On the Dilution of On-Policy Gradients in Direct Preference Optimization

Anonymous ACL submission

Abstract

On-policy preference data is a natural remedy for distribution mismatch in Direct Preference Optimization (DPO), yet empirical reports disagree on whether it improves or degrades alignment. We revisit this puzzle and find that the apparent contradiction is largely moderated by the DPO coefficient β: the same on-policy pool hurts at high β but helps at low β, with population-level sign flips observed on both Qwen3-4B and Llama-3.1-8B across multiple seeds. We trace this behavior to *gradient dilution*: at standard β, on-policy examples retain high gradient weight but develop negligible preference margins, so their gradients remain large while contributing little useful signal. Motivated by this mechanism, β-split DPO decouples the coefficients for on- and off-policy sources and reverses the on-policy harm across all four seeds in the primary setting. We measure preference discrimination via held-out pairwise accuracy; an important caveat is that these gains trade off against generation quality, with our best 4-seed-validated recipes (0.6419 ± 0.0058 standard DPO; 0.655 ± 0.007 β-split) both losing on MT-Bench. The contribution is therefore best read as a diagnosis of preference-discrimination training dynamics, accompanied by a preference–generation Pareto frontier that practitioners must navigate.

1 Introduction

Direct Preference Optimization (DPO) has become a default recipe for preference tuning because it replaces reward-model training and PPO with a direct classification objective (Rafailov et al., 2023; Schulman et al., 2017). The next common step is to add on-policy preferences: sample responses from the model being trained, label the sampled pairs with a stronger judge, and train DPO on the enlarged mixture (Xiong et al., 2024; Guo et al., 2024; Chen et al., 2024). This practice should reduce distribution shift, but the empirical record is inconsistent. Some online and iterative DPO studies report clear gains; other settings find weak gains or degradation when more model-generated preference pairs are added (Xu et al., 2024; Wang et al., 2025). Throughout, our main metric is held-out pairwise preference discrimination; we treat the resulting models as preference discriminators (relevant to reranking and reward scoring (Lambert et al., 2024)) and report MT-Bench generation quality as a separately tracked deployment caveat.

The disagreement is not explained by whether data is on-policy alone. It is moderated by β, the scalar in DPO that controls the reference-model log-ratio entering the sigmoid. In a controlled 3×2 ablation on Qwen3-4B (Yang et al., 2025) and UltraFeedback (Cui et al., 2024), on-policy data hurts at β=0.02, moving pairwise accuracy from $.5767 \pm .0026$ to $.5690 \pm .0023$ (-0.8 pp). The same data helps at β=0.002, moving from $.6117 \pm .0064$ to $.6294 \pm .0038$ ($+1.8$ pp). The interaction is large relative to seed noise: the difference-in-differences estimate is $+0.025$ with 95% CI $[0.012, 0.034]$ and Cohen's d=3.36. Every one of the 4 seeds shows the same sign pattern at the two endpoints.

The mechanism is *gradient dilution*. DPO weights each example's gradient by $\beta\sigma(-\beta\Delta_i)$, where Δ_i is the training-time margin (Eq. 1). Training-time diagnostics at β=0.02 reveal a source-conditioned asymmetry: off-policy examples develop large preference margins ($\beta\Delta \approx 0.9$, $\sigma \approx 0.35$) as the model learns them, while on-policy examples remain near $\beta\Delta \approx 0.05$ with near-maximal gradient weight ($\sigma \approx 0.49$). On-policy examples are not gradient-suppressed; they receive full weight but carry little directional preference signal because the model cannot distinguish the two on-policy candidates. At high β, these large, low-effective-signal gradients introduce optimization behavior consistent with gradient dilution. A pre-training proxy ($\sigma(-\beta\delta^{\text{ref}})$, Eq. 4) identifies an

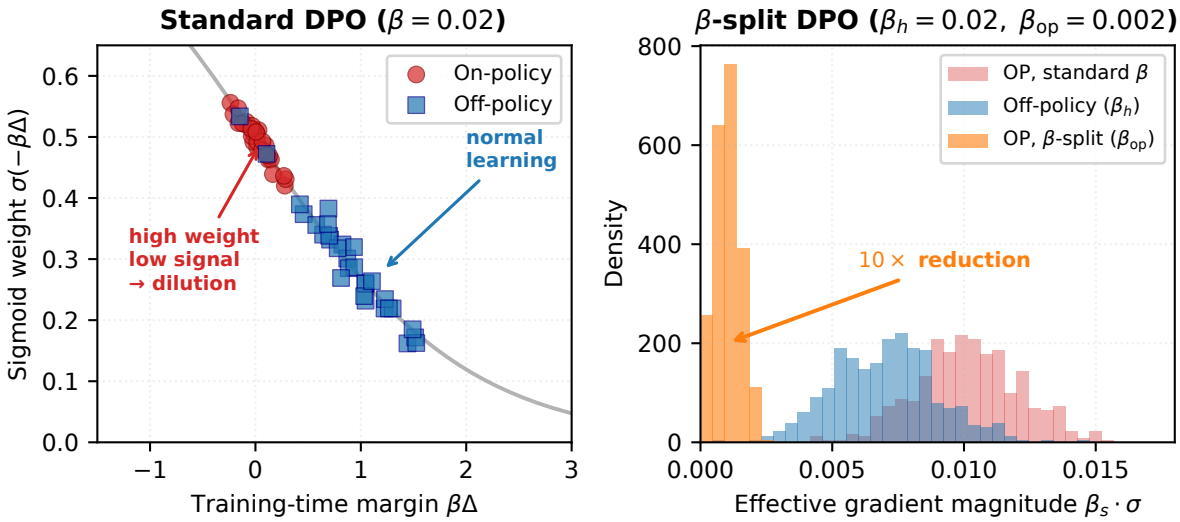


Figure 1: **Gradient dilution mechanism and β-split fix.** *Left*: under standard DPO (β=0.02), on-policy examples (red) cluster near $\beta\Delta\approx0$ with maximal sigmoid weight $\sigma\approx0.5$, contributing large but low-effective-signal gradients, while off-policy examples (blue) develop larger margins as the model learns. *Right*: β-split reduces on-policy gradient magnitude by 10× via a smaller β_{op} prefactor (orange), suppressing dilution while keeping off-policy pairs at the standard β_h regime.

apparently harmful OP subset, and a subset intervention is consistent with this: training on only proxy-flagged OP examples produces the worst outcome (Figure 1).

This mechanism yields a targeted fix. β-split DPO keeps the standard β_h=0.02 for human and AI-judge off-policy pairs and assigns a smaller β_{op}=0.002 to on-policy pairs. Because the per-example gradient magnitude scales as $\beta_{\text{op}} \cdot \sigma(\cdot)$, this reduces the on-policy gradient contribution by roughly 10×, suppressing dilution while keeping off-policy pairs at the standard β_h regime where reference regularization is intact. β-split reverses the high-β on-policy effect from −0.77 pp under standard DPO to +0.31 pp over the no-on-policy baseline, with a positive reversal in all 4 seeds.

The best preference discriminator comes from pushing the same interaction harder. With global β=0.002, on-policy data, and LoRA (Hu et al., 2022) rank r=64, Qwen3-4B reaches 0.6419±0.0058 pairwise accuracy, a +6.5 pp gain over the β=0.02 no-on-policy r=32 baseline. The gain decomposes into low β (+3.5 pp), on-policy data (+1.8 pp), and higher LoRA rank (+1.3 pp). This number is not a free improvement in general generation. Position-debiased MT-Bench (Zheng et al., 2023) against the baseline shows that the best recipe wins 6 of 80 questions, loses 53, ties 21, and produces shorter responses on average (1,706 vs. 2,936 characters). First, we identify and quantify the β–data interaction in DPO on Qwen3-4B with UltraFeedback (4 seeds, DiD with CI), with directionally consistent cross-model (Qwen3-0.6B) and cross-dataset (DPO-Mix-7k) checks. Second, we connect the interaction to gradient dilution and validate the mechanism with training-time diagnostics, a data-subset intervention, and threshold sensitivity analysis. Third, we introduce β-split DPO as a mechanism-derived intervention and map the preference–generation frontier in our setting, showing where on-policy data is Pareto-useful and where capacity-matched β tuning without on-policy data is equally efficient.

