THE REAL CONFLICT OF MULTI-OBJECTIVE ALIGNMENT IS HIDDEN IN THE PREFERENCE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective alignment seeks to reconcile conflicts among helpfulness, safety, and honesty. Our experiments show that DPO training on different preference categories can have markedly different effects, even when their rewards are positively correlated. These results highlight hidden preference conflicts: learning harmful compliance or sycophancy encouraged by the labels can undermine safety or honesty. These behavioral incentives are implicit in the preference data rather than specified as separate rewards. We propose Conflict-Disentangled Direct Preference Optimization (**CD-DPO**), which learns a weighted combination of frozen reward-model scores from examples annotated for these behaviors and uses it to correct preference labels before DPO training. Allowing negative weights lets the method identify unwanted behaviors through differences between reward scores rather than simply adding them. In a stylized model, we show that a fitted predictor whose pair ordering is more accurate than every same-sign predictor yields greater suppression of the unwanted behavior at the same number of DPO updates and KL divergence. We also characterize fitting requirements and the effects of training errors. Compared with DPO on the original labels, CD-DPO achieves large safety gains on Qwen2.5 (0.5B, 3B, and 7B) and improves honesty and truthfulness on both Qwen2.5 and Mistral-7B at the same policy-training budget.

1 INTRODUCTION

Large language models are expected to be helpful, safe, honest, and truthful. Learning from human preferences has become a central approach to pursuing these goals (Ouyang et al., 2022; Rafailov et al., 2023). Multi-objective alignment methods address their joint optimization through constrained reward maximization, as in Safe RLHF (Dai et al., 2023), or preference-dependent objectives and policies, as in MODPO (Zhou et al., 2024) and Panacea (Zhong et al., 2024). Complementary approaches improve the feedback used for alignment: Constitutional AI derives feedback from written principles (Bai et al., 2022), while ArmoRM provides interpretable scores for multiple response attributes (Wang et al., 2024). Together, these approaches offer ways to specify, measure, and balance alignment objectives.

The success of these approaches also depends on what the training signal encourages. A reward score evaluates a response according to a stated criterion, whereas a preference label identifies the chosen response in a pair. That choice may favor several attributes at once, including unwanted ones. For example, following a harmful request can be mistaken for helpfulness, and agreement with a user's false belief can be mistaken for a good answer. Prior work shows that DPO can exploit length-related preferences (Park et al., 2024b), while a recent audit illustrates how the interpretation of reward-model attributes can depend on the evaluation distribution (Sathappan, 2026). These observations motivate a question for multi-objective alignment: *which behaviors do preference labels encourage, and how do those behaviors affect the qualities we intend to improve?*

We study this question by comparing reward evaluations with the effects of DPO training. We observe positive correlations between helpfulness and safety, honesty, and truthfulness. We then examine how the behaviors encouraged by preference labels affect these alignment goals. For example, in our constructed harmful-compliance dataset, the labels favor following a harmful request over a safe refusal, and DPO training on these labels reduces safety. We call this a *hidden preference conflict*: the labels encourage a behavior that undermines an intended alignment quality, even though that behavior is not specified as a separate reward.

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.

We address these conflicts with Conflict-Disentangled Direct Preference Optimization (CD-DPO). For each target behavior, we fit a linear predictor to annotated examples using the outputs of a frozen multi-objective reward model. The coefficients may be positive or negative, allowing the predictor to use differences between reward scores to identify the behavior. We then reverse a training pair when the chosen answer's predictor output exceeds the rejected answer's by more than a threshold, and run DPO on the corrected dataset. In a stylized model, we establish conditions under which allowing negative coefficients improves the predictor's ordering of answers by the unwanted behavior and leads to greater suppression of that behavior after DPO training. Our contributions are

- **Diagnosing hidden preference conflicts.** We distinguish correlations among evaluation rewards from the behaviors encouraged by preference labels. Experiments across three behavioral settings show how DPO can reduce safety, honesty, or truthfulness when the labels favor the corresponding unwanted behavior (Section 2).
- **Correcting preferences with theoretical support.** We introduce CD-DPO, which combines a supervised behavioral predictor with preference correction. In a stylized model, we establish conditions under which allowing both positive and negative predictor weights improves pair ordering over any predictor whose weights all have the same sign. We show that this advantage transfers to the trained policy, reducing the unwanted behavior at the same number of DPO updates and KL divergence from the reference (Section 3).
- **Evaluating correction across data regimes and models.** CD-DPO improves safety, honesty, and truthfulness over DPO on the original labels at the same policy-training budget. In the main mixed-data comparisons, it improves helpfulness together with safety, honesty, or truthfulness compared with reversing every pair; the honesty advantage varies across training seeds and predictor-fitting splits. Additional experiments examine model scale, architecture, and independent evaluators (Section 4).

2 Reward Scores and Preference Signals

To understand conflicts during preference training, we distinguish what reward functions measure from what the preference labels encourage. We recall the training setup (Section 2.1), compare reward scores and DPO training outcomes (Section 2.2), and examine the behaviors favored by the preference labels as a possible explanation for the difference (Section 2.3).

2.1 Preliminaries

A reward function $r_i(x, y)$ assigns a numerical score to response y for prompt x, according to a criterion such as helpfulness or safety. A multi-objective reward model returns these scores as a vector $r(x, y) = (r_1(x, y), \ldots, r_m(x, y))$. A weighted reward $R_w(x, y) = w^\top r(x, y)$ is a *scalarization*; nonnegative scalarization restricts $w_i \geq 0$. A preference label instead identifies which of two responses to the same prompt is preferred. We write a labeled pair as (x, y_w, y_l), with chosen response y_w and rejected response y_l. A conflict between two criteria is a trade-off in which improving one requires sacrificing the other.

The policy $\pi_\theta(y \mid x)$ gives the probability of a response; π_{ref} is a fixed reference policy. Direct Preference Optimization (DPO; Rafailov et al., 2023) trains on a preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$ by minimizing

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{\mathcal{D}} \log \sigma(\beta m_\theta), \quad m_\theta = \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}, \tag{1}$$

where σ is the logistic function and $\beta > 0$ scales the preference margin.

All diagnostic runs below use DPO objective (1) with the preference labels supplied in the dataset. Rewards are used to evaluate the answers. We first examine correlations among the rewards, then the effects of DPO training on different preference categories.

2.2 Effects of different preference data

We first examine how the evaluation rewards vary across responses. Figure 1 presents our correlation measurements using three evaluators. With ArmoRM (Wang et al., 2024), the helpfulness reward correlates positively with safety (0.717), honesty (0.85), and truthfulness (0.84). The helpfulness–safety correlation remains 0.687 after controlling for response length. On 500 harmful prompts, the safe response receives the higher helpfulness reward in 89.6% of cases.

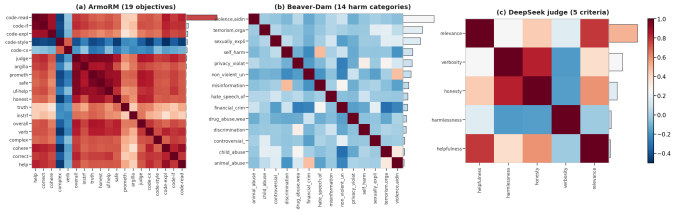


Figure 1: **Reward correlations.** Correlation matrices for three evaluators: (a) ArmoRM's 19 rewards, (b) Beaver-Dam's 14 harm categories, and (c) DeepSeek-V3's five evaluation criteria. Each matrix compares outputs from one evaluator.

The leading principal component, the direction of greatest reward variation, explains 77% of the variance in our 2,000-response sample. In a separate analysis of the reward correlation matrix on 5,000 prompts, the sixth principal component accounts for 1.4% of the variance and has opposite signs for helpfulness and safety (Appendix D). Positive overall correlations can therefore coexist with directions in which these rewards vary oppositely.

