When Does DPO Add Value over Continued SFT? Evidence from Correction–Preservation Frontiers

Anonymous authors
Paper under double-blind review

Abstract

After supervised fine-tuning (SFT), when does Direct Preference Optimization (DPO) add value over chosen-only continued SFT (CSFT)? Starting from shared SFT checkpoints, we track correction of inherited errors and preservation of initially correct decisions. Across two models and two ranking datasets, controlled increases in training disagreement widen the DPO–CSFT accuracy gap by 2.62–4.08 percentage points, mainly through CSFT decline rather than increased DPO gain over SFT. DPO incurs less corruption over overlapping correction ranges; the tested CSFT schedule and KL variants on MS MARCO do not recover its high-correction, low-corruption region. In naturally sampled ESCI and TL;DR, DPO instead gains mainly through additional correction despite greater corruption. These contrasting results show how DPO's relative value shifts between correction and preservation across training conditions. Exploratory analysis suggests that high training disagreement combined with a small SFT-error pool can signal stringent preservation requirements before further training; predicting realized gains remains open.

1 Introduction

Supervised fine-tuning (SFT) followed by preference optimization is a standard way to adapt language models (Ouyang et al., 2022; Rafailov et al., 2023). Once preference pairs are available, Stage-2 training poses a practical choice: continue cross-entropy training on preferred responses alone, or use Direct Preference Optimization (DPO), which contrasts preferred and rejected responses relative to the SFT policy (Rafailov et al., 2023). Existing comparisons vary with the model, data, and task, but aggregate outcomes do not reveal which inherited decisions drive that variation (Thakkar et al., 2024; Feng & Yang, 2026). We ask when DPO adds value over chosen-only continued SFT (CSFT), whether that value comes from correcting errors or preserving correct decisions, and what the starting SFT policy can tell us before further training.

A controlled comparison shows why aggregate performance is insufficient. In a constructed high-disagreement MS MARCO arm, DPO and CSFT correct nearly the same fraction of SFT's inherited errors: 56.1% and 56.3%. Yet they corrupt 1.52% and 6.29% of the decisions that SFT initially made correctly. Similar correction can carry different preservation costs, while similar accuracy can hide additional corrections canceled by corruptions. We therefore characterize an objective by its joint correction–preservation operating region rather than by its endpoint score alone.

We study policy–preference disagreement as a candidate driver of this separation. A Stage-2 pair exhibits disagreement when SFT assigns higher likelihood to the rejected response than to the preferred response. Such pairs indicate correction demand in the training data, distinct from the errors available to correct on the target distribution. We vary their share while fixing the SFT checkpoint and held-out examples. This tests how training composition changes the correction–preservation balance while the inherited evaluation error pool stays fixed.

The experiments proceed in three steps. First, 13 natural ranking settings across models and datasets establish how correction and preservation combine under naturally sampled Stage-2 distributions. Second, controlled sweeps vary intermediate-disagreement mass for Qwen3-4B and Gemma-3-4B on MS MARCO and ESCI. A CSFT learning-rate–checkpoint grid and targeted reverse-KL check

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.

on MS MARCO test simple optimization and anchoring explanations. Third, TL;DR extends the analysis to sequence preferences (Stiennon et al., 2020); generation and supplementary TREC ranking evaluate task utility. All comparisons start from shared SFT checkpoints, share Stage-2 examples within an arm, and use disjoint held-out examples. Pairwise ranking is the primary testbed because its margins expose inherited decisions and boundary crossings directly (Liu et al., 2025; 2024; Cai et al., 2025).

The results reveal two sources of DPO's value. In natural ranking settings, DPO consistently corrects more inherited errors; in naturally sampled ESCI and TL;DR, its advantage is mainly correction-driven despite greater corruption than CSFT. Under controlled stress, the gap instead widens in all four model–dataset cells mainly through CSFT decline, while DPO's gain over SFT remains roughly stable or declines. One additional independently trained Qwen/MS MARCO SFT start reproduces this response. The frontier remains task-dependent: correction converges at the strongest MS MARCO arms, whereas on ESCI CSFT reaches higher correction at greater corruption. DPO has lower corruption where correction ranges overlap. The tested CSFT schedule grid and one reverse-KL formulation do not recover its high-correction, low-corruption region, but do not exhaust CSFT alternatives. Task utility also requires separate evaluation: TL;DR generation favors DPO under two model judges while exposing criterion-level trade-offs (Section 6.3).

A relative advantage over CSFT can coexist with little gain over retaining SFT. We therefore also explore whether further training offers a useful gain. An exploratory disagreement-only screen identifies whether DPO gains at least one accuracy point over SFT in 9/11 natural settings under leave-one-setting-out evaluation. Reading disagreement alongside the SFT-error fraction adds a different question: for an assumed correction rate and target gain over SFT, how much corruption can be afforded? A small error pool can make that budget stringent even when training disagreement is high. Our retrospective analysis illustrates this diagnostic; estimating the error pool on a separate calibration set and forecasting realized correction and corruption remain open (Section 7.1; Appendix G).

The main contributions are:

- Evidence that DPO's additional correction in natural ranking settings is often partly offset by preservation loss, which aggregate accuracy conceals.
- A two-model, two-dataset controlled intervention revealing lower DPO corruption over overlapping correction ranges and a CSFT-decline stress response, supported by an additional SFT start and bounded schedule and KL checks.
- Cross-task evidence and exploratory diagnostics linking training correction demand to preservation requirements, motivating calibration-based assessment before Stage-2 training while keeping fixed-pair gains and task utility separate.

2 Related Work

2.1 SFT, Preference Optimization, and Post-SFT Dynamics

Direct Preference Optimization (DPO) contrasts preferred and rejected responses relative to a fixed reference policy (Rafailov et al., 2023). DPO and continued SFT share positive supervision but constrain departures from the inherited policy differently. A unified account connects them through implicit rewards and argues that conventional SFT lacks an effective KL constraint (Wang et al., 2025). This motivates our bounded check of whether explicit KL regularization recovers DPO's observed correction–preservation region.

Empirical comparisons show task-dependent advantages. Across more than 300 parameter-efficient runs, Thakkar et al. (2024) find conditions where SFT outperforms DPO; a small-model study likewise reports task-dependent marginal DPO gains (Feng & Yang, 2026). Such comparisons motivate examining which inherited decisions produce the aggregate performance difference.

Finer-grained analyses examine preference errors and optimization dynamics. Preference objectives may fail to repair ranking errors inherited from the reference policy (Chen et al., 2024), and preference distinguishability changes update rates (Im & Li, 2024). Other work studies rejected-response suppression, preferred-response likelihood displacement, and sensitivity to rejected responses and initialization (Feng et al., 2024; Razin et al., 2025; Cho et al., 2025; Jian et al., 2025). Our focus is

the held-out correction–preservation response to training composition. We branch from shared SFT checkpoints, match Stage-2 examples between objectives, and vary disagreement mass while keeping the evaluation error and correct pools fixed. This identifies differences in observed operating regions; it does not isolate their optimization mechanism.

2.2 Preference Optimization for Ranking

Ranking-oriented preference methods learn from ranked responses, connect listwise optimization to learning-to-rank, optimize ranking metrics, or target reranking and top-K consistency (Yuan et al., 2023; Song et al., 2024; Liu et al., 2025; Zhou et al., 2025; Liu et al., 2024; Cai et al., 2025). These methods develop objectives for ranking performance. We use pairwise ranking to compare how DPO and chosen-only continuation correct inherited errors and preserve correct decisions. TREC provides a separate downstream evaluation because these pairwise transitions do not directly determine listwise utility.

3 Experimental Framework

3.1 Shared SFT Starting Points and Stage-2 Objectives

The pipeline is Base $\rightarrow$ SFT $\rightarrow$ $\{\text{CSFT}, \text{DPO}\}$. Within each model–dataset–seed comparison, CSFT and DPO branch from the same SFT checkpoint and receive the same ordered Stage-2 examples. Stage-1, Stage-2, and held-out inputs are disjoint.

For input x with preferred and rejected outputs y^+ and y^-, define

$$\ell_\theta(y \mid x) = \sum_{t=1}^{|y|} \log P_\theta(y_t \mid x, y_{<t}), \qquad m_\theta(x) = \ell_\theta(y^+ \mid x) - \ell_\theta(y^- \mid x).$$

For the ranking tasks, y^+ and y^- are single-token labels; for TL;DR they are complete summaries. CSFT continues cross-entropy training on y^+ alone. DPO minimizes

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta[m_\theta(x) - m_{\text{ref}}(x)]),$$

where both the policy and frozen reference are initialized from the shared SFT checkpoint. The primary DPO setting uses $\beta = .1$; Appendix B.4 reports ranking sensitivity.

Unless varied explicitly, all stages use LoRA with rank 16, $\alpha = 32$, dropout .05, attention and MLP projections, three epochs, learning rate 10^{-4}, effective batch size 8, and bfloat16 precision.

We use seeds $7, 42, 123$. Natural comparisons use their corresponding seed-specific SFT checkpoints. Within each controlled sweep, the three Stage-2 seeds share a fixed SFT start. The additional independent-start replication is reported separately (Appendix C.2).

3.2 Tasks, Models, and Evaluation Units

The primary experiments cast passage and product ranking as single-token `A`/`B` decisions. MS MARCO uses BM25 hard negatives; U.S. ESCI pairs Exact products with Complement or Irrelevant products and excludes ambiguous Substitute labels (Bajaj et al., 2016; Reddy et al., 2022). Splits are query-disjoint and answer positions randomized. We evaluate Qwen3-1.7B/4B/8B, Gemma-3-1B/4B, Llama-3.2-3B-Instruct, and Llama-3.1-8B-Instruct on both datasets (Yang et al., 2025; Gemma Team et al., 2025; Grattafiori et al., 2024). This gives 14 model–dataset settings; the primary aggregate retains 13 after a post-hoc stability exclusion of Gemma-3-1B/ESCI. Each retained setting has three Stage-2 seeds and a common 10,000-pair held-out set. Appendix B.3 reports the exclusion sensitivity.

TL;DR provides a sequence-preference setting (Stiennon et al., 2020): 4,000 Stage-1 posts, 1,600 Stage-2 pairs, and 1,500 held-out pairs, disjoint by post with at most one pair per post. Qwen3-4B-Instruct-2507 and Gemma-3-4B-it use the same three seeds. Sequence-sum margins matching the DPO objective are primary; token averaging is a length-normalized sensitivity (Appendix D).

Task utility is evaluated separately. Two model judges score greedy TL;DR summaries of the same 500 held-out posts; paired total-score comparisons give ties half credit. Appendix F specifies the scoring protocol and audits. Appendix E.1 reports the supplementary listwise TREC evaluation.

3.3 Correction, Preservation, and Aggregate Performance

For held-out example i, a decision is target-consistent when $m_i(\theta) > 0$. Let

$$W = \{i : m_i(\theta_{\mathrm{SFT}}) \le 0\}, \qquad R = \{i : m_i(\theta_{\mathrm{SFT}}) > 0\}$$

be the inherited SFT-error and SFT-correct pools. Both methods use these fixed pools. For checkpoint θ, define correction, preservation, and corruption as

$$c_\theta = \Pr(m_i(\theta) > 0 \mid i \in W), \qquad p_\theta = \Pr(m_i(\theta) > 0 \mid i \in R), \qquad q_\theta = 1 - p_\theta.$$

Aggregate accuracy is $(|W|c_\theta + |R|p_\theta)/N$. Writing $e = |W|/N$, the accuracy gain over SFT is $ec_\theta - (1-e)q_\theta$. Thus, a small error pool limits correction gains and gives preservation greater weight. Appendix G.3 uses this identity to derive corruption budgets under assumed correction rates and target gains.

For the DPO–CSFT comparison, let $\Delta c = c_{\mathrm{DPO}} - c_{\mathrm{CSFT}}$ and $\Delta p = p_{\mathrm{DPO}} - p_{\mathrm{CSFT}}$. Then

$$\mathrm{Acc}(\mathrm{DPO}) - \mathrm{Acc}(\mathrm{CSFT}) = \frac{|W|}{N}\Delta c + \frac{|R|}{N}\Delta p.$$

We call the two terms the correction and preservation contributions. Preservation differences receive greater weight when $|R| \gg |W|$.

Each checkpoint defines an operating point (c_θ, q_θ), with higher correction and lower corruption preferred. A point Pareto-dominates another if it has at least as much correction and no more corruption, with one strict inequality. The observed frontier covers only tested checkpoints and hyperparameters.

In ranking, target consistency is dataset-designated correctness. In TL;DR, correction means crossing the annotated preference boundary; it does not establish factual summary quality, which is evaluated separately.

Multi-seed summaries use the mean and sample standard deviation unless stated otherwise. Resampling preserves pairing and the evaluation unit—examples, TREC queries, or generation posts—and aggregate intervals preserve their stated setting and seed structure. Intervals condition on the observed checkpoints.

3.4 Controlled Policy–Preference Disagreement

A Stage-2 pair with $m_{\mathrm{SFT}}(x) < 0$ is a policy–preference disagreement: the SFT policy favors the rejected output. This training-set property is distinct from the disjoint held-out error pool W. An exploratory location study found the clearest late frontier separation in the intermediate-margin stratum, which was then fixed as the focal population; the exploratory comparison is not pooled with the intervention.

