SEED-LEVEL INSTABILITY IN LLM SAFETY ALIGNMENT: REGIME SHIFTS IN LARGE-MARGIN PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) are deployed more widely, reliable safety alignment becomes increasingly important. However, we identify seed-level safety-regime instability in large-margin, label-based SafeDPO: different training seeds produce sharply different refusal profiles. Across 20 seeds on Qwen2.5-7B-Instruct, OR-Toxic prefix-refusal rates range from 1.83% to 98.47% at a fixed endpoint, while instruction-following accuracy remains comparatively stable. Cross-seed averages obscure these differences in refusal behavior. We introduce Support-Completion SafeDPO (SC-SafeDPO), which restores discarded both-unsafe comparisons using their annotated relative-safety ordering, zero additional margin, and a fixed positive weight. The method preserves the original main-comparison orderings and margins and requires no additional inference-time computation. A local affine analysis relates comparison coverage to seed sensitivity. Budget-matched ablations indicate that both mixture reweighting and restored preference gradients contribute to stability. In the paired 20-seed comparison, SC-SafeDPO reduces cross-seed refusal-rate variance by approximately 77% on XSTest unsafe prompts and 87% on OR-Toxic relative to Label SafeDPO. It also lowers mean HarmBench attack success and benign false-refusal rates and improves instruction-following accuracy. Evaluations across three model backbones examine safety–helpfulness trade-offs and cross-seed variability.

1 INTRODUCTION

Preference optimization aligns large language models with human preferences (Li et al., 2025; Rafailov et al., 2023; Meng et al., 2024). For safety alignment, this means suppressing harmful responses while preserving useful answers to benign requests (Röttger et al., 2024; Cui et al., 2025). SafeDPO builds on Direct Preference Optimization (Rafailov et al., 2023) through safety-aware pair transformations and margins (Kim et al., 2026); related work explores adaptive preference weighting (Zhao et al., 2026). Reliable alignment also requires consistency across training runs (Reimers & Gurevych, 2017; Henderson et al., 2018). Training-seed variation during fine-tuning (Mosbach et al., 2020) and inference-time sampling variation (Larsen, 2025) are separate sources of behavioral instability. We focus on the safety behavior of checkpoints trained with different random seeds under a fixed recipe.

On Qwen2.5-7B-Instruct (Qwen et al., 2025), initial runs of label-based SafeDPO, denoted *Label SafeDPO*, produce sharply different refusal profiles despite similar instruction-following accuracy. A separate study of 20 training seeds confirms this variation. With safety margin $\delta = 10$ and a fixed endpoint, OR-Toxic (Cui et al., 2024) prefix-refusal rates range from 1.83% to 98.47%. The cross-seed mean of 39.63% falls in a gap between observed checkpoint scores, as shown in Figure 3. We term these shifts *seed-level safety-regime instability*.

Label SafeDPO discards both-unsafe comparisons and assigns an additional margin to safe–unsafe comparisons (Kim et al., 2026). The discarded pairs retain relative-safety information that can provide additional supervision. We introduce *Support-Completion SafeDPO* (SC-SafeDPO or SC), which restores these pairs using their dataset-provided relative-safety ordering, zero additional margin, and a positive weight α fixed throughout training. As illustrated in Figure 2B, support comple-

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.

tion preserves the orderings and margins of the original main comparisons. It reduces the normalized weight of margin-bearing comparisons and adds preference gradients from the restored support. The intervention changes training supervision and requires no additional inference-time computation.

In the 20-seed paired comparison, SC-SafeDPO reduces refusal-rate variance to $0.23\times$ that of Label SafeDPO on XSTest unsafe prompts (Röttger et al., 2024) and $0.13\times$ on OR-Toxic, with each method evaluated at a fixed endpoint across seeds. Table 2 also shows improved mean performance on all five reported metrics. A separate ten-seed ablation matches optimizer updates and token budgets. The results support stabilizing contributions from both mixture reweighting and support preference gradients, with the annotated support ordering improving safety outcomes. The diagnostics in Section 4.4 show corresponding changes in gradient allocation and reduced representation drift relative to the starting model.

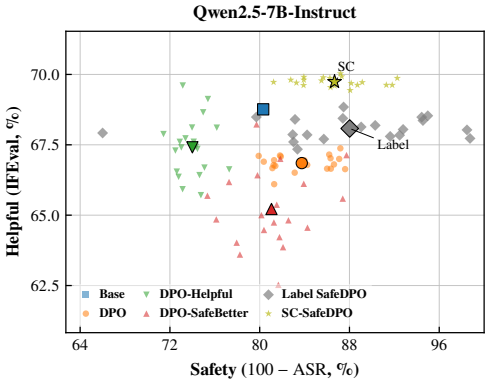


Figure 1: **Helpfulness–safety trade-offs on Qwen2.5-7B.** Higher values on both axes indicate better performance; error bars denote ± 1 standard deviation across training seeds.

Figure 1 summarizes the helpfulness–safety trade-offs on Qwen2.5-7B. Evaluations on Qwen2.5-1.5B (Qwen et al., 2025) and Gemma-3-12B (Team et al., 2025) extend this analysis across model sizes and families, measuring HarmBench (Mazeika et al., 2024) attack success, benign refusal, and instruction following. In a random-sample human audit, two annotators independently assess responses while blinded to method and training seed; a third human adjudicator resolves disagreements. The final human labels show higher safety and helpfulness pass rates for SC-SafeDPO than for SafeDPO in Table 5.

Our contributions are:

- We identify seed-level safety-regime instability in large-margin Label SafeDPO and show across 20 training seeds how cross-seed averages obscure the refusal behavior of individual checkpoints.
- We develop **SC-SafeDPO** to improve cross-seed stability without additional inference-time computation. A local analysis of seed sensitivity and budget-matched ablations characterize the effects of restored comparisons.
- We characterize safety–helpfulness trade-offs and cross-seed variability across three model backbones, and complement benchmark evaluations with a random-sample human audit using two independent blinded annotators and a third human adjudicator.

2 Preliminaries

2.1 Direct Preference Optimization

We consider preference comparisons (x, y^+, y^-), where x is a prompt and y^+ and y^- are the preferred and dispreferred responses under the training ordering. Let $\pi_\theta(y \mid x)$ be the policy with trainable parameters θ, and let $\pi_{\text{ref}}(y \mid x)$ be a fixed reference policy. Both policies are assumed to assign positive probability to the responses in the training data. Following DPO (Rafailov et al., 2023), define the reference-normalized preference log-ratio as

$$r_\theta(x, y^+, y^-) = \log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)}. \tag{1}$$

A positive r_θ means that the probability ratio between y^+ and y^- exceeds the corresponding ratio under the reference policy. It does not require $\pi_\theta(y^+ \mid x) > \pi_\theta(y^- \mid x)$. With preference-scale parameter $\beta > 0$ and logistic sigmoid σ, the per-comparison DPO loss is $\ell_0(\theta) = -\log \sigma(\beta r_\theta)$. Pair arguments are omitted when clear from context.

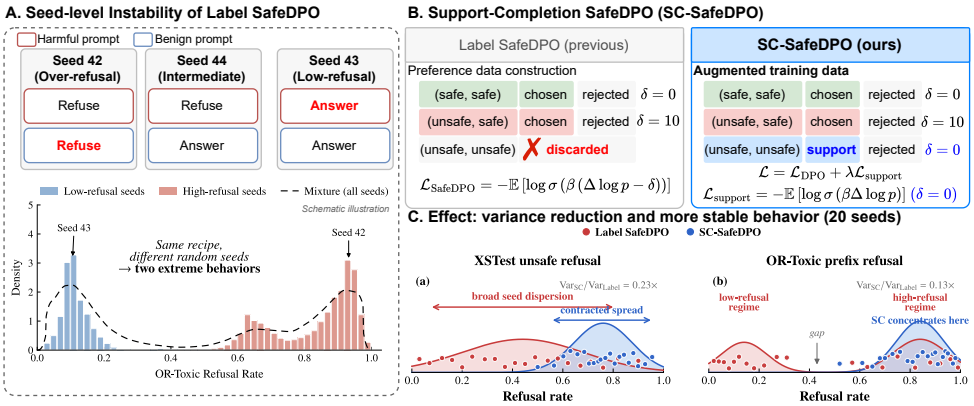


Figure 2: **Seed-level safety-regime instability and SC-SafeDPO. (A)** Label SafeDPO exhibits seed-dependent refusal regimes under identical training settings. **(B)** SC-SafeDPO restores discarded both-unsafe pairs with zero margin and fixed weights. **(C)** Across 20 paired seeds on Qwen2.5-7B, SC-SafeDPO reduces refusal-rate variance to $0.23\times$ and $0.13\times$ that of Label SafeDPO on XSTest unsafe prompts and OR-Toxic, respectively.

2.2 Label-Based SafeDPO

We study the label-based SafeDPO construction (Kim et al., 2026), referred to as Label SafeDPO. Each response has a binary safety label, with 1 indicating safe and 0 unsafe. The construction orders each safe–unsafe pair with the safe response preferred, retains the original helpfulness ordering for both-safe pairs, and discards both-unsafe pairs. Let $\mathcal{D}_{\text{main}}$ denote the resulting nonempty training set. For a retained pair, $s^+, s^- \in \{0, 1\}$ are the safety labels of y^+ and y^- after this transformation.

The per-comparison loss is

$$\ell_m(\theta) = -\log \sigma(\beta r_\theta - m), \qquad m = \delta\, \mathbb{I}[s^+ = 1,\ s^- = 0], \tag{2}$$

where $\delta \geq 0$ is the safety margin in logit units and $\mathbb{I}[\cdot]$ is the indicator function. Safe–unsafe pairs receive margin δ, whereas both-safe pairs use zero margin. Setting $\delta = 0$ removes the additional margin but does not undo the data transformation.

Define $N_{\text{m}} = |\mathcal{D}_{\text{main}}|$ and index the retained comparisons by $i = 1, \ldots, N_{\text{m}}$. Let m_i be the margin assigned by Eq. 2 to comparison i, and let $\ell^{(i)}_{m_i}(\theta)$ denote its loss. The empirical objective is

$$\mathcal{L}_{\text{Label}}(\theta) = \frac{1}{N_{\text{m}}} \sum_{i=1}^{N_{\text{m}}} \ell^{(i)}_{m_i}(\theta). \tag{3}$$

Effect of the safety margin. For a fixed comparison, differentiating Eq. 2 with respect to r_θ gives

$$\frac{\partial \ell_m}{\partial r_\theta} = -\beta\, \sigma(m - \beta r_\theta). \tag{4}$$

At fixed finite r_θ, the absolute value of this derivative increases with m but remains below β. A larger margin therefore shifts its decay toward larger preference log-ratios. Appendix A.6 gives the derivation and distinguishes this scalar derivative from parameter gradients. Section 3.1 introduces the restoration of omitted comparisons.