2 Related Work

DPO and β sensitivity. DPO turns preference optimization into a binary classification loss against a fixed reference policy (Rafailov et al., 2023), and many later objectives adjust the loss geometry or reference dependence (Azar et al., 2024; Meng et al., 2024; Ethayarajh et al., 2024). The original DPO work used larger β values for full fine-tuning, while LoRA-based recipes often use β=0.02 (Tunstall et al., 2023; Hu et al., 2022). Recent studies show that optimal β depends on scale and data quality (Xu et al., 2024; Tang et al., 2024); β-DPO adapts β dynamically from per-example quality signals (Wu et al., 2024a). These works treat β as a tuning target for overall performance. Our result is orthogonal: a *static* β choice flips the *sign* of

on-policy data's effect, making β a moderator of data strategy, not only of regularization strength.

On-policy and online preference data. Iterative DPO, online DPO, and self-play fine-tuning sample from the current or recent policy to reduce distribution mismatch (Xiong et al., 2024; Guo et al., 2024; Chen et al., 2024; Wu et al., 2024b; Yuan et al., 2024). WPO directly addresses on-policy/off-policy distribution shift via per-example reweighting (Zhou et al., 2024), and InCo-DPO frames the on-policy/off-policy choice as a quality–distribution trade-off (Wang et al., 2025); a recent survey (Liu et al., 2025) catalogs the broader DPO design space. These methods vary β, scale, and data pipeline simultaneously, so they cannot isolate whether the on-policy gain comes from better coverage or from the β regime. Our controlled ablation holds everything else fixed and shows the same on-policy pool is harmful at β=0.02 and helpful at β=0.002, offering a candidate reconciliation for the conflicting reports.

Gradient imbalance and dilution. Ma et al. (2025) identify gradient imbalance as a general source of DPO instability and propose gradient-level corrections. Our work identifies a specific, empirically measurable instance: at standard β, on-policy examples maintain near-maximal gradient weight but develop negligible preference margins, introducing low-effective-signal updates consistent with gradient dilution. Where Ma et al. (2025) characterize the problem broadly, we isolate a source-conditioned pathway, validate it with training-time diagnostics, and test a minimal source-conditioned fix (β-split).

Reference-free objectives. SimPO removes the reference model and uses length-normalized margins (Meng et al., 2024); IPO and KTO make different utility assumptions (Azar et al., 2024; Ethayarajh et al., 2024); ORPO combines preference and SFT in a single reference-free objective (Hong et al., 2024). Our SimPO check (3 seeds) shows a weaker but directionally similar β–data interaction (DiD $+0.007$, roughly $4\times$ smaller than DPO at the same scale). The length-normalized SimPO reward $r(y) = (\beta/|y|)\sum \log \pi(y|\cdot)$ implicitly source-conditions the per-example coefficient when $|y|$ differs across sources: in our data, OP responses have $|\bar{y}| = 37.7$ tokens (harmonic mean) vs. 42.2 for off-policy, so OP examples receive an effective $1.12\times$ per-token coefficient under SimPO even without a reference model. The residual SimPO interaction is therefore consistent with this implicit length-induced source weighting; the larger DPO interaction reflects the additional, dominant role of the reference log-ratio margin in the gradient-dilution pathway. Preference labels in our study follow standard LLM-as-judge practice with position debiasing (Zheng et al., 2023).

3 Method

Preference setup. Each training example is a tuple (x, y_w, y_l, s) with prompt x, preferred response y_w, rejected response y_l, and source $s \in \{\text{off}, \text{op}\}$. The off-policy set contains 764 human UltraFeedback pairs and 764 AI-judge pairs; the on-policy set contains 2,292 Kimi K2-labeled pairs generated by the Qwen3-4B pre-DPO checkpoint. We use "on-policy" in the reference-policy sampled sense: responses are drawn once from the same checkpoint that DPO uses as π_{ref}, and are not refreshed during training. The no-on-policy condition therefore has 1,528 pairs, and the full mixture has 3,820 pairs. Pair labels are position-debiased where judges are used; the Kimi confidence is the fraction of identical verdicts across two prompt orderings, and pairs with confidence ≥ 0.5 are retained. The primary model is Qwen3-4B (Yang et al., 2025); additional cross-validation checks include Qwen3-0.6B, DPO-Mix-7k (Argilla, 2024), Llama-3.1-8B (Grattafiori et al., 2024), Llama-3.2-1B, and a SimPO (Meng et al., 2024) replication.

DPO as a dilution-sensitive loss. For a pair i, define the *training-time margin*

$$\Delta_i = \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}. \quad (1)$$

At initialization ($\pi_\theta = \pi_{\text{ref}}$), $\Delta_i = 0$ for every example and the sigmoid weight $\sigma(-\beta\Delta_i) = 0.5$. As training progresses, Δ_i grows for examples the model learns to discriminate, and their sigmoid weight decreases. Standard DPO minimizes

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_i \log \sigma(\beta\Delta_i). \quad (2)$$

The gradient of the loss with respect to θ for a single pair i is

$$\nabla_\theta \mathcal{L}_i = -\underbrace{\beta\,\sigma(-\beta\Delta_i)}_{\text{effective weight}} \nabla_\theta \Delta_i. \quad (3)$$

The effective weight $\beta\,\sigma(-\beta\Delta_i)$ controls how much each example influences the update. For

learned examples ($\Delta_i \gg 0$), $\sigma \to 0$ and the gradient vanishes (normal convergence). For unlearned examples ($\Delta_i \approx 0$), $\sigma \approx 0.5$ and the full $\beta/2$ weight is retained. If such an example's $\nabla_\theta \Delta_i$ carries no consistent preference signal, the retained weight propagates a low-signal update; we refer to this pathway as gradient dilution (Figure 1).

The *reference preference gap* is a static, pre-training data property:

$$\delta_i^{\text{ref}} = \log \pi_{\text{ref}}(y_w|x) - \log \pi_{\text{ref}}(y_l|x). \quad (4)$$

We use $\sigma(-\beta\delta_i^{\text{ref}})$ as a pre-training proxy for identifying examples at risk of contributing low-effective-signal gradients during training: examples with large positive δ^{ref} are ones the reference model already "knows the answer to," and their gradient steps tend to reinforce an existing preference direction rather than build new discrimination.

Why on-policy data interacts with β. On-policy responses are sampled from the model being trained and then externally labeled. Because both candidate responses lie near the current model distribution, the policy cannot easily separate them: both are roughly equally likely under π_θ. This means Δ_i develops slowly for on-policy pairs during training, keeping their effective weight $\beta\,\sigma(-\beta\Delta_i) \approx \beta/2$ near the maximum. Training-time diagnostics (Section 4.3) confirm this: at β=0.02, on-policy examples maintain $\beta\Delta \approx 0.05$ and $\sigma \approx 0.49$ throughout training, while off-policy examples develop $\beta\Delta \approx 0.9$ and $\sigma \approx 0.35$. On-policy data is therefore not gradient-*suppressed*; it is gradient-*diluting*: large weight, no directional signal.

The interaction with β follows directly from Eq. 3. At high β (e.g. 0.02), the on-policy effective weight is $0.02 \times 0.5 = 0.01$, comparable to off-policy's $0.02 \times 0.35 = 0.007$—low-signal updates carry weight comparable to high-signal ones. At low β (e.g. 0.002), on-policy's effective weight drops to $0.002 \times 0.5 = 0.001$, an order of magnitude smaller, making the dilution negligible relative to the coverage benefit of on-policy diversity.