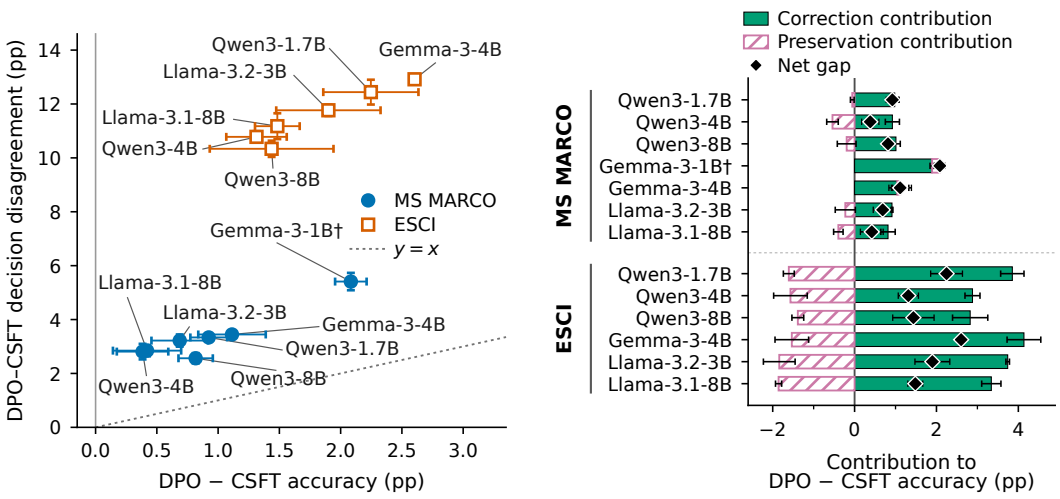
Controlled sweeps cover Qwen3-4B and Gemma-3-4B on MS MARCO and ESCI. Each arm contains 2,000 pairs, replacing nested, position-balanced focal subsets with SFT-correct anchors as disagreement mass decreases. Within each setting, the starting SFT and 10,000 held-out pairs remain fixed; within each arm and seed, CSFT and DPO share examples and order. We evaluate updates 75, 300, and 750 using position-symmetrized margins.

The primary Qwen/MS MARCO sweep uses masses $0/2/5/10/20/40\%$; its canonical intermediate-disagreement share is 1.2%. Appendix C.1 reports the replication grids and construction audits.

Focal cohorts are reconstructed for each SFT start. High-mass arms are stress tests, and focal pairs and anchors may differ beyond margin sign; the intervention therefore tests constructed mixtures rather than an isolated disagreement mechanism. Section 6.2 specifies the CSFT schedule and KL robustness checks.

4 Natural Training Distributions Reveal a Hidden Correction–Preservation Trade-off

CSFT and DPO can reach similar aggregate accuracy while changing different held-out decisions. We first expose this trade-off relative to their shared SFT starting point, then test policy–preference disagreement as one candidate driver in Section 5.



(a) Net accuracy gap versus decision disagreement.

(b) Correction and preservation contributions.

Figure 1: Natural-setting transitions across 13 settings. Points and bars are three-seed means; error bars show one standard deviation. (a) Decision disagreement exceeds the net accuracy gap. (b) The gap decomposes into correction and preservation contributions; black diamonds show their sum. The dagger marks the additional complete Gemma-3-1B/MS MARCO setting.

4.1 Similar Aggregate Performance Conceals Different Transitions

Let $n_{D\bar{C}}$ count held-out examples that DPO classifies correctly and CSFT incorrectly, and let $n_{C\bar{D}}$ count the reverse. Then

$$\mathrm{Acc}(\mathrm{DPO}) - \mathrm{Acc}(\mathrm{CSFT}) = \frac{n_{D\bar{C}} - n_{C\bar{D}}}{N},$$
$$\mathrm{Disagree}(\mathrm{DPO}, \mathrm{CSFT}) = \frac{n_{D\bar{C}} + n_{C\bar{D}}}{N}.$$

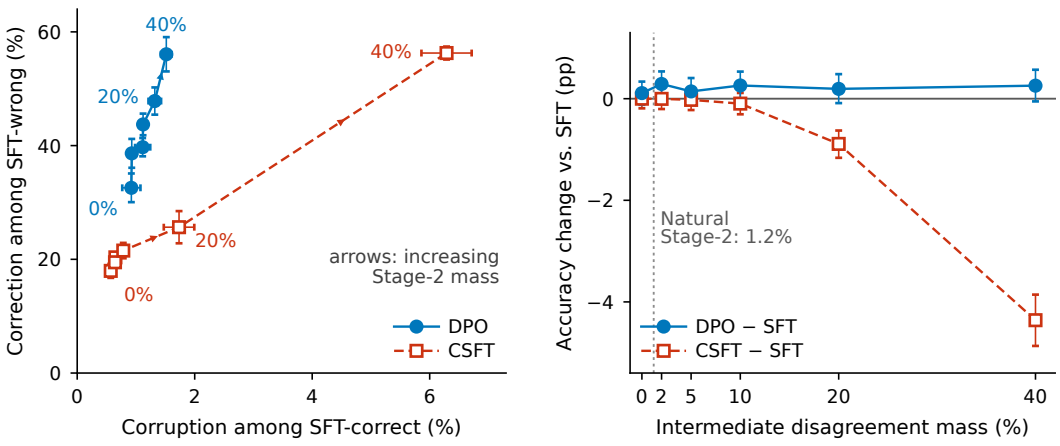
The accuracy gap can therefore be small even when the methods change many different decisions. Conditioning these disagreements on the inherited SFT state distinguishes method-specific corrections on SFT-wrong examples from corruptions of SFT-correct examples.

This cancellation appears throughout the natural ranking runs (Figure 1a). Appendix B.1 reports the paired-outcome audits. For Qwen3-4B on MS MARCO, DPO exceeds CSFT accuracy by only 0.4 ± 0.2 percentage points across three seeds, yet their decision status differs on $2.8 \pm 0.3\%$ of examples. On ESCI with the same model, the corresponding values are 1.3 ± 0.2 points and $10.8 \pm 0.2\%$. Across five MS MARCO settings with an accuracy gap below one point, disagreement is 2.6–3.3%, approximately three to seven times the net gap.

4.2 DPO Corrects More, but Preservation Shapes the Net Gain

Applying the exact decomposition from Section 3.3, we separate DPO's net advantage into correction and preservation contributions. The preservation term is negative when DPO corrupts more initially correct decisions than CSFT. Because the inherited-correct pool is much larger, a small conditional preservation difference can offset a larger correction advantage.

Across the 39 retained runs, DPO always corrects more than CSFT, with a 13-setting hierarchical mean difference of 21.9 percentage points (95% CI $[19.9, 24.2]$). Including the post-hoc excluded Gemma-3-1B/ESCI setting, correction remains higher in 41 of 42 runs, but the 14-setting net-gap interval includes zero (Appendix B.3). DPO's additional conditional corruption relative to CSFT averages 1.0 point (95% CI $[0.5, 1.5]$), but its weighted effect differs sharply across datasets (Figure 1b). On MS MARCO, correction and preservation contribute $+1.1$ and -0.2 accuracy points on average, leaving a net gain of $+0.9$. On ESCI, they contribute $+3.5$ and -1.6 points, leaving $+1.8$; preservation is negative for every model and offsets approximately 48% of the correction gain.



(a) Correction–corruption paths at update 750.

(b) Accuracy change relative to the starting SFT.

Figure 2: Controlled response to intermediate disagreement mass in Qwen3-4B/MS MARCO. Points are three-seed means. (a) Each path connects masses 0, 2, 5, 10, 20, and 40% at update 750; error bars show one standard deviation across Stage-2 seeds. (b) Error bars are 95% joint-bootstrap intervals for the three-seed mean. The dotted line marks the canonical Stage-2 intermediate share (1.2%, 24 of 2,000 pairs) as descriptive context, not a threshold.

Thus, in the retained settings, additional correction overcomes a task-dependent preservation cost; DPO does not uniformly preserve better. Section 5 tests training disagreement as a candidate driver using controlled mixtures.

5 Controlled Disagreement Reveals When the Objectives Diverge

Natural distributions cannot identify what changes the correction–preservation balance because they differ in inherited error mass, pair composition, and difficulty. We therefore intervene on one candidate driver: the share of Stage-2 pairs on which the SFT policy disagrees with the preference label. Holding the model, task, starting SFT, training-set size, and optimization budget fixed, we trace how CSFT and DPO move through the correction–preservation plane. This intervention measures responses to our constructed mixtures; it does not define a task-independent threshold.

5.1 A Controlled Disagreement-Mass Intervention

Using the controlled design in Section 3.4, we vary the intermediate-disagreement share in Qwen3-4B/MS MARCO over $\{0, 2, 5, 10, 20, 40\}\%$. The canonical share is 1.2%, so the 20–40% arms are stress tests rather than deployment-prevalence estimates. Appendix C.1 reports construction details and execution audits.

5.2 Disagreement-Enriched Mixtures Separate the Correction–Preservation Frontier

At low disagreement mass, the objectives occupy different operating regions but retain similar aggregate accuracy. From 0% through 10%, DPO corrects 32.6–43.7% of inherited errors at update 750, compared with 18.0–21.5% for CSFT. This correction advantage is accompanied by slightly more corruption: 0.9–1.1% for DPO versus 0.6–0.8% for CSFT. The opposing terms leave only a 0.1–0.4 percentage-point accuracy gap.

As the focal-pair share increases, the paths separate (Figure 2a). At 20% mass, DPO reaches 47.8% correction with 1.32% corruption, while CSFT reaches 25.6% correction with 1.73% corruption. At 40%, the correction rates converge (56.1% for DPO and 56.3% for CSFT), but their preservation costs do not: DPO corruption remains at 1.52%, whereas CSFT corruption rises to 6.29%. Increas-

Table 1: Controlled high-minus-low stress decomposition at update 750. Values are percentage-point changes in the DPO–CSFT gap and each method's gain over its shared SFT.

Setting	ΔGap	ΔDPO	ΔCSFT
Qwen / MS MARCO	$+2.62$	$+0.04$	-2.58
Gemma / MS MARCO	$+4.08$	$+0.14$	-3.94
Qwen / ESCI	$+4.02$	-0.14	-4.15
Gemma / ESCI	$+3.54$	-0.65	-4.19

ing the focal-pair share moves DPO mainly toward higher correction, while the high-mass CSFT path bends sharply toward higher corruption.

The checkpoint comparison shows how this separation develops. At update 75, DPO corrects 12.3–24.9 percentage points more inherited errors across the six masses, while also corrupting 0.6–0.9 points more initially correct examples. In the 40% arm, CSFT has nearly closed the correction gap by update 300, but now corrupts 7.0 points more than DPO. The relative role of DPO thus changes over training: its early distinction is additional correction, whereas its late high-mass distinction is reaching comparable correction with substantially less preservation loss.

5.3 The Relative Advantage Is Driven by CSFT Preservation Loss

Measuring both methods relative to their shared starting point reveals where the widening gap comes from (Figure 2b). Across all six masses, DPO remains only 0.1–0.3 percentage points above SFT at update 750, with no clear mass response. CSFT instead moves from no change at 0% to -0.9 points at 20% and -4.4 points at 40%. Defining the high-mass contrast as the mean of 20% and 40% minus the mean of 0% and 10%, the DPO–CSFT gap increases by 2.62 points (95% CI $[2.30, 2.93]$). Only 0.04 points (CI $[-0.13, 0.21]$) come from a larger DPO gain over SFT, whereas a 2.58-point CSFT decline (CI $[-2.88, -2.27]$) accounts for a point estimate of 98.5% of the widening gap.

The intervention does not make DPO progressively better than the starting policy. Instead, CSFT becomes increasingly costly as it moves toward comparable correction. At low disagreement mass, CSFT remains near a conservative low-change point. In the 40% mixture, the methods reach comparable correction, but DPO incurs substantially less preservation loss. Section 6 next tests whether the separation survives changes in model family, CSFT learning rate and duration, and data distribution.

6 Robustness and Scope of the Correction–Preservation Frontier

Section 5 establishes the controlled response in Qwen3-4B/MS MARCO. We now ask which parts replicate across models and datasets, which survive CSFT tuning, and which remain task-dependent.

6.1 The Stress Response Replicates Across Models and Datasets

The stress decomposition has the same direction in all four model–dataset cells (Table 1). From the mean low-mass arms to the mean high-mass arms, the DPO–CSFT gap widens by 2.62–4.08 points. CSFT's gain over the shared SFT falls by 2.58–4.19 points, while DPO's changes by only -0.65 to $+0.14$ points. The replicated result is therefore CSFT decline, not increasing DPO gain with mass. Appendix C.4 reports the intervals, arm definitions, and audits.

One additional independently trained Qwen/MS MARCO SFT start reproduces this pattern: the gap widens by 3.31 points, with CSFT's gain falling by 3.27 and DPO's changing by $+0.05$ (Appendix C.2).

The frontier shape remains task-dependent. At the strongest MS MARCO arms, the methods differ by at most 1.2 correction points, while CSFT corruption is 6.29–8.25% versus DPO's 1.52–1.59%. On ESCI at 35%, CSFT instead reaches higher correction than DPO in both models (45.4–46.8% versus 36.7–42.2%), but with much greater corruption (16.3–17.6% versus 6.1–8.0%). DPO occu-

pies the lower-corruption side where correction ranges overlap, but does not uniformly dominate or uniquely reach the highest correction. Natural-scale contrasts remain small and protocol-sensitive.

6.2 Tested CSFT Tuning Does Not Recover the High-Correction Frontier

On MS MARCO at 20% and 40% mass, we test whether the separation is a CSFT schedule artifact. The CSFT envelope combines learning rates $\{10^{-4}, 3 \times 10^{-5}, 10^{-5}\}$ and updates $\{75, 300, 750\}$ on the same frozen arms, starting SFT, and within-seed order. DPO retains its canonical 10^{-4} rate; this tests whether CSFT tuning closes the gap, not whether DPO is less tuning-sensitive.