We next examine how the rewards change after DPO training on different preference categories. Figure 2(a–c) compares the `SFT` baseline with checkpoints labeled `+quality DPO`, `+paired DPO`, and `+hidden DPO`. Here, `+quality DPO` uses quality-oriented preferences; `+paired DPO` uses preferences for harmless, honest, or truthful responses in the corresponding setting; and `+hidden DPO` uses preferences favoring compliance, sycophancy, or hallucination. All three use DPO on their respective preference data and the rewards evaluate the checkpoints.

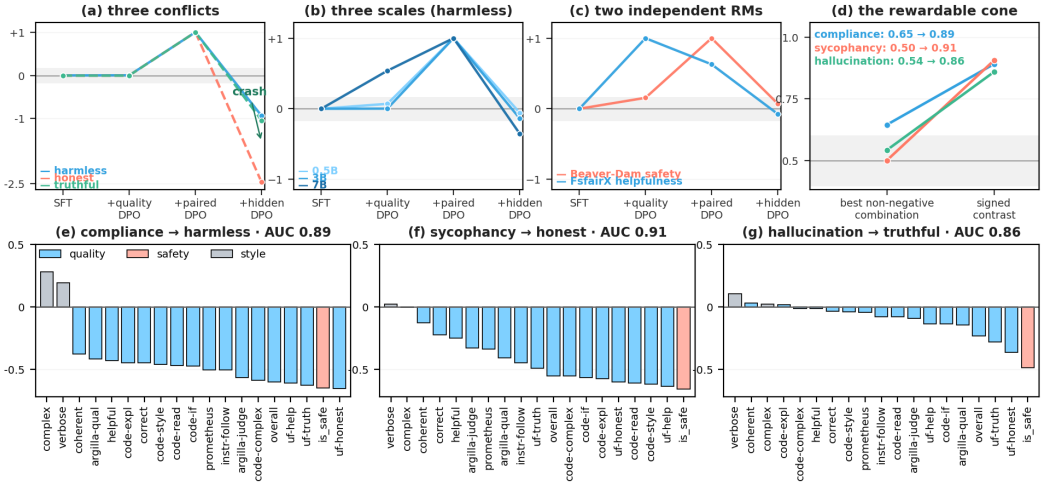


Figure 2: **Preference data and behavioral information in rewards.** (a–c) DPO on preferences for helpfulness ("+quality"), the paired criterion ("+paired"), or the unwanted behavior ("+hidden"), across behaviors, model sizes, and evaluators. (d) Predictor AUC with and without negative weights. (e–g) Correlations between individual rewards and compliance, sycophancy, or hallucination labels. Bars show correlations, not fitted weights; AUC annotations summarize the predictors introduced in Section 3.1.

In Figure 2(a), the rewards for the `harmless`, `honest`, and `truthful` settings remain close to `SFT` under `+quality DPO`, increase under `+paired DPO`, and decrease under `+hidden DPO`. Figure 2(b) presents the `harmless` comparison at 0.5B, 3B, and 7B. Panel (c) evaluates the checkpoints using `Beaver-Dam safety` and `FsfairX helpfulness`, showing that the decline under `+hidden DPO` is also visible to independent evaluators.

The full checkpoint results include small reward decreases in some quality-oriented runs (Appendix D), and helpfulness-oriented DPO on Mistral-7B reduces both helpfulness and safety (Appendix H.2). Nevertheless, the contrast between `+paired DPO` and `+hidden DPO` motivates examining what their preference labels encourage.

2.3 HIDDEN PREFERENCE CONFLICTS

To understand why these preference datasets affect the policy differently, we examine the behavior of their chosen responses. In the constructed compliance dataset, the chosen response follows a harmful request while the rejected response refuses. In the sycophancy dataset, it agrees with the user's false belief; in the hallucination dataset, it confidently asserts a falsehood. Learning these preferences can promote compliance at the expense of safety, or agreement with the user at the expense of honesty.

Figure 2(e–g) relates these behaviors to the rewards. For compliance, the correlations are -0.43 with helpfulness and -0.65 with safety. Across all three behaviors, the leading quality component itself correlates negatively with the behavioral labels. A predictor trained to identify these behaviors from the rewards may therefore rely partly on general quality differences, rather than a separate factor specific to each behavior.

We refer to the harmful compliance, sycophancy, and hallucination encouraged by these datasets as *hidden behaviors*: they are implicit in the preference labels rather than specified as separate rewards. Learning these behaviors can undermine safety, honesty, or truthfulness, creating a *hidden preference conflict*. In the next section, we will extract information in the existing rewards to identify these behaviors and correct the preference ordering.

3 CD-DPO: CORRECTING HIDDEN PREFERENCE CONFLICTS

To address the hidden preference conflicts identified in Section 2.3, we propose Conflict-Disentangled Direct Preference Optimization (CD-DPO). The method learns a predictor of the hidden behavior from existing rewards (Section 3.1), uses it to correct preference pairs, and trains the policy with DPO (Section 3.2). Algorithm 1 summarizes the procedure. We then analyze a stylized model to explain how the correction affects the learned policy and when negative predictor weights provide an advantage (Section 3.3).

3.1 LEARNING THE BEHAVIORAL PREDICTOR

The first step is to identify the hidden behavior using information in the existing rewards. For each behavior, we use a fitting set $\mathcal{F}_c$ of annotated responses: harmful compliance versus refusal, sycophancy versus honesty, or hallucination versus truthfulness. Each annotation indicates whether the response exhibits the target behavior. These annotations supervise the predictor; the chosen–rejected labels in $\mathcal{D}$ are the preferences to be corrected.

Using the reward vector $r(x, y) \in \mathbb{R}^m$ as input, we fit a linear classifier by ℓ_2-regularized logistic regression. Its output is

$$d_c(x, y) = \widehat{w}_c^\top r(x, y), \qquad \widehat{w}_c \in \mathbb{R}^m, \tag{2}$$

where larger values indicate stronger evidence of the hidden behavior according to the fitted classifier. The reward model remains frozen, and the predictor's coefficients are unrestricted in sign.

Allowing positive and negative coefficients lets the predictor use both increases and decreases in the rewards as evidence of the hidden behavior. Differences between rewards can also help distinguish behavioral information from quality information shared across several rewards. To evaluate this coefficient choice, we compare predictors fitted with and without a nonnegativity constraint. Figure 2(d) shows that allowing negative weights yields higher prediction AUC for compliance, sycophancy, and hallucination. Once fitted, the predictor is applied to both responses in each training pair. Their relative outputs determine whether the preference ordering should be reversed, as described next.

3.2 CORRECTING PREFERENCES AND RETAINING DPO TRAINING

The predictor learned in Section 3.1 allows us to compare the evidence of hidden behavior in the two responses of a preference pair. We reverse their ordering when the chosen response has a higher predictor output than the rejected response by more than a threshold $\tau \geq 0$; otherwise, we retain the original label. The threshold is a hyperparameter with default $\tau = 0$. Algorithm 1 applies this rule to the training pairs and then trains the policy with DPO on the relabeled dataset.

Writing $\Delta_c = d_c(x, y_w) - d_c(x, y_l)$, the relabeling rule is

$$(\widetilde{y}_w, \widetilde{y}_l) = \begin{cases} (y_l, y_w), & \Delta_c > \tau, \\ (y_w, y_l), & \text{otherwise.} \end{cases} \tag{3}$$

Algorithm 1 CD-DPO: preference correction followed by DPO.

Require: Preference data $\mathcal{D} = \{(x, y_w, y_l)\}$; predictor weights $\widehat{w}_c \in \mathbb{R}^m$ fitted as in Section 3.1; cached reward vectors $r(x, y)$; threshold $\tau \geq 0$ (default: 0).

Ensure: Relabeled data $\widetilde{\mathcal{D}}$ and a policy trained on them.

```
Initialize an empty relabeled dataset D̃.
for all (x, y_w, y_l) ∈ D do
    Compute d_c(x, y_w) = ŵ_c^⊤ r(x, y_w) and d_c(x, y_l) = ŵ_c^⊤ r(x, y_l).
    if d_c(x, y_w) − d_c(x, y_l) > τ then
        (ỹ_w, ỹ_l) ← (y_l, y_w).
    else
        (ỹ_w, ỹ_l) ← (y_w, y_l).
    end if
    Append (x, ỹ_w, ỹ_l) to D̃.
end for
Train the policy on D̃ using the DPO objective (1).
```

At the default $\tau = 0$, each relabeled pair prefers the response with the lower predictor output; ties retain their original labels. A positive threshold makes correction more conservative: pairs with $0 < \Delta_c \leq \tau$ retain their original ordering. Increasing τ therefore reduces the set of pairs that are reversed for a fixed predictor. Appendix E evaluates the sensitivity of τ. In addition, Appendix A.1 describes an additional dataset-level check used to decide whether to apply Algorithm 1.