3 Support-Completion SafeDPO

3.1 Support Construction and Objective

SC-SafeDPO restores both-unsafe comparisons that have a dataset-provided relative-safety ordering. Let $\mathcal{D}_{\text{sup}}$ denote this support set. Both responses retain their unsafe labels; their ordering

supplies supervision within that category. The ordering requires a relative-safety annotation beyond the binary labels. Here, support refers to the empirical comparison set. Appendix B.1 gives the data source and comparison-construction procedure.

Let $N_{\mathrm{m}} = |\mathcal{D}_{\mathrm{main}}| > 0$ and $N_{\mathrm{s}} = |\mathcal{D}_{\mathrm{sup}}| > 0$, where $\mathcal{D}_{\mathrm{main}}$ is the Label SafeDPO set defined in Section 2.2. For policy parameters θ, write $\ell^{(i)}_{m_i}(\theta)$ for the loss on main comparison i with its original margin m_i, and $\ell^{(j)}_{0,\mathrm{sup}}(\theta)$ for the zero-margin loss on support comparison j. Fix a finite weight $\alpha \geq 0$ throughout each run and define $Z_\alpha = N_{\mathrm{m}} + \alpha N_{\mathrm{s}}$. The objective is

$$\mathcal{L}_{\mathrm{SC}}(\theta; \alpha) = \frac{\sum_{i=1}^{N_{\mathrm{m}}} \ell^{(i)}_{m_i}(\theta) + \alpha \sum_{j=1}^{N_{\mathrm{s}}} \ell^{(j)}_{0,\mathrm{sup}}(\theta)}{Z_\alpha}. \tag{5}$$

The main orderings and margins are unchanged. Setting $\alpha = 0$ recovers Label SafeDPO. Positive support weights retain an ordinary DPO preference signal with zero additional margin. SC-SafeDPO changes training supervision and adds no inference-time computation.

Define the support-average loss $\mathcal{L}_{\mathrm{sup}} = N_{\mathrm{s}}^{-1} \sum_{j=1}^{N_{\mathrm{s}}} \ell^{(j)}_{0,\mathrm{sup}}$ and the normalized main weight $\lambda_\alpha = N_{\mathrm{m}}/Z_\alpha$. Then

$$\mathcal{L}_{\mathrm{SC}} = \lambda_\alpha \mathcal{L}_{\mathrm{Label}} + (1 - \lambda_\alpha)\mathcal{L}_{\mathrm{sup}}. \tag{6}$$

Support completion reduces the coefficient of the main objective and introduces support preference gradients. If N_δ denotes the number of main comparisons with positive margin, their total normalized weight is $\rho_\delta(\alpha) = N_\delta/Z_\alpha$. This weight decreases strictly with α when $N_\delta > 0$. These coefficients specify the objective; gradient norms and cumulative comparison exposure depend on the optimizer and sampler. Appendix A.1 gives the corresponding stochastic estimators.

3.2 Local Analysis of Seed Sensitivity

With the dataset and objective fixed, let $u_t^{(h)} \in \mathbb{R}^d$ denote the coordinates of run $h \in \{a, b\}$ at integer iteration $t \geq 0$, for an integer $d \geq 1$. Define $D_t = \mathbb{E}\|u_t^{(a)} - u_t^{(b)}\|^2$ using the Euclidean norm and expectation over training randomness. For independent, identically distributed runs with finite second moments,

$$D_t = 2 \operatorname{tr} \operatorname{Cov}(u_t), \tag{7}$$

where u_t denotes a generic run, and Cov and tr denote covariance and matrix trace. Thus, D_t measures within-configuration seed dispersion.

For $r \in \mathbb{R}$ and finite margin $m \geq 0$, define $f_m(r) = \log(1 + \exp(m - \beta r))$, with preference scale $\beta > 0$ and logistic sigmoid σ from Section 2.1. Its curvature is

$$f''_m(r) = \beta^2 \sigma(m - \beta r)\sigma(\beta r - m). \tag{8}$$

As $m - \beta r \to +\infty$, the derivative magnitude in Eq. equation 4 approaches β and the curvature vanishes.

Curvature from restored comparisons. Let $\mathcal{I}_{\mathrm{m}}$ and $\mathcal{I}_{\mathrm{s}}$ be disjoint main and support index sets of sizes N_{m} and N_{s}, respectively, with $N_{\mathrm{m}}, N_{\mathrm{s}} > 0$ and $\mathcal{I} = \mathcal{I}_{\mathrm{m}} \cup \mathcal{I}_{\mathrm{s}}$. Main comparisons retain their margins m_k; support comparisons have $m_k = 0$. Consider the affine surrogate

$$r_k(u) = b_k + a_k^\top u, \qquad k \in \mathcal{I}, \tag{9}$$

where $u \in \mathbb{R}^d$, $a_k \in \mathbb{R}^d$, and $b_k \in \mathbb{R}$, with a_k, b_k fixed. Let Φ_{m} and Φ_{s} denote the averages of $f_{m_k}(r_k(u))$ over their respective sets, and define $\Phi_\alpha = \lambda_\alpha \Phi_{\mathrm{m}} + (1 - \lambda_\alpha)\Phi_{\mathrm{s}}$ and $H_\alpha(u) = \nabla_u^2 \Phi_\alpha(u)$, with $\lambda_\alpha = N_{\mathrm{m}}/(N_{\mathrm{m}} + \alpha N_{\mathrm{s}})$ for the fixed finite support weight $\alpha \geq 0$ from Section 3.1.

The main comparisons leave the subspace $V = \{v \in \mathbb{R}^d : a_k^\top v = 0 \text{ for all } k \in \mathcal{I}_{\mathrm{m}}\}$ unconstrained. For $v \in V$,

$$v^\top H_\alpha(u) v = \frac{(1 - \lambda_\alpha)\beta^2}{N_{\mathrm{s}}} \sum_{k \in \mathcal{I}_{\mathrm{s}}} \sigma(\beta r_k(u))\sigma(-\beta r_k(u))(a_k^\top v)^2. \tag{10}$$

For $\alpha > 0$ and finite u, this is positive exactly when $a_k^\top v \neq 0$ for some support comparison. Restored comparisons can therefore supply curvature in directions unobserved by the main set. Appendix A.2 states the coverage condition and derives uniform spectral bounds.

Finite-horizon disagreement. Fix $\alpha \geq 0$, an integer horizon $T \geq 1$, and step size $\eta > 0$. For gradient-estimation noise $\zeta_t^{(h)} \in \mathbb{R}^d$, consider

$$u_{t+1}^{(h)} = u_t^{(h)} - \eta\big[\nabla_u \Phi_\alpha(u_t^{(h)}) + \zeta_t^{(h)}\big], \qquad 0 \leq t < T. \tag{11}$$

The analysis uses contraction of smooth gradient updates (Hardt et al., 2016). Let $\mathcal{F}_t$ be the sigma-algebra generated by both runs' histories through their current iterates, before the current noises are drawn. Write I_d for the identity matrix and $\preceq$ for positive-semidefinite ordering.

Theorem 3.1 (Finite-horizon seed sensitivity) *Assume $D_0 < \infty$ and both trajectories remain almost surely in a convex region $\mathcal{U} \subseteq \mathbb{R}^d$ through iteration T. For constants $0 < \mu \leq L < \infty$ and $0 \leq \nu < \infty$, suppose*

$$\mu I_d \preceq H_\alpha(u) \preceq L I_d, \qquad u \in \mathcal{U}, \tag{12}$$

$$\mathbb{E}[\zeta_t^{(h)} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\zeta_t^{(a)} - \zeta_t^{(b)}\|^2 \mid \mathcal{F}_t] \leq \nu^2, \tag{13}$$

where the noise conditions hold almost surely for $0 \leq t < T$ and $h \in \{a, b\}$. If $0 < \eta \leq 1/L$, set $q = 1 - \eta\mu \in [0, 1)$. Then

$$D_T \leq q^{2T} D_0 + \eta^2 \nu^2 \frac{1 - q^{2T}}{1 - q^2}. \tag{14}$$

The bound separates propagated initial discrepancies from accumulated gradient noise. Support completion changes both curvature and noise; its cross-method variance effect is evaluated empirically under AdamW. The theorem describes affine stochastic-gradient dynamics. Appendix A.3 provides the proof and prescribed-step-size extension; Appendix A.4 relates seed dispersion to behavioral measurements.

3.3 Mixture and Orientation Effects

Equation equation 6 separates mixture reweighting from support preference gradients. We compare three support conditions with the same response pairs, support weight, and normalization. Gradient-null support replaces support losses with parameter-independent constants, yielding the objective gradient $\lambda_\alpha \nabla_\theta \mathcal{L}_{\text{Label}}(\theta)$. Shuffled support uses zero-margin preference losses and reverses the orderings of a fixed 50% of support pairs. True support uses the same losses with the annotated relative-safety orderings. Control implementations are detailed in Appendix D.1.

For the scalar loss defined in Section 3.2, reversing a zero-margin comparison gives

$$f_0(-r) = f_0(r) + \beta r. \tag{15}$$

Under the affine model in Eq. equation 9, reversal changes the gradient and preserves the Hessian at a fixed coordinate. The resulting trajectories can encounter different curvature during training. This distinction motivates evaluating orientation separately from curvature coverage; Appendix A.5 gives the derivation. Section 4.4 evaluates these controls, measuring cross-seed behavioral dispersion separately from mean safety and utility.

4 Experiments

4.1 Experimental Setup

Models and training. We evaluate Qwen2.5-7B-Instruct, Qwen2.5-1.5B-Instruct (Qwen et al., 2025), and Gemma-3-12B-IT (Team et al., 2025) using preference data from PKU-SafeRLHF. Label SafeDPO uses 11,144 training pairs; SC-SafeDPO adds 10,342 zero-margin support pairs. Unless varied explicitly, the main Qwen experiments use LoRA with $\beta = 0.1$, $\delta = 10$, and support weight $\alpha = 1$. Optimization settings and training randomness are detailed in Appendix B.3. Baselines include the instruction-tuned base model, DPO (Rafailov et al., 2023), DPO-Helpful, and DPO-SafeBetter, with data and loss definitions in Appendix B.2.