β-split DPO. The dilution mechanism suggests a source-conditioned coefficient:

$$\mathcal{L}_{\beta\text{-split}}(\theta) = -\mathbb{E}_i \log \sigma(\beta_{s_i}\Delta_i), \quad (5)$$

$$\beta_{s_i} = \begin{cases} \beta_h, & s_i = \text{off}, \\ \beta_{\text{op}}, & s_i = \text{op}. \end{cases} \quad (6)$$

The default intervention sets β_h=0.02 and β_{op}=0.002 ($= 0.1\,\beta_h$). This ratio is chosen so that on-policy effective weight ($\beta_{\text{op}} \cdot \sigma \approx 0.001$) is roughly 7× smaller than off-policy's ($\beta_h \cdot \sigma \approx 0.007$), effectively downweighting the diluting source while keeping off-policy pairs at the standard β_h regime. β-split adds no parameters, auxiliary networks, or new labels. The β_{op} sensitivity analysis in Section 4.3 shows that the intervention is robust to the exact ratio: a plateau exists in $\beta_{\text{op}}/\beta_h \in [0.1, 0.25]$.

4 Experiments

4.1 Setup

The primary ablation crosses $\beta \in \{0.02, 0.005, 0.002\}$ with data composition (no on-policy vs. +on-policy) on Qwen3-4B, using 4 seeds per cell (Table 1). The no-on-policy condition contains 764 human and 764 AI-judge pairs; the +on-policy condition adds 2,292 Kimi K2-labeled on-policy pairs. The reference policy is the pre-DPO checkpoint. Unless noted, multi-seed values are reported as mean ± sample standard deviation (Bessel-corrected).

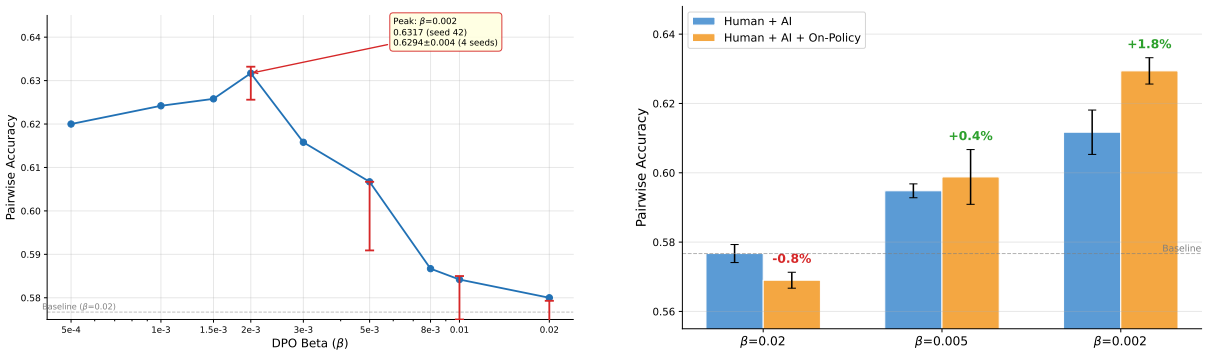
Training details. All runs use LoRA (Hu et al., 2022) with target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`; AdamW (β_1=0.9, β_2=0.999, weight decay 0), cosine schedule, no warmup, per-device batch 4, gradient accumulation 4, max length 512, lr 2×10^{-5}, 2 epochs, `bf16` on A800-SXM4-80GB. Baseline: r=32, α=64; best recipe: r=64, α=128; dropout 0.05. On-policy responses are sampled from the pre-DPO checkpoint (temperature 0.7, top-p 0.9, max length 512) and labeled by Kimi K2 with position-debiased pairwise judging; pairs with confidence ≥ 0.5 are retained.

Evaluation. Pairwise accuracy on 1,200 held-out UltraFeedback pairs, scored by answer-token mean log probability. Generation quality: position-debiased MT-Bench (80 questions × 2 orderings, Qwen3-14B judge, `enable_thinking=False`, `max_tokens=2048`, verbose system prompt) (Zheng et al., 2023).

4.2 The β–Data Interaction

On-policy data changes sign as β decreases. Table 2 shows the primary 3×2 ablation.[1] At β=0.02,

[1]Throughout, Δ columns report absolute accuracy differences in percentage points (pp), not relative percent change.



(a) Full-mixture β sweep.

(b) $\beta\times$data interaction.

Figure 2: Low β unlocks on-policy gains. The full-mixture sweep peaks near β=0.002, and the controlled 3×2 ablation shows a sign flip in the on-policy effect.

Component	Setting
Primary model	Qwen3-4B
Cross-scale model	Qwen3-0.6B
Primary data	UltraFeedback
Cross-dataset	DPO-Mix-7k
No-OP mixture	764 human + 764 AI-judge
+OP mixture	no-OP + 2,292 Kimi K2 OP
Training	LoRA, 2 epochs, lr 2×10^{-5}
Batching	grad. acc. 4, max len 512
Baseline / best rank	r=32 / r=64
Pair eval.	1,200 held-out pairs
Generation eval.	80 MT-Bench × 2 orderings

Table 1: Experimental constants. The main ablation changes only β and OP inclusion unless otherwise stated.

β	No OP	+OP	Δ
0.02	.5767 ± .0026	.5690 ± .0023	**−0.8**
0.005	.5948 ± .0020	.5988 ± .0079	+0.4
0.002	**.6117 ± .0064**	**.6294 ± .0038**	**+1.8**

Table 2: Primary Qwen3-4B UltraFeedback ablation (4 seeds). β determines whether on-policy data helps or hurts.

β	Seed 42	Multi-seed	Δ
0.02 (no-OP ref)	.5800	.5767 ± .0026	—
0.01	.5842	.5800 ± .0050	+0.4
0.008	.5867	—	+0.7
0.005	.6067	.5988 ± .0079	+2.2
0.003	.6158	—	+3.6
0.002	**.6317**	**.6294 ± .0038**	**+5.2**
0.0015	.6258	—	+4.6
0.001	.6242	—	+4.4
0.0005	.6200	—	+4.0

Table 3: Full-mixture β sweep on Qwen3-4B, r=32 (+OP at each $\beta \leq 0.01$). β=0.02 row is the no-OP reference; Δ in pp relative to that baseline. Multi-seed for the three validated cells.

adding on-policy data reduces accuracy by −0.8 pp. At β=0.005, the effect is near neutral at +0.4 pp. At β=0.002, the same data gives +1.8 pp. The individual seeds match the aggregate pattern: all 4 seeds are negative at β=0.02 and positive at β=0.002 (Appendix A).

A difference-in-differences statistic isolates the moderation:

$$\mathrm{DiD} = (\bar{a}^{+\mathrm{OP}}_{0.002} - \bar{a}^{\mathrm{no}}_{0.002}) - (\bar{a}^{+\mathrm{OP}}_{0.02} - \bar{a}^{\mathrm{no}}_{0.02}). \quad (7)$$

The estimate is +0.025 with 95% CI $[0.012, 0.034]$ (paired bootstrap over 4 seeds, 10,000 resamples) and Cohen's d=3.36. Size-matched, same-judge, and equal-volume off-policy controls (Table 4; full results in Appendix B) preserve the pattern; the equal-volume control further indicates that at high β extra data of either source slightly hurts, while at low β both extra-volume conditions help. The source-specific OP contribution is therefore most clearly attributable at low β, while the β moderation itself does not require an on-policy source.

The broader full-mixture sweep explains why β=0.002 becomes the low-β operating point. Accuracy improves as β falls from 0.02 to 0.002, then declines below that value (Table 3). This inverted-U profile is consistent with the β prefactor in Eq. 2: lowering β shrinks the magnitude of every per-example gradient (including the dilution-prone on-policy ones), but excessively small β also weakens the off-policy contribution and loosens the reference anchor.