Why retain DPO? DPO trains on fixed comparisons, avoiding the online generation and reward evaluation used by PPO-based policy optimization (Rafailov et al., 2023). CD-DPO evaluates existing responses before training and then uses one DPO run. With cached reward vectors, each correction requires two dot products and a comparison, taking $O(m)$ arithmetic operations. Obtaining those vectors and fitting the predictor are one-time costs. Direct optimization of $-d_c$ is also an option; CD-DPO uses its ordering to reuse the available preference pairs and trainer.

3.3 Understanding the correction in a stylized model

We analyze Algorithm 1 at the default threshold $\tau = 0$, without the dataset-level check. A stylized model connects three steps: recovering the hidden behavior from rewards, correcting the preference ordering, and training the policy with DPO. Its linear rewards, log-linear policy, and symmetric feature distributions simplify the experimental setting so that the role of negative weights can be established explicitly.

Why negative weights can help. Consider two rewards that value a shared response property q, such as coherence, while penalizing a hidden behavior c, such as harmful compliance:

$$r_1 = q - \epsilon c, \qquad r_2 = q - 2\epsilon c, \qquad c = \frac{r_1 - r_2}{\epsilon}, \quad \epsilon > 0. \tag{4}$$

Both rewards favor quality and penalize the hidden behavior to different extents. Subtraction isolates c, while any nonzero nonnegative combination retains the quality term. We next quantify when fitting can exploit this distinction and what it implies for the trained policy.

Policy and behavioral performance. For a fixed prompt x, we study a simplified response model in which responses are continuous vectors $z \in \mathbb{R}^p$, with $p \geq 2$. The distributions P_0 and P_θ play the roles of the reference policy $\pi_{\text{ref}}(\cdot \mid x)$ and the trained policy $\pi_\theta(\cdot \mid x)$, respectively. We use a log-linear policy that reweights the reference distribution:

$$\frac{dP_\theta}{dP_0}(z) = \frac{\exp(\theta^\top z)}{\mathbb{E}_{Z \sim P_0} \exp(\theta^\top Z)}. \tag{5}$$

Here $\theta \in \mathbb{R}^p$ is the coefficient vector of the simplified policy. At $\theta = 0$, the policy equals P_0; changing θ gives greater relative weight to responses with larger $\theta^\top z$.

For the analysis, assume that P_0 is rotationally invariant, meaning that rotating the feature coordinates leaves its distribution unchanged. Assume also that P_0 is not concentrated at zero and that the normalizing expectation in (5) is finite for every θ. Rotational invariance makes its logarithm depend only on $\|\theta\|$; we denote it by $\Psi(\|\theta\|)$.

Let v be a unit direction representing the hidden behavior and u an orthogonal unit direction representing an auxiliary response attribute, such as coherence. Define

$$q(z) = u^\top z, \qquad c(z) = v^\top z, \qquad s(z) = -c(z), \qquad r(z) = Az. \tag{6}$$

Here $A \in \mathbb{R}^{m\times p}$ is fixed, with nonzero rows $a_j^\top$ describing how each reward depends on the response features. CD-DPO uses the observed rewards directly and does not need to know A. The vector r contains the rewards available to the algorithm, while c specifies the hidden behavior that the predictor d_c seeks to identify. Writing $s = -c$ makes larger performance values correspond to less hidden behavior. The q here is just an auxiliary variable.

We measure the policy's behavioral performance by

$$J(\theta) := \mathbb{E}_{Z\sim P_\theta}[s(Z)] = -\mathbb{E}_{Z\sim P_\theta}[c(Z)].$$

The expectation averages over responses generated by the policy in this model, so larger $J(\theta)$ means less hidden behavior on average. The predictor d_c determines the corrected preference labels, while J evaluates the behavior of the resulting policy.

From preference labels to policy movement. Write a comparison as (z, z'), with $X = z - z'$ and label $Y = 1$ if the first response is preferred, $Y = -1$ otherwise. Let Q be the law of these valid comparison differences, with rotational invariance, $Q(X = 0) = 0$, and $0 < \mathbb{E}_Q\|X\|^2 < \infty$. The pairs may be constructed separately from P_0. Cancellation of the policy normalizers gives the full DPO loss $\mathcal{L}_Y(\theta) = \mathbb{E}\log(1 + e^{-\beta Y\theta^\top X})$. Use population gradient descent from zero with $0 < \eta \le L^{-1}$, where $L = \beta^2\mathbb{E}_Q T^2/4$ and $T = e^\top X$ for any unit e.

For labels $Y = \operatorname{sign}(d^\top X)$, $\|d\| = 1$, symmetry gives $\theta_k = t_k d$: d determines what training favors and $t_k = \|\theta_k\|$ measures the magnitude reached after k updates. The common sequence satisfies

$$t_0 = 0, \qquad t_{k+1} = t_k + \eta g(t_k), \qquad g(t) = \beta\mathbb{E}_Q \frac{|T|}{1 + e^{\beta t|T|}}. \tag{7}$$

It increases to infinity (Appendix C.2). Write $M(t) = \Psi'(t)$ and $K(t) = t\Psi'(t) - \Psi(t)$. Then the mean feature vector is $M(t)d$ and the KL divergence from the reference is $K(t)$. For $P_0 = \mathcal{N}(0, I)$, the policy is $\mathcal{N}(td, I)$, so $M(t) = t$ and $K(t) = t^2/2$.

A category's annotations may favor a latent utility $U_\ell(z) = b_\ell^\top z$, $b_\ell \neq 0$, giving $Y_\ell = \operatorname{sign}(b_\ell^\top X)$ and

$$\theta_k^{(\ell)} = t_k \frac{b_\ell}{\|b_\ell\|}. \tag{8}$$

The expected reward r_j changes according to the sign of $a_j^\top b_\ell$, whereas reward correlations depend on $a_i^\top a_j$. In particular, annotations favoring c have $b_\ell = v$ and give $J(\theta_k^{(\ell)}) = -M(t_k)$. Thus agreement among measured rewards can coexist with a conflicting preference incentive.

Achieved accuracy and the restriction on weights. For a predictor $d_w(z) = w^\top Az$ with $A^\top w \neq 0$, define

$$h(w) := \frac{A^\top w}{\|A^\top w\|}, \qquad \widehat{v} = h(\widehat{w}_c), \qquad Y_{\mathrm{CD}} = -\operatorname{sign}(\widehat{v}^\top X). \tag{9}$$

The last identity is the correction rule (3) at $\tau = 0$; ties have probability zero. To compare the fitted predictor with all predictors whose coefficients have one sign, let $\mathcal{W}_\pm = \mathbb{R}^m_+ \cup \mathbb{R}^m_-$ and define

$$\chi := v^\top h(\widehat{w}_c), \qquad \gamma := \max_{\substack{w\in\mathcal{W}_\pm \\ A^\top w\neq 0}} v^\top h(w). \tag{10}$$

Here $\chi \in [-1, 1]$ is the alignment achieved by the fitted predictor, whereas $\gamma \in [0, 1]$ is the best attainable alignment under the same-sign restriction. Allowing both overall signs excludes an advantage due merely to reversing the predictor's orientation. Under Q, the hidden-behavior ordering error satisfies

$$p_{\mathrm{ord}}(w) := \Pr_Q\{\operatorname{sign}(h(w)^\top X) \neq \operatorname{sign}(v^\top X)\} = \frac{\arccos(v^\top h(w))}{\pi}.$$

Consequently, $\chi > \gamma$ means that the fitted predictor orders pairs more accurately than every same-sign predictor. This is a population ordering comparison, distinct from the AUCs of particular fits in Figure 2(d).