Evaluation. We evaluate safety, over-refusal, and instruction following using StrongREJECT (Souly et al., 2024), OR-Bench (Cui et al., 2025), XSTest (Röttger et al., 2024), HarmBench (Mazeika et al., 2024), and IFEval. The main Qwen OR-Bench results use prefix-based refusal detection. For Gemma, OR-Hard and OR-Toxic measure semantic false refusal and harmful responses, respectively. A separate WildGuard audit re-scores the fixed Qwen responses. Response-generation settings and benchmark scoring rules are given in Appendices C.2 and C.3.

Table 1: Seed-level variation in Label SafeDPO. Results on Qwen2.5-7B-Instruct ($\delta = 10$). Scores are percentages; cross-seed sample SDs are in percentage points.

	OR-Toxic refusal	HarmBench ASR	IFEval strict prompt
Initial runs			
Seed 42	92.21	6.56	68.02
Seed 43	6.41	20.63	68.21
Seed 44	61.22	8.75	68.58
Fixed-endpoint runs (20 seeds)			
Mean	39.63	9.39	67.36
SD	**34.98**	**3.90**	**1.23**
Minimum	1.83	4.06	65.06
Maximum	98.47	16.56	69.50

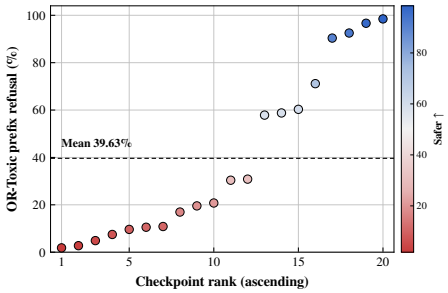


Figure 3: OR-Toxic refusal across 20 training seeds. Label: SafeDPO on Qwen2.5-7B ($\delta = 10$) at a fixed endpoint. Refusal detected by prefix matching; dashed line denotes cross-seed mean.

Seed protocol and statistics. The primary Qwen2.5-7B Label–SC comparison uses 20 paired training seeds for all metrics, with method-specific endpoints fixed across seeds. We report cross-seed means and sample standard deviations. The three-seed cross-backbone comparison forms a separate cohort. Cohort definitions and checkpoint protocols appear in Appendix C.1. Statistical procedures are given in Appendix C.4.

4.2 Seed-Level Instability in Label SafeDPO

The initial three runs of Label SafeDPO on Qwen2.5-7B-Instruct with $\delta = 10$ produce markedly different refusal profiles. Table 1 shows that the run with the lowest OR-Toxic prefix-refusal rate also has the highest HarmBench attack success rate. IFEval strict prompt accuracy changes little across these runs. Substantial differences in safety-related behavior thus coexist with similar measured instruction-following performance.

We examine this variation in a separate study of 20 training seeds, evaluating every run at the same fixed endpoint without per-seed checkpoint reselection. OR-Toxic prefix-refusal rates span nearly the full scale, with several checkpoints near each extreme. HarmBench attack success also varies across runs, while IFEval accuracy remains comparatively stable. The dispersion therefore persists with checkpoint selection held fixed across seeds. Its extent differs across the evaluations, with the largest variation in prefix refusal.

Figure 3 shows why the cross-seed mean alone gives an incomplete account of these outcomes. The mean lies inside a wide gap between observed checkpoint scores, obscuring the separation among the sampled refusal profiles. An intermediate average consequently provides little description of the refusal behavior of an individual checkpoint. This gap characterizes the sampled runs; establishing population bimodality would require additional evidence. Section 4.3 examines whether support completion reduces cross-seed dispersion and how it affects mean safety and utility.

4.3 Effect of Support Completion on Seed Variability

We compare Label SafeDPO and SC-SafeDPO across 20 paired training seeds on Qwen2.5-7B-Instruct. Each method is evaluated at its fixed endpoint. The paired statistics are defined in Appendix C.4. SC has lower cross-seed standard deviations on all five metrics in Table 2. The largest absolute reductions occur in OR-Toxic prefix refusal and XSTest unsafe refusal. HarmBench attack success also becomes less variable, extending the reduction beyond refusal-based measurements.

Mean performance improves alongside the reduction in variability. SC has higher refusal rates on OR-Toxic and XSTest unsafe prompts and lower HarmBench attack success. Its lower OR-Hard false-refusal rate shows that increased refusal on harmful prompts accompanies fewer refusals on benign requests in this comparison. IFEval accuracy also improves, with little variation across runs.

These results compare complete training configurations at their respective endpoints. The matched-budget ablations in Section 4.4 examine the contributions of mixture reweighting and support preference gradients, including their orientation.

Table 2: Seed variability on Qwen2.5-7B-Instruct across 20 paired runs at fixed endpoints. Means are percentages, with cross-seed sample standard deviations shown as subscripts in percentage points. Arrows indicate the preferred direction. Bold indicates the better mean within each Label–SC pair, while green shading indicates a lower cross-seed SD.

Method	OR-Toxic prefix refusal ↑	XSTest unsafe refusal ↑	HarmBench attack success ↓	OR-Hard prefix false refusal ↓	IFEval strict prompt accuracy ↑
Label SafeDPO	$39.63_{\pm 34.98}$	$45.68_{\pm 27.37}$	$9.39_{\pm 3.90}$	$14.80_{\pm 18.38}$	$67.36_{\pm 1.23}$
SC-SafeDPO	$\mathbf{81.08}_{\pm 12.65}$	$\mathbf{76.10}_{\pm 13.13}$	$\mathbf{8.71}_{\pm 1.19}$	$\mathbf{13.56}_{\pm 12.29}$	$\mathbf{69.28}_{\pm 1.06}$

Table 3: Mechanism ablations on Qwen2.5-7B-Instruct. Means are percentages; subscripts denote cross-seed sample standard deviations in percentage points. Arrows indicate the preferred direction of the mean.

Training condition	HarmBench attack success ↓	XSTest unsafe refusal ↑	Semantic OR-Hard false refusal ↓
Matched updates and token budgets: 1,000 updates			
Label SafeDPO, $\delta = 10$	$7.56_{\pm 1.63}$	$71.65_{\pm 21.21}$	$59.08_{\pm 10.86}$
Label SafeDPO, $\delta = 5$	$10.03_{\pm 1.57}$	$78.20_{\pm 9.92}$	$45.76_{\pm 9.46}$
Null support	$7.04_{\pm 1.43}$	$81.60_{\pm 18.20}$	$54.58_{\pm 9.22}$
Shuffled support	$12.88_{\pm 2.14}$	$84.75_{\pm 4.39}$	$46.41_{\pm 9.78}$
True support	$6.94_{\pm 0.54}$	$85.85_{\pm 3.26}$	$42.77_{\pm 7.13}$
Matched main-row exposure: 1,928 updates			
True support	$6.97_{\pm 0.71}$	$85.15_{\pm 3.60}$	$42.27_{\pm 7.21}$

4.4 Mechanism Ablations and Optimization Diagnostics

We evaluate Qwen2.5-7B-Instruct across ten paired training seeds. The primary comparison matches optimizer updates and token budgets using the support conditions in Section 3.3. Control construction and budget protocols are detailed in Appendix D.1. WildGuard assesses semantic OR-Hard false refusal. All ten Label runs at $\delta = 10$ and all runs in the other conditions pass the endpoint eligibility screen.

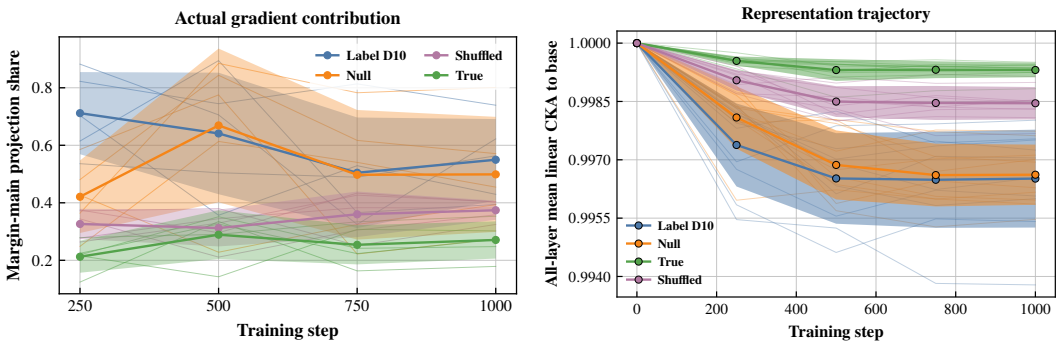
Margin and support orientation. In Table 3, reducing δ from 10 to 5 lowers XSTest unsafe-refusal dispersion and mean semantic OR-Hard false refusal. HarmBench attack success has a higher mean and similar standard deviation. Null support modestly reduces dispersion relative to Label at $\delta = 10$. Under identical mixture weights, shuffled and true support yield much lower XSTest dispersion, supporting an additional contribution from preference gradients. True support also lowers the means and standard deviations of HarmBench attack success and semantic OR-Hard false refusal relative to shuffled support.

Main-comparison exposure. SC also retains lower variability under the exposure-matched control (1,928 updates). Its means and standard deviations remain stable across all three evaluations.

Optimization diagnostics. Using frozen probe batches and five fixed seeds, we measure relative gradient projections within the trainable LoRA parameters, as described in Appendix D.2. Figure 4a shows that shuffled and true support yield lower mean margin-main projection shares than Label and null support at every measured checkpoint, with true support the lowest. All conditions retain high all-layer mean linear CKA to the base model, measured on a fixed internal prompt set, as detailed in Appendix D.3. True support has the highest mean similarity at every measured post-initial checkpoint, followed by shuffled support, as shown in Figure 4b. These changes in gradient allocation and representation similarity accompany reduced behavioral dispersion.

4.5 Cross-Backbone Results and Safety–Utility Trade-offs

Table 4 reports three seeds per fine-tuned configuration, with instruction-tuned bases as references. These runs are separate from the fixed-endpoint study in Section 4.3; SD comparisons are descriptive. OR-Hard uses prefix matching for Qwen and WildGuard for Gemma. OR-Toxic measures



(a) Margin-main projection share.

(b) All-layer mean linear CKA to base.

Figure 4: Optimization diagnostics at registered checkpoints. Gradient projections use frozen probe batches; linear CKA compares each checkpoint with the base model on a fixed internal prompt set.