4.3 Gradient Dilution Mechanism

We use "gradient dilution" descriptively for the observed margin/gradient pattern (not as a fully causal proof across architectures). We measure training-time gradient diagnostics by logging $\beta\Delta_i$

Control	β	+OP	No OP
Size-matched OP	0.002	$.639 \pm .006$	$.612 \pm .006$
Size-matched OP	0.02	$.572 \pm .005$	$.577 \pm .003$
Same-judge OP	0.002	$.621 \pm .008$	$.612 \pm .006$
Same-judge OP	0.02	$.564 \pm .004$	$.577 \pm .003$

Table 4: Confound controls (2 seeds). The interaction persists when OP size is matched to AI-judge data and when OP pairs are labeled by the same Qwen3-8B judge.

Phase	OP $\beta\Delta$	OP σ	offp $\beta\Delta$	offp σ
Early ($\leq$100 steps)	-0.02	0.505	0.16	0.461
Mid (100–500)	0.11	0.473	0.90	0.346
Late ($>$500)	**0.05**	**0.488**	**0.81**	**0.388**

Table 5: Training-time gradient diagnostics at β=0.02, standard DPO with +OP. Off-policy examples develop large $\beta\Delta$ (the model learns them); on-policy examples stay near $\beta\Delta\approx 0$ with maximal sigmoid weight $\sigma\approx 0.5$.

and $\sigma(-\beta\Delta_i)$ separately for on-policy and off-policy examples every 10 steps during a standard DPO run at β=0.02 with +OP data (Table 5).

The asymmetry is clear: off-policy examples are *learned* during training ($\beta\Delta$ rises to 0.9, σ drops to 0.35), while on-policy examples remain *unlearned* ($\beta\Delta \approx 0.05$, $\sigma \approx 0.49$). On-policy examples are not gradient-suppressed—their sigmoid weight stays near the maximum—but their gradients empirically fail to produce margin growth because the model cannot build a clear preference between two responses it generated itself.

Gradient norms confirm OP gradients are not magnitude-suppressed ($4-8$ vs. $6-17$ for off-policy; Appendix B). OP-offp gradient cosine is near zero, but so are offp-offp and OP-OP cosines under null baselines, so orthogonality is a high-dimensional default. The dilution evidence rests on the *margin asymmetry*: off-policy gradients cumulatively build $\beta\Delta$; OP gradients do not.

Comparing off-policy $\beta\Delta$ trajectories under no-OP vs. +OP at β=0.02 (Appendix B), adding OP does not slow off-policy margin growth (late-training $\beta\Delta$: 0.81 with OP vs. 0.67 without); the harm is therefore consistent with optimization-trajectory noise rather than direct opposition to off-policy learning.

Why low β helps. Low β has two effects. First, it changes the DPO margin geometry for all sources: even without OP, β=0.002 improves over β=0.02 by $+3.5$ pp (Table 3), keeping more examples in the sigmoid's sensitive region longer. Second, it reduces absolute OP gradient magnitude: magnitudes shrink $10\times$ ($\beta \cdot \sigma \approx 0.001$), making low-effective-signal OP gradients proportionally less disruptive. The on-policy gain at low β ($+1.8$ pp) reflects both effects.

Condition	Accuracy	Δ vs. no OP
No OP	$.577 \pm .003$	—
All OP (34.8% flagged)	$.569 \pm .002$	-0.8
Flagged-only OP	$\mathbf{.566 \pm .001}$	$\mathbf{-1.1}$
Non-flagged OP	$\mathbf{.576 \pm .001}$	$\mathbf{-0.1}$

Table 6: Subset intervention at β=0.02. The proxy-flagged OP subset carries most of the harm. Threshold sensitivity in Appendix B shows the gap grows monotonically from τ=0.2 to 0.5.

Pre-training proxy and subset intervention. The reference preference gap δ_i^{ref} (Eq. 4) provides a pre-training proxy for identifying on-policy examples that contribute most to dilution. Per-source mean absolute gaps: human $|\delta^{\text{ref}}|$=67.1, AI 66.0, OP 35.7 (signed means in Appendix: 39.6, 21.7, 14.0). We partition on-policy examples by $\sigma(-\beta\delta_i^{\text{ref}}) < 0.4$ and retrain at β=0.02. Training only on proxy-flagged OP gives the worst result; non-flagged OP nearly recovers the no-OP baseline.

Two proxy checks rule out simpler interpretations. Flagged and non-flagged on-policy examples differ only modestly in response length (412 vs. 447 tokens), too small to explain the accuracy gap. Applying the same $\sigma < 0.4$ rule to off-policy pairs at β=0.02 flags only 3.2% of them, and removing those pairs changes accuracy by ± 0.1 pp. The dilution effect is therefore concentrated in the on-policy pool under standard β.

Linking the proxy to training-time behavior. Although OP has *lower* mean $|\delta^{\text{ref}}|$ than off-policy overall (35.7 vs. 67), the proxy flags a specific subset: OP examples where δ^{ref} is large and positive (the reference already strongly prefers the chosen response). These "already-known" pairs contribute gradient steps that reinforce an established preference without adding new signal—consistent with the margin-stagnation profile in Table 5. Non-flagged OP (small or negative δ^{ref}) nearly recovers the no-OP baseline (Table 6). The harm is subset-driven, not uniform across the OP source.

A threshold sensitivity analysis over $\tau \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ (Table 7) shows the proxy is not threshold-dependent: the accuracy gap grows

τ	Flagged-only	Non-flagged	Δ (pp)
0.2	.5692	.5708	+0.17
0.3	.5650	.5708	+0.58
0.4	.5667	.5767	+1.00
0.5	**.5583**	**.5775**	**+1.92**
0.6	.5642	.5767	+1.25

Table 7: Threshold sensitivity at β=0.02 (seed 42). The gap grows monotonically to τ=0.5, confirming τ=0.4 is not cherry-picked.

Config	Accuracy	Δ vs. no OP
DPO no-OP	.5767 ± .0026	—
DPO +OP	.5690 ± .0023	**−0.77** pp
β-split +OP	**.5798 ± .0032**	**+0.31** pp

Table 8: β-split reverses high-β on-policy harm at β_h=0.02, β_{op}=0.002, r=32 (4 seeds).

monotonically from +0.17 pp (τ=0.2) to +1.92 pp (τ=0.5); at τ=0.6 it decreases as the non-flagged subset becomes small.

4.4 β-Split as a Mechanism Probe

We use β-split DPO as a mechanism probe: it tests whether the high-β failure is source-conditioned or a global regularization need. With β_h=0.02 for off-policy and β_{op}=0.002 for OP, the same +OP data that hurts standard DPO becomes mildly beneficial (Table 8): −0.77 pp → +0.31 pp over no-OP, +1.08 pp over standard +OP (4 seeds; per-seed in Appendix B). The no-OP gain (+0.31 pp) is modest and should be read as supporting source-specificity, not as a strong method claim.

Two controls test alternative explanations, now upgraded to 4 seeds each (Table 9). Matched-weight DPO sets an on-policy loss multiplier α=0.113 so that the average OP gradient magnitude matches β-split; it reaches .5762 ± .0022, while β-split reaches .5798 ± .0032, a +0.36 pp gap (4 seeds each). β-split therefore primarily reduces OP gradient magnitude via the β prefactor, but the source-conditioned β also changes per-example sigmoid weighting dynamics, which matched-weight scaling does not replicate. Margin-conditioned DPO applies the small β to examples with δ^{ref} in the OP interquartile range and obtains .5737 ± .0032 vs. .5798 ± .0032 for source-conditioned β-split (4 seeds), a +0.60 pp gap. Under β-split, OP gradient magnitude drops to $\beta_{\text{op}} \cdot \sigma \approx 0.001$ while off-policy remains at $\beta_h \cdot \sigma \approx 0.005$, a 5× ratio that suppresses dilution while keeping off-policy pairs near the standard β_h regime (off-policy training-time $\beta\Delta$ comparable to no-OP; Appendix B).