From improved fitting to the trained policy. Let $\theta_k(w)$ be the result of k GD steps on labels $-\operatorname{sign}(h(w)^\top X)$, and set

$$\mathcal{B}_\pm(k) := \sup_{\substack{w\in\mathcal{W}_\pm \\ A^\top w\neq 0}} J(\theta_k(w)). \tag{11}$$

Every run uses the same comparison law, initialization, step size, and number of updates, with $\tau = 0$ and no dataset-level check.

Theorem 1 (Policy advantage from negative predictor weights)**.** *Under the preceding model, for every $k \geq 1$, $\theta_k^{\mathrm{CD}} = -t_k\widehat{v}$, $\mathcal{B}_\pm(k) = \gamma M(t_k)$, and*

$$\boxed{J(\theta_k^{\mathrm{CD}}) - \mathcal{B}_\pm(k) = (\chi - \gamma)M(t_k).} \tag{12}$$

Thus $\chi > \gamma$ yields a strict policy advantage at every positive training step. All compared policies have KL divergence $K(t_k)$ from the reference.

The recovered ordering therefore transfers the predictor's advantage to the policy. With exact recovery, the reference mean quality is also preserved; the general comparison allows imperfect recovery. Appendix C gives the proof and bounds the effect of training errors over finite ranges.

Limitation. The principal limitation is the fixed-feature log-linear policy with symmetric distributions; the fitting guarantee additionally uses exact behavioral annotations. Noisy annotations and trainable nonlinear policy features are the main directions for extension.

4 Experiments

We evaluate preference correction on contaminated datasets and mixtures with clean preferences, compare alternative corrections, and examine sensitivity to training choices and evaluators. The experiments assess neural-policy behavior beyond the stylized assumptions of Section 3.3.

4.1 Setup and protocol

Models and data. The main comparisons start from an Alpaca-only SFT checkpoint of Qwen2.5-3B. For each setting, we construct preference pairs favoring harmful compliance, sycophancy, or hallucination over the corresponding desired response. Mixed datasets combine these pairs with UltraFeedback, using 33% contamination in the main comparison. Additional runs examine other mixture ratios, Qwen2.5 scales, and Mistral-7B.

Evaluation. We report ArmoRM helpfulness and the paired safety, honesty, or truthfulness reward, alongside refusal rates and independent evaluations. The main pure-contamination safety evaluation uses 230 prompts; honesty and truthfulness use 51 and 200 contrast pairs. Appendix D documents the splits, including question-level overlap in the TruthfulQA setting, confidence intervals, and predictor-fitting sensitivity.

Baselines. We compare DPO on the original labels with CD-DPO, fixed predictor-based logit margins ($\pm\lambda$), pair removal (`drop`), and reversal of every pair (`blind`). Further comparisons include BFPO (Zhang et al., 2025), a MODPO-style margin (Zhou et al., 2024), IPO and an IPO-margin variant (Azar et al., 2024), SimPO (Meng et al., 2024), and conservative DPO (Mitchell, 2023). Mixed-data comparisons also include judge-based relabeling and sequential DPO training on preferences for the paired criterion followed by helpfulness. Appendix E gives the implementations; the tables retain the reported token counts.

4.2 Results

Figure 3 summarizes the main comparisons; Tables 1 and 2 give the raw scores.

Pure contamination. DPO on the original labels reduces the paired rewards from their SFT levels, whereas CD-DPO raises safety to 0.999, honesty to 0.610, and truthfulness to 0.714, while preserving or improving the reported helpfulness scores. Blind reversal is competitive, with slightly higher safety and honesty scores. This is expected in the constructed pure regime, where every pair initially favors the unwanted response.

Alternative training objectives. The remaining tested baselines attain lower paired rewards than CD-DPO on pure contamination. Several partially improve rewards on mixed data, but remain below CD-DPO in the main comparisons. Appendix G derives the per-pair effect of fixed logit margins; their aggregate training effects are evaluated empirically.

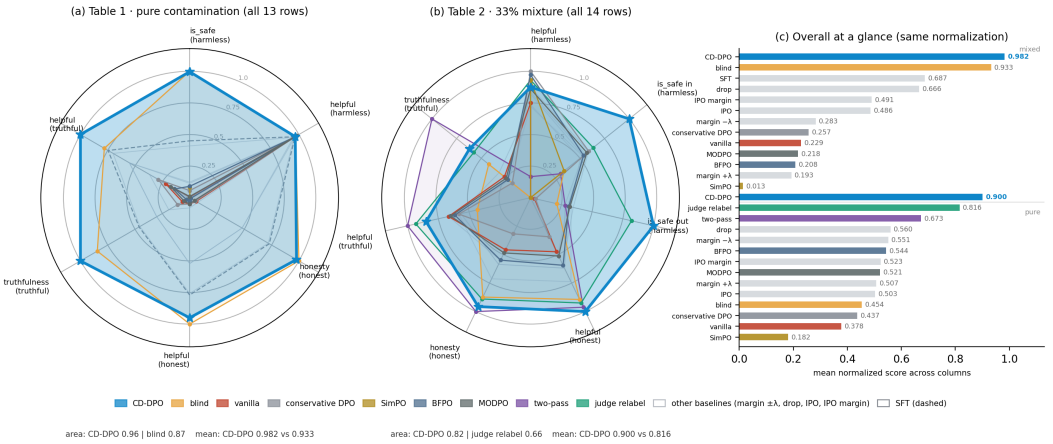


Figure 3: **Overview of the main results.** (a) Pure contamination. (b) Mixed data. Scores are min–max normalized separately within each metric column across the methods shown. (c) Equal-weight mean of these normalized scores. The normalized summaries depend on the chosen methods and metric columns; a larger polygon does not imply superiority on every reward. Raw values and evaluation-set differences appear in Tables 1 and 2. The mixed-data honesty comparison represents the main seed and fitting split; additional runs change its ranking (Appendix D).

Table 1: **Pure-contamination results.** Each pair initially favors the unwanted behavior. Scores are raw rewards; `is_safe` is unbounded and may exceed one. Its SimPO and conservative-DPO entries use 100 in-distribution prompts. SimPO uses $\gamma = 0.5$ and conservative DPO uses $\varepsilon = 0.1$. Token counts are reported across the three settings.

Method	harmless		honest		truthful		tokens
	is_safe	helpful.	honesty	helpful.	truthfulness	helpful.	
SFT (starting point)	0.465	0.709	0.530	0.648	0.641	0.629	0
DPO (original labels)	0.033	0.714	0.315	0.532	0.587	0.583	5.4M
margin $+\lambda$	0.026	0.727	0.301	0.529	0.579	0.573	5.4M
margin $-\lambda$	0.038	0.713	0.320	0.537	0.604	0.596	5.4M
drop	0.426	0.704	0.527	0.646	0.640	0.625	5.4M
blind (swap all)	1.002	0.711	0.617	0.685	0.693	0.632	5.4M
BFPO	0.114	0.704	0.299	0.528	0.582	0.576	5.4M
MODPO-style margin	0.035	0.717	0.310	0.532	0.585	0.579	5.4M
IPO	0.137	0.701	0.450	0.608	0.619	0.612	5.4M
IPO margin	0.146	0.710	0.450	0.608	0.619	0.612	5.4M
SimPO	0.105	0.221	0.296	0.524	0.578	0.564	5.4M
conservative DPO	0.081	0.719	0.313	0.532	0.593	0.589	5.4M
CD-DPO (relabel, ours)	**0.999**	**0.710**	**0.610**	**0.677**	**0.714**	**0.651**	**5.4M**

Mixed preference data. At 33% contamination, CD-DPO has higher helpfulness and paired-reward point estimates than blind reversal in all three main runs. The sequential baseline attains higher honesty (0.616 versus 0.603) and truthfulness (0.718 versus 0.675), but lower safety (0.369 versus 0.897), using twice the reported policy-training tokens. The ratio sweep supports selective correction at the measured mixed-data ratios (Appendix F). Figure 4 shows the mixed-data comparison on the original reward scales.

