}$ at the live point of the learning-rate schedule, with Q_k measuring seed-to-seed scatter of 100-step probe increments. Repeating the same readout at other learning rates need not produce an exact $1/\eta_{\max}$ law unless the innovation scale Q_k, the effective mobility $H_k g_{\text{eff}}^2(k)$, and the position within the decay schedule scale in the corresponding way. Agreement with the sweep value therefore supports a noise-limited interpretation, while a systematic difference should be read as a diagnostic of window scale, learning-rate decay, Jacobian evolution, or seed-level noise, not as a failure of the cross-β estimate.

Left and right noise-dominated regimes. The same operational tolerance can act on both sides of the sweep. At initialization, $a_0 = \beta/2$; hence the left dead-zone condition is $\beta/2 \lesssim \tau$, so noise dominates before a directed transit develops. For larger β, a_t starts above tolerance and decreases as the margin grows, eventually crossing the same level on the right. This does not make the dead zone absolute: stochastic updates and finite-sample effects can still produce nonzero displacement, consistent with the measured KL at $\beta = 5 \times 10^{-4}$ in Table G1.

The τ inferred in this way is specifically a noise-scale tolerance at the chosen temporal resolution. The effective τ fitted in Figure 2 also absorbs finite-horizon and learning-rate-decay effects. Agreement between the two would indicate a noise-limited regime; a systematic difference would diagnose an additional horizon-limited contribution. If the apparent plateau occurs only after η_t has become negligible, τ is not identifiable from that plateau alone, because any trajectory freezes when the learning rate vanishes.

Takeaway. This optional appendix is intended to clarify, rather than support, the principal analytical and empirical claims of the paper. Its practical proposal is to treat τ as an operational drift–noise crossover: estimate the smooth margin drift and stochastic innovations from aligned trajectories, identify where noise becomes comparable to directed motion, and evaluate the DPO marginal-gradient factor at that point. No equilibrium distribution is assumed. The procedure can provide an exploratory estimate from one well-resolved trajectory; multiple seeds or repeated fixed-state minibatch probes, together with a fixed probe set, provide a more reliable and independently checkable estimate. Figure M1 is the latter on Orca-DPO at $\beta = 0.008$, $\eta_{\max} = 5 \times 10^{-5}$.

N Training Details and Compute Resources

Datasets and preprocessing. For HelpSteer3-Preference, we keep non-zero pairwise labels and exclude indifference cases ($\text{overall_preference} = 0$). For UltraFeedback Binarized, we use the provided chosen/rejected pairs directly. For HH-RLHF, we use the PKU-Alignment processed release with explicit context/chosen/rejected fields (PKU-Alignment, 2023) and cite the original Anthropic corpus (Anthropic, 2022; Bai et al., 2022). Prior data-centric and reward-modeling analyses report that HH-RLHF is comparatively noisy, less informative, and prone to weak or ambiguous pairwise margins (Shen et al., 2024; Wang et al., 2024); we discuss the resulting training dynamics in Appendix C. For Orca-DPO we use the public Hugging Face release `Intel/orca_dpo_pairs` (Intel, 2024): chosen responses are OpenOrca-style GPT-4 / GPT-3.5 traces (Lian et al., 2023; Mukherjee et al., 2023) and rejected responses are from Llama-2-13B-chat. We keep the provided chosen/rejected fields and hold out a 10% validation split (seed 42), yielding $11{,}573/1{,}286$ training/validation pairs. This follows common preference-learning practice that emphasizes unambiguous winner/loser supervision.

Splits, sequence lengths, and training times. Dataset splits are as follows: HelpSteer3-Preference has 36,299/1,920 training/validation pairs, UltraFeedback has 61,054/1,997, HH-RLHF has 159,700/8,492, and Orca-DPO has 11,573/1,286. For all our experiments, the input context is constrained to 768 prompt tokens, with a 1,536-token cap on the full prompt-plus-response sequence. Typical one-epoch training times are 12.4 h for HelpSteer3 with Qwen3-4B-Instruct-2507, 15.5 h for UltraFeedback Binarized with Qwen3-4B-Instruct-2507, and 19.9 h for HH-RLHF with Qwen2.5-3B-Instruct. Depending on the experiment, runs are trained for three or six epochs (see figure captions and per-run configurations); for UltraFeedback and HH-RLHF, validation is also performed at mid-epoch.