Control / diagnostic	Result
Matched-weight DPO, α=0.113 (4s)	.5762 ± .0022
β-split (4 seeds)	**.5798 ± .0032**
Margin-conditioned, $\delta^{\text{ref}} \in$ OP IQR (4s)	.5737 ± .0032
Human mean $\lvert\delta^{\text{ref}}\rvert$	67.1
AI mean $\lvert\delta^{\text{ref}}\rvert$	66.0
OP mean $\lvert\delta^{\text{ref}}\rvert$	35.7
OP mean $\beta\Delta$ (training-time, β-split)	0.023
OP $\sigma(-\beta\Delta)$ (training-time, β-split)	0.494

Table 9: β-split controls and diagnostics (all 3 conditions at 4 seeds; Appendix B gives per-seed values). Matched loss magnitude alone is insufficient; source identity proxies the dilution-prone regime. Training-time $\beta\Delta$ and σ are consistent with the gradient asymmetry under β-split.

β_{op}	Ratio to β_h	Accuracy
0.002	0.1×	.5767
0.005	0.25×	**.5792**
0.01	0.5×	.5742
0.02	1.0×	≈.5690

Table 10: β_{op} sensitivity at seed 42 with β_h=0.02. Single-seed sweep; the 0.5767 value at β_{op}=0.002 here is seed-42-only and should not be compared directly to the 4-seed β-split mean .5798±.0032 in Table 8. The useful range is a plateau rather than a single sharp value.

Sensitivity to β_{op} is forgiving around the target scale. With seed 42 and β_h=0.02, $\beta_{\text{op}} \in [0.002, 0.005]$ forms a plateau, while performance degrades as β_{op} approaches β_h. The operating rule is therefore simple: lower the on-policy coefficient enough to keep OP examples near the sigmoid's sensitive region, but do not treat the exact value as a brittle constant.

4.5 Cross-Validation

The interaction direction repeats beyond Qwen3-4B+UltraFeedback, with a second clear population-level sign flip on Llama-3.1-8B at β=0.05 (Table 11). Llama-3.1-8B (r=64, 4 seeds): +OP gives **−0.46** pp at β=0.05, −0.06 pp at β=0.02, +0.40 pp at β=0.002; 3 of 4 seeds flip from β=0.05 to β=0.002 (DiD +0.0085). β-split at β_h=0.02 (4s): .5960±.0019, +0.23/+0.29 pp over no-OP/+OP. On Qwen3-0.6B, DPO-Mix-7k (2s each), β moderation is weaker without flip. Llama-3.2-1B (4s) has +OP mildly harmful at both β; a Qwen3-4B r=8 ablation (4s) preserves moderation (DiD +0.010) but with +OP helpful at both β—sign flip requires sufficient β *and* capacity. SimPO (3s) shows a 4×

Setting	β	No OP	+OP
Qwen3-0.6B (2s)	0.02	$.526 \pm .002$	.529
Qwen3-0.6B (2s)	0.002	.534	$\mathbf{.548 \pm .002}$
Qwen3-4B r=8 (4s)	0.02	$.5560 \pm .0017$	$\mathbf{.5635 \pm .0008}$
Qwen3-4B r=8 (4s)	0.002	$.5779 \pm .0030$	$\mathbf{.5950 \pm .0033}$
Llama-3.1-8B (r=64, 4s)	0.05	$\mathbf{.5900 \pm .0006}$	$\mathbf{.5854 \pm .0014}$
Llama-3.1-8B (r=64, 4s)	0.02	$.5938 \pm .0017$	$.5931 \pm .0034$
Llama-3.1-8B (r=64, 4s)	0.002	$.5860 \pm .0028$	$\mathbf{.5900 \pm .0006}$
Llama-3.2-1B (4s)	0.02	$\mathbf{.545 \pm .001}$	$.542 \pm .002$
Llama-3.2-1B (4s)	0.002	$.539 \pm .002$	$.534 \pm .002$
DPO-Mix-7k (2s)	0.02	$.630 \pm .003$	$.632 \pm .001$
DPO-Mix-7k (2s)	0.002	$.673 \pm .002$	$\mathbf{.681 \pm .001}$
SimPO (3s)	2.0	$.5658 \pm .0018$	$.5689 \pm .0004$
SimPO (3s)	0.2	$.5533 \pm .0018$	$\mathbf{.5631 \pm .0010}$

Table 11: Cross-checks. Low-β OP advantage is larger for DPO across model scale and dataset; SimPO shows a smaller, directionally similar interaction (DiD $+0.007$ vs. DPO $+0.025$). Seed counts in parentheses (1s/2s/3s/4s).

Capacity / validation	Accuracy	Δ
r=16, β=0.002, +OP	.6175	-1.4 vs. r=32
r=32, β=0.002, +OP	.6317	—
r=64, β=0.002, +OP	**.6458**	$\mathbf{+1.4}$
r=128, β=0.002, +OP	.6425	$+1.1$
Best recipe, 4 seeds	$\mathbf{.6419 \pm .0058}$	$\mathbf{+6.5}$ pp

Table 12: Low-β OP training benefits from added LoRA capacity. The final best recipe uses r=64.

Pareto recipe	pair acc	Δ_{MT}
Baseline (β=.02, no OP, r=32, 4s)	$.577 \pm .003$	0
β-split β_h=.02, r=64 [OP] (4s)	$.586 \pm .006$	-30.0
No OP, β=.005, r=64 (4s)	$.601 \pm .003$	-32.5
No OP, β=.002, r=64 (4s)	$.630 \pm .006$	-51.2
Best (β=.002, +OP, r=64, 4s)	$\mathbf{.642 \pm .006}$	-58.8
β-split β_h=.002, r=64 [OP] (4s)	$\mathbf{.655 \pm .007}$	-63.8

Table 13: Pareto-optimal recipes (6 of 15 tested; all now 4-seed validated, with Δ_{MT} from the original single-seed MT-Bench evaluation). Dominated recipes in Appendix C.

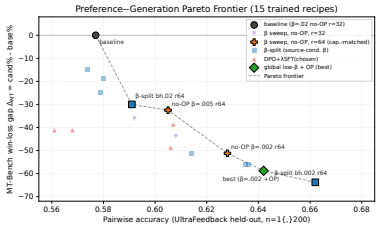


Figure 3: Preference–generation Pareto map. Higher pair accuracy comes with lower MT-Bench win-loss gap against the baseline.

smaller directional interaction (DiD $+0.007$), consistent with length-normalized rewards implicitly source-conditioning per-token coefficients.

Population-level sign flip is established in two settings: Qwen3-4B+UltraFeedback ($\beta \in \{0.02, 0.002\}$, 4/4 seeds) and Llama-3.1-8B ($\beta \in \{0.05, 0.002\}$, 3/4 seeds; the narrower $\beta \in \{0.02, 0.002\}$ window yields weaker moderation). Qwen3-0.6B and DPO-Mix-7k show moderation without a high-β negative cell. The Llama-3.2-1B and Qwen3-4B r=8 settings jointly indicate capacity also gates whether OP harms or helps. SimPO shows a much smaller, directionally similar interaction, consistent with length-normalized rewards implicitly source-conditioning per-token coefficients.