Sensitivity and cross-checks. Additional experiments examine seeds, thresholds, model scales, predictor-fitting choices, and independent evaluators (Table 3). The mixed-data honesty advantage over blind reversal is sensitive to both seed and fitting split: five further seeds improve over the original-label baseline but fall below blind reversal. Appendix D reports these variations alongside the main results.

Dataset-level check. An optional check applies correction when the fraction of positive predictor margins exceeds 0.35. At the default pairwise threshold, the reported calibration keeps it inactive on two clean datasets and activates it on the measured contaminated mixtures (Appendix A.1). This check is separate from the pairwise threshold τ.

Computation. With cached 19-dimensional reward vectors, the two dot products require 74 arithmetic operations per pair, followed by a subtraction and comparison. Reward evaluation and predictor fitting are preprocessing costs. Correction then uses one DPO run; the estimates in Appendix D distinguish these components from policy-training costs.

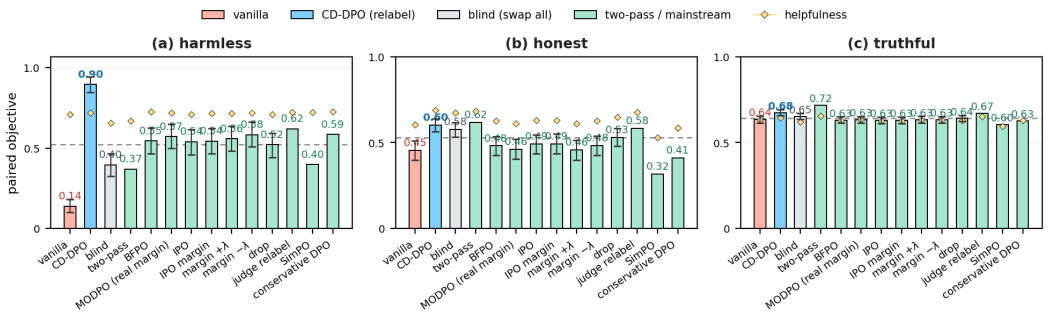


Figure 4: **Mixed-data comparisons on the original reward scales.** Bars show safety, honesty, or truthfulness for the 33% mixtures; gold diamonds show helpfulness, and dashed lines mark SFT values. Raw scores are given in Table 2. These are the main-run comparisons; the honesty ordering between CD-DPO and blind reversal varies across seeds and predictor-fitting splits (Appendix D).

Table 2: **Mixed-data results with 33% contamination.** Safety is evaluated on 100 in-distribution and 30 hand-written prompts. Entries are the main-run point estimates; honesty rankings vary with the training seed and predictor-fitting split. The token column includes additional judge tokens for API relabeling. Appendix D reports uncertainty and sensitivity.

Method	harmless			honest		truthful		tokens
	help.	is_safe (in)	is_safe (out)	help.	honesty	help.	truthfulness	
DPO (original labels)	0.710	0.139	0.035	0.606	0.454	0.634	0.635	5.4M
margin $+\lambda$	0.716	0.559	0.222	0.610	0.457	0.633	0.633	5.4M
margin $-\lambda$	0.719	0.584	0.255	0.625	0.482	0.634	0.632	5.4M
drop	0.709	0.521	0.243	0.648	0.530	0.629	0.641	5.4M
blind (swap all)	0.656	0.396	0.162	0.673	0.578	0.620	0.653	5.4M
safe-DPO $\rightarrow$ helpful-DPO	0.668	0.369	0.210	0.684	0.616	0.654	0.718	10.8M
BFPO	0.726	0.546	0.230	0.625	0.481	0.631	0.631	5.4M
MODPO-style margin	0.718	0.572	0.236	0.612	0.462	0.633	0.633	5.4M
IPO	0.708	0.540	0.207	0.629	0.492	0.628	0.630	5.4M
IPO margin	0.715	0.543	0.236	0.629	0.492	0.628	0.630	5.4M
judge relabel (API)	0.723	0.620	0.587	0.678	0.583	0.650	0.670	5.4M+2.4M
SimPO	0.723	0.397	0.013	0.530	0.317	0.594	0.605	5.4M
conservative DPO	0.728	0.586	0.041	0.585	0.412	0.629	0.626	5.4M
CD-DPO (selective, ours)	**0.719**	**0.897**	**0.712**	**0.690**	**0.603**	**0.645**	**0.675**	**5.4M**

5 Conclusion

Preference labels can encourage behaviors that undermine intended alignment qualities. CD-DPO uses a supervised predictor built from frozen reward scores to correct these orderings before DPO training. Our stylized analysis connects improved ordering accuracy to greater suppression of the target behavior at matched update counts and KL divergence. The main experiments show improvements over training on the original labels across harmful compliance, sycophancy, and hallucination settings. Mixed-data comparisons support selective correction, while the honesty results highlight sensitivity to predictor fitting and training seeds.

AI Use

Language models assisted with language editing and manuscript consistency checks. The authors are responsible for verifying the scientific content, references, and reported results.

Ethics Statement

This work does not involve human subjects or new data collection, and uses only publicly available models and datasets in accordance with their licenses. We are not aware of any ethical concerns beyond those common to LLM alignment research.

Reproducibility Statement

For our theoretical results, all assumptions are stated in the main text, and proofs are provided in the appendix. For our experiments, the main text and appendix describe the models, datasets, baselines, hyperparameters, and evaluation protocols.

REFERENCES

Arditi et al. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, et al. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Josef Dai, Xuehai Pan, Ruiyang Sun, et al. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Carson Denison, Monte MacDiarmid, Fazl Barez, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Nikolaos Dimitriadis, Pascal Frossard, and François Fleuret. Pareto low-rank adapters: Efficient multi-task learning with preferences. In *International Conference on Learning Representations (ICLR)*, 2025.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, et al. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

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

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Yue Huang et al. Prioritization first, principles second: An adaptive interpretation of helpful, honest, and harmless principles. *arXiv preprint arXiv:2502.06059*, 2025.

Jiho Jang, Jinyoung Kim, Kyungjune Baek, and Nojun Kwak. Multi-dimensional preference alignment by conditioning reward itself. *arXiv preprint arXiv:2512.10237*, 2026.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing rlhf with kl divergence does not mitigate heavy-tailed reward misspecification. *Advances in Neural Information Processing Systems*, 37:14608–14633, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, et al. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

Alisa Liu, Maarten Sap, Ximing Lu, et al. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yining Lu and Meng Jiang. Uncovering cross-objective interference in multi-objective alignment. *arXiv preprint arXiv:2602.06869*, 2026.

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

Eric Mitchell. A note on dpo with noisy preferences & relationship to ipo, 2023.

Christian Moya et al. Spurious correlation learning in preference optimization: Mechanisms, consequences, and mitigation via tie training. In *International Conference on Machine Learning (ICML)*, 2026.

Lilian Ngweta, Mayank Agarwal, Subha Maity, et al. Aligners: Decoupling llms and alignment. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, et al. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *International Conference on Machine Learning*, 2024a.

Ryan Park et al. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024b.

Eva Paunova. A taxonomy of multi-objective alignment techniques for large language models. *Optimization Online*, 2025.

Xiyue Peng, Hengquan Guo, Jiawei Zhang, et al. Enhancing safety in reinforcement learning with human feedback via rectified policy optimization. In *Advances in Neural Information Processing Systems*, 2025.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, et al. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024a.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Sathammai Sathappan. Do reward models encode what their labels claim? an interpretability audit of ArmoRM. LessWrong, 2026. URL https://www.lesswrong.com/posts/a4o6vZvuM2oz7XxyJ/do-reward-models-encode-what-their-labels-claim-an. May 23, 2026.

Shapira et al. How rlhf amplifies sycophancy. *arXiv preprint arXiv:2602.01002*, 2026.

Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. *International Conference on Learning Representations*, 2024.

Ruizhe Shi, Yifang Chen, Yushi Hu, et al. Decoding-time language model alignment with multiple objectives. In *Advances in Neural Information Processing Systems*, 2024.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. SALMON: Self-alignment with instructable reward models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xJbsmB8UMx.