Mamba-2 backbone. For HH-RLHF experiments with a Mamba-2 backbone (including the TRL baseline runs in Appendix C and the β_{crit} analysis in Appendix B), we use the public Hugging Face checkpoint `AntonV/mamba2-2.7b-hf` (AntonV, 2024). The full model has $2{,}716{,}050{,}944$ parameters. We fine-tune with LoRA ($r = 16$, $\alpha = 32$, dropout 0.05, target modules `in_proj`, `x_proj`, `dt_proj`), leaving $13{,}451{,}264$ trainable parameters ($\approx 0.50\%$ of the total).

Phi-4-mini backbone. For Orca-DPO experiments (the centered-softplus (η, β) grid in Figure 5 and Appendix H; the standard-DPO slices in Figure G2 and Appendix L; the five-seed probe readout in Figure M1), we use the public Hugging Face checkpoint `microsoft/Phi-4-mini-instruct` (Microsoft, 2025a;b). The full model has $3{,}836{,}021{,}760$ parameters. We fine-tune with LoRA ($r = 16$, $\alpha = 32$, dropout 0.05, target modules `qkv_proj`, `o_proj`, `gate_up_proj`, `down_proj`), leaving $23{,}068{,}672$ trainable parameters ($\approx 0.60\%$ of the total). These runs use SGD, batch size 22, and six epochs, with seed 42 except the Appendix M readout, which uses seeds $\{42, 43, 44, 45, 46\}$ on a frozen 256-pair validation probe.

Additional backbones and evaluation models. Appendix I additionally uses the community checkpoint `ministral/Ministral-3b-instruct` (ministral, 2024) (not an official Mistral AI release) on UltraFeedback Binarized. For downstream checks in Tables J2 and J1, AlpacaEval 2 pairwise scoring uses `Qwen/Qwen2.5-14B-Instruct` (Qwen Team, 2024) as a fixed judge model; IFEval uses the official verifiable benchmark without a judge (Zhou et al., 2023).

Optimization. Optimizer choices and key hyperparameters are reported separately for each run, including learning rate, β, batch size, gradient accumulation, random seed, and gradient clipping. In the TRL baseline (`DPOTrainer`), we use per-device batch size 4 with gradient accumulation 4 (effective batch size 16). In our code runs, we use dataset-specific batch sizes: UltraFeedback 26, HelpSteer3-Preference 24, HH-RLHF 32, and Orca-DPO 22. The learning-rate schedule is linear warmup from 0 to $\eta_{\max}$ over 5% of the planned optimizer steps, then linear decay to 0 (a triangular schedule). For the HelpSteer3 SGD sweep this is ≈ 450 of $T_{\text{budget}} = 9078$ steps. We optimize with either AdamW or SGD, depending on the experiment (see per-run settings). AdamW uses the PyTorch defaults $(\rho_1, \rho_2) = (0.9, 0.999)$ for the moment decays and $\epsilon = 10^{-8}$. Weight decay is recorded in the released run configurations; we do not retune it as a function of β.

Monte Carlo KL estimation. We compute the reported KL ourselves as a Monte Carlo estimate of the policy-to-reference divergence $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})$ (sometimes called reverse KL when π_{ref} is treated as the target): for each selected validation prompt, the policy generates multiple responses, and we average $\log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$ over samples $y \sim \pi_\theta(\cdot \mid x)$. In this experiment, KL is estimated on the first 256 validation prompts with 8 samples per prompt and reported both per sequence and per token.

Hardware and software. All reported experiments were run in the cloud on NVIDIA A100-SXM4-80GB GPUs with driver 560.35.03. The software environment used Python 3.10.12, PyTorch 2.5.1+cu121 with CUDA 12.1 and cuDNN 90100, TRL 0.18.2, Transformers 4.53.2, Accelerate 1.7.0, and PEFT 0.18.1.