For Qwen at 40% mass, lowering the rate reduces CSFT corruption from 6.29% to 0.52%, but also reduces correction from 56.3% to 12.2%. Gemma retains correction at lower rates, yet its corruption remains high. Reduced CSFT corruption therefore either accompanies much less correction or fails to recover DPO's preservation at comparable correction.

In the tested learning-rate grid, DPO has lower minimum corruption in every shared correction bin across all three presentation protocols. Yet CSFT supplies non-dominated low-correction, low-corruption points, so DPO does not uniformly dominate. Within this grid, CSFT does not recover DPO's high-correction, low-corruption region. Appendix C.5 gives the complete grid and frontier audit; Section 7.2 discusses untested alternatives.

A separate Qwen 40% check adds token-normalized, answer-position reverse-KL to CSFT. Within two correction points of DPO, corruption remains 5.38–7.36%, near vanilla CSFT's 6.29% and far from DPO's 1.52%. Stronger anchoring reaches low corruption only as correction falls to 14–15% (Table 15).

6.3 Across Tasks, Correction Demand and Preservation Differ

Controlled ESCI stress arms reproduce the CSFT-decline pattern, whereas natural ESCI favors DPO through additional correction despite greater corruption (Section 4.2). Natural disagreement rates count all disagreement and are not on the intervention's numerical axis. Across 11 ranking settings, train disagreement is nearly collinear with held-out SFT error rate ($r = .995$), so these data cannot separate disagreement from correction opportunity or other task differences (Appendix B.2).

TL;DR provides a more distant boundary condition. Across two models, three seeds, and two scoring rules, total train disagreement is 41.3–43.8%. DPO corrects 27.6–32.9 points more inherited pairwise errors than CSFT while corrupting 1.26–4.87 points more initially correct pairs. Most of the net gap comes from DPO improving over SFT by 8.9–11.3 points; CSFT degradation contributes only 6–20%, compared with approximately 97–99% of the high-minus-low gap widening in controlled MS MARCO. DPO can therefore be valuable through correction capacity without a preservation advantage (Appendix D.2).

DPO also scores better in a separate TL;DR generation check: across both models, three seeds, and two model judges, its tie-adjusted rates against SFT and CSFT range from $.638$ to $.747$ (Appendix F). The gains favor faithfulness and coverage but reduce conciseness and hallucination-avoidance scores. Model judges do not replace human evaluation, and this check does not link particular pairwise corrections to generation quality. Training disagreement therefore marks correction demand, while the inherited error pool determines how correction and preservation enter accuracy. Section 7.1 considers their joint use for assessing preservation requirements before further training.

7 Discussion

7.1 When Does DPO Add Value over Continued SFT?

In natural ESCI and TL;DR, DPO's advantage over CSFT comes mainly from additional correction. Under controlled stress, the widening gap instead comes mainly from CSFT's preservation loss, while DPO's gain over SFT remains roughly stable or declines. An advantage over CSFT therefore need not imply a large benefit from further training over retaining SFT.

To explore the conditions for useful gains over retaining SFT, we consider two signals available before Stage-2 training. Training disagreement indicates correction demand: an exploratory disagreement-only screen identifies whether DPO's gain over SFT reaches one accuracy point in 9/11 natural settings under leave-one-setting-out evaluation (Appendix G.1). With the SFT checkpoint and evaluation distribution fixed, increasing training disagreement does not enlarge the held-out SFT-error pool. A representative labeled calibration set could estimate this pool and the correct decisions to preserve. When errors are scarce, correction contributes little to overall accuracy, while losses from the larger correct pool can offset that contribution.

Reading these signals together makes the preservation requirement explicit. For an assumed correction rate and target gain over SFT, the decomposition gives the maximum affordable corruption. At 50% correction and a one-point gain, this budget is 0.56% for Qwen/MS MARCO but 9.19% for Gemma/ESCI (Appendix G.3). These retrospective scenarios suggest when high training disagreement should prompt closer attention to preservation. They use existing benchmark scores; a separate calibration-based diagnostic remains to be evaluated. The opportunity is to anticipate which settings require especially strong preservation, while forecasting actual correction and corruption remains open.

7.2 Scope, Limitations, and Open Questions

The controlled claim is deliberately narrow. The sweeps cover MS MARCO and ESCI, Qwen3-4B and Gemma-3-4B under LoRA, model-specific cohorts, and different mass grids. One additional independently trained SFT start replicates the response only in Qwen/MS MARCO; the other cells retain one canonical start. High-mass arms are stress tests, not natural prevalence estimates. The Gemma pipelines are not bit-exact and show run-to-run variation. We therefore replicate the CSFT-decline stress response, not a common frontier shape, maximum correction, onset, effect size, or threshold. Broader starting-state and task coverage remain open.

The optimization comparison is also bounded. The learning-rate and checkpoint grid covers MS MARCO, while the reverse-KL check covers only Qwen/MS MARCO at 40% mass and one regularizer formulation. DPO's learning rate and β remain fixed in the controlled sweeps. These results weaken explanations based solely on the canonical CSFT schedule or lack of explicit anchoring; they neither show that every CSFT variant is inferior nor isolate the contribution of rejected responses. Pareto claims use observed point estimates, and matched-correction comparisons use prespecified bins or a descriptive tolerance rather than exact equivalence.

Finally, disagreement and utility remain task-dependent. Natural disagreement covaries with SFT error rate, label structure, and dataset, and most preference pairs have one annotation. TL;DR is the only open-ended task; its generation evaluation uses model judges rather than humans, includes a same-family policy–judge pairing, and exhibits length and criterion trade-offs. Fixed-pair corrections therefore cannot be treated as causes of generation gains. Future validation should compare calibration-based preservation requirements with realized correction, corruption, and gains on a separate evaluation set after Stage-2 training.

8 Conclusion

Under controlled stress across two models and two datasets, DPO occupies lower-corruption regions over overlapping correction ranges. The widening accuracy gap comes mainly from CSFT decline, while DPO's gain over SFT remains roughly stable or declines. An additional Qwen/MS MARCO SFT start reproduces this response, and the tested MS MARCO CSFT schedule grid and one reverse-KL formulation do not recover DPO's high-correction, low-corruption region. Natural ESCI and TL;DR show a different source of value: additional DPO correction outweighs greater corruption.

Our exploratory analysis suggests that high training disagreement can coexist with few target-distribution errors to correct, making preservation especially important. Under assumed correction rates, SFT error mass determines the corruption budget compatible with a target gain. Estimating this requirement on a representative calibration set is a practical next step; predicting realized correction and corruption remains open.

AI USE STATEMENT

Generative AI tools assisted with literature search, code development, figure and table preparation, and manuscript editing. Quantitative claims draw on the experimental results and audits documented in the paper and appendices. Language models also served as judges in the TL;DR generation evaluation, as described in Appendix F. The authors take responsibility for the methods, interpretation, claims, and final manuscript.

ETHICS STATEMENT

This work uses existing benchmark datasets and publicly available model checkpoints for offline experiments. No new participants were recruited, no new individual-level data were collected, and no model was deployed in a high-stakes setting. TL;DR includes public Reddit posts that may contain personal narratives. We report aggregate analyses, do not attempt re-identification, and do not reproduce individual posts. Public availability does not remove privacy or consent concerns.

Benchmark-designated correctness and preference agreement do not establish factual truth, safety, or broader model quality. Generated-summary scores are model-judge assessments rather than human validation; the observed trade-offs include lower hallucination-avoidance scores despite higher total scores. The findings characterize correction and preservation in the studied settings and do not establish deployment readiness.

REPRODUCIBILITY STATEMENT

The source-code repository is available anonymously at `https://anonymous.4open.science/r/ICLR2027_sftdpo_sourcecode-F6C5/`. The main paper specifies tasks, models, objectives, training settings, metrics, and statistical units. Appendix A records shared protocols and checkpoint reuse. Appendices B and C document natural-setting results, exclusion sensitivity, controlled cohort construction, execution audits, the additional independent SFT start, and hyperparameter checks. Appendices D, E.1, and F specify fixed-pair and task-utility evaluations, including scoring sensitivities and model-judge protocols. Appendix G documents the exploratory screening rules and preservation-budget assumptions.

The independent-start replication covers Qwen/MS MARCO only; other controlled cells retain one canonical SFT start. Confidence intervals condition on the observed checkpoints, and the Gemma pipelines are not bit-exact.

REFERENCES

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Shihao Cai, Chongming Gao, Yang Zhang, Wentao Shi, Jizhi Zhang, Keqin Bao, Qifan Wang, and Fuli Feng. K-order ranking preference optimization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4844–4859. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.250.

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking DPO: The role of rejected responses in preference misalignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference*. National Institute of Standards and Technology, 2020.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference*. National Institute of Standards and Technology, 2021.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of DPO: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024. URL https://arxiv.org/abs/2404.04626.

Yuming Feng and Christy Yang. An empirical study of SFT-DPO interaction and parameterization in small language models. *arXiv preprint arXiv:2603.20100*, 2026. doi: 10.48550/arXiv.2603.20100. URL https://arxiv.org/abs/2603.20100.

Gemma Team, Aishwarya Kamath, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20983–21006. PMLR, 2024. URL https://proceedings.mlr.press/v235/im24a.html.

Chengtao Jian, Kai Yang, Tianhao Gao, Wuguang Ni, Keying Yang, Bowen Xiao, Jiajun Liu, and Ye Ouyang. Stable preference optimization: A bilevel approach to catastrophic preference shift. *arXiv preprint arXiv:2507.07723*, 2025.

Haowei Liu, Xuyang Wu, Guohao Sun, Zhiqiang Tao, and Yi Fang. ChainRank-DPO: Chain rank direct preference optimization for llm rankers. *arXiv preprint arXiv:2412.14405*, 2024.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2404–2420. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.121.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588*, 2022.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18990–18998, 2024. doi: 10.1609/aaai.v38i17.29865.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, and Sarath Chandar. A deep dive into the trade-offs of parameter-efficient preference alignment techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5732–5745. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.311. URL https://aclanthology.org/2024.acl-long.311/.

Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. Implicit reward as the bridge: A unified view of SFT and DPO connections. *arXiv preprint arXiv:2507.00018*, 2025. doi: 10.48550/arXiv.2507.00018. URL https://arxiv.org/abs/2507.00018.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0482.

Jiacong Zhou, Xianyun Wang, Min Zhang, and Jun Yu. Holistic utility preference learning for listwise alignment. *arXiv preprint arXiv:2410.18127*, 2025. doi: 10.48550/arXiv.2410.18127. URL https://arxiv.org/abs/2410.18127.

A Experimental Scope and Reporting Protocols

This section identifies the evidence retained for the paper's correction–preservation argument and states the conventions shared across analyses. Counts distinguish matched comparisons from trained checkpoints and repeated evaluations: evaluating one checkpoint under another scoring rule, benchmark, or decoding protocol does not constitute an independent training run.

A.1 Retained Evidence and Analysis Inventory

Table 2 summarizes the analysis units that support Sections 4–6. A setting fixes a model and dataset, while a matched comparison additionally fixes the training seed and inherited SFT checkpoint. Rows are not additive because later evaluations reuse checkpoints from earlier training comparisons.

Appendix C.2 adds one independently trained Qwen/MS MARCO SFT start and 12 matched Stage-2 comparisons (24 method runs). Appendix G reuses 11 natural settings and 18 controlled arms for exploratory screening and preservation-requirement analyses. Its additional Gemma-3-12B checks compare DPO with SFT: MS MARCO uses one SFT start and three Stage-2 seeds, whereas ESCI uses three independently trained SFT starts, each followed by one DPO run. Repeated analysis of existing runs is not counted as new training.

A.2 Shared Training and Evaluation Protocol

Unless a task-specific subsection states otherwise, each comparison follows the shared Base $\rightarrow$ SFT $\rightarrow \{\text{CSFT}, \text{DPO}\}$ design from Section 3.1. Within a model–dataset–seed comparison, CSFT and DPO inherit the same SFT checkpoint and receive the same ordered Stage-2 examples. Stage-1, Stage-2, and evaluation inputs are disjoint at the query or post level. CSFT optimizes the preferred output only, whereas DPO uses the same preferred output together with its rejected counterpart and a frozen copy of the inherited SFT policy.

The shared training configuration uses LoRA rank 16, $\alpha = 32$, dropout .05, attention and MLP projection adapters, three epochs, learning rate 10^{-4}, effective batch size 8, and bfloat16 precision. The primary DPO value is $\beta = .1$. Seeds 7, 42, and 123 vary data order and dropout while preserving the paired comparison. Deliberate deviations—the controlled CSFT learning-rate grid, KL-CSFT robustness grid, DPO β sensitivity, and task-specific sequence lengths—are reported with the corresponding experiment.

All decision analyses recompute margins from the named checkpoint and fix the inherited-error and inherited-correct pools using its corresponding SFT policy. A zero margin is treated as target-inconsistent. When an evaluation uses both candidate orders, the two oriented margins are combined

Table 2: Core analysis inventory. Counts describe the comparison or evaluation unit named in the middle column; shared checkpoints, examples, posts, and queries are noted explicitly.