4.6 Pareto Trade-off

The best multi-seed standard-DPO recipe uses global β=0.002, +OP, r=64, reaching 0.6419 ± 0.0058 over 4 seeds; an even higher 4-seed-validated β-split recipe with β_h=0.002, β_{op}=0.0002, r=64 reaches 0.655 ± 0.007 at the cost of further generation degradation (Table 13). The standard-DPO recipe's gain over the β=0.02 no-OP r=32 baseline decomposes into low β ($+3.5$ pp), OP ($+1.8$ pp), and r=64 ($+1.3$ pp; Table 12). The same recipe loses heavily on MT-Bench: best wins 6 questions (7.5%), baseline wins 53 (66.2%), ties 21 (26.2%), and the best recipe is shorter on average (1,706 vs. 2,936 characters). The trade-off is systematic rather than a single failure case.

Table 13 and Figure 3 show the 15-recipe Pareto sweep (6 Pareto-optimal, all 4-seed validated). OP variants occupy both extremes; the middle is Pareto-optimal without OP. DPO+λ NLL does not bridge the trade-off (Appendix D); practitioner regimes in Appendix E.

5 Conclusion

The effect of on-policy data in DPO is conditional on β: the same data pool hurts at high β and helps at low β, with sign flips on Qwen3-4B (DiD $+0.025$, 4/4 seeds) and Llama-3.1-8B (DiD $+0.009$, 3/4 seeds). The mechanism is consistent with gradient dilution: on-policy examples retain near-maximal gradient weight but fail to develop preference margins. β-split DPO reverses the high-β harm by source-conditioning the coefficient. However, pairwise gains trade off against generation quality, so the contribution is a diagnosis of preference-discrimination dynamics rather than a general alignment improvement.

Limitations

The full population-level sign flip is established on Qwen3-4B+UltraFeedback (4/4 seeds, DiD $+0.025$) and Llama-3.1-8B (3/4 seeds, DiD $+0.0085$ at $\beta \in \{0.05, 0.002\}$). Other settings show directional β moderation without a high-β negative cell; a Qwen3-4B $r{=}8$ ablation shows β moderation persists at low capacity but without sign flip. The on-policy condition changes both data source and volume: an equal-volume off-policy control reveals partial volume coupling, so the strongest supported claim is β moderation of additional preference data, with a modest OP-source-specific component.

The "gradient dilution" label is descriptive: it is supported by training-time margin stagnation, gradient norms, a held-out alignment measurement, subset interventions, and threshold sensitivity, but the most direct diagnostics (SNR, alignment) are single-seed, short-run, and use small cosine magnitudes. The β sweep covers $[0.0005, 0.05]$ in the LoRA-DPO regime; full fine-tuning settings may differ. β-split is evaluated as a mechanism probe, not a fully optimized algorithm.

The generation-quality analysis uses single-seed MT-Bench with a Qwen3-14B judge; complementary evaluations such as length-controlled AlpacaEval (Dubois et al., 2024) or RewardBench (Lambert et al., 2024) could strengthen the trade-off characterization.

References

Argilla. 2024. DPO-Mix-7k: A mixed-source preference dataset. https://huggingface.co/datasets/argilla/dpo-mix-7k.

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

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

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. 2024. UltraFeedback: Boosting language models with scaled AI feedback. *International Conference on Machine Learning*.

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

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

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 2024. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. 2025. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

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

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

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

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

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *International Conference on Machine Learning*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*.

Yunan Wang, Jijie Li, Bo-Wen Zhang, Liangdong Wang, and Guang Liu. 2025. InCo-DPO: Balancing distribution shift and data quality for enhanced preference optimization. *arXiv preprint arXiv:2503.15880*.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024a. β-DPO: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024b. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. *International Conference on Machine Learning*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? a comprehensive study. *International Conference on Machine Learning*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024. WPO: Enhancing RLHF with weighted preference optimization. *arXiv preprint arXiv:2406.11827*.

A Experimental Details

On-policy data construction. On-policy responses are generated from the pre-DPO Qwen3-4B checkpoint using vLLM with temperature 0.7, top-p 0.9, and `max_tokens=512`. For each UltraFeedback prompt, 2 candidate responses are sampled. Training prompts and evaluation prompts are strictly disjoint: the 1,200 held-out evaluation pairs use a separate prompt set that is never seen during training or on-policy generation.

Kimi K2 labels each pair with position-debiased pairwise judging (both orderings; majority vote). The judge prompt is:

```
Please act as an impartial judge and evaluate
the quality of the responses provided by two
AI assistants to the user question below.
Choose the assistant that better follows the
user's instructions and answers the question.
Output your verdict strictly as: "[[A]]" if
assistant A is better, "[[B]]" if B is better,
"[[C]]" for a tie.

[User Question]  {question}
[Assistant A's Answer]  {answer_a}
[Assistant B's Answer]  {answer_b}
```

Pairs where both orderings agree are assigned that label; one-win-one-tie pairs take the winner; disagreements are excluded. Of the original candidate pool, approximately 78% had consistent labels across both orderings, 14% had one-win-one-tie outcomes, and 8% were discarded as disagreements. Pairs with Kimi confidence ≥ 0.5 are retained, yielding 2,292 on-policy preference pairs. The judge prompt and filtering procedure are identical for the Qwen3-8B same-judge control.

Per-seed reproducibility. A key concern is whether the β–data sign flip is robust or driven by a single lucky seed. Table 14 reports the full per-seed breakdown for all 24 runs in the 3×2 ablation. At β=0.02, every seed shows a negative on-policy effect; at β=0.002, every seed shows a positive effect. Table 15 shows that the best recipe (β=0.002, r=64) is stable across seeds with low variance ($\pm$0.0058).

B Supplementary Diagnostics and Validation

Gradient norms and cosine similarity. The main text claims that on-policy gradients are not magnitude-suppressed. Table 17 is consistent with this: OP gradient norms (4–8) are comparable to off-policy (6–17) throughout training at β=0.02. We also report cosine similarity between OP and

β	Data	s42	s123	s456	s789
0.02	no OP	**.5800**	**.5742**	**.5742**	**.5783**
0.02	+OP	**.5700**	**.5708**	**.5700**	**.5650**
0.005	no OP	.5975	.5925	.5933	.5958
0.005	+OP	.6067	.5992	.5858	.6033
0.002	no OP	**.6100**	**.6075**	**.6225**	**.6067**
0.002	+OP	**.6317**	**.6333**	**.6233**	**.6292**

Table 14: Per-seed pairwise accuracy for the 3×2 ablation. The sign flip holds for every seed at the β=0.02 and β=0.002 endpoints.

Seed	Accuracy	Δ vs. Baseline
42	0.6458	+0.0658
123	0.6450	+0.0650
456	0.6333	+0.0533
789	0.6433	+0.0633
Mean	$\mathbf{0.6419 \pm 0.0058}$	$\mathbf{+0.0652}$

Table 15: Multi-seed validation of the best configuration (β=0.002, r=64, full data).

off-policy gradients, alongside null baselines (offp-offp and OP-OP half-batch cosines). All three cosine columns are near zero, indicating that orthogonality is a property of the high-dimensional LoRA parameter space rather than an OP-specific signal. The dilution evidence therefore rests on the margin asymmetry reported in Table 5 of the main text.

β-split per-seed accuracy. Table 16 reports per-seed pairwise accuracy for the β-split intervention (β_h=0.02, β_{op}=0.002, r=32). All four seeds reverse the harm relative to standard DPO +OP (.5690±.0023); the improvement relative to the no-OP baseline (.5767±.0026) is positive on 3 of 4 seeds, with the fourth (s42) within the no-OP standard deviation of the baseline.

Seed	β-split **+OP**	Δ vs. standard DPO +OP
42	.5767	+0.77 pp
43	.5775	+0.85 pp
44	.5850	+1.60 pp
45	.5800	+1.10 pp
Mean	$\mathbf{.5798 \pm .0032}$	$\mathbf{+1.08}$ pp

Table 16: Per-seed β-split results (β_h=0.02, β_{op}=0.002, r=32, 4 seeds).