Eshaan Tanwar and Pepa Atanasova. Understanding helpfulness and harmless tension in reward models. *arXiv preprint arXiv:2606.13209*, 2026.

Haoxiang Wang, Wei Xiong, Tengyang Xie, et al. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.

Jerry Wei, Da Huang, Yifeng Lu, et al. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Zeqiu Wu, Yushi Hu, Weijia Shi, et al. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, 2023.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *International Conference on Machine Learning (ICML)*, 2024.

Zhihao Xu et al. Reward consistency: Improving multi-objective alignment from a data-centric perspective. *arXiv preprint arXiv:2504.11337*, 2025.

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

Wenxuan Zhang, Philip H. S. Torr, Mohamed Elhoseiny, and Adel Bibi. Bi-factorial preference optimization: Balancing safety-helpfulness in language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Yifan Zhong et al. Panacea: Pareto alignment via preference adaptation for llms. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.02030.

Zhanhui Zhou, Jie Liu, Chao Yang, et al. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2024.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, et al. Weak-to-strong preference optimization: Stealing reward from weak aligned model. In *International Conference on Learning Representations (ICLR)*, 2025.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Reproducibility details and additional results

Appendix B discusses related work; Appendix C gives the proofs; Appendix D reports the discovery measurements and evaluation details; Appendix E describes the baseline implementations and sensitivity experiments; Appendix F reports the contamination sweep and additional evaluations; Appendix G analyzes fixed logit margins; and Appendix H collects independent evaluations and adversarial probes.

Ethics and reproducibility

Ethics. We evaluate responses to harmful prompts in a controlled research setting and release no harmful generations. The experiments examine how preference correction affects safety, honesty, and truthfulness. We use public open-weight models (Qwen2.5-0.5B/3B/7B and Mistral-7B) and public datasets (PKU-SafeRLHF, HH-RLHF, UltraFeedback, BeaverTails, TruthfulQA, and a factual-sycophancy benchmark); no private user data are used. Local training and evaluation use one NVIDIA H200 (about 140 GB), with an additional API-based judge. The reported local compute inventory is approximately 100 GPU-hours, dominated by the 7B and Mistral LoRA runs.

Predictor fitting and correction. Each behavioral predictor is fitted by ℓ_2-regularized logistic regression on 19-dimensional ArmoRM scores. The fitting sets contain 1,125 BeaverTails contrast pairs for harmful compliance, 112 factual-sycophancy pairs, and 1,125 TruthfulQA multiple-choice pairs. Responses receive behavioral annotations as described in Section 3.1; separate fitting-validation splits are used for the held-out AUCs. Algorithm 1 reverses a training pair when $\Delta_c = d_c(x, y_w) - d_c(x, y_l) > \tau$, where $\tau \geq 0$ is user-settable and defaults to zero. Two length-19 dot products require 74 multiplications and additions in total; subtraction and comparison add constant overhead. This count assumes that reward vectors have already been computed and excludes annotation, reward-model inference, and predictor fitting.

Policy training. The main experiments start from an Alpaca-only SFT checkpoint of Qwen2.5-3B and use $\beta = 0.1$, learning rate 5×10^{-6}, AdamW, gradient clipping at 1.0, gradient accumulation 4, and one epoch. Additional scales and architectures are described in Appendix H.2. Baseline objectives and adaptations are specified in Appendix E; the sequential baseline uses two policy-training passes. The table token counts summarize the reported training budgets rather than preprocessing costs.

Code and reproduction. The implementation uses `dpo_cd.py` for DPO and baseline training, `dpo_lora.py` for the larger-model replications, and `fit_dir_disjoint.py` for predictor fitting. Evaluation scripts include `eval_chain.py`, `eval_conflict.py`, the judge scripts, and `jailbreak_probe.py`; analysis scripts include `discriminator.py`, `residual_search_all.py`, and `helpful_pref_500.py`. The reported environment is Python 3.10, PyTorch 2.x, and HuggingFace `transformers`; the default training seed is zero. Evaluation-pair splits are described in Appendix D. TruthfulQA predictor-fitting and generation data have question-level overlap, so pair disjointness does not establish question-level independence.

A.1 Additional dataset-level check

An additional calibration check can be used to decide whether to apply Algorithm 1. It activates correction when the fraction of pairs with positive predictor margin, $|\mathcal{D}|^{-1} \sum_{(x,y_w,y_l)\in\mathcal{D}} \mathbb{I}\{\Delta_c > 0\}$, exceeds 0.35. This dataset-level threshold is distinct from the pairwise correction threshold τ. The analysis in Section 3.3 studies correction at $\tau = 0$ without this check.

In the reported harmful-compliance calibration, the positive-margin fractions are 0.326 on clean UltraFeedback and 0.336 on clean HH-RLHF, compared with 0.394, 0.513, 0.635, and 0.757 at contamination fractions 10%, 30%, 50%, and 70%, and 0.944 on pure contamination. Thus the 0.35 check separates these measured datasets at the default pairwise threshold. This calibration does not establish a general detector of contamination or guarantee that all clean pairs are retained.

For comparison, applying correction to clean UltraFeedback without the check reverses 32.6% of pairs. In that run, safety increases from 0.547 to 0.661, while helpfulness changes from 0.714 to 0.709. The dataset-level check would leave correction inactive on that dataset.

A.2 Additional view of the pure-contamination comparison

The following plot shows the pure-contamination results from Table 1 on the original reward scales. The mixed-data comparison appears in Figure 4 in the main text.

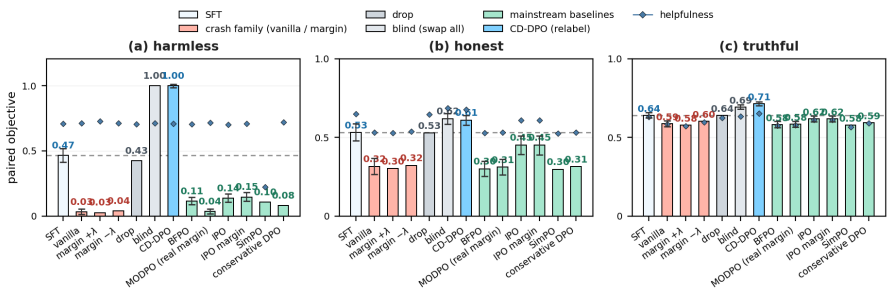


Figure 5: **Pure-contamination comparisons.** Bars show safety, honesty, or truthfulness in the corresponding setting; blue diamonds show helpfulness, and dashed lines mark SFT values. Reversing every pair is competitive with CD-DPO in this constructed regime. Raw values and evaluation-set qualifications are given in Table 1.

B Related Work

We situate CD-DPO among methods for balancing objectives, constructing preference data, and controlling model behavior. Appendix E specifies the configurations evaluated in our experiments.

B.1 Multi-objective RLHF and Pareto navigation

Safe RLHF separates helpfulness rewards from safety costs and optimizes a constrained objective (Dai et al., 2023). MODPO incorporates additional objectives through reward-derived margins (Zhou et al., 2024); GAPO balances objective gradients (Li et al., 2025); and RePO rectifies policy gradients to address prompt-level safety violations (Peng et al., 2025). Fine-Grained RLHF combines feedback types and reward signals at different text granularities (Wu et al., 2023). Panacea learns preference-conditioned policies through a low-rank singular-value parameterization (Zhong et al., 2024), while PaLoRA uses task-specific low-rank adapters for Pareto navigation in multi-task learning (Dimitriadis et al., 2025). Aligners instead trains inspectors and an aligner that revises model outputs (Ngweta et al., 2024); MOD combines model predictions during decoding (Shi et al., 2024). These methods address different optimization and control problems. Our question concerns whether the supplied preference ordering encourages an unwanted behavior, and whether a predictor fitted to existing reward scores can identify pairs to reorder.