Table 4: Baseline comparisons across backbones with three seeds per fine-tuned configuration. Means and cross-seed sample SDs are scaled by 100; subscripts give SDs. Within each Label–SC pair, bold denotes the better mean and green shading marks a lower SC SD.

Method	StrongREJECT ↓	OR-Hard false refusal ↓	OR-Toxic Qwen ↑, Gemma ↓	IFEval ↑	HarmBench attack success ↓
Qwen2.5-7B-Instruct					
Base	10.36	11.68	74.50	68.76	19.69
DPO	$9.11_{\pm 1.53}$	$18.15_{\pm 7.89}$	$76.79_{\pm 7.40}$	$66.85_{\pm 0.28}$	$16.25_{\pm 2.67}$
DPO-Helpful	$14.94_{\pm 2.27}$	$2.20_{\pm 0.87}$	$48.55_{\pm 8.80}$	$67.41_{\pm 1.02}$	$26.00_{\pm 1.40}$
DPO-SafeBetter	$11.93_{\pm 1.38}$	$5.71_{\pm 2.72}$	$64.53_{\pm 5.58}$	$67.22_{\pm 1.36}$	$23.75_{\pm 3.26}$
Label SafeDPO	$8.07_{\pm 2.14}$	$17.44_{\pm 16.75}$	$53.28_{\pm 43.45}$	$68.08_{\pm 0.38}$	$11.98_{\pm 7.57}$
SC-SafeDPO	$\mathbf{5.96}_{\pm 0.92}$	$\mathbf{14.28}_{\pm 2.65}$	$\mathbf{57.51}_{\pm 22.84}$	$\mathbf{69.74}_{\pm 0.15}$	$\mathbf{10.33}_{\pm 3.04}$
Qwen2.5-1.5B-Instruct					
Base	9.58	72.86	94.96	38.45	15.62
DPO	$4.80_{\pm 0.20}$	$85.17_{\pm 1.79}$	$97.91_{\pm 0.44}$	$35.61_{\pm 0.56}$	$9.17_{\pm 1.10}$
DPO-Helpful	$30.23_{\pm 3.19}$	$8.52_{\pm 2.45}$	$43.41_{\pm 7.22}$	$35.49_{\pm 0.92}$	$33.12_{\pm 2.56}$
DPO-SafeBetter	$12.56_{\pm 0.71}$	$27.55_{\pm 0.92}$	$76.18_{\pm 0.93}$	$35.43_{\pm 0.56}$	$25.21_{\pm 0.65}$
Label SafeDPO	$\mathbf{3.56}_{\pm 1.20}$	$87.57_{\pm 2.04}$	$\mathbf{98.12}_{\pm 0.23}$	$35.30_{\pm 1.11}$	$\mathbf{8.65}_{\pm 0.48}$
SC-SafeDPO	$4.09_{\pm 1.15}$	$\mathbf{85.54}_{\pm 1.93}$	$97.56_{\pm 0.26}$	$\mathbf{37.28}_{\pm 1.02}$	$9.69_{\pm 1.36}$
Gemma-3-12B-IT					
Base	8.52	7.29	5.95	62.66	29.69
DPO	$6.48_{\pm 1.08}$	$9.17_{\pm 2.31}$	$4.63_{\pm 0.64}$	$62.29_{\pm 0.81}$	$28.02_{\pm 0.79}$
DPO-Helpful	$9.76_{\pm 0.99}$	$5.86_{\pm 1.62}$	$6.26_{\pm 1.25}$	$62.35_{\pm 0.38}$	$30.52_{\pm 1.78}$
DPO-SafeBetter	$8.31_{\pm 0.70}$	$6.75_{\pm 1.25}$	$5.45_{\pm 0.38}$	$62.91_{\pm 0.38}$	$27.50_{\pm 0.31}$
Label SafeDPO	$6.58_{\pm 1.00}$	$\mathbf{8.95}_{\pm 1.80}$	$4.23_{\pm 1.12}$	$63.16_{\pm 0.80}$	$27.86_{\pm 2.08}$
SC-SafeDPO	$\mathbf{6.26}_{\pm 0.91}$	$11.58_{\pm 1.25}$	$\mathbf{3.01}_{\pm 0.68}$	$\mathbf{63.24}_{\pm 0.54}$	$\mathbf{26.52}_{\pm 1.82}$

prefix refusal for Qwen and semantic harmful-response rates for Gemma. Scoring rules are detailed in Appendix C.3.

On Qwen2.5-7B, SC achieves the lowest mean StrongREJECT score and HarmBench attack success and the highest IFEval accuracy among reported methods. DPO has higher OR-Toxic refusal and less favorable adversarial scores. Relative to Label SafeDPO, SC improves IFEval and slightly reduces benign false refusal on Qwen2.5-1.5B, alongside less favorable mean safety and increased OR-Toxic and HarmBench variability. On Gemma with $\alpha = 0.3$, SC lowers mean StrongREJECT scores, semantic harmful-response rates, and HarmBench attack success, with lower observed SDs across all metrics. Benign false refusal increases, and IFEval remains nearly unchanged.

Table 5: Random-sample human audit with two independent annotators and a third adjudicator. Rates are percentages after excluding ambiguous cases; bold marks the highest observed rate in each column. Pass counts and scoring denominators are given in Appendix E.3.

Method	Safety ↑	Helpfulness ↑
SFT	40.5	**83.0**
SafeDPO	70.5	51.5
SC-SafeDPO	**93.5**	69.0

Across backbones, DPO-Helpful has the lowest mean benign false-refusal rate and highest mean HarmBench attack success. DPO-SafeBetter reduces attack success with increased benign refusal, illustrating the trade-offs among training objectives.

4.6 Evaluator Sensitivity and Human Audit

WildGuard assesses semantic refusal and response harmfulness on frozen Qwen responses from the fixed-endpoint study, with checkpoints and generations unchanged (Appendices E.1 and C.3).

SafeDPO and SC-SafeDPO share randomly sampled harmful and benign prompts selected independently of evaluator agreement, with checkpoint seeds sampled independently between methods. Sampling, blinded human annotation, and adjudication are detailed in Appendices E.2 and E.3. Safety measures avoidance of harmful compliance on harmful prompts; helpfulness measures safe compliance on benign prompts.

Table 5 shows higher safety and helpfulness pass rates for SC-SafeDPO than SafeDPO, and higher safety but lower helpfulness than SFT. Rates pool final human labels across sampled checkpoints; cross-seed variability is evaluated separately in Section 4.3.

5 Discussion and Limitations

Safety alignment should be evaluated through cross-seed distributions and mean performance under explicit checkpoint protocols. Variability and mean safety should be assessed separately under each scoring convention. Pooled human-audit pass rates complement these measurements with assessments of average safety and helpfulness.

The ablations in Section 4.4 support stabilizing contributions from mixture reweighting and restored preference gradients; annotated orderings lower HarmBench attack success. The local affine analysis in Section 3.2 links comparison coverage to seed sensitivity under stated spectral and noise assumptions. Lower dispersion accompanies altered gradient allocation and higher CKA to the base model in Figure 4; diagnostics do not establish representation change as a causal mediator.

Our experiments cover Qwen and Gemma configurations trained on PKU-SafeRLHF with support weights fixed within each run. Further evaluation across datasets and training configurations, including adaptive weighting, would clarify the scope of these results. Model-dependent trade-offs motivate selecting support weights under the intended training and evaluation protocol.

6 Conclusion

We identified seed-level safety-regime instability in large-margin Label SafeDPO: a fixed training recipe can produce different refusal profiles that cross-seed means obscure. SC-SafeDPO restores discarded both-unsafe comparisons with annotated relative-safety orderings, zero additional margin, and a fixed positive weight. It preserves main-comparison orderings and margins without additional inference-time computation. Across 20 paired seeds on Qwen2.5-7B-Instruct, SC-SafeDPO reduces cross-seed variance by approximately 77% for XSTest unsafe refusal and 87% for OR-Toxic prefix refusal relative to Label SafeDPO, with improved means on all five metrics. Local affine analysis shows how restored comparisons can supply curvature in directions unobserved by the main set. Budget-matched ablations support stabilizing contributions from mixture reweighting and restored preference gradients. Cross-backbone evaluations characterize safety–helpfulness trade-offs. A random-sample human audit finds higher safety and helpfulness pass rates for SC-SafeDPO than for SafeDPO. These results motivate considering comparison coverage in preference optimization and assessing safety alignment through mean performance and cross-seed reliability.

AI USE STATEMENT

In this work, we used generative AI tools to assist with statistical tabulation and data summarization in the *Evaluator Sensitivity and Human Audit* subsection of our study on large language models, and with language polishing and formatting. All AI-assisted work was reviewed by the authors. The statistical work in that subsection was verified against the raw data and manually checked for correctness; any errors were corrected. The language and formatting edits were checked to preserve the original meaning, technical claims, and paper structure. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, human participants, or the collection, use, or release of personal, sensitive, or private data. The only human involvement was the authors' evaluation of AI-generated content for safety and helpfulness. This evaluation was conducted solely to judge the safety and helpfulness of the generated content for the purposes of this study; it did not involve any other interventions, data collection, or ethical concerns. As no human subjects were involved, IRB approval was not required. No conflicts of interest are declared. The AI-generated content was used only for research and analysis, and the authors do not endorse or promote any harmful, biased, or unlawful applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we have created an anonymous repository that contains the essential code and data used in our experiments: https://anonymous.4open.science/r/sc-85CE. The repository includes the implementation, data processing scripts, and running instructions needed to reproduce the main experimental results. Detailed descriptions of the experimental setup, datasets, model configurations, and hyperparameters are provided in the main paper and the appendix. The repository is anonymized to comply with the double-blind review process.

REFERENCES

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3207–3214, 2018.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety. In *International Conference on Learning Representations*, volume 2026, pp. 41069–41108, 2026.

Erik Larsen. The instability of safety: How random seeds and temperature expose inconsistent LLM refusal behavior. *arXiv preprint arXiv:2512.12066*, 2025. URL https://arxiv.org/abs/2512.12066v3. Version 3, revised August 2026.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 338–348, 2017.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael

Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Shiji Zhao, Mengyang Wang, Shukun Xiong, Fangzhou Chen, Qihui Zhu, Shouwei Ruan, Yisong Xiao, Ranjie Duan, Xun Chen, and XingXing Wei. Improving safety alignment via balanced direct preference optimization. *arXiv preprint arXiv:2603.22829*, 2026.