Direct per-source gradient SNR. We additionally measure per-source gradient direction stability by re-running training for 100 steps at β=0.02 and β=0.002 (Qwen3-4B, LoRA r=32, seed 42, batch-per-source 2), computing on each step the

Phase	OP $\|\nabla\|$	offp $\|\nabla\|$	OP-offp cos.	offp-offp cos.	OP-OP cos.
Early	4.3	6.1	+0.035	+0.040	−0.029
Mid	8.3	9.1	−0.021	−0.021	+0.007
Late	6.9	17.0	−0.022	+0.006	+0.035

Table 17: Gradient norms and cosine similarity at β=0.02 with +OP data, with null baselines (offp-offp and OP-OP half-batch cosines).

cosine of the current per-source gradient with its previous-step counterpart (Table 18, Figure 4). On-policy gradients show lower cross-step cosine consistency than off-policy ($\cos(g_{\text{op}}^{(t)}, g_{\text{op}}^{(t-1)})$=0.007 vs. $\cos(g_{\text{offp}}^{(t)}, g_{\text{offp}}^{(t-1)})$=0.011 at β=0.02, a 30–44% ratio gap), and a correspondingly lower SNR proxy $|\mu|/\sigma$ (0.143 for OP vs. 0.179 for offp at β=0.02, about 20–26% lower). The pattern is unchanged at β=0.002 (OP 0.132 vs. offp 0.178); gradient norms scale roughly 10× smaller (consistent with the β prefactor), but the per-source SNR ratio is preserved. This indicates that β moderates the *magnitude* of dilution-prone OP gradients via the prefactor rather than altering their per-source directional consistency.

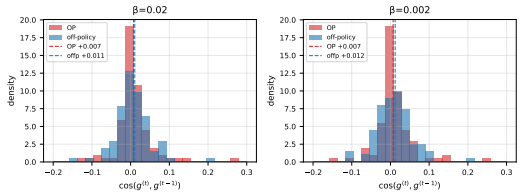


Figure 4: Distribution of per-step cosine consistency $\cos(g^{(t)}, g^{(t-1)})$ for OP vs. off-policy gradients. OP gradients have lower cross-step directional consistency at both β values.

Metric	β=0.02	β=0.002
mean $\|g_{\text{op}}\|$	4.30	0.48
mean $\|g_{\text{offp}}\|$	5.28	0.67
$\cos(g_{\text{op}}^{(t)}, g_{\text{op}}^{(t-1)})$	+0.0073	+0.0068
$\cos(g_{\text{offp}}^{(t)}, g_{\text{offp}}^{(t-1)})$	+0.0107	+0.0122
$\cos(g_{\text{op}}, g_{\text{offp}})$	−0.0007	+0.0011
SNR proxy OP ($\|\mu\|/\sigma$)	**0.143**	**0.132**
SNR proxy offp ($\|\mu\|/\sigma$)	**0.179**	**0.178**
OP/offp cross-step cosine ratio	0.688	0.561
OP/offp SNR proxy ratio	**0.798**	**0.742**

Table 18: Direct per-source gradient SNR over 100 training steps (Qwen3-4B, r=32, seed 42). OP gradients have 20–26% lower SNR proxy ($|\mu|/\sigma$) and 30–44% lower cross-step cosine consistency than off-policy at both β values; gradient magnitudes scale ∼ 10× with β.

Alignment with held-out preference direction. The cross-step SNR above measures internal consistency but not whether OP gradients point toward a useful direction. We additionally compute, on each step, the cosine of g_{op} and g_{offp} with a per-step "ground-truth" direction g_{ref}, defined as the DPO-loss gradient on a held-out batch of 16 high-confidence preference pairs (drawn from the test set, disjoint from training; 80 steps, seed 42, Qwen3-4B, r=32). Off-policy gradients are weakly but consistently aligned with the held-out direction (mean $\cos(g_{\text{offp}}, g_{\text{ref}}) = +0.005$ at β=0.02, +0.020 at β=0.002; 24% and 28% of steps with $\cos > 0.05$). On-policy gradients show *no* positive alignment at either β (mean $\cos(g_{\text{op}}, g_{\text{ref}}) = -0.013$ at β=0.02, -0.005 at β=0.002; only 10% and 7% of steps with $\cos > 0.05$). The offp-minus-OP alignment gap is +0.018 at β=0.02 and +0.025 at β=0.002. Two implications: (i) OP gradients carry essentially no useful directional preference signal regardless of β, supporting the dilution interpretation; (ii) the gap between offp and OP useful signal is larger at lower β, consistent with offp gradients being able to "win" the optimization step when the β prefactor de-emphasizes OP magnitude.

Off-policy margin trajectory. A natural question is whether adding on-policy data slows off-policy margin growth. Table 19 compares off-policy $\beta\Delta$ under the no-OP condition (all examples are off-policy) and the +OP condition (off-policy subset only). Late-training off-policy $\beta\Delta$ is 0.81 with OP vs. 0.67 without, indicating that OP does not slow off-policy margin development. The harm from OP therefore acts through optimization-trajectory noise rather than by opposing off-policy learning.

Phase	no-OP $\beta\Delta$	+OP offp $\beta\Delta$
Early	−0.006	0.164
Mid	0.091	0.899
Late	**0.669**	**0.811**

Table 19: Off-policy margin trajectory: adding OP does not slow off-policy $\beta\Delta$ growth.

Threshold sensitivity. See Table 7 in the main text. Figure 5 provides a visual summary.

Confound controls. The main text reports that the interaction persists under size-matching and same-judge relabeling. Table 20 provides the full

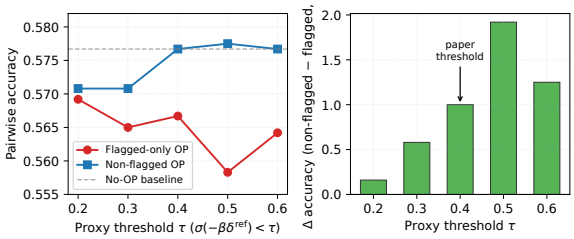


Figure 5: Flagged-only vs. non-flagged OP training across proxy thresholds.

2-seed results. Subsampling on-policy data to 764 pairs (matching the AI-judge pool) preserves the interaction at both β levels. Re-labeling on-policy pairs with the same Qwen3-8B judge (929 consistent pairs after position debiasing) also preserves the pattern, ruling out judge quality as the primary driver.

We additionally run an equal-volume off-policy control (4 seeds): resampling 2,292 additional off-policy pairs in place of on-policy, so total training volume matches the +OP condition. Results: at β=0.02, .5683 ± .0010 (vs. +OP .5690, baseline no-OP .5767); at β=0.002, .6394±.0067 (vs. +OP .6294, baseline no-OP .6117). At high β, extra volume of either source slightly underperforms the no-OP baseline, suggesting a volume component to the high-β harm. At low β, both extra-volume conditions improve over no-OP, with extra off-policy modestly above +OP. We therefore interpret the high-β harm as partially explained by volume and partially by source, and the low-β improvement as primarily driven by the lower β rather than by on-policy source specifically.

Control	β	+OP	No OP
Size-matched (764 OP)	0.002	**.639 ± .006**	.612 ± .006
Size-matched	0.02	.572 ± .005	.577 ± .003
Same-judge (8B)	0.002	**.621 ± .008**	.612 ± .006
Same-judge	0.02	.564 ± .004	.577 ± .003

Table 20: Confound controls (2 seeds each). The β–data interaction persists under size-matching and same-judge relabeling.

Cross-setting DiD and SimPO. Table 22 summarizes the difference-in-differences statistic across all settings. DiD is positive in four of five DPO settings, indicating directionally consistent β moderation. The Llama-3.2-1B DiD is mildly negative, reflecting on-policy data being harmful at both β values rather than reversing the moderation direction. SimPO (γ=1.0, 3 seeds at SimPO

$\beta \in \{2.0, 0.2\}$) shows a weaker but directionally similar interaction (high-β $\Delta = +0.31$ pp, low-β $\Delta = +0.97$ pp, DiD $+0.007$), about $4\times$ smaller than the DPO Qwen3-4B effect.