Evidence	Unit and scope	Dependence and reuse
Natural ranking	13 retained model–dataset settings; 39 seed-level comparisons	CSFT and DPO share the inherited SFT checkpoint and 10,000 held-out pairs within each comparison. The excluded three-seed setting is reported separately.
MS MARCO intervention	2 models; 10 model–mass cells; 30 seed–mass comparisons	Qwen contributes 6 masses and Gemma 4. Each comparison contains matched CSFT/DPO trajectories on the same 2,000 Stage-2 pairs; three checkpoints reuse each trajectory.
ESCI intervention	2 models; 8 model–mass cells; 24 seed–mass comparisons	Each model contributes 4 masses. CSFT and DPO share examples and order within a cell, but focal pairs and anchors are reconstructed for each model; three checkpoints reuse each trajectory.
CSFT tuning	2 models, 2 MS MARCO high-mass arms, and 3 seeds; 12 model–mass–seed cells	Three CSFT learning rates and three checkpoints define the tested envelope. Canonical CSFT and DPO trajectories are reused.
KL-CSFT robustness	1 Qwen/MS MARCO 40% cell; 5 strengths and 3 seeds; 15 trajectories	Seven checkpoints reuse each KL-CSFT trajectory. Vanilla CSFT and DPO are reused; the $\lambda_{\mathrm{TN}} = 0$ fidelity run is a gate, not an additional comparison.
TL;DR fixed pairs	2 policy models and 3 seeds on 1,500 shared posts	SFT, CSFT, and DPO checkpoints are evaluated under sequence-sum and token-average scoring; the scoring rules do not create new training runs.
TL;DR generation	18 checkpoints and 9,000 summaries from 500 shared posts	Two model judges score the same generated outputs; judge assessments are reported separately rather than treated as independent task replications.
TREC ranking	9 checkpoints on 43 DL19 and 54 DL20 queries	Methods share BM25 candidates within each benchmark; temperature and numerical audits reuse the same checkpoints and query sets.

before assigning decision status; results from a single orientation are not substituted for the symmetrized result. Canonical endpoints, intermediate checkpoints, and fidelity reruns remain distinct unless an analysis explicitly identifies them as the same artifact.

A.3 Statistical Units, Dependence, and Uncertainty

Multi-seed summaries report the arithmetic mean and sample standard deviation unless stated otherwise. Uncertainty calculations retain the pairing induced by shared inputs and use the task's inference unit: examples for fixed-pair analyses, queries for TREC, and posts for free generation. Aggregates spanning seeds or settings preserve the stated grouping rather than treating all evaluation rows as independent observations. Their intervals therefore condition on the trained checkpoints that were actually observed; three training seeds do not define a broader population of possible training runs.

Repeated scoring rules, judges, checkpoints along a trajectory, and benchmark evaluations of one checkpoint are dependent views of shared artifacts. They are reported as sensitivity analyses or repeated measurements, not additional replications. Multiple-testing corrections are defined within the family stated by each analysis. Pareto comparisons use observed correction and corruption point estimates over the tested checkpoint and hyperparameter grid; shared correction bins summarize nearby operating regions without interpolating unobserved points or establishing population-level dominance.

B Natural-Distribution Results and Sensitivity

This section provides the setting-level results behind Section 4, reports the post-hoc exclusion transparently, and tests whether the natural-distribution comparison is sensitive to the primary DPO β.

All rates and contributions are recomputed from the saved per-example margins; no row is inferred from aggregate accuracy.

B.1 Paired Outcomes across Retained Ranking Settings

The audit covers 39 model–dataset–seed comparisons. Within every comparison, SFT, CSFT, and DPO have the same 10,000 stable IDs in the same order, with no duplicate, missing, or non-finite margins. Directly counted status disagreements reproduce the paired-cell identity, and the maximum residual between the observed accuracy gap and the weighted correction–preservation decomposition is 1.3×10^{-16}.

Table 3 reports the conditional differences that underlie the aggregate gap. DPO corrects more inherited errors than CSFT in all 39 retained runs. Across the 13 settings, the hierarchical mean correction-rate difference is 21.92 percentage points (95% CI $[19.87, 24.16]$); the mean conditional preservation cost is 1.01 points (CI $[0.50, 1.53]$). These comparatively small conditional preservation differences still matter because the inherited-correct pool is substantially larger than the inherited-error pool.

Table 3: Paired natural-distribution outcomes. Values are three-seed mean±sample SD in percent or percentage points. Err. is the inherited SFT-error rate; Gap is DPO minus CSFT accuracy; Dis. is their decision-status disagreement; Δc and Δp are the DPO–CSFT correction and preservation differences. M/E denote MS MARCO/ESCI.

Setting	Err.	Gap	Dis.	Δc	Δp
Qwen3-4B / M	3.53 ± 0.06	0.38 ± 0.21	2.81 ± 0.29	26.1 ± 4.6	-0.56 ± 0.15
Qwen3-4B / E	17.54 ± 0.28	1.31 ± 0.25	10.78 ± 0.21	16.4 ± 0.8	-1.90 ± 0.50
Qwen3-1.7B / M	4.45 ± 0.03	0.92 ± 0.15	3.34 ± 0.11	22.0 ± 2.5	-0.06 ± 0.05
Qwen3-1.7B / E	20.97 ± 0.11	2.25 ± 0.39	12.44 ± 0.46	18.4 ± 1.3	-2.03 ± 0.17
Qwen3-8B / M	3.47 ± 0.04	0.82 ± 0.14	2.56 ± 0.19	29.1 ± 2.9	-0.20 ± 0.24
Qwen3-8B / E	17.06 ± 0.22	1.44 ± 0.51	10.34 ± 0.30	16.6 ± 2.7	-1.67 ± 0.18
Gemma-3-4B / M	4.40 ± 0.11	1.11 ± 0.28	3.45 ± 0.11	23.7 ± 2.1	0.07 ± 0.22
Gemma-3-4B / E	18.55 ± 0.23	2.60 ± 0.01	12.92 ± 0.15	22.3 ± 2.0	-1.88 ± 0.51
Llama-3.2-3B / M	3.87 ± 0.16	0.69 ± 0.23	3.22 ± 0.23	23.6 ± 1.4	-0.24 ± 0.26
Llama-3.2-3B / E	18.07 ± 0.10	1.90 ± 0.43	11.77 ± 0.21	20.7 ± 0.2	-2.25 ± 0.47
Llama-3.1-8B / M	3.38 ± 0.12	0.42 ± 0.28	2.85 ± 0.21	24.0 ± 4.5	-0.41 ± 0.12
Llama-3.1-8B / E	16.36 ± 0.22	1.48 ± 0.18	11.18 ± 0.48	20.4 ± 1.2	-2.22 ± 0.09
Gemma-3-1B / M	8.80 ± 0.22	2.08 ± 0.13	5.41 ± 0.32	21.5 ± 1.0	0.21 ± 0.11

Table 4 restores the inherited-pool weights. The correction contribution is positive in every setting, whereas the preservation contribution is negative in 11 of 13 setting means. On MS MARCO, the preservation term is small and its hierarchical interval includes zero; on ESCI, it is consistently negative and offsets approximately 48% of the correction gain. The positive net gap therefore has the same sign across retained settings but arises from different balances of the two terms.

Table 4: Accuracy decomposition in percentage points. Setting rows are three-seed mean±sample SD. Aggregate rows are hierarchical mean [95% CI] over the stated settings. The two contributions sum to the DPO–CSFT accuracy gap.

Setting	Correction contribution	Preservation contribution	Sum
Qwen3-4B / M	0.92 ± 0.17	-0.54 ± 0.14	0.38 ± 0.21
Qwen3-4B / E	2.88 ± 0.18	-1.57 ± 0.41	1.31 ± 0.25
Qwen3-1.7B / M	0.98 ± 0.11	-0.06 ± 0.05	0.92 ± 0.15
Qwen3-1.7B / E	3.85 ± 0.29	-1.61 ± 0.13	2.25 ± 0.39
Qwen3-8B / M	1.01 ± 0.11	-0.19 ± 0.23	0.82 ± 0.14
Qwen3-8B / E	2.82 ± 0.43	-1.39 ± 0.14	1.44 ± 0.51
Gemma-3-4B / M	1.04 ± 0.11	0.07 ± 0.21	1.11 ± 0.28
Gemma-3-4B / E	4.14 ± 0.41	-1.53 ± 0.41	2.60 ± 0.01
Llama-3.2-3B / M	0.91 ± 0.02	-0.23 ± 0.25	0.69 ± 0.23
Llama-3.2-3B / E	3.74 ± 0.04	-1.84 ± 0.39	1.90 ± 0.43
Llama-3.1-8B / M	0.82 ± 0.18	-0.40 ± 0.12	0.42 ± 0.28
Llama-3.1-8B / E	3.34 ± 0.24	-1.86 ± 0.08	1.48 ± 0.18
Gemma-3-1B / M	1.89 ± 0.05	0.19 ± 0.10	2.08 ± 0.13
MS MARCO (7)	$1.08\,[0.88, 1.38]$	$-0.16\,[-0.36, 0.03]$	$0.92\,[0.56, 1.36]$
ESCI (6)	$3.46\,[3.04, 3.88]$	$-1.63\,[-1.84, -1.43]$	$1.83\,[1.42, 2.27]$
All (13)	$2.18\,[1.51, 2.87]$	$-0.84\,[-1.26, -0.42]$	$1.34\,[0.97, 1.73]$

B.2 Natural Train-Disagreement Audit

The train-side audit scores each setting's canonical 2,000 Stage-2 pairs with the corresponding canonical SFT policy and counts a disagreement when the preference-directed margin is negative. Exact stored margins for six runs were reproduced bit for bit before the remaining margins were regenerated. For all 33 accepted runs, an additional identity check against the first 320 stored held-out margins had maximum absolute error at most 4.8×10^{-7}, with no sign or zero-margin disagreements.

Train-side identity was established for 11 of the 13 retained natural settings. The two omitted settings are Llama-3.2-3B on MS MARCO and ESCI. The local base model used by the canonical runs was no longer available, and a newly downloaded checkpoint did not reproduce the stored held-out margins, even under the recorded prompt-template variants. We therefore exclude those six runs rather than substitute non-identical train margins. This artifact exclusion is separate from the outcome-based Gemma-3-1B/ESCI exclusion in Appendix B.3: the former affects only this train-side audit, whereas the latter defines the 13-setting primary outcome aggregate.

Table 5: Natural train disagreement and held-out correction opportunity. Values are three-seed setting means in percent. Train dis. is the fraction of the 2,000 Stage-2 pairs with a negative preference-directed SFT margin; held-out err. is the inherited SFT-error rate on the disjoint evaluation set. M/E denote MS MARCO/ESCI.

Setting	Train dis.	Held-out err.
Qwen3-8B / M	3.28	3.47
Llama-3.1-8B / M	3.53	3.38
Qwen3-4B / M	3.78	3.53
Gemma-3-4B / M	4.23	4.40
Qwen3-1.7B / M	4.62	4.45
Gemma-3-1B / M	10.13	8.80
Llama-3.1-8B / E	14.67	16.36
Qwen3-8B / E	15.03	17.06
Qwen3-4B / E	15.65	17.54
Gemma-3-4B / E	16.52	18.55
Qwen3-1.7B / E	18.37	20.97

Across these 11 setting means, train disagreement and the held-out SFT-error rate are nearly collinear (Pearson $r = .995$, Spearman $\rho = .991$); their maximum absolute difference is 2.6 percentage points. At the run level, the DPO–CSFT accuracy gap and correction contribution are positive in all 33 runs. DPO corruption exceeds CSFT corruption in 29 of 33 runs, including all 15 ESCI runs, and the preservation contribution is positive in only 4 of 33 runs, all on MS MARCO. For the five models with both datasets, ESCI exceeds MS MARCO in train disagreement by 11.1–13.8 points, but dataset, inherited-error mass, and disagreement move together. These observations therefore support a task-dependent correction–preservation difference, not a pooled disagreement threshold or a causal effect of natural disagreement mass.

B.3 Post-hoc Exclusion and Inclusion Sensitivity

Gemma-3-1B/ESCI was excluded after inspecting its three-seed stability; no numerical threshold was prespecified. The reason is setting-level heterogeneity rather than a uniform Stage-1 failure. Seed 123 has below-chance SFT accuracy, a zero median margin, and 2,956 zero-margin examples. Seed 42 begins from an SFT accuracy similar to seed 7 but its Stage-2 DPO endpoint falls to 50.58% accuracy and preserves only 60.57% of inherited-correct decisions. Seed 7 does not exhibit either failure mode but is excluded with the setting rather than selected individually.

Including this setting changes the universal retained-run statement: DPO corrects more than CSFT in 41 of 42 rather than 39 of 39 runs. The average correction–preservation sign pattern remains, but uncertainty grows materially (Table 7). In particular, the 14-setting net gap is (0.80) points (95% CI $[-0.99, 1.75]$), and the seven-setting ESCI net gap is (0.69) points (CI $[-2.94, 2.36]$); neither interval excludes zero. The exclusion therefore does not create the correction advantage, but it does delimit where the positive net-gain claim is statistically supported.

Table 6: Excluded Gemma-3-1B/ESCI runs. Acc. and Corr./Pres. are percentages; C/P contributions and Gap are percentage points. C/D denotes CSFT/DPO. The exclusion was post hoc and used no prespecified threshold.

Seed	SFT Acc.	Acc. C/D	Corr. C/D	Pres. C/D	C contr.	P contr.	Gap
7	61.64	69.40/78.14	39.31/64.89	88.12/86.39	$+9.81$	-1.07	$+8.74$
42	61.48	70.77/50.58	38.86/34.63	90.76/60.57	-1.63	-18.56	-20.19
123	38.82	57.35/50.31	49.46/50.28	69.78/50.36	$+0.50$	-7.54	-7.04

Table 7: Hierarchical means [95% CIs] in percentage points before and after including Gemma-3-1B/ESCI. Preservation cost is $-\Delta p$; preservation contribution retains its signed contribution to accuracy.