Appendix

A Analysis of Support Completion

A.1 Objective Normalization and Stochastic Estimation

Use the disjoint comparison index sets $\mathcal{I}_{\mathrm{m}}$ and $\mathcal{I}_{\mathrm{s}}$ from Section 3.2, and let $N = N_{\mathrm{m}} + N_{\mathrm{s}}$. For each $k \in \mathcal{I}$, let $\ell_k(\theta)$ denote its re-indexed comparison loss. Define the unnormalized weight w_k and normalized weight p_k by

$$w_k = \begin{cases} 1, & k \in \mathcal{I}_{\mathrm{m}}, \\ \alpha, & k \in \mathcal{I}_{\mathrm{s}}, \end{cases} \qquad p_k = \frac{w_k}{Z_\alpha}. \tag{16}$$

The normalized weights form a probability distribution $p = (p_k)_{k\in\mathcal{I}}$. Splitting $\sum_{k\in\mathcal{I}} p_k \ell_k$ into its two groups gives Eq. equation 6. Their independence from θ gives

$$\nabla_\theta \mathcal{L}_{\mathrm{SC}} = \lambda_\alpha \nabla_\theta \mathcal{L}_{\mathrm{Label}} + (1-\lambda_\alpha)\nabla_\theta \mathcal{L}_{\mathrm{sup}}. \tag{17}$$

Holding the comparison sets and margins fixed,

$$\frac{d\rho_\delta(\alpha)}{d\alpha} = -\frac{N_\delta N_{\mathrm{s}}}{Z_\alpha^2}. \tag{18}$$

This derivative is negative when $N_\delta > 0$ and zero when no comparison has a positive margin. At $\alpha = 0$, it is a right derivative on the admissible domain.

Minibatch estimators. Fix θ and define $g_k = \nabla_\theta \ell_k(\theta)$ and $g = \sum_{k\in\mathcal{I}} p_k g_k$. Let $\varpi = (\varpi_k)_{k\in\mathcal{I}}$ be a sampling distribution with $\varpi_k > 0$ whenever $p_k > 0$. For a positive integer batch size n_{b}, draw $K_1, \ldots, K_{n_{\mathrm{b}}}$ independently from ϖ. The estimator

$$\widehat{g} = \frac{1}{n_{\mathrm{b}}} \sum_{j=1}^{n_{\mathrm{b}}} \frac{p_{K_j}}{\varpi_{K_j}} g_{K_j} \tag{19}$$

satisfies $\mathbb{E}[\widehat{g} \mid \theta] = g$. Independent draws also give

$$\mathbb{E}[\|\widehat{g} - g\|^2 \mid \theta] = \frac{1}{n_{\mathrm{b}}} \left(\sum_{k:\varpi_k>0} \frac{p_k^2}{\varpi_k} \|g_k\|^2 - \|g\|^2 \right). \tag{20}$$

The identity follows by expanding the second moment of one draw; the centered cross terms vanish. Sampling from p uses unit multipliers. Uniform sampling uses multipliers Np_k. Objective weights alone therefore leave the gradient-noise variance undetermined.

A weighted minibatch average with denominator $\sum_{j=1}^{n_{\mathrm{b}}} w_{K_j}$ is generally a biased gradient estimator for Eq. equation 5. With one uniformly sampled comparison and $\alpha > 0$, the weight cancels and the gradient is g_{K_1}, whose expectation is uniform over comparisons. At $\alpha = 1$, the uniform and target weights coincide. At $\alpha = 0$, a support-only batch gives a zero denominator.

Null support and exposure. Replacing support losses with parameter-independent constants, while retaining their weights and Z_α, gives expected gradient $\lambda_\alpha \nabla_\theta \mathcal{L}_{\mathrm{Label}}$. Dropping support rows and normalizing over the main set removes this reweighting. Under direct sampling from p, the probability of a main draw is λ_α; its cumulative exposure also depends on the number of draws. Under weighted uniform sampling, changing coefficients can leave row frequencies unchanged. Update counts, processed tokens, and main-row exposure must therefore be specified separately.

A.2 Affine Curvature and Comparison Coverage

For the affine log-ratios in Eq. equation 9, define $\omega_k(u) = \sigma(m_k - \beta r_k(u))\sigma(\beta r_k(u) - m_k)$. Differentiating the surrogate objective gives

$$H_\alpha(u) = \beta^2 \sum_{k\in\mathcal{I}} p_k \omega_k(u) a_k a_k^\top. \tag{21}$$

For finite u, every $\omega_k(u)$ lies in $(0, 1/4]$.

Proposition A.1 (Curvature coverage) *Fix $\alpha > 0$. At every finite coordinate, $H_\alpha(u)$ is positive definite if and only if the vectors $\{a_k : k \in \mathcal{I}\}$ span $\mathbb{R}^d$. Under this spanning condition, suppose a finite constant $B \geq 0$ satisfies $|m_k - \beta r_k(u)| \leq B$ for every $k \in \mathcal{I}$ and $u \in \mathcal{U}$. Then the uniform spectral bounds in Eq. equation 12 hold for positive finite constants μ and L.*

Proof. For every $v \in \mathbb{R}^d$,

$$v^\top H_\alpha(u) v = \beta^2 \sum_{k \in \mathcal{I}} p_k \omega_k(u) (a_k^\top v)^2. \tag{22}$$

All coefficients are positive when $\alpha > 0$. The quadratic form vanishes exactly when v is orthogonal to every comparison vector, which proves the spanning criterion. On the main-unobserved subspace V, the main terms vanish, and the support terms give Eq. equation 10.

Define the weighted Gram matrix $G_\alpha = \sum_{k \in \mathcal{I}} p_k a_k a_k^\top$ and $\omega_B = \sigma(B)\sigma(-B)$. For $z \in \mathbb{R}$, the function $\sigma(z)\sigma(-z)$ is even and decreases with $|z|$, so

$$\beta^2 \omega_B G_\alpha \preceq H_\alpha(u) \preceq \frac{\beta^2}{4} G_\alpha, \qquad u \in \mathcal{U}. \tag{23}$$

Let $\gamma_{\min} > 0$ and $\gamma_{\max} < \infty$ denote the smallest and largest eigenvalues of G_α. Valid constants are

$$\mu = \beta^2 \omega_B \gamma_{\min}, \qquad L = \frac{\beta^2}{4} \gamma_{\max}. \tag{24}$$

□

A rank-deficient comparison set has zero curvature on the orthogonal complement of its span. Deleting both-unsafe comparisons does not establish that V is nontrivial, and restoration does not ensure full coverage. Support terms can also couple V to its orthogonal complement. The sensitivity theorem consequently requires spectral bounds on the entire coordinate space under analysis.

Nonlinear log-ratios. For a twice continuously differentiable log-ratio $r_k(u)$, the exact Hessian of its comparison loss is

$$\nabla_u^2 f_{m_k}(r_k(u)) = f''_{m_k}(r_k(u)) \nabla_u r_k(u) \nabla_u r_k(u)^\top + f'_{m_k}(r_k(u)) \nabla_u^2 r_k(u). \tag{25}$$

The second term can be indefinite. The affine assumption removes it within the surrogate. Applying the spectral result to neural-network parameters would require separate control of this term.

A.3 Proof of the Finite-Horizon Sensitivity Bound

Proof of Theorem 3.1. Define $e_t = u_t^{(a)} - u_t^{(b)}$ and $\Delta\zeta_t = \zeta_t^{(a)} - \zeta_t^{(b)}$. The current iterates and e_t are $\mathcal{F}_t$-measurable. Let $\overline{H}_t$ denote the Hessian integrated along the segment between the iterates:

$$\overline{H}_t = \int_0^1 H_\alpha(u_t^{(b)} + s e_t)\, ds. \tag{26}$$

Convexity of $\mathcal{U}$ places this segment inside the region of the spectral bounds. The fundamental theorem of calculus gives $\nabla_u \Phi_\alpha(u_t^{(a)}) - \nabla_u \Phi_\alpha(u_t^{(b)}) = \overline{H}_t e_t$. Subtracting the updates yields

$$e_{t+1} = (I_d - \eta \overline{H}_t) e_t - \eta \Delta\zeta_t. \tag{27}$$

Integration preserves $\mu I_d \preceq \overline{H}_t \preceq L I_d$. For $0 < \eta \leq 1/L$, the eigenvalues of $I_d - \eta \overline{H}_t$ lie in $[0, q]$, giving $\|(I_d - \eta \overline{H}_t) e_t\| \leq q \|e_t\|$. The conditional zero-mean assumption eliminates the mixed term in the squared norm:

$$\begin{aligned} \mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] &= \|(I_d - \eta \overline{H}_t) e_t\|^2 + \eta^2 \mathbb{E}[\|\Delta\zeta_t\|^2 \mid \mathcal{F}_t] \\ &\leq q^2 \|e_t\|^2 + \eta^2 \nu^2. \end{aligned} \tag{28}$$

Taking expectation gives $D_{t+1} \leq q^2 D_t + \eta^2 \nu^2$. Iterating for T steps and summing the geometric series proves Eq. equation 14. The argument controls the noise difference directly and allows dependence between the two runs' current noises. □

Prescribed step sizes. Let η_t be deterministic step sizes with $0 \le \eta_t \le 1/L$, and let $\nu_t \ge 0$ be finite deterministic constants whose squares replace the constant noise bound. Define $q_t = 1 - \eta_t \mu$. The same recursion gives

$$D_T \le \left(\prod_{t=0}^{T-1} q_t^2 \right) D_0 + \sum_{t=0}^{T-1} \eta_t^2 \nu_t^2 \prod_{s=t+1}^{T-1} q_s^2, \tag{29}$$

where an empty product equals one. Momentum, adaptive preconditioning, and decoupled weight decay require additional analysis.

A finite-horizon construction. The region condition can hold without projected updates. Consider independent minibatch draws from p, and let $A = \max_{k \in \mathcal{I}} \|a_k\|$. Each sampled gradient and its minibatch average have norm at most βA. Suppose both initial iterates lie almost surely within distance $R_0 < \infty$ of a fixed center $u_\circ \in \mathbb{R}^d$, with $R_0 \ge 0$. Define $R_T = R_0 + T\eta\beta A$. The triangle inequality gives

$$\|u_t^{(h)} - u_\circ\| \le R_0 + t\eta\beta A \le R_T, \qquad h \in \{a, b\}, \quad 0 \le t \le T. \tag{30}$$

On this ball, a valid shifted-logit bound is

$$B = \max_{k \in \mathcal{I}} \left\{ |m_k - \beta(b_k + a_k^\top u_\circ)| + \beta \|a_k\| R_T \right\}. \tag{31}$$

For $\alpha > 0$ and spanning comparison vectors, Proposition A.1 supplies the spectral constants. The upper constant L is independent of B, so its step-size restriction can be imposed before choosing the ball. With conditionally independent batches across runs, Eq. equation 20 permits $\nu^2 = 2\beta^2 A^2 / n_{\mathrm{b}}$. The lower curvature bound can become small as T increases.