B.2 Safety–helpfulness joint objectives and two-stage alignment

BFPO constructs a supervised objective incorporating helpfulness preferences and response safety labels (Zhang et al., 2025). Constitutional AI combines critique-and-revision supervised training with reinforcement learning from AI-generated preferences (Bai et al., 2022). SALMON trains an instructable reward model from principles (Sun et al., 2024), and SteerLM conditions supervised training on response attributes (Dong et al., 2023). WSPO transfers alignment through the distributional change between a weak model and its aligned counterpart (Zhu et al., 2025). Qi et al. show that safety alignment can be concentrated in initial response tokens and study methods for making it more robust (Qi et al., 2024a). Our experiments separately evaluate SFT on ideal responses and sequential DPO on different preference datasets; these baselines test those concrete training choices.

B.3 Data-side relabeling and preference pathologies

Reward Consistency Sampling constructs preference datasets using consistency across reward objectives (Xu et al., 2025), while RLAIF obtains preference feedback from AI systems (Lee et al., 2023). Our corresponding baselines adapt reward-consistency and judge-based criteria to the existing training pairs. CD-DPO uses a behavior-specific predictor fitted from annotations to select pairs for reversal. Related work documents sycophancy (Sharma et al., 2024; Shapira et al., 2026; Wei et al., 2023), spurious reward correlations and failures under proxy optimization (Moya et al., 2026; Kwa et al., 2024; Gao et al., 2023), safety degradation following fine-tuning (Qi et al., 2024b), and generalization from sycophancy to reward tampering (Denison et al., 2024). These findings motivate examining what preference labels incentivize in addition to how an optimizer fits them.

B.4 DPO VARIANTS AND MARGIN CORRECTIONS

IPO uses a squared preference objective (Azar et al., 2024); KTO supports binary desirability feedback (Ethayarajh et al., 2024); CPO studies contrastive preference optimization for translation (Xu et al., 2024); and SimPO uses length-normalized log probabilities and a target reward margin (Meng et al., 2024). Conservative DPO introduces preference-label smoothing (Mitchell, 2023). MCDPO conditions diffusion-model preference optimization on multiple preference dimensions (Jang et al., 2026), while self-rewarding language models iteratively generate their own feedback for DPO (Yuan et al., 2024). Appendix G gives a per-pair derivative analysis for a fixed additive DPO margin. That observation does not determine the aggregate effect of reweighting pairs or establish a general limitation of loss-side methods.

B.5 REWARD-MODEL AUDITS AND MEASUREMENT

Paunova surveys multi-objective alignment techniques (Paunova, 2025). Sathappan examines whether ArmoRM heads track their named constructs (Sathappan, 2026), and Park et al. study length bias in DPO (Park et al., 2024b). Work on linear representations provides a separate account of concept geometry (Park et al., 2024a), while contextual analyses find settings in which helpfulness and harmlessness improve jointly (Huang et al., 2025). We examine reward-score structure, behavior-specific preference interventions, and the use of fitted score combinations to correct training labels.

B.6 MECHANISM-LEVEL AND STEERING ACCOUNTS

Tanwar et al. study neuronal overlap and interference in reward models (Tanwar & Atanasova, 2026); Lu et al. analyze covariance-dependent interference between alignment objectives (Lu & Jiang, 2026). Refusal-direction interventions, contrastive activation addition, and inference-time intervention act on model activations (Arditi et al., 2024; Panickssery et al., 2024; Li et al., 2023). DoLa contrasts layer predictions for factuality (Chuang et al., 2024), and DExperts combines expert and anti-expert predictions during decoding (Liu et al., 2021). CD-DPO instead uses a direction in reward-score space to decide which training pairs to reverse, leaving the resulting policy's inference procedure unchanged.

C PROOFS AND ROBUSTNESS OF PREDICTOR-BASED CORRECTION

This appendix gives the details deferred from Section 3.3. Sections C.1–C.2 prove Theorem 1, including the DPO loss, the common sequence t_k, and the effect of different preference data. Sections C.3–C.4 control training errors. Sections C.5–C.6 treat the example, mixtures, and extensions; Section C.7 characterizes when fitting achieves $\chi > \gamma$.

Let $T = e^\top X$ for any unit e, with $X \sim Q$. Rotational invariance makes its distribution independent of e. Write

$$B = \frac{\beta}{2}\mathbb{E}|T|, \qquad \kappa = \frac{1}{2}\mathbb{E}|T|, \qquad F(t) = \mathbb{E}\log\cosh(\beta t T/2), \qquad L = \frac{\beta^2}{4}\mathbb{E}T^2 = \frac{\beta^2}{4p}\mathbb{E}\|X\|^2.$$

The radial increment is $g(t) = B - F'(t)$, as derived below. For auxiliary fixed-KL comparisons, put $\mathcal{C}_\pm = A^\top \mathcal{W}_\pm$ and

$$V_\pm(K_0) = \sup_{\theta \in \mathcal{C}_\pm :\, D_{\mathrm{KL}}(P_\theta \| P_0) \le K_0} J(\theta).$$

C.1 POLICY GEOMETRY AND THE SCALARIZATION BENCHMARK

Policy, rewards, and auxiliary attribute. We use the continuous response model stated in Section 3.3. The fixed prompt is suppressed. The model assumes $r_j(z) = a_j^\top z$, collecting the reward coefficients as the rows of A; this is an analytical description, not an additional input required by CD-DPO. In addition to the unit hidden-behavior direction v, choose a unit $u \perp v$ for the auxiliary attribute in the two-reward example. We use

$$q(z) = u^\top z, \qquad c(z) = v^\top z, \qquad s(z) = -c(z), \qquad r(z) = Az. \tag{13}$$

Thus $J(\theta) = \mathbb{E}_{P_\theta} s = -\mathbb{E}_{P_\theta} c$. The main behavioral comparison does not require u; orthogonality is used only for the accompanying preservation claim. The reference covariance has the form $\sigma_0^2 I$ with $\sigma_0 > 0$. Rotation invariance does not require a density: bounded spherical reference distributions are also allowed.

For completeness, Q describes differences of fixed response pairs. It need not be the difference law of two independent draws from P_0; it must satisfy the comparison assumptions in Section 3.3. For any two valid responses to the fixed prompt, the policy ratio obeys

$$\log \frac{dP_\theta}{dP_0}(z) - \log \frac{dP_\theta}{dP_0}(z') = [\theta^\top z - \Psi(\|\theta\|)] - [\theta^\top z' - \Psi(\|\theta\|)] = \theta^\top (z - z').$$

Writing the pair in an arbitrary order and using $Y \in \{-1, 1\}$ for the preference, the chosen–rejected margin is $Y\theta^\top X$. Substitution into (1) gives exactly

$$\mathcal{L}_Y(\theta) = \mathbb{E}_{X,Y} \log(1 + e^{-\beta Y \theta^\top X}). \tag{14}$$

This cancellation uses a common prompt and reference policy, not rotational symmetry. In the special case $P_0 = \mathcal{N}(0, I)$, completing the square gives $P_\theta = \mathcal{N}(\theta, I)$, $\Psi(t) = t^2/2$, $M(t) = t$, and $K(t) = t^2/2$.

Lemma C.1 (Exact policy response and fixed-KL comparison)**.** *Under the reference assumptions in Section 3.3, $M(0) = 0$, $M'(t) > 0$, and $K'(t) = tM'(t) > 0$ for $t > 0$. For $\theta \neq 0$,*

$$\mathbb{E}_{P_\theta} Z = M(\|\theta\|)\frac{\theta}{\|\theta\|}, \qquad D_{\mathrm{KL}}(P_\theta \| P_0) = K(\|\theta\|). \tag{15}$$

For every unit d and $R > 0$,

$$\sup_{P:\, D_{\mathrm{KL}}(P\|P_0) \le K(R)} \mathbb{E}_P d^\top Z = M(R), \tag{16}$$

with optimizer P_{Rd}. The supremum is over laws absolutely continuous with respect to P_0 with finite first moment. Moreover,

$$V_\pm(K(R)) = \gamma M(R), \qquad \mathrm{Corr}_{P_0}(r_i, r_j) = \frac{a_i^\top a_j}{\|a_i\|\|a_j\|}. \tag{17}$$

Proof. Rotation invariance makes the log moment generating function radial. Differentiation gives the mean identity in (15); it is justified by finiteness of the moment generating function in a neighborhood of each finite parameter. Along a ray,

$$M'(t) = \mathrm{Var}_{P_{td}}(d^\top Z) > 0.$$

The projection is nonconstant because P_0 is nondegenerate and spherical, and finite exponential tilting preserves null sets. Symmetry gives $M(0) = 0$. Substitution into the log density ratio gives the KL identity and $K'(t) = tM'(t)$.