Quantity	13 retained settings	14 settings
Correction-rate difference	$21.92\,[19.87, 24.16]$	$20.88\,[17.68, 23.53]$
Conditional preservation cost	$1.01\,[0.50, 1.53]$	$2.16\,[0.63, 5.04]$
Correction contribution	$2.18\,[1.51, 2.87]$	$2.23\,[1.50, 3.04]$
Preservation contribution	$-0.84\,[-1.26, -0.42]$	$-1.43\,[-3.00, -0.53]$
Net accuracy gap	$1.34\,[0.97, 1.73]$	$0.80\,[-0.99, 1.75]$

B.4 Sensitivity to the DPO β Parameter

We vary $\beta \in \{0.05, 0.1, 0.2\}$ for Qwen3-4B/MS MARCO while holding the inherited SFT checkpoint, Stage-2 data and order, optimizer, learning rate, duration, and held-out examples fixed. Each DPO run is compared with the same canonical CSFT checkpoint for its seed.

Table 8: DPO β sensitivity on Qwen3-4B/MS MARCO. Values are three-seed mean$\pm$sample SD in percent or percentage points. Contributions are weighted by the inherited SFT-error and SFT-correct pool sizes.

Quantity	$\beta = .05$	$\beta = .1$	$\beta = .2$
DPO accuracy	97.04 ± 0.05	97.00 ± 0.22	97.07 ± 0.03
DPO correction	48.77 ± 0.80	48.08 ± 4.05	48.28 ± 2.85
DPO preservation	98.81 ± 0.08	98.79 ± 0.20	98.86 ± 0.06
DPO–CSFT accuracy gap	0.43 ± 0.10	0.38 ± 0.21	0.46 ± 0.08
Correction-rate difference	26.79 ± 0.66	26.11 ± 4.65	26.31 ± 1.49
Preservation-rate difference	-0.54 ± 0.08	-0.56 ± 0.15	-0.49 ± 0.03
Correction contribution	0.95 ± 0.02	0.92 ± 0.17	0.93 ± 0.07
Preservation contribution	-0.52 ± 0.08	-0.54 ± 0.14	-0.47 ± 0.03

Across this fourfold β range, the correction contribution remains approximately $+0.9$ points, the preservation contribution approximately -0.5 points, and the net gap approximately $+0.4$ points. The largest gap difference across β is 0.08 points, smaller than the seed SD at $\beta = .1$. Thus, this setting shows no resolvable β dependence in the correction–preservation decomposition with three seeds; it does not establish β-invariance beyond the tested model, task, and range.

C Controlled Disagreement: Full Results and Robustness

This section provides the construction audits and complete observed grids behind Sections 5–6.2. The intervention varies the share of intermediate policy–preference disagreements in a fixed 2,000-pair Stage-2 set. It does not identify disagreement sign as the only causal difference between focal and anchor examples, and it does not estimate a deployment threshold. MS MARCO supplies the primary intervention and CSFT tuning grid; ESCI completes a qualitative two-model, two-dataset replication.

C.1 Intervention Construction and Integrity Checks

For Qwen, the focal master contains 800 intermediate-disagreement pairs, balanced between the two chosen positions. The 2, 5, 10, and 20% focal sets are nested prefixes of this master after stratification by margin and length tertiles; the 40% arm contains all 800 focal pairs. Removing focal pairs exposes matched anchors in the same slots and chosen position. Every arm therefore contains 2,000 unique queries, 1,000 examples in each chosen position, the same 1,200 common anchors, and the requested focal count. The reconstructed 40% arm is byte-identical to the earlier frozen intervention set. Across focal subsets, the largest absolute standardized mean difference from the focal master is .096; between all anchors and the focal master it is .107.

Gemma repeats this procedure with margins from its own inherited SFT policy and with no use of Qwen cohort membership. Its first focal-selection rule exceeded the declared .20 balance cap on rejected-response length (absolute SMD .243), so no arm was written and no model was trained from that attempt. A second, covariate-only pooled-centroid rule was declared and then applied, yielding 800 balanced focal pairs. The resulting 0, 10, 20, and 40% arms have the same size, position balance, nested construction, and common-anchor logic as the Qwen arms; the maximum focal-subset imbalance is .073 and the maximum extra-versus-common anchor imbalance is .052. Because the cohort rules and model-specific margins differ, the cross-model comparison is a protocol replication rather than a one-variable model swap.

The ESCI replications use a common pre-scoring candidate universe but reconstruct focal membership from each model's SFT margins. Exact ties are excluded from the focal pool, equal-margin blocks are kept intact, and the 5, 20, and 35% arms use nested subsets drawn from a 700-pair intermediate-disagreement master, with matched correct anchors. Both models pass the frozen-arm, balance, duplication, and query-leakage checks. Their 35% focal sets overlap on only 137 of 700 IDs (Jaccard .109), and their three-way disagreement-status agreement is moderate (Cohen's $\kappa = .48$). Nominal mass therefore denotes the same construction rule, not the same examples or a shared margin scale.

Table 9: Controlled-intervention construction and execution audit. Runs count matched Stage-2 method trainings; Qwen/MS MARCO 40% runs were reused from the frozen precursor experiment.

Setting	Masses (%)	Arms	CSFT/DPO runs	Scores	Failures
Qwen / MS MARCO	0/2/5/10/20/40	$6 \times 2{,}000$	18/18	216	0
Gemma / MS MARCO	0/10/20/40	$4 \times 2{,}000$	12/12	144	0
Qwen / ESCI	0/5/20/35	$4 \times 2{,}000$	12/12	144	0
Gemma / ESCI	0/5/20/35	$4 \times 2{,}000$	12/12	144	0

Within each setting and seed, CSFT and DPO use the same frozen arm and epoch order; all runs start from the named SFT adapter, complete 750 updates over three epochs, expose each focal example exactly three times, and contain no non-finite updates. All original- and swapped-order score files contain the same 10,000 ordered held-out IDs, and the SFT-relative accuracy identity has maximum residual 1.0×10^{-16}. The "Scores" column counts checkpoint–presentation evaluations: three checkpoints and two presentations for every trained method run.

The Qwen seed-42 fidelity rerun reproduced both method adapters exactly at update 75 on MS MARCO. Gemma/MS MARCO did not reproduce the stored endpoint bit-for-bit. Its controlled rerun disagreed with the stored canonical DPO on 216 of 10,000 held-out signs; an independent canonical-script rerun disagreed on 214. The acceptance criterion was set after the exact-reproduction failure, not prespecified. We therefore treat the Gemma experiment as a three-run protocol replication and do not interpret a single Gemma endpoint as exact reproduction of the earlier adapter.

On ESCI, Qwen's DPO fidelity run is bit-exact; its CSFT rerun differs from the canonical trainer because the intervention deliberately matches CSFT and DPO example order. Gemma/ESCI score identity is exact, but GPU reruns are not bit-exact even within the controlled DPO pipeline. The prespecified relative-L2 criterion passes, all 24 new method runs match order and exposure, and no run is excluded. These checks support qualitative replication while cautioning against interpreting small endpoint differences as deterministic model effects.

C.2 Replication from an Independently Trained SFT Start

We train one additional Qwen3-4B/MS MARCO SFT adapter using the canonical recipe with an explicit initialization seed of 20260922; the canonical run did not explicitly seed LoRA initialization. Its symmetrized held-out accuracy is 96.74%, versus 96.92% for the canonical start. Cohorts are reconstructed from the new SFT margins: the 800 focal pairs and 1,200 common anchors have zero within-role ID overlap with their canonical counterparts. Thus this is a replication with a new starting state and reconstructed cohorts, not an isolated checkpoint swap on identical training examples.

The $0/10/20/40\%$ arms each contain 2,000 pairs. CSFT and DPO share arm files and within-seed example order for Stage-2 seeds 7, 42, and 123, yielding 24 method runs. All complete 750 updates over three epochs with no non-finite steps; the reported audit confirms frozen-arm hashes, matched order, and three exposures per focal example. The primary contrast is the mean of the 20/40% arms minus the mean of the 0/10% arms at update 750 under symmetrized scoring. Table 10 gives three-seed means and 95% intervals from 10,000 joint example-bootstrap draws. These intervals condition on the observed checkpoints; three Stage-2 seeds are not three independently trained SFT starts.

Table 10: High-minus-low stress contrasts for the additional Qwen/MS MARCO SFT start (percentage points). Each method's gain is measured against this new SFT.

Quantity	Change	95% interval
DPO–CSFT accuracy gap	$+3.31$	$[+2.92, +3.71]$
CSFT gain over SFT	-3.27	$[-3.67, -2.87]$
DPO gain over SFT	$+0.05$	$[-0.11, +0.19]$

The gap widens and CSFT's gain declines in all three Stage-2 seeds; both directions also hold under original- and swapped-order scoring. The canonical and new starts therefore agree on the CSFT-decline stress response, while their gap changes differ ($+2.62$ versus $+3.31$ points). We retain separate within-start contrasts rather than pool them. The small DPO contrast does not establish a zero effect, and this single additional start does not establish starting-state invariance or a common frontier or threshold.

C.3 Full MS MARCO Mass-Sweep Trajectories across Models

Table 11 reports every prespecified checkpoint in the primary position-symmetrized evaluation. Qwen's low-mass arms remain in distinct operating regions: through 10%, DPO has substantially higher correction and only slightly higher corruption. At 20%, DPO has both higher correction and lower corruption; at 40%, correction converges while CSFT corruption rises sharply. Gemma reaches the same qualitative separation earlier in the tested grid. At 20%, its DPO correction is 7.7 points higher and corruption 2.9 points lower; at 40%, correction differs by only 1.2 points while corruption differs by 6.7 points in DPO's favor.

Table 11: Complete MS MARCO controlled mass trajectories. Each cell is three-seed mean correction/corruption (%) on the fixed held-out set under position-symmetrized scoring.

		CSFT update			DPO update		
Model	Mass	75	300	750	75	300	750
Qwen	0%	16.9/0.56	18.0/0.48	18.0/0.57	29.2/1.24	32.8/1.11	32.6/0.92
	2%	15.7/0.48	18.7/0.51	20.3/0.65	32.7/1.30	37.6/1.07	38.6/0.93
	5%	16.7/0.55	19.9/0.63	19.5/0.64	34.6/1.20	41.6/1.03	39.7/1.11
	10%	16.3/0.61	21.1/0.72	21.5/0.78	39.3/1.19	45.0/1.21	43.7/1.12
	20%	18.4/0.71	19.9/0.85	25.6/1.73	41.1/1.32	48.6/1.21	47.8/1.32
	40%	18.3/0.88	54.4/8.39	56.3/6.29	43.2/1.76	55.5/1.43	56.1/1.52
Gemma	0%	18.6/0.62	21.9/0.73	22.5/0.84	39.1/2.07	41.8/1.34	43.8/1.18
	10%	18.7/0.74	24.6/1.15	24.5/1.23	37.4/2.06	46.1/1.34	46.7/1.36
	20%	18.3/1.09	35.7/3.73	43.1/4.30	44.4/2.73	54.4/1.89	50.8/1.41
	40%	20.4/2.09	53.1/7.48	58.9/8.25	47.0/2.59	57.9/1.87	57.7/1.59

The endpoint accuracy decomposition confirms that the widening gap is not a growing DPO improvement over the shared starting policy. Comparing the mean of 20/40% with the mean of 0/10%,

the DPO–CSFT gap widens by 2.62 points for Qwen (95% CI [2.30, 2.93]) and 4.08 points for Gemma ([3.66, 4.51]). Changes in DPO's gain over SFT are 0.04 points ([-0.13, 0.21]) and 0.14 points ([-0.03, 0.31]), whereas CSFT declines by 2.58 points ([-2.88, -2.27]) and 3.94 points ([-4.37, -3.52]), respectively. The corresponding point-estimate shares of gap widening are 98.5% and 96.6%; because these are ratios of estimated contrasts, they are descriptive rather than bounded proportions.

Presentation order does not explain the high-mass result. At Qwen 20 and 40%, the endpoint DPO–CSFT accuracy gap is positive for every seed under original, swapped, and symmetrized scoring. The original and swapped Gemma high-versus-low contrasts also have the same directions as the symmetrized contrast. Smaller Qwen effects at 0–10% vary more across seeds and presentation protocols, which is why the paper does not infer an onset between adjacent grid points.

C.4 Controlled ESCI Replication and the Two-by-Two Comparison

Gemma/ESCI completes the controlled comparison across Qwen and Gemma on MS MARCO and ESCI. Table 12 reports the ESCI endpoints for both models. The 0–5% arms retain the familiar split between conservative CSFT and more corrective DPO. At 35%, CSFT reaches higher correction than DPO in both models, but its corruption more than doubles. The resulting accuracy gap is therefore not evidence that DPO reaches a correction level unavailable to CSFT; it reflects a lower preservation cost in the range where the methods overlap.

Table 12: ESCI endpoints at update 750 (three-seed, position-symmetrized means). C/R is correction/corruption in percent; gains and PairGain are percentage points relative to the shared SFT or between methods.

Mass	CSFT C/R	DPO C/R	PairGain	CSFT–SFT	DPO–SFT
Qwen3-4B					
0%	15.70/2.87	28.30/4.36	+0.82	+0.16	+0.97
5%	18.10/3.10	31.00/4.46	+0.97	+0.35	+1.32
20%	16.90/5.46	35.80/5.64	+2.92	-1.82	+1.10
35%	46.80/16.26	36.70/6.05	+6.90	-5.98	+0.91
Gemma-3-4B					
0%	17.35/3.01	31.34/4.67	+1.04	+0.49	+1.53
5%	18.48/3.25	33.04/4.59	+1.40	+0.49	+1.89
20%	21.85/5.35	37.63/6.04	+2.15	-0.67	+1.48
35%	45.38/17.55	42.15/7.98	+7.36	-6.72	+0.64

The prespecified stress contrast compares the mean of the two high masses with the mean of the two low masses in each setting. As Table 13 shows, the DPO–CSFT gap widens in all four model–dataset cells, primarily because CSFT falls relative to the common SFT start. DPO's gain is approximately stable on MS MARCO and Qwen/ESCI, but declines by .65 points on Gemma/ESCI; the latter remains much smaller than the 4.19-point CSFT decline. Because mass grids and model-specific cohorts differ, the four rows are qualitative replications rather than a pooled factorial estimate.