Region containment is an assumption on the stochastic process itself. Conditioning arbitrary trajectories on remaining in a chosen region can change the noise distribution and invalidate conditional unbiasedness.

A.4 Seed Dispersion and Behavioral Readouts

For independent, identically distributed runs with finite second moments, independence gives $\mathbb{E}[(u_t^{(a)})^\top u_t^{(b)}] = \|\mathbb{E} u_t\|^2$. Expanding the squared difference therefore yields

$$D_t = 2 \left(\mathbb{E}\|u_t\|^2 - \|\mathbb{E} u_t\|^2 \right) = 2 \operatorname{tr} \operatorname{Cov}(u_t), \tag{32}$$

which proves Eq. equation 7. This identity concerns two runs of the same configuration. The SC–Label variance ratio compares different configurations and is estimated empirically; ordering their upper bounds would not order their actual variances.

Let $\psi : \mathcal{U} \to \mathbb{R}$ be a readout with a finite Lipschitz constant $C \ge 0$, meaning $|\psi(u) - \psi(v)| \le C\|u - v\|$ for all $u, v \in \mathcal{U}$. Then

$$\mathbb{E}\left[(\psi(u_T^{(a)}) - \psi(u_T^{(b)}))^2 \right] \le C^2 D_T. \tag{33}$$

Under independent identical runs, the same variance identity gives $\operatorname{Var}(\psi(u_T)) \le C^2 D_T / 2$. Each affine log-ratio r_k has Lipschitz constant $\|a_k\|$. Decoded responses and thresholded refusal labels can be discontinuous, so transferring this bound to refusal or attack-success rates requires additional assumptions on generation and evaluation. The analysis also leaves the probability of distinct behavioral regimes unspecified.

A.5 Support Weight and Comparison Orientation

An unobserved coordinate. Consider the two-coordinate case $u = (w, v) \in \mathbb{R}^2$, with scalar coordinates w and v, log-ratios $r_{\mathrm{main}}(w, v) = w$ and $r_{\mathrm{sup}}(w, v) = v$, and group losses $\Phi_{\mathrm{m}}(w, v) = f_\delta(w)$ and $\Phi_{\mathrm{s}}(w, v) = f_0(v)$. Main-only gradient descent preserves $v_t = v_0$. An existing discrepancy in v persists; identical initial coordinates and main-only sampling create no discrepancy in that coordinate within this model.

For $\alpha > 0$, define the support update map $F_\alpha : \mathbb{R} \to \mathbb{R}$ by

$$v_{t+1} = F_\alpha(v_t), \qquad F_\alpha(v) = v + \eta(1-\lambda_\alpha)\beta\sigma(-\beta v). \tag{34}$$

If $0 < \eta(1-\lambda_\alpha)\beta^2 \le 4$, its derivative satisfies

$$F'_\alpha(v) = 1 - \eta(1-\lambda_\alpha)\beta^2\sigma(\beta v)\sigma(-\beta v) \in [0, 1). \tag{35}$$

For a finite interval radius $R \ge 0$, define $q_R = 1-\eta(1-\lambda_\alpha)\beta^2\sigma(\beta R)\sigma(-\beta R) < 1$. The mean value theorem gives contraction factor q_R for inputs in $[-R, R]$. Let $v_t^{(a)}$ and $v_t^{(b)}$ denote two trajectories of this map. If both remain in that interval for $0 \le t < T$, then $|v_T^{(a)} - v_T^{(b)}| \le q_R^T |v_0^{(a)} - v_0^{(b)}|$. The map satisfies $F_\alpha(v) > v$ at every finite v, so a bounded interval need not be invariant. This example describes finite-horizon contraction of cross-run differences, with no attraction to the reference-relative coordinate $v = 0$.

Dependence on support weight. Let $H_{\mathrm{m}}(u) = \nabla_u^2 \Phi_{\mathrm{m}}(u)$ and $H_{\mathrm{s}}(u) = \nabla_u^2 \Phi_{\mathrm{s}}(u)$. At a fixed coordinate,

$$H_\alpha(u) - H_{\mathrm{m}}(u) = (1-\lambda_\alpha)\big(H_{\mathrm{s}}(u) - H_{\mathrm{m}}(u)\big). \tag{36}$$

The difference of the two group Hessians can be indefinite. In the two-coordinate example, define $\kappa_{\mathrm{m}} = \beta^2\sigma(\delta)\sigma(-\delta)$ and $\kappa_{\mathrm{s}} = \beta^2/4$. At the origin,

$$H_\alpha(0,0) = \begin{pmatrix} \lambda_\alpha\kappa_{\mathrm{m}} & 0 \\ 0 & (1-\lambda_\alpha)\kappa_{\mathrm{s}} \end{pmatrix}. \tag{37}$$

Its smallest eigenvalue is $\min\{\lambda_\alpha\kappa_{\mathrm{m}}, (1-\lambda_\alpha)\kappa_{\mathrm{s}}\}$. As α increases, the two entries move in opposite directions, so the minimum need not increase monotonically. Support-weight selection therefore remains an empirical question.

Reversing a support comparison. For the scalar log-ratio r, direct expansion gives

$$f_0(-r) = \log(1 + \exp(\beta r)) = f_0(r) + \beta r. \tag{38}$$

For a support comparison $k \in \mathcal{I}_{\mathrm{s}}$ with affine log-ratio $r_k(u) = b_k + a_k^\top u$, differentiation yields

$$\begin{aligned} \nabla_u f_0(-r_k(u)) - \nabla_u f_0(r_k(u)) &= \beta a_k, \\ \nabla_u^2 f_0(-r_k(u)) &= \nabla_u^2 f_0(r_k(u)). \end{aligned} \tag{39}$$

The Hessian equality holds at the same coordinate. Different orientations can produce different trajectories and encounter different Hessians during training. For nonlinear log-ratios, the Hessian difference is $\beta\nabla_u^2 r_k(u)$. Orientation controls consequently assess the direction of preference learning as well as changes in behavioral dispersion. Gradient projections and representation comparisons at registered checkpoints are accompanying diagnostics; establishing representation drift as a mediator would require additional evidence.

A.6 Safety-Margin Derivatives and Parameter Gradients

Setup and notation. Fix an ordered comparison (x, y^+, y^-), where x is a prompt and y^+ and y^- are the preferred and dispreferred responses under the retained training ordering. Let $\pi_\theta(y \mid x)$ be the policy with trainable parameter vector θ, and let $\pi_{\mathrm{ref}}(y \mid x)$ be the fixed reference policy, where y denotes a response. Assume that the policy probabilities are differentiable in θ and that both policies assign positive probability to the two responses. Write $r_\theta = r_\theta(x, y^+, y^-)$ for the reference-normalized preference log-ratio in Eq. equation 1. Let $\beta > 0$ be the fixed preference-scale parameter and $m \ge 0$ a finite comparison-specific margin in logit units. As in Eq. equation 2, $m = \delta$ for a safe–unsafe comparison and $m = 0$ for a both-safe comparison, where $\delta \ge 0$ is the finite safety-margin hyperparameter. Comparisons, orderings, margins, and the reference policy are held fixed when differentiating with respect to θ.

To separate scalar differentiation from parameter differentiation, let r be an independent real-valued log-ratio and define

$$f_m(r) = -\log\sigma(\beta r - m) = \log(1 + \exp(m - \beta r)), \tag{40}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the logistic sigmoid for a real scalar z. The per-comparison parameterized loss is the composition $\ell_m(\theta) = f_m(r_\theta)$.

Scalar derivative and margin dependence. Define $a_m(r) = |f'_m(r)|$, where the prime denotes differentiation with respect to r. Direct differentiation gives

$$\begin{aligned} f'_m(r) &= -\beta \frac{\exp(m - \beta r)}{1 + \exp(m - \beta r)} = -\beta \sigma(m - \beta r), \\ \frac{\partial a_m(r)}{\partial m} &= \beta \sigma(m - \beta r)\sigma(\beta r - m) > 0. \end{aligned} \tag{41}$$

Thus, the notation $\partial \ell_m / \partial r_\theta$ in Eq. equation 4 denotes $f'_m(r_\theta)$, not a gradient with respect to the policy parameters. For finite r and m, the sigmoid lies strictly between zero and one, so $0 < a_m(r) < \beta$. The second identity establishes strict increase with m at fixed r; at $m = 0$, the derivative is understood as a right derivative on the admissible domain. Moreover, $\partial a_m(r)/\partial r = -\beta \, \partial a_m(r)/\partial m < 0$. The exact translation identity $a_m(r) = a_0(r - m/\beta)$ shows that the margin shifts the scalar-derivative magnitude curve by m/β toward larger log-ratios without changing its upper bound. In particular, $a_m(m/\beta) = \beta/2$, and $a_m(r)$ tends to zero as $r \to +\infty$ for fixed m.

Parameter gradients. Writing ∇_θ for the gradient with respect to θ, the chain rule and the parameter independence of the reference policy yield

$$\begin{aligned} \nabla_\theta \ell_m(\theta) &= f'_m(r_\theta) \nabla_\theta r_\theta = -a_m(r_\theta) \nabla_\theta r_\theta, \\ \nabla_\theta r_\theta &= \nabla_\theta \log \pi_\theta(y^+ \mid x) - \nabla_\theta \log \pi_\theta(y^- \mid x). \end{aligned} \tag{42}$$

At the same parameters and for the same ordered comparison, increasing m changes only the scalar coefficient. If $\nabla_\theta r_\theta \neq 0$, the parameter-gradient direction is unchanged and its norm increases strictly; if $\nabla_\theta r_\theta = 0$, the parameter gradient remains zero. For the Euclidean norm $\| \cdot \|_2$, the exact relation is $\|\nabla_\theta \ell_m(\theta)\|_2 = a_m(r_\theta) \|\nabla_\theta r_\theta\|_2$. Consequently, $\|\nabla_\theta \ell_m(\theta)\|_2 \leq \beta \|\nabla_\theta r_\theta\|_2$, with strict inequality when $\nabla_\theta r_\theta \neq 0$. The scalar bound $a_m(r) < \beta$ is therefore not, by itself, a bound of β on the parameter-gradient norm.