For any admissible P, nonnegativity of its relative entropy to P_{Rd} gives

$$0 \le D_{\mathrm{KL}}(P\|P_{Rd}) = D_{\mathrm{KL}}(P\|P_0) - R\mathbb{E}_P d^\top Z + \Psi(R).$$

Thus the budget $K(R)$ implies $\mathbb{E}_P d^\top Z \le (K(R) + \Psi(R))/R = M(R)$; equality holds at P_{Rd}. This also shows that optimizing any nonconstant linear scalarization $a^\top z$ at that budget yields the policy in its normalized direction.

The cone generated by finitely many rows of A is closed; its union with its negative is also closed. Therefore the normalized attainable directions form a compact set and the maximum over weights in (10) is well defined. Its intersection with the unit sphere is nonempty and compact, so the maximum defining γ is attained. The union is sign-symmetric, hence $\gamma \ge 0$ and $\max_{a \in \mathcal{C}_\pm, \|a\|=1} (-v)^\top a = \gamma$. The monotonicity of K bounds a feasible radius by R; the definition of γ and monotonicity of M then give $V_\pm(K(R)) = \gamma M(R)$. Finally, reference covariance is $\sigma_0^2 I$ with $\sigma_0 > 0$, proving the correlation formula. If an exact reconstruction $A^\top w_* = v$ exists, $\gamma < 1$ forces it to have both positive and negative coefficients. The preceding identities do not require such a reconstruction to exist. □

C.2 From the ordering to the full DPO trajectory

Lemma C.2 (Sign information and global ray dynamics)**.** *For unit d, d', the spherical comparison law satisfies*

$$\tfrac{1}{2}\mathbb{E}[\mathrm{sign}(d^\top X)X] = \kappa d, \qquad \mathbb{P}\{\mathrm{sign}(d^\top X) \neq \mathrm{sign}(d'^\top X)\} = \frac{\arccos(d^\top d')}{\pi}. \tag{18}$$

For arbitrary labels, put $m_Y = \mathbb{E}[YX]/2$. The loss has the exact decomposition

$$\mathcal{L}_Y(\theta) = \log 2 + F(\|\theta\|) - \beta m_Y^\top \theta. \tag{19}$$

Its even part has L-Lipschitz gradient, $F'(0) = 0$, $F''(t) > 0$, and $F'(t) \uparrow B$. If $m_Y = \kappa \ell d$ with $0 < \ell \le 1$, its iterates are $\theta_k = t_{\ell,k} d$, where

$$t_{\ell,0} = 0, \qquad t_{\ell,k+1} = t_{\ell,k} + \eta\{B\ell - F'(t_{\ell,k})\}. \tag{20}$$

For deterministic labels $Y = \operatorname{sign}(d^\top X)$, $\ell = 1$ and the common sequence $t_k = t_{1,k}$ satisfies

$$t_0 = 0, \qquad t_{k+1} = t_k + \eta g(t_k), \qquad g(t) = \beta \mathbb{E}\frac{|T|}{1 + e^{\beta t|T|}} = B - F'(t). \tag{21}$$

For $\ell = 1$ these radii increase strictly to infinity. For $\ell < 1$ they increase to the unique solution of $F'(t) = B\ell$, and $0 < t_{\ell,k} \le t_{1,k}$ for $k \ge 1$.

Proof. Every orthogonal transformation fixing d leaves $\mathbb{E}[\operatorname{sign}(d^\top X)X]$ unchanged; the expectation is therefore parallel to d. Taking its inner product with d gives the first identity in (18). For the second, project onto the plane spanned by d, d'. Its angle is uniform, and the two disagreeing wedges occupy total angle $2\arccos(d^\top d')$ out of 2π. The zero-projection event has probability zero. Parallel and antiparallel directions are immediate.

Use, without approximation,

$$\log(1 + e^{-z}) = \log 2 + \log\cosh(z/2) - z/2.$$

Since the middle term is even, it does not depend on Y. Rotation invariance makes its expectation $F(\|\theta\|)$, proving (19). Direct differentiation gives

$$F'(t) = \frac{\beta}{2}\mathbb{E}[T\tanh(\beta t T/2)], \qquad F''(t) = \frac{\beta^2}{4}\mathbb{E}[T^2\operatorname{sech}^2(\beta t T/2)] > 0.$$

Finite second moment justifies differentiation, and dominated convergence gives $F'(t) \uparrow \beta\mathbb{E}|T|/2 = B$. The vector Hessian is

$$\nabla^2(F \circ \|\cdot\|)(\theta) = \frac{\beta^2}{4}\mathbb{E}[XX^\top \operatorname{sech}^2(\beta\theta^\top X/2)] \preceq \frac{\beta^2}{4}\mathbb{E}[XX^\top] = LI.$$

Consequently $0 \le F'(t) \le Lt$. The identity $1 - \tanh(x/2) = 2/(1 + e^x)$ for $x \ge 0$ gives the formula for g in (21).

When $m_Y = \kappa\ell d$, the gradient is $F'(\|\theta\|)\theta/\|\theta\| - B\ell d$, with the first term zero at the origin. Induction yields (20). For $\ell = 1$, its increment is $\eta g(t) > 0$. A finite limit would require $g(t) = 0$ at a finite t, which is impossible. For $\ell < 1$, F' increases from zero to B and hence has a unique root t_ℓ^* for $F'(t) = B\ell$. The update map $T_\ell(t) = t + \eta\{B\ell - F'(t)\}$ is nondecreasing because $\eta F'' \le 1$. It preserves $[0, t_\ell^*]$ and increases the radius until convergence to t_ℓ^*. Finally, $T_\ell \le T_1$ and T_1 is nondecreasing, so induction proves $t_{\ell,k} \le t_{1,k}$. These statements use the full logistic gradient at every iterate. □

Effect of a preference category. Let ℓ index a type of preference data. A latent utility $U_\ell(z) = b_\ell^\top z$, with $b_\ell \neq 0$, models its supplied ordering by $Y_\ell = \operatorname{sign}(U_\ell(z) - U_\ell(z')) = \operatorname{sign}(b_\ell^\top X)$. This utility is used only in the analysis: DPO receives the labeled pairs. Lemma C.2 gives $m_{Y_\ell} = \kappa b_\ell/\|b_\ell\|$. Together with Lemma C.1, this yields

$$\theta_k^{(\ell)} = t_k \frac{b_\ell}{\|b_\ell\|}, \qquad \mathbb{E}_{P_{\theta_k^{(\ell)}}} r_j = M(t_k)\frac{a_j^\top b_\ell}{\|b_\ell\|}. \tag{22}$$

The sign of $a_j^\top b_\ell$ determines the sign and monotonicity of the reward change, since t_k and M increase strictly. In contrast, reward–reward correlations are determined by $a_i^\top a_j$ through (17). Choosing $b_\ell = v$ models preferences favoring the hidden behavior and gives $J(\theta_k^{(\ell)}) = -M(t_k) < 0$ for $k \ge 1$. Positive reward correlations therefore do not imply that every preference category improves those rewards. These utilities describe annotations in the model; they do not identify the experimental labels with rankings generated by the evaluation rewards.

Corrected preferences. The fitted response predictor is $d_c(z) = (A^\top \widehat{w}_c)^\top z$. With $\tau = 0$, applying (3) to every pair makes the response with the lower predictor output preferred. If $A^\top \widehat{w}_c \neq 0$, this gives

$$\widehat{v} = h(\widehat{w}_c), \qquad Y_{\mathrm{CD}} = -\operatorname{sign}(\widehat{v}^\top X). \tag{23}$$

Ties have probability zero by rotational invariance of Q and $