Table 13: Controlled two-by-two stress decomposition at update 750. Entries are changes in percentage points [95% joint-bootstrap CI] from the mean low-mass arms to the mean high-mass arms. ΔGap is the change in DPO minus CSFT accuracy; ΔDPO and ΔCSFT are changes in each method's gain over SFT.

Setting	ΔGap	ΔDPO	ΔCSFT
Qwen / MS MARCO	$+2.62\,[2.30, 2.93]$	$+0.04\,[-0.13, 0.21]$	$-2.58\,[-2.88, -2.27]$
Gemma / MS MARCO	$+4.08\,[3.66, 4.51]$	$+0.14\,[-0.03, 0.31]$	$-3.94\,[-4.37, -3.52]$
Qwen / ESCI	$+4.02\,[3.46, 4.58]$	$-0.14\,[-0.45, 0.18]$	$-4.15\,[-4.65, -3.66]$
Gemma / ESCI	$+3.54\,[2.98, 4.09]$	$-0.65\,[-0.97, -0.33]$	$-4.19\,[-4.69, -3.67]$

The ESCI frontier has the same bounded interpretation in both models. In their overlapping correction ranges, the minimum observed corruption is 4.4% for DPO versus 11.8% for CSFT under Qwen, and 4.59% versus 11.18% under Gemma. Yet at 35%, CSFT reaches the highest correction point: 46.8% versus DPO's 36.7% for Qwen and 45.4% versus 42.1% for Gemma, at corruption costs of 16.3% versus 6.1% and 17.6% versus 8.0%, respectively. Thus the replication supports a

lower-corruption DPO frontier where correction overlaps, not uniform DPO dominance or exclusive access to high correction.

Tie handling does not reverse the stress result. Qwen's 20% CSFT corruption is partly composed of exact-zero margins, and the corresponding share is also material for Gemma, but strict non-tie and half-credit sensitivities preserve the high-versus-low gap direction. The natural-scale 5–0% contrast remains small: Qwen's interval includes zero, while Gemma's primary interval barely excludes it and becomes inconclusive under some presentation and tie conventions. We therefore generalize the stress response, not a natural-scale benefit, onset, or threshold.

C.5 CSFT Tuning and Frontier Audit

The tuning audit adds CSFT learning rates 3×10^{-5} and 10^{-5} to the canonical 10^{-4} trajectory at masses 20 and 40% for both models. The frozen arms, starting SFT, seed-specific order, three checkpoints, and held-out set are unchanged. DPO is reused at 10^{-4} and $\beta = .1$; the experiment therefore tests whether CSFT tuning reaches DPO's observed region, not whether DPO is less tuning-sensitive.

Table 14: Learning-rate and duration grid. Cells are three-seed mean correction/corruption (%) under position-symmetrized scoring. DPO uses learning rate 10^{-4}.

Setting	Method	Update 75	Update 300	Update 750
Qwen 20%	CSFT 10^{-4}	18.4/0.71	19.9/0.85	25.6/1.73
	CSFT 3×10^{-5}	12.3/0.47	15.0/0.47	15.4/0.48
	CSFT 10^{-5}	7.8/0.33	12.9/0.44	13.1/0.44
	DPO	41.1/1.32	48.6/1.21	47.8/1.32
Qwen 40%	CSFT 10^{-4}	18.3/0.88	54.4/8.39	56.3/6.29
	CSFT 3×10^{-5}	14.4/0.64	14.0/0.65	26.3/2.76
	CSFT 10^{-5}	10.8/0.42	13.4/0.63	12.2/0.52
	DPO	43.2/1.76	55.5/1.43	56.1/1.52
Gemma 20%	CSFT 10^{-4}	18.3/1.09	35.7/3.73	43.1/4.30
	CSFT 3×10^{-5}	12.7/0.58	33.3/3.21	43.0/4.04
	CSFT 10^{-5}	7.5/0.33	12.1/0.52	11.3/0.66
	DPO	44.4/2.73	54.4/1.89	50.8/1.41
Gemma 40%	CSFT 10^{-4}	20.4/2.09	53.1/7.48	58.9/8.25
	CSFT 3×10^{-5}	16.0/0.92	56.9/7.31	60.7/7.13
	CSFT 10^{-5}	9.2/0.52	31.8/3.86	59.7/6.79
	DPO	47.0/2.59	57.9/1.87	57.7/1.59

Lower learning rates do not produce one uniform CSFT response. For Qwen they mostly retain a low-correction, low-corruption policy; for Gemma at 40%, they eventually recover high correction but retain 6.79–7.13% corruption. This distinction prevents a simple claim that the canonical CSFT rate alone causes the separation.

Explicit KL robustness. A separate Qwen/MS MARCO 40% experiment adds reverse KL to the chosen-only loss at the answer position, with its coefficient normalized by the mean supervised token count. The prespecified grid uses token-normalized strengths $\lambda_{\mathrm{TN}} \in \{.03, .1, .3, 1, 3\}$, three seeds, and seven checkpoints through update 750. Table 15 shows the primary endpoints and the point selected by the prespecified matched-correction rule. Across five observed points within two correction points of DPO, corruption is 5.38–7.36%, compared with 6.29% for vanilla CSFT and 1.52% for DPO. Stronger anchoring lowers corruption but holds correction near 14–15%. Thus, within this formulation and grid, explicit KL strength alone does not reproduce DPO's high-correction, low-corruption point. This check does not identify the causal contribution of the rejected response or rule out other regularizers.

For each model, mass, seed (and the three-seed mean), and presentation protocol, we form the observed frontier over nine CSFT points and three DPO points. No DPO point is Pareto-dominated by a CSFT point in any of the 16 rows per protocol (48 rows across symmetrized, original, and swapped scoring). In every correction bin containing both methods, DPO has the lower minimum corruption: 14/14 symmetrized, 13/13 original, and 14/14 swapped comparisons. These are replicated point-estimate audits, not confidence statements, and the bins compare nearby rather than exactly matched

Table 15: Qwen/MS MARCO 40% KL-CSFT robustness. Values are three-seed position-symmetrized means in percent. The $\lambda_{\mathrm{TN}} = .1$, update-600 row is selected by the prespecified matched-correction rule. The table shows endpoints and that selected point; the five-point near-match range in the text uses the full seven-checkpoint grid.

Method	Update	Correction	Corruption
Vanilla CSFT	750	56.28	6.29
DPO	750	56.06	1.52
KL-CSFT $\lambda_{\mathrm{TN}} = .03$	750	56.49	5.95
KL-CSFT $\lambda_{\mathrm{TN}} = .1$	600	55.95	6.29
KL-CSFT $\lambda_{\mathrm{TN}} = .1$	750	54.87	5.38
KL-CSFT $\lambda_{\mathrm{TN}} = .3$	750	48.48	4.36
KL-CSFT $\lambda_{\mathrm{TN}} = 1$	750	14.94	1.15
KL-CSFT $\lambda_{\mathrm{TN}} = 3$	750	14.18	0.43

correction levels. CSFT still contributes non-dominated low-correction, low-corruption points that DPO does not occupy. The conclusion is therefore a bounded frontier difference over the tested learning rates and checkpoints, not uniform DPO dominance or an exhaustive search over CSFT recipes.

D TL;DR AS A CORRECTION-DRIVEN BOUNDARY CONDITION

The controlled ranking intervention shows a regime in which DPO's relative advantage is produced primarily by CSFT preservation loss. TL;DR supplies a different boundary condition: inherited preference disagreement is common, and DPO's fixed-pair advantage comes mainly from improving over the shared SFT policy. This appendix documents that comparison. A positive margin denotes agreement with the annotated preference, not factual correctness or absolute summary quality.

D.1 DATA CONSTRUCTION AND PREFERENCE-MARGIN EVALUATION

The fixed prompt is `Summarize the following Reddit post in one concise TL;DR.` followed by the post. Stage 1 contains 4,000 unique training posts paired with the official `policy=ref` summaries. Stage 2 contains 1,600 human-preference pairs from the training split, and evaluation contains 1,500 pairs from `valid2`. The split seed is 20240; posts are unique within each split and disjoint across splits, and at most one pair is retained per post.

Pairs require annotator confidence of at least 7, distinct nonempty responses, at least 16 dataset-native tokens per response, a longer-to-shorter response length ratio no greater than 1.25, and a prompt–response length no greater than 1,024 tokens for either candidate. Sampling is stratified by confidence, policy combination, and single- versus multi-annotation status. Most retained pairs have one annotation (96.88% of Stage 2 and 96.80% of held-out pairs); only 48 held-out pairs have multiple agreeing annotations. We therefore evaluate consistency with the retained label rather than treating it as unanimous preference evidence. Neither policy tokenizer produces a sequence longer than 1,024 tokens.

Qwen3-4B-Instruct-2507 and Gemma-3-4B-it are trained with seeds 7, 42, and 123 on the same fixed splits. Within each model and seed, CSFT and DPO inherit the same SFT checkpoint and use the same Stage-2 order. Standard DPO uses $\beta = .1$. Final checkpoints are evaluated without selection on the held-out pairs. Re-scoring the first 480 held-out pairs of all six model–seed runs reproduces the stored sequence-sum and token-average margins exactly, with no sign or token-length disagreement.

For response tokens $y = (y_1, \ldots, y_T)$, the primary score matches the DPO objective:

$$\ell_{\mathrm{sum}}(y \mid x) = \sum_{t=1}^{T} \log P(y_t \mid x, y_{<t}), \qquad m_{\mathrm{sum}} = \ell_{\mathrm{sum}}(y^+ \mid x) - \ell_{\mathrm{sum}}(y^- \mid x).$$

The length-normalized sensitivity uses $\ell_{\mathrm{avg}} = \ell_{\mathrm{sum}}/T$ and defines m_{avg} analogously. Inherited correct/error pools and all transitions are recomputed separately under each score; membership is not transferred between scoring rules.

D.2 DPO's Advantage Comes Primarily from Additional Correction

The inherited SFT disagrees with the Stage-2 preference label on 41.69–43.52% of pairs across models and scoring rules. On held-out pairs, DPO corrects 27.6–32.9 percentage points more inherited preference errors than CSFT in every one of the 12 model–seed–score comparisons. It also corrupts 1.26–4.87 points more inherited preference-consistent pairs in all 12 comparisons. The net advantage is therefore correction-driven, not a preservation advantage.

Table 16: TL;DR fixed-pair decomposition at standard DPO $\beta = .1$. Values are three-seed means. Train disagreement, Δc, and Δq are conditional percentage-point rates; correction/preservation contributions, DPO–CSFT gap, and gains over SFT are accuracy points. The final column is DPO/CSFT gain over their shared SFT.

Model	Score	Train dis.	Δc	Δq	Corr.	Pres.	Gap	Gain/SFT
Qwen	Sum	43.52	+30.0	+3.37	+13.40	-1.87	+11.53	+9.62/-1.91
Qwen	Avg	41.88	+32.5	+3.30	+13.89	-1.89	+12.00	+10.22/-1.78
Gemma	Sum	42.81	+28.4	+3.22	+12.31	-1.82	+10.49	+9.16/-1.33
Gemma	Avg	41.69	+30.6	+2.18	+12.82	-1.27	+11.56	+10.13/-1.42

The correction contribution is $+11.9$ to $+14.1$ points across individual runs, while the preservation contribution is negative (-0.7 to -2.8 points). Consequently, DPO exceeds CSFT by 10.2–12.6 points in every run and scoring rule. CSFT is below SFT by 0.7–2.3 points, but this accounts for only about 6–20% of the DPO–CSFT gap at the run level (12–17% across the four setting means). Most of the gap comes from DPO itself improving over SFT by 8.9–11.3 points.

This source of advantage differs from the controlled high-disagreement MS MARCO contrast. There, the high-versus-low widening was attributed almost entirely to CSFT decline (96.6–98.5% as a point-estimate ratio); here, additional DPO correction supplies most of the fixed-pair gain. High disagreement can therefore mark training correction demand without implying a preservation advantage or a single cross-task mechanism.

D.3 Scoring Sensitivity and the Boundary of Fixed-Pair Evidence

Sequence-sum scoring is meaningfully length-sensitive. On Stage-2 pairs, longer chosen responses receive lower sequence-sum margins, and the disagreement rate is substantially higher when the chosen response is longer. Token averaging largely removes this association and changes the identity of 11.56% of Qwen pairs and 11.83% of Gemma pairs, even though its aggregate disagreement mass is similar.

Table 17: Training-pair length sensitivity. ρ is Spearman correlation between chosen-minus-rejected response length and the SFT margin. Disagreement entries compare pairs whose chosen response is longer versus shorter; values are three-seed means (%).

	Length–margin ρ		Disagreement: longer/shorter	
Model	Sum	Avg	Sum	Avg
Qwen	-.340	-.021	56.4/29.2	43.5/41.1
Gemma	-.310	+.022	54.7/28.9	41.9/41.5

The main conclusion is nevertheless stable to this scoring change. Under both scores, both models, and all three seeds, DPO has higher correction, higher corruption, and a positive net gap relative to CSFT. Under sequence-sum scoring, the DPO–CSFT gap also remains positive within every response-length group ($+6.4$ to $+18.2$ points across seeds and groups); no evaluated pair is truncated. The corruption cost is larger when the chosen response is longer, so length normalization is a robustness check rather than evidence that length is irrelevant.