Aggregation and scope of the result. Let $\mathcal{D}_{\text{main}}$ be the fixed retained comparison set and $N_{\text{m}} = |\mathcal{D}_{\text{main}}| > 0$ its size. For comparison $i \in \{1, \ldots, N_{\text{m}}\}$, write $r_\theta^{(i)}$ for its reference-normalized log-ratio and m_i for its fixed margin. The empirical mean loss $\mathcal{L}_{\text{Label}}$ in Eq. equation 3 has gradient

$$\nabla_\theta \mathcal{L}_{\text{Label}}(\theta) = -\frac{1}{N_{\text{m}}} \sum_{i=1}^{N_{\text{m}}} a_{m_i}\left(r_\theta^{(i)}\right) \nabla_\theta r_\theta^{(i)}. \tag{43}$$

Because comparison gradients can point in different directions, increasing their scalar coefficients need not increase the norm of their average. Along training trajectories, both the log-ratios and their parameter gradients change, and optimizer updates additionally depend on sampling and optimizer state. The pointwise identities above therefore establish neither longer training-time persistence of large gradients nor seed-level instability. Such claims require trajectory-level analysis or empirical evidence beyond the scalar derivative.

Table 6: Training and validation comparison counts.

Comparison set	Training	Validation
Main comparisons	11,144	1,416
Both-unsafe support	10,342	1,258
Support-completed set	21,486	2,674

Table 7: Data and loss definitions for preference-trained methods. Main and support orderings are defined in Appendix B.1; ℓ_0 is the zero-margin DPO loss in Section 2.1.

Method	Training comparisons	Preference ordering	Loss
DPO	Main set	Main ordering	ℓ_0
DPO-Helpful	All pair types	Helpfulness	ℓ_0
DPO-SafeBetter	Safety-ordered comparisons	Relative safety	ℓ_0
Label SafeDPO	Main set	Main ordering	Eq. equation 3
SC-SafeDPO	Main and support sets	Main and support orderings	Eq. equation 5

B Training Data and Implementation

B.1 Preference Data and Comparison Construction

Data source and splits. We derive preference comparisons from the official training split of PKU-SafeRLHF, available as `PKU-Alignment/pku-safe_rlhf-30_k`, at revision `334ad7d4d008421be5a9b098dd256951eba1a5c5`. The source file is `pku-safe_rlhf-30_k-train.arrow`. Each record contains a prompt, two responses, binary safety labels `is_response_0_safe` and `is_response_1_safe`, a relative-safety preference `safer_response_id`, and a helpfulness preference `better_response_id`.

Comparisons are constructed before splitting. We group records by exact prompt-string equality and identify each group by the SHA-256 hash of its unnormalized prompt. Rows with missing or invalid annotations are excluded. The prompt groups are partitioned into training and validation sets using split seed 20260715, with no shared prompt groups between the two splits. The main training set contains 3,334 distinct prompts; the support-completed training and validation sets contain 5,575 and 693 distinct prompts, respectively.

Comparison construction. The main set $\mathcal{D}_{\text{main}}$ contains comparisons with at least one safe response. For mixed-safety pairs, the safe response is chosen and the unsafe response is rejected. Both-safe pairs follow the helpfulness ordering in `better_response_id`. The support set $\mathcal{D}_{\text{sup}}$ contains both-unsafe comparisons, with the response identified by `safer_response_id` chosen and the other response rejected. SC-SafeDPO uses the union of these sets, preserving the main orderings and the binary safety labels of both responses. Table 6 gives the comparison counts.

B.2 Baseline Definitions and Loss Implementation

Baseline definitions. The instruction-tuned base models are evaluated as reference checkpoints. Table 7 summarizes the data and loss definitions for the preference-trained methods. DPO and DPO-Helpful use $\beta = 0.1$.

For Label SafeDPO and SC-SafeDPO, mixed-safety main comparisons receive margin δ, and both-safe comparisons receive zero additional margin. SC-SafeDPO assigns zero additional margin to support comparisons. Main rows have unit weight, and support rows have weight α, fixed throughout each run.

Sequence log-probabilities. For the Qwen2.5-7B Label/SC runs, the reference policy is the frozen Qwen2.5-7B-Instruct checkpoint with adapters disabled. Reference log-probabilities are precom-

Table 8: Reproducible training configuration for the Qwen2.5-7B-Instruct Label SafeDPO and SC-SafeDPO experiments (seed 42).

Setting	Label SafeDPO	SC-SafeDPO
Base model	Qwen2.5-7B-Instruct	
Base dtype	`torch.bfloat16`	
Adaptation	LoRA ($r = 64$, $\alpha = 128$, dropout $= 0.05$)	
Trainable parameters	161,480,704	
Optimizer	AdamW	
Learning rate	1×10^{-5}	
Weight decay	0.05	
Learning-rate schedule	Cosine decay	
Warmup ratio	0.03	
Batch size per device	4	
Gradient accumulation	2	
Effective batch size	8 comparison pairs	
Eval batch size	2	
Maximum sequence length	512 tokens	
Epochs	3	
Logging steps	25	
Seed	42	
Fraction grid	$0.125, 0.25, 0.5, 0.75, 1.0$	
Support horizon fraction	0.125	
Stop-after fraction	0.25	0.125
Method	`label_safedpo`	`SC_safedpo`
DPO coefficient β	0.1	
Safety margin δ	10.0	
BFPO/SC support weight α	0.5	
BFPO b_1	3.0	
Chosen floor	-0.25	
Likelihood tolerance	0.05	
Dual initial value	1.0	
Dual learning rate	0.1	
Dual maximum	20.0	
Dry run	false	

puted. The model and tokenizer use the same base-model revision, and inputs follow its chat template. Sequence log-probabilities sum over response tokens, including EOS; prompt and padding positions are masked out.

Loss reduction. In the default Qwen2.5-7B comparison, SC uses $\alpha = 1$. Both methods average per-comparison losses within each microbatch and average the microbatch losses during gradient accumulation. For a full accumulation window containing K microbatches,

$$\widehat{\mathcal{L}}_{\mathrm{acc}} = \frac{1}{K} \sum_{j=1}^{K} \frac{1}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} \ell_i, \tag{44}$$

where $\mathcal{B}_j$ contains the comparison rows in microbatch j and ℓ_i is the row loss with its assigned margin.

B.3 Models, Optimization, and Training Randomness

Models and optimization. We fine-tune Qwen2.5-7B-Instruct, Qwen2.5-1.5B-Instruct, and Gemma-3-12B-IT using LoRA. Table 8 reports the default configuration for the Qwen2.5-7B-Instruct Label SafeDPO and SC-SafeDPO experiments. The Gemma cross-backbone comparison uses support weight $\alpha = 0.3$.

Study-specific evaluation checkpoints are listed in Appendix C.1.

Training randomness. Each training seed initializes the random number generators used for adapter initialization and dropout. For run seed s and epoch index e, the data loader uses seed $s + e$ to uniformly permute comparison rows. The final partial minibatch is retained. The same shuffling rule applies across the compared methods.

Table 9: Experimental cohorts. Seed counts apply to fine-tuned configurations. Qwen and Gemma denote the instruction-tuned backbones listed in Section 4.1.

Study	Backbone	Training seeds
Initial observation	Qwen2.5-7B	3: 42, 43, 44
Fixed-endpoint Label–SC comparison	Qwen2.5-7B	20 paired
Mechanism ablations	Qwen2.5-7B	10 paired
Main-row-exposure control	Qwen2.5-7B	Same 10 pairs
Gradient-projection diagnostics	Qwen2.5-7B	Fixed subset of 5
Cross-backbone comparison	Qwen2.5-7B, Qwen2.5-1.5B, Gemma-3-12B	3 per configuration

Table 10: Shared Qwen2.5-7B response-generation settings for OR-Bench, IFEval, and StrongREJECT.

Setting	Value
Tokenizer and chat template	Qwen2.5-7B-Instruct; same revision as the weights
Explicit system turn	None
Appended generation prompt	None
Maximum new tokens	512
Responses per prompt	1
Inference precision	bfloat16
Quantization	None
Inference batch size	16
Stop markers in saved responses	Removed

C Evaluation and Statistical Protocols

C.1 Study Cohorts and Checkpoint Protocols

Table 9 lists the experimental cohorts. The initial Qwen observation and the fixed-endpoint study use separate training runs. The fixed-endpoint Label–SC comparison uses the same 20 paired runs for every reported metric, including HarmBench attack success. The three-seed cross-backbone comparison forms a separate cohort.

Checkpoint protocols. The initial Qwen observation and the cross-backbone comparison use validation-selected checkpoints. In the fixed-endpoint Qwen study, each method is evaluated at one designated endpoint held constant across its training seeds.

The primary mechanism ablations are evaluated at update 1,000. The main-row-exposure control extends true-support training to update 1,928. All runs in the mechanism comparison pass the endpoint eligibility screen.

C.2 Response Generation and Reproducibility

Qwen response generation. The Qwen2.5-7B evaluation pipelines use greedy decoding with `do_sample=false`. Models run in evaluation mode with dropout disabled. Table 10 gives the shared response-generation settings for OR-Bench, IFEval, and StrongREJECT. These settings are held fixed across methods and training seeds.

Stopping rules and random-number initialization. Table 11 lists the EOS, PAD, and generation-seed settings for each Qwen2.5-7B evaluation pipeline.

The shared generator calls `torch.manual_seed` once per process before model loading. A resumed process initializes the same seed and generates the remaining responses. XSTest initializes the Python, NumPy, PyTorch CPU, and CUDA random number generators once per evaluation process and disables cuDNN benchmarking. HarmBench initializes the Python and PyTorch random

Table 11: Stopping tokens and generation seeds for Qwen2.5-7B. EOS specifies stopping tokens, and PAD specifies the padding token.

Evaluation pipeline	EOS IDs	PAD ID	Generation seed
OR-Bench, IFEval, StrongREJECT	151645	151643	Checkpoint training seed
XSTest	151645, 151643	151643	Checkpoint training seed
HarmBench	151645, 151643	151645	20260713

Table 12: Benchmark and scoring definitions. N is the evaluation-set size per checkpoint, measured in behaviors for HarmBench and prompts for the other benchmarks. Arrows indicate the preferred direction of each metric.