Table 21 reports the per-seed accuracy for the Llama-3.1-8B (r=64) replication at $\beta \in \{0.05, 0.02, 0.002\}$. At β=0.05, the +OP cell is negative on all 4 seeds (mean -0.46 pp); at β=0.02, it is negative on 2 of 4 (mean -0.06 pp); at β=0.002, it is non-negative on 4 of 4 (mean $+0.40$ pp). The β=0.05 $\rightarrow$ 0.002 comparison gives DiD $+0.0085$ with 3 of 4 seeds showing per-seed sign flip, establishing a second population-level sign flip beyond the Qwen3-4B+UltraFeedback setting.

β	Data	s42	s123	s456	s789
0.05	no OP	.5908	.5900	.5900	.5892
0.05	+OP	**.5842**	**.5875**	**.5858**	**.5842**
0.02	no OP	.5917	.5958	.5950	.5925
0.02	+OP	**.5875**	**.5933**	.5958	.5958
0.002	no OP	.5850	.5850	.5833	.5908
0.002	+OP	**.5900**	**.5892**	.5900	.5908

Table 21: Per-seed pairwise accuracy for Llama-3.1-8B with LoRA r=64 at $\beta \in \{0.05, 0.02, 0.002\}$ (4 seeds each). At β=0.05 +OP is negative on 4/4 seeds; at β=0.002 +OP is non-negative on 4/4 seeds; 3/4 seeds flip from β=0.05 to β=0.002.

Setting	DiD	95% CI	Seeds
4B (UltraFeedback)	$\mathbf{+0.025}$	$[\mathbf{0.012}, \mathbf{0.034}]$	4
Llama-3.1-8B ($\beta \in \{0.05, 0.002\}$)	$\mathbf{+0.0085}$	—	4
4B-UF r=8 ($\beta \in \{0.02, 0.002\}$)	$+0.010$	—	4
0.6B (UltraFeedback)	$+0.009$	—	2
SimPO ($\beta \in \{2.0, 0.2\}$)	$+0.007$	—	3
4B (DPO-Mix-7k)	$+0.006$	—	2
Llama-3.1-8B ($\beta \in \{0.02, 0.002\}$)	$+0.005$	$[-0.001, +0.008]$	4
Llama-3.2-1B	-0.002	—	4

Table 22: DiD across all settings. Two settings show full population-level sign flip (Qwen3-4B+UltraFeedback at r=32 and Llama-3.1-8B at r=64 with $\beta \in \{0.05, 0.002\}$); the rest show β moderation without sign flip. Llama-3.2-1B's negative DiD reflects mildly harmful on-policy data at both β values, indicating capacity is also a gate.

Pre-training δ^{ref} by source. The pre-training reference preference gap δ^{ref} differs systematically across data sources (500-pair sample per source). Mean absolute $|\delta^{\text{ref}}|$: human 67.1, AI 66.0, OP 35.7. Signed mean (σ_{std}, IQR): human 39.6 (106.7, $[-7.6, 73.0]$), AI 21.7 (95.3, $[-21.3, 66.1]$), OP 14.0 (52.8, $[-9.0, 36.1]$). On-policy gaps are smaller in absolute magnitude and tighter than off-policy, consistent with the observation that both OP candidates are drawn from the same model distribution.

C MT-Bench and Pareto Frontier

Table 23 breaks down the position-debiased MT-Bench evaluation by category. The best pairwise recipe wins only 7.5% of the 80 comparisons, losing most heavily in stem (1/10), roleplay (0/10), and humanities (0/10). Mean response length for the best recipe is 1,706 characters vs. 2,936 for the baseline, consistent with reduced KL regularization allowing drift toward shorter responses.

Category	Best	Base	Ties	Disagree
coding	0	8	2	2
extraction	1	4	5	1
humanities	0	8	2	2
math	0	4	6	0
reasoning	2	4	4	1
roleplay	0	9	1	1
stem	1	9	0	0
writing	2	7	1	1
Total	**6**	**53**	**21**	**8**

Table 23: Position-debiased MT-Bench: 80 questions $\times$ 2 orderings collapsed to 80 verdicts via per-question majority over the two orderings (Qwen3-14B judge).

Table 24 lists the full 15-recipe Pareto sweep. Six recipes are Pareto-optimal: the standard baseline, two capacity-matched no-OP points, two source-conditioned β variants, and the best recipe. On-policy data defines the extremes of the frontier (lowest and highest pair accuracy among Pareto points), while capacity-matched no-OP β tuning is Pareto-optimal in the intermediate band. DPO+λSFT contributes no Pareto-optimal point. SFT$\rightarrow$DPO (r=64, β=0.002, 4 seeds) achieves mean 0.6354 ± 0.0158 vs. pure DPO 0.6419 ± 0.0058—lower mean and $3\times$ higher variance.

Recipe (r=64 unless noted)	pair acc	cand%	base%	tie%	Δ_{MT}	Pareto
Baseline (β=0.02, no-OP, r=32, 4s)	.577±.003	0	0	100	0	✓
DPO no-OP β=0.005 (4s)	.601±.003	11.2	43.8	45.0	−32.5	✓
DPO no-OP β=0.002 (4s)	.630±.006	6.2	57.5	36.2	−51.2	✓
β-split β_h=0.02, β_{op}=0.002 (4s)	.586±.006	8.8	38.8	52.5	−30.0	✓
β-split β_h=0.002, β_{op}=0.0002 (4s)	**.655±.007**	3.8	67.5	28.7	−63.8	✓
Best (β=0.002, +OP, 4s)	**.642±.006**	7.5	66.2	26.2	−58.8	✓
Selected dominated recipes						
β-split β_h=0.02, r=32 (4s)	.580	12.8	31.6	55.6	−18.8	dom.
DPO+SFT λ=5×10^{-4}	.607	10.0	48.8	41.2	−38.7	dom.
DPO no-OP β=0.005, r=32	.592	6.2	42.5	51.2	−36.2	dom.

Table 24: Pareto frontier across 15 trained recipes. All 6 Pareto-optimal recipes are now 4-seed validated (mean $\pm$ std); MT-Bench values are from the original single-seed evaluation per recipe. Dominated recipes remain single-seed exploratory.

D DPO+SFT does not bridge the trade-off

We tested joint DPO+λ NLL(y_w) at β=0.002 with $\lambda \in \{5\times10^{-4}, 10^{-3}, 2\times10^{-3}, 5\times10^{-3}\}$ as a candidate mechanism to decouple pair-accuracy gains from generation drift. None of these recipes is Pareto-optimal: the best variant (λ=5×10^{-4}, pair acc .607, Δ_{MT}=-38.7) is dominated by the β-split β_h=0.02 point (.586/-30 on the high-MT side) and by no-OP β=0.002 (.628/-51.2, higher pair acc with lower MT cost) on opposite sides; no other variant achieves Pareto-optimality either. The SFT-on-chosen anchor fails because chosen responses in the on-policy mixture inherit the same margin distribution that drives gradient dilution—anchoring on them reinforces rather than counterbalances the drift.

E Practitioner Deployment Regimes

The Pareto frontier maps onto three deployment regimes: (i) **generation-paramount**: use standard β=0.02 no-OP; (ii) **moderate gain**: use β-split at β_h=0.02 (r=64), $+1.4$ pp pair acc at Δ_{MT}=-30, or no-OP β tuning at r=64 if OP is unavailable; (iii) **maximum preference discrimination**: low β with OP and r=64, accepting $\Delta_{\text{MT}} < -50$, suited for reranking and preference-scoring tasks where generation quality is controlled separately.