These results concern the likelihood ordering of two supplied responses. They do not show that DPO generates more factual, concise, or useful summaries, and they inherit noise from predominantly single-annotator preference labels. Appendix F therefore evaluates free generation separately under standard DPO. That evaluation remains a model-judge utility check, not human validation or a causal link between particular fixed-pair corrections and generated-summary quality.

E TREC RANKING UTILITY AND POSITION-AWARE TRANSITIONS

E.1 EVALUATION PROTOCOL AND LISTWISE UTILITY

The fixed-pair analyses measure whether a checkpoint prefers the designated item, but retrieval utility depends on the ordering of a candidate list. We therefore evaluate the Qwen3-4B SFT, standard CSFT, and DPO checkpoints from all three training seeds on TREC Deep Learning 2019 and 2020 (Craswell et al., 2020; 2021). This is a supplementary cross-benchmark evaluation of existing checkpoints, not another training replication or a test of the controlled disagreement intervention.

A common full-corpus BM25 retriever with $k_1 = 0.9$ and $b = 0.4$ supplies the top 100 candidates for every method. Each document pair is evaluated in both A/B orientations, and the two margins are combined before constructing the ranking. Symmetric margins are mapped to soft win probabilities, clipped to $[10^{-4}, 1 - 10^{-4}]$, and aggregated by Bradley–Terry maximum likelihood. The Bradley–Terry model is fitted with batch gradient descent. The evaluation contains 43 judged queries for DL19 and 54 for DL20. Table 18 reports nDCG@10, MRR@10, MAP, and directly computed log-probability margins, summarized across training seeds. Query-level paired resampling is used for inference.

Table 18: Standardized TREC DL evaluation across training seeds $\{7, 42, 123\}$. Values are mean±population SD across seeds, preserving the convention used by the original TREC analysis.

Dataset	Method	nDCG@10	MRR@10	MAP	Mean $\lvert m \rvert$
DL19	SFT	0.6532 ± 0.0035	0.9793 ± 0.0081	0.3911 ± 0.0023	3.53 ± 0.24
	Standard CSFT	0.6396 ± 0.0076	0.9832 ± 0.0090	0.3875 ± 0.0022	4.00 ± 0.12
	DPO	0.6528 ± 0.0096	0.9845 ± 0.0178	0.3939 ± 0.0053	69.13 ± 11.04
DL20	SFT	0.6302 ± 0.0026	0.9476 ± 0.0049	0.4106 ± 0.0045	3.49 ± 0.30
	Standard CSFT	0.6184 ± 0.0039	0.9344 ± 0.0074	0.4052 ± 0.0012	4.06 ± 0.17
	DPO	0.6463 ± 0.0039	0.9495 ± 0.0056	0.4189 ± 0.0065	69.89 ± 10.71

Statistical comparisons are performed separately within each benchmark and training seed. Confidence intervals use $10{,}000$ query-level paired bootstrap resamples, and p-values use a $10{,}000$-sample paired sign-flip permutation test. DPO exceeds standard CSFT in nDCG@10 for all three seeds on both benchmarks. On DL20, the unadjusted p-values are .013, .016, and .031; all three remain significant under Holm correction across the three seed-level tests, and two remain significant under Bonferroni correction. On DL19, all three differences favor DPO, but none is individually significant.

The comparison with the inherited SFT checkpoint is benchmark-dependent. DPO exceeds SFT in all three DL20 seeds, although only one seed-level comparison is individually significant. On DL19, the direction varies across seeds and the three-seed means are nearly identical. Thus the TREC evaluation supports a consistent point-estimate advantage over continued SFT in all six benchmark–seed comparisons, with inferential support concentrated on DL20. It does not establish a benchmark-independent improvement over the inherited SFT policy. The result also shows why fixed-pair gains and downstream utility should remain separate estimands: the pairwise correction–preservation decomposition does not by itself determine listwise ranking quality.

E.2 DECISION TRANSITIONS AND RANKING-SENSITIVE POSITIONS

To clarify why pairwise decision changes need not translate directly into listwise improvement, we examine SFT-to-CSFT and SFT-to-DPO transitions among judged pairs with unequal relevance grades. Each pair is oriented so that the higher-grade document is positive, evaluated in both A/B orders, and assigned the average symmetric margin. A correction is a wrong-to-correct transition (WC); a corruption is a correct-to-wrong transition (CW); zero is treated as wrong. These transitions use the inherited SFT decision as the reference, but they are a descriptive utility diagnostic rather than the direct DPO–CSFT decomposition used in Sections 4.2 and 6.3.

For SFT-to-CSFT, corruption rates average $4.24 \pm 0.32\%$ on DL19 and $4.41 \pm 0.36\%$ on DL20, while correction rates are $14.65 \pm 0.69\%$ and $15.54 \pm 0.69\%$, respectively; these summaries use

population SD across the three seeds. The rates use different denominators—SFT-correct pairs for corruption and SFT-wrong pairs for correction—and therefore do not determine the raw count balance. Corruption is concentrated among less-confident SFT decisions: no high-confidence tertile pair is corrupted in any of the six dataset–seed combinations.

The unweighted transition balance is

$$N_{\mathrm{unw}} = n_{WC} - n_{CW}.$$

To account for where a changed pair occurs in the inherited ranking, each transition is also assigned the proxy weight

$$w_{\mathrm{dcg}}(i,j) = |G(g_i) - G(g_j)| \left| \frac{1}{\log_2(r_i^{\mathrm{SFT}}+1)} - \frac{1}{\log_2(r_j^{\mathrm{SFT}}+1)} \right|, \qquad G(g) = 2^g - 1,$$

where g_i and g_j are relevance grades and r_i^{SFT} and r_j^{SFT} are ranks under SFT. This weight combines the relevance-gain gap with the DCG discount gap at the fixed SFT ranking; it is an explicit ranking-impact proxy rather than an exact causal decomposition of per-pair DCG. The corresponding weighted balance is

$$N_{\mathrm{dcg}} = \sum_{WC} w_{\mathrm{dcg}} - \sum_{CW} w_{\mathrm{dcg}}.$$

Table 19: Unweighted and position-aware transition diagnostics. Correlations relate query-normalized transition balance to the corresponding query-level change in nDCG@10.

Dataset	Seed	N_{unw}	N_{dcg}	ρ_{unw}	ρ_{dcg}
SFT to CSFT					
DL19	7	−311	−53.75	.047	.529
DL19	42	+7	−35.15	.575	.815
DL19	123	−85	−21.81	.508	.695
DL20	7	−287	−49.84	.265	.512
DL20	42	+145	−12.48	.368	.555
DL20	123	−205	−58.89	.282	.547
SFT to DPO					
DL19	7	−1386	−322.83	.303	.475
DL19	42	+308	−82.85	.315	.409
DL19	123	−396	−267.19	.450	.632
DL20	7	−2582	−655.44	.304	.487
DL20	42	+77	−80.96	.545	.638
DL20	123	−2265	−330.85	.329	.653

For both target methods, the unweighted balance changes sign across seeds, whereas the DCG-impact balance is negative in all six dataset–seed runs. Position-aware weighting reverses every seed-42 case, where corrections outnumber corruptions on both benchmarks. In the SFT-to-CSFT analysis, this difference is not explained by relevance-grade separation alone: corrupted and corrected pairs have similar mean grade-gap ranges (1.23–1.49 and 1.35–1.51), but corrupted pairs have larger mean DCG-impact weight in all six runs (0.086–0.106 versus 0.070–0.089). The top-10-restricted balance is also negative in every SFT-to-CSFT run. The sign of this proxy does not reproduce the aggregate utility result: all three DPO DL20 runs have positive nDCG@10 differences from SFT even though their N_{dcg} values are negative. The balance is therefore used only to describe where transitions occur, not to decompose or predict the net ranking change.

For the query-level analysis, the unweighted balance is normalized by the number of eligible pairs in each query, while the weighted balance is normalized by the total DCG-impact weight:

$$z_q^{\mathrm{unw}} = \frac{n_{WC,q} - n_{CW,q}}{n_q}, \qquad z_q^{\mathrm{dcg}} = \frac{\sum_{WC,q} w_{\mathrm{dcg}} - \sum_{CW,q} w_{\mathrm{dcg}}}{\sum_{i \in q} w_{\mathrm{dcg},i}}.$$

Table 19 reports their Spearman correlations with $\Delta\mathrm{nDCG@10}_q = \mathrm{nDCG@10}_{\mathrm{target},q} - \mathrm{nDCG@10}_{\mathrm{SFT},q}$, where the target is CSFT or DPO. All 43 DL19 and 54 DL20 queries have valid

values. The position-aware correlation exceeds the unweighted correlation in all six runs for each target.

These diagnostics are descriptive rather than causal because the weighting function explicitly incorporates relevance and inherited rank position. The uniformly larger position-aware correlations are consistent with transition location carrying information beyond transition count, but they do not explain the observed method-level nDCG gains. Aggregate correction and corruption rates therefore cannot be read directly as ranking metrics.

E.3 Numerical and Aggregation Audits

The Bradley–Terry aggregation is potentially sensitive to DPO's much larger pairwise margins. DPO increases mean $|m|$ from approximately 4 under standard CSFT to approximately 69 on both benchmarks. At the primary temperature $\tau = 1$, 80.0% of symmetric DPO win probabilities are clipped. A saturation audit divides each orientation margin by $\tau \in \{1, 2, 5, 10, 20\}$ before repeating the same symmetric mapping and Bradley–Terry fit. As τ increases from 1 to 20, the mean DPO clipping rate falls from 80.0% to zero and mean nDCG@10 changes only from .6494 to .6512, averaging the six DPO benchmark–seed evaluations. The direction of all 12 DPO-versus-SFT and DPO-versus-CSFT nDCG@10 comparisons is unchanged across the five temperatures. This audit tests the stability of the ranking result to probability saturation; it does not treat margin magnitude as a predictor of utility.

The reported mean $|m|$ values are recomputed directly as log-probability differences. Applying a probability-space floor of 10^{-12} would instead compress DPO's mean $|m|$ from 69.13 to 22.48 on DL19 and from 69.89 to 22.29 on DL20, while leaving SFT and CSFT unchanged. The floor does not enter the win probabilities or Bradley–Terry aggregation. An independent objective audit confirmed that the implemented gradient is the standard Bradley–Terry log-likelihood gradient; an independently fitted maximum-likelihood solution matched nDCG@10 on all eight audited queries and achieved rank correlation above $.998$.

A separate numerical-robustness audit changed the evaluation batch grouping while holding checkpoints and inputs fixed. Fifteen of 18 TREC benchmark–method–seed conditions reproduced every query-level nDCG@10 and MRR@10 value exactly. The other three contained four query-level nDCG@10 discrepancies, no MRR@10 discrepancy, and condition-level aggregate changes of at most 0.0005. Thus rare boundary decisions can depend on batch composition, but the reported aggregate ranking conclusions are stable to this check.

F TL;DR Generation Utility under Model Judges

The fixed-pair analysis in Appendix D measures consistency with a supplied preference label. This section asks a separate downstream question: whether DPO also receives more favorable assessments when each checkpoint must generate a summary. The evaluation does not identify particular fixed-pair corrections as causes of generation quality.

F.1 Evaluation Design and Judge Validation

A fixed random sample (seed 2026091528) selects 500 posts from the 1,450 held-out posts remaining after exclusion of 50 evaluation-development posts. Qwen3-4B-Instruct-2507 and Gemma-3-4B-it SFT, CSFT, and DPO checkpoints for training seeds 7, 42, and 123 summarize the same posts. The resulting design contains 18 checkpoints and 9,000 generated summaries, but only 500 unique evaluation posts. Within each model and seed, CSFT and DPO branch from the same SFT checkpoint; DPO uses $\beta = .1$.

The task prompt and greedy decoding are fixed across methods. Outputs first receive a 128-token limit; outputs reaching that limit are regenerated from the original prompt with a common 192-token limit. Model-specific termination and residual truncation counts are reported in Appendix F.4.

Llama-3.1-8B-Instruct and Gemma-3-12B-it each receive the source post and one generated summary, without method identity or a competing output. Each judge separately scores all 9,000 summaries under the same rubric, producing 18,000 assessments. Greedy judging uses a 256-token

output limit and batch size 16. The requested JSON contains scores from 1 to 5 for faithfulness, coverage, conciseness, and hallucination avoidance. The four equally weighted scores are summed, and DPO is compared with each baseline on the same post and training seed.

Before the evaluation, the scoring protocol was checked on 1,600 Stage-2 human-preference pairs disjoint by post from the generation set. Conditional on unequal parsed total scores, agreement with the human preference label was 63.2% for Llama and 69.5% for Gemma; approximately 25% of pairs tied. These labels calibrate the judge protocol only; none of the generated outputs in the 500-post evaluation receives a human rating. Repeated scoring of 320 summaries per judge reproduced all four scores exactly under the same batch composition. This establishes repeatability only for that configuration, not invariance to arbitrary batching. A pairwise reason-first alternative exhibited substantial order inconsistency and was not used. The retained judges share the rubric and generated outputs, so their assessments are reported separately rather than treated as independent task replications.

F.2 Cross-Policy Generation Results and Seed-Level Uncertainty

The primary statistic is $(W + 0.5T)/(W + L + T)$, where W, L, and T are DPO wins, losses, and ties against the named baseline. Parsing failures are excluded from the denominator and reported as F. Each seed-level interval uses 5,000 paired post-bootstrap samples and conditions on the corresponding observed checkpoint. The three-seed rates summarized below are descriptive aggregates recomputed from the displayed counts; no joint interval is reported for the full-post aggregate range.