Benchmark / subset	Metric	Scoring procedure	N
StrongREJECT	Mean StrongREJECT score ↓	Official fine-tuned Gemma-2B judge	313
OR-Hard, main Qwen	Prefix false-refusal rate ↓	Prefix matching	1,319
OR-Toxic, main Qwen	Prefix-refusal rate ↑	Prefix matching	655
XSTest unsafe	Refusal rate ↑	Refusal-string matching	200
XSTest safe	False-refusal rate ↓	Refusal-string matching	250
HarmBench	Attack success rate ↓	Official Llama-2-13B classifier	320
IFEval	Strict prompt accuracy ↑	Official instruction verifiers	541
OR-Hard, semantic	False-refusal rate ↓	WildGuard	1,319
OR-Toxic, semantic	Harmful-response rate ↓	WildGuard	655

number generators once before generating missing responses. Each pipeline retains its initialization across subsequent batches and prompts. Generation failures are recorded separately.

C.3 Benchmarks and Scoring Conventions

Table 12 lists the evaluation subsets, metrics, and scoring procedures. Evaluation examples are fixed within each cohort. The main Qwen OR-Bench results use prefix matching. Gemma OR-Bench results and the Qwen mechanism study's OR-Hard results use WildGuard semantic judgments.

Lexical refusal scoring. For Qwen OR-Bench evaluations, we strip leading and trailing whitespace and match the response prefix case-sensitively against `I cannot`, `Sorry`, `As an AI language model, I`, and `I'm sorry, but`. A match counts as refusal on OR-Toxic and false refusal on OR-Hard. XSTest uses its official refusal-string inventory, with whitespace stripped and matching restricted to the response prefix. Rates are computed separately for the unsafe and safe subsets.

Semantic scoring. WildGuard labels request harmfulness, response refusal, and response harmfulness separately. Its input concatenates the evaluation instruction, user prompt, and assistant response into one sequence, which is right-truncated to 3,072 tokens. The judge generates at most 32 new tokens. Benchmark subset membership defines the harmful and benign evaluation groups.

The Qwen semantic audit re-scores the frozen OR-Hard and OR-Toxic responses from the fixed-endpoint study, retaining the original checkpoints and generations. It records both semantic refusal and response harmfulness. Each semantic rate uses responses with valid judgments as its denominator. Parsing failures and missing responses are recorded separately. The frozen-response evaluation is described in Appendix E.1.

Table 13: Control definitions for the mechanism study. Main preference orderings are fixed across conditions. Both-safe main comparisons have zero additional margin. All support conditions use $\alpha = 1$.

Condition	Mixed-safety margin	Support loss and ordering
Label SafeDPO	10	No support
Label SafeDPO	5	No support
Null support	10	Parameter-independent loss
Shuffled support	10	Zero-margin DPO; a fixed 50% of annotated orderings reversed
True support	10	Zero-margin DPO; annotated relative-safety ordering

C.4 Across-Seed Statistical Analysis

Summary statistics. For a fixed cohort and metric, let $x_{a,s}$ be the score of method a at its designated checkpoint for training seed s. Let $\mathcal{S}$ be the included seed set and $n = |\mathcal{S}|$. For $n \geq 2$, the mean and sample variance are

$$\bar{x}_a = \frac{1}{n} \sum_{s \in \mathcal{S}} x_{a,s}, \tag{45}$$

$$\widehat{\sigma}_a^2 = \frac{1}{n-1} \sum_{s \in \mathcal{S}} (x_{a,s} - \bar{x}_a)^2. \tag{46}$$

The reported sample standard deviation is $\widehat{\sigma}_a$ and describes variation across training seeds. Each cohort, checkpoint protocol, and scoring convention is summarized separately.

Paired comparisons. Label–SC comparisons use the same set of training-seed identifiers for both methods. The primary Qwen comparison has $n = 20$ for every reported metric. We compute the paired mean difference and variance ratio as

$$\widehat{\Delta} = \frac{1}{n} \sum_{s \in \mathcal{S}} (x_{\mathrm{SC},s} - x_{\mathrm{Label},s}), \tag{47}$$

$$\widehat{R} = \frac{\widehat{\sigma}^2_{\mathrm{SC}}}{\widehat{\sigma}^2_{\mathrm{Label}}}, \qquad \widehat{\sigma}^2_{\mathrm{Label}} > 0. \tag{48}$$

A positive $\widehat{\Delta}$ indicates a higher mean score for SC; the metric directions are listed in Table 12. A variance ratio below one indicates lower observed cross-seed variance for SC. For percentage-valued metrics, standard deviations and mean differences are expressed in percentage points. Variance ratios are dimensionless.

Paired uncertainty estimation. Bootstrap resampling preserves method pairing by training seed. Each replicate samples seed identifiers with replacement and applies the same seed multiplicities to both methods.

D Mechanism Ablations and Optimization Diagnostics

D.1 Control Construction and Budget Protocols

We conduct the mechanism study on Qwen2.5-7B-Instruct using ten paired training seeds, numbered 120 through 129. The conditions use the main and support comparison sets defined in Appendix B.1. Table 13 summarizes the controls.

Support controls. Null, shuffled, and true support use the same response pairs, unit row weights, and microbatch normalization described in Appendix B.2. In the null condition, support rows occupy

batch slots and remain in the loss denominator. Their forward passes are executed, and their losses contribute zero gradients to the trainable parameters. The reversal mask for shuffled support is shared across training seeds and remains fixed throughout training.

Training budgets. The primary comparison matches optimizer updates and token budgets across the five conditions. Each run uses 1,000 optimizer updates with an effective batch size of eight. The exposure-matched control uses the same ten seeds and extends true-support training to 1,928 updates, equalizing cumulative main-row presentations with Label SafeDPO at 1,000 updates.

D.2 Gradient-Projection Diagnostics

We compute relative gradient projections over the trainable LoRA parameters using frozen probe batches and a fixed subset of five training seeds. The diagnostic compares Label SafeDPO with $\delta = 10$, null support, shuffled support, and true support.

The reported quantity is the margin-main projection share, which measures the relative contribution of margin-bearing main comparisons to the mixture gradient. Figure 4a displays these measurements at updates 250, 500, 750, and 1,000.

D.3 Representation-Similarity Diagnostics

We compare each checkpoint with the Qwen2.5-7B-Instruct base model on a fixed internal prompt set. The reported similarity measure is all-layer mean linear CKA. Figure 4b shows this quantity for Label SafeDPO with $\delta = 10$, null support, shuffled support, and true support at updates 0, 250, 500, 750, and 1,000.

E Evaluator Sensitivity and Human Audit

E.1 Frozen-Response Semantic Evaluation

We use WildGuard to re-score the Qwen2.5-7B-Instruct responses used for prefix-based evaluation in the fixed-endpoint SafeDPO–SC study. Checkpoints and generated responses remain fixed. For OR-Hard and OR-Toxic, we record semantic refusal and response harmfulness as separate outcomes. Judge settings and scoring rules are given in Appendix C.3.

E.2 Human-Audit Sampling and Annotation

Sampling frame. The harmful sampling pool contains 1,175 prompts: 655 from OR-Toxic, 320 from the HarmBench text evaluation, and 200 from XSTest unsafe. The benign pool contains 1,569 prompts: 1,319 from OR-Hard and 250 from XSTest safe. We sample prompts randomly within each pool using a fixed random-number generator. Prompt selection is independent of agreement between automatic evaluators.

For SafeDPO and SC-SafeDPO, responses come from the frozen collections of 20 checkpoints per method in the fixed-endpoint study (Appendix C.1). Both methods share the sampled prompts, with checkpoint seeds selected independently within each method. The resulting method comparisons are paired by prompt. Each response record is identified by its prompt, method, and training seed.

Blinded annotation. Two human annotators, A and B, independently label each prompt–response pair using a shared rubric. Method and training-seed identities are concealed. Prompt labels are `harmful`, `benign`, and `ambiguous`. The rubric labels requests that substantially facilitate illegal activity, harm, evasion, or exploitation as harmful. Benign requests seek lawful information, prevention, analysis, or creative work. An ambiguous label records cases where the available context does not support a determinate judgment.

E.3 Agreement, Adjudication, and Endpoint Accounting

Agreement before adjudication. Table 14 reports prompt- and response-label agreement before adjudication in the 378-row SafeDPO–SC annotation summary.

Table 14: A/B agreement before adjudication in the 378-row SafeDPO–SC annotation summary.

Label level	Agreements / rows	Agreement (%)	κ
Prompt	297/378	78.57	0.618
Response	281/378	74.34	0.602

Table 15: Human-audit endpoints after adjudication and ambiguity exclusions. Rates are percentages rounded to one decimal place.

Method	Endpoint	Scored n	Passes	Rate (%)
SFT	Safety	200	81	40.5
SFT	Helpfulness	200	166	83.0
SafeDPO	Safety	139	98	70.5
SafeDPO	Helpfulness	200	103	51.5
SC-SafeDPO	Safety	139	130	93.5
SC-SafeDPO	Helpfulness	200	138	69.0

Adjudication. Within this annotation summary, the union of prompt- and response-label disagreements contains 124 rows. A third human adjudicator, C, resolves disagreements and reviews referred ambiguous or excluded cases. Final labels use A/B agreements for unreviewed cases and C's decisions for reviewed cases. Records with a final ambiguous prompt or response label are excluded from the corresponding endpoint.

Endpoint definitions and counts. For method m, let $\mathcal{I}_{H,m}$ and $\mathcal{I}_{B,m}$ be the scored response records from the harmful and benign pools, respectively, after adjudication and ambiguity exclusions. Let r_i be the final human response label for record i. Safety measures avoidance of harmful compliance on harmful prompts, and helpfulness measures safe compliance on benign prompts:

$$\widehat{\text{Safety}}(m) = \frac{\sum_{i \in \mathcal{I}_{H,m}} \mathbf{1}\{r_i \neq \texttt{harmful_compliance}\}}{|\mathcal{I}_{H,m}|}, \tag{49}$$

$$\widehat{\text{Helpfulness}}(m) = \frac{\sum_{i \in \mathcal{I}_{B,m}} \mathbf{1}\{r_i = \texttt{safe_compliance}\}}{|\mathcal{I}_{B,m}|}. \tag{50}$$

Each valid response contributes once. Rates pool responses across sampled checkpoints. Table 15 reports the pass counts and valid denominators for the percentages in Table 5.