Table 20 reports all 24 policy–judge–baseline–seed comparisons. Under the equal-weight total-score statistic, every seed-level point estimate favors DPO. Across the eight descriptive three-seed policy–judge–baseline aggregates, tie-adjusted rates range from .638 to .747. The common direction across the two policy models is a replication of the model-judge utility result, not an independent replication at the post or judge level: the same 500 posts, rubric, and generated outputs are reused within their stated comparisons.

Table 20: DPO generation comparisons, each on the same 500 posts. $W/L/T/F$ are wins, losses, ties, and parsing failures. Rates count ties as half a win; intervals use paired post resampling.

Policy	Judge	Baseline	Seed	W	L	T	F	Rate	95% CI
Qwen	Llama	SFT	7	271	110	118	1	.661	[.626, .696]
	Llama	SFT	42	276	118	106	0	.658	[.622, .695]
	Llama	SFT	123	273	128	98	1	.645	[.607, .682]
	Llama	CSFT	7	281	112	106	1	.669	[.633, .705]
	Llama	CSFT	42	258	137	105	0	.621	[.583, .658]
	Llama	CSFT	123	260	136	103	1	.624	[.587, .660]
	Gemma	SFT	7	300	128	72	0	.672	[.633, .709]
	Gemma	SFT	42	307	118	75	0	.689	[.650, .726]
	Gemma	SFT	123	344	88	68	0	.756	[.721, .790]
	Gemma	CSFT	7	266	152	82	0	.614	[.575, .653]
	Gemma	CSFT	42	270	149	81	0	.621	[.583, .660]
	Gemma	CSFT	123	305	109	86	0	.696	[.659, .731]
Gemma	Llama	SFT	7	280	108	111	1	.672	[.635, .706]
	Llama	SFT	42	305	110	84	1	.695	[.660, .731]
	Llama	SFT	123	301	99	100	0	.702	[.667, .737]
	Llama	CSFT	7	294	108	97	1	.686	[.650, .722]
	Llama	CSFT	42	293	104	103	0	.689	[.653, .725]
	Llama	CSFT	123	295	90	113	2	.706	[.671, .740]
	Gemma	SFT	7	340	95	65	0	.745	[.708, .779]
	Gemma	SFT	42	342	103	55	0	.739	[.703, .774]
	Gemma	SFT	123	349	91	60	0	.758	[.724, .792]
	Gemma	CSFT	7	295	114	91	0	.681	[.644, .717]
	Gemma	CSFT	42	312	117	71	0	.695	[.658, .732]
	Gemma	CSFT	123	302	125	73	0	.677	[.639, .714]

F.3 Criterion-Level Trade-offs

The favorable total-score comparison is not a uniform improvement across rubric dimensions. For both policy models, judges, and baselines, DPO receives higher faithfulness and coverage scores but lower conciseness and hallucination- avoidance scores (Table 21). These are judge-assigned dimen-

sions that need not be statistically or conceptually independent. Their equal-weight sum encodes one particular quality trade-off.

Table 21: Mean paired rubric-score differences (DPO minus baseline), averaged over three seeds. Each criterion is scored from 1 to 5. Hall. avoid. denotes hallucination avoidance.

Policy	Judge	Baseline	Faithfulness	Coverage	Conciseness	Hall. avoid.
Qwen	Llama	SFT	+.449	+.507	-.072	-.080
	Llama	CSFT	+.414	+.479	-.127	-.099
	Gemma	SFT	+.810	+.768	-.122	-.175
	Gemma	CSFT	+.496	+.605	-.142	-.210
Gemma	Llama	SFT	+.509	+.714	-.105	-.061
	Llama	CSFT	+.525	+.677	-.073	-.105
	Gemma	SFT	+1.051	+1.102	-.266	-.254
	Gemma	CSFT	+.695	+.886	-.277	-.316

F.4 Length, Termination, and Scope Sensitivities

Generation length differs across methods and could influence the rubric. For Qwen, mean output lengths are 29.446, 33.420, and 41.187 tokens for SFT, CSFT, and DPO. A descriptive sensitivity therefore retains comparisons whose longer-to-shorter output-length ratio is at most 1.25. Table 22 shows favorable DPO point estimates for both policy models, judges, and baselines in this subset. Counts pool post–seed comparisons rather than unique posts, and no subset intervals are reported. Because inclusion is determined by generated length, this restriction is not a causal control for length.

Table 22: Similar-length sensitivity across policy models. n counts valid post–seed comparisons with length ratio at most 1.25 across three seeds, not unique posts. Rates are descriptive point estimates.

Policy	Judge	Baseline	n	Rate
Qwen	Llama	SFT	429	.624
	Llama	CSFT	627	.647
	Gemma	SFT	430	.626
	Gemma	CSFT	628	.611
Gemma	Llama	SFT	228	.647
	Llama	CSFT	333	.638
	Gemma	SFT	228	.662
	Gemma	CSFT	333	.634

For Qwen, seven outputs reach the initial 128-token cap and are regenerated; three still reach 192 tokens (two SFT and one CSFT), and none is empty. Excluding comparisons containing these residual truncations changes each affected seed-level rate by less than .0008. Gemma uses EOS IDs $\{1, 106\}$. Of 56 outputs reaching 128 tokens, 27 still reach 192 tokens: two SFT and 25 DPO, with no CSFT output. Residual Gemma DPO truncations number 1/13/11 by seed 7/42/123, and SFT truncations number 1/1/0. Truncated summaries are scored as generated, not assigned automatic losses or excluded.

Restricting Gemma comparisons to outputs where both methods terminate normally retains favorable three-seed point estimates and joint post-bootstrap intervals (Table 23). The full-post comparison remains primary.

The generation evaluation therefore provides a scoped model-judge utility check, not human validation. Gemma policy outputs are evaluated by a same-family Gemma judge as well as by Llama, and both judges share the rubric and posts. Neither the total scores nor the criterion labels establish factual correctness or safety, and the results do not equate generated-summary quality with the fixed-pair correction and preservation categories.

Table 23: Gemma policy sensitivity restricted to comparisons where both outputs terminate normally. Rates are three-seed means with joint post-bootstrap 95% CIs.

Judge	Baseline	Rate [95% CI]
Llama	SFT	0.695 [0.674, 0.717]
Llama	CSFT	0.700 [0.677, 0.723]
Gemma	SFT	0.756 [0.733, 0.778]
Gemma	CSFT	0.694 [0.670, 0.718]

G Exploratory Diagnostics of Correction Demand and Preservation Requirements

This section develops the diagnostic direction in Section 7.1. Most analyses reuse existing training and evaluation artifacts; we also report Gemma-3-12B checks on MS MARCO and ESCI whose screening threshold was frozen before Stage-2 training. We explore whether SFT-visible correction demand helps anticipate DPO's gain and when preservation requirements qualify that signal.

G.1 Screening with SFT Training Disagreement

We reuse the 11 natural ranking settings with verified SFT train margins (Appendix B.2). For each seed, training disagreement is $d = N_{\text{train}}^{-1} \sum_i \mathbf{1}[m_i < 0]$ over the 2,000 Stage-2 pairs. Features and outcomes are averaged across the three seeds within each setting. The outcome is whether DPO improves held-out pair accuracy over SFT by at least $\delta = 1$ percentage point. This scenario yields seven positive and four negative settings; a negative label can still represent a positive gain below the chosen target. At $\delta = 0.5$, all 11 settings are positive, so this cohort cannot evaluate two-class discrimination at that target.

The initial screen recommends DPO when d exceeds a threshold. In leave-one-setting-out evaluation, the threshold is selected from training-fold feature midpoints to maximize balanced accuracy, breaking ties by fewer false negatives, more SFT-retention recommendations, and then the lower threshold. The held-out setting is not used for fitting. The screen correctly classifies 9/11 settings (Table 24); leaving out an entire model family gives the same two errors. It misses Qwen3-1.7B/MS MARCO (gain 1.02 points) and recommends DPO for Gemma-3-4B/MS MARCO (0.91 points). It recommends retaining SFT in four settings while preserving 83.7% of the summed gain obtained by applying DPO everywhere in this cohort. These are retrospective recommendations, not measured compute savings.

A subsequent exploratory comparison uses L2-regularized logistic regression (regularization parameter $C = 1$, an intercept, no class weighting, and training-fold standardization). At a probability cutoff of 0.5, disagreement alone classifies 10/11 settings correctly. Table 24 also reports the preservation-proxy comparison below. These model-form comparisons reuse the same small cohort and do not constitute independent validation.

Table 24: Exploratory leave-one-setting-out screening at a one-point DPO–SFT gain target. All rows reuse the same 11 settings. Missed gains and extra recommendations count false negatives and false positives, respectively.

Input and rule	Correct	Balanced acc.	Missed gains	Extra rec.
Disagreement threshold	9/11	80.4%	1	1
Disagreement logistic	10/11	92.9%	1	0
Disagreement + fragility logistic	10/11	92.9%	1	0

One additional check froze the full-cohort threshold, 0.04425, before Stage-2 training on Gemma-3-12B/MS MARCO. Its SFT-retention prediction agrees with the subsequently observed mean gain of 0.363 points. The three Stage-2 seeds share one SFT checkpoint: this is one additional model–dataset setting, not three independent starting-state validations. The full-cohort threshold is distinct from the fold-specific thresholds used for the 9/11 result.

A subsequent Gemma-3-12B/ESCI check used the same threshold without refitting. Before Stage-2 training, mean training disagreement of 14.78% predicted a DPO gain of at least one point over SFT. Unlike the MS MARCO check, seeds 7, 42, and 123 each use an independently trained SFT start. On the same 10,000 held-out pairs, their DPO gains are 2.12, 1.25, and 1.85 points, respectively; mean accuracy rises from 83.66% to 85.40%, a 1.74-point gain. Thus the frozen recommendation agrees with the outcome in both additional settings, in opposite recommendation directions. These checks remain separate from the 9/11 leave-one-setting-out result. They support the screen's exploratory promise but do not separate disagreement from dataset differences or validate a combined correction–preservation predictor.

G.2 A Candidate Preservation-Fragility Signal

We tested whether small positive SFT train margins add information about preservation. For each seed, let $s = \mathrm{median}(m_i \mid m_i > 0)$ and $v = N_{\mathrm{train}}^{-1} \sum_i \mathbf{1}[0 < m_i \le 0.25s]$. Zero margins are excluded from both d and v. This proxy describes boundary proximity among currently agreeing training pairs; it does not measure actual corruption risk.

Across settings, v correlates with the realized DPO corruption cost $K = 100(1-e)q$ (Spearman $\rho = 0.70$), but the relationship is near zero within MS MARCO (0.09) and reverses within ESCI (-0.70). Cutoffs of $0.10s$ and $0.50s$ also yield positive pooled associations. Adding v to the logistic model leaves every held-out action unchanged; the improvement over the threshold rule is also obtained without v. Thus this particular proxy offers no additional action-level discrimination here. Its pooled association motivates examining preservation, but does not establish a transferable risk predictor.

G.3 Separating Training Demand from Preservation Requirements

Natural training disagreement and held-out SFT error mass e are nearly collinear (Spearman $\rho = 0.991$). The controlled sweeps separate their roles: training disagreement varies while the SFT checkpoint and held-out error pool remain fixed within each cell. Applying the natural screen's fixed 0.04425 threshold to the 18 constructed arms produces nine recommendations whose realized gain is below one point and misses one gain above it. Seven of those nine over-recommendations occur in the two MS MARCO cells, where the inherited error pool is only 3–4%. The arms are dependent views of four starting checkpoints, not 18 independent settings.

For correction rate c, corruption rate q, and inherited error fraction e, the net accuracy gain in percentage points is

$$G = 100\{ec - (1-e)q\}.$$

For $e < 1$, an assumed correction level $c_\star$ and target δ therefore require

$$q \le q_{\max}(e, c_\star, \delta) = \frac{ec_\star - \delta/100}{1-e}. \tag{1}$$

A small positive bound imposes strict preservation; a negative bound means the target is unattainable at the assumed correction even with zero corruption. Table 25 uses common correction scenarios, not predicted or fitted correction rates.

Table 25: Maximum corruption (%) compatible with a one-point gain, under assumed correction rates. The error fractions come from the canonical controlled SFT checkpoints. Negative entries denote infeasibility under the stated assumption, not a measured negative corruption rate.

Starting setting	Error (%)	$c_\star = 25\%$	$c_\star = 50\%$	$c_\star = 75\%$
Qwen / MS MARCO	3.08	-0.24	0.56	1.35
Gemma / MS MARCO	3.98	-0.01	1.03	2.07
Qwen / ESCI	16.30	3.67	8.54	13.41
Gemma / ESCI	17.21	3.99	9.19	14.38

At 50% correction, Qwen/MS MARCO requires corruption below 0.56%, versus 9.19% for Gemma/ESCI. An illustrative check combining $d \ge 0.04425$ with $q_{\max}(e, 0.5, 1) \le 2\%$ marks the seven MS MARCO over-recommendations. The 2% cutoff is an analyst-chosen explanatory

value after outcomes were available, not a validated decision boundary. It misses the two ESCI over-recommendations at 35% mass (gains 0.91 and 0.64 points), and also marks natural Qwen3-1.7B/MS MARCO, which does achieve 1.02 points. A strict preservation requirement is therefore a reason for closer examination, not an automatic instruction to retain SFT.

This analysis reuses held-out benchmark SFT scores; no independent calibration/outcome split was evaluated. Within a controlled cell, e cannot distinguish mass-specific outcomes, and the illustrative flags coincide with dataset membership among recommended controlled arms. The useful implication is a diagnostic question: can a representative labeled calibration set reveal when high training disagreement is accompanied by a small correction opportunity and a strict preservation requirement? Such a check would complement training disagreement while leaving prediction of realized correction and corruption as an open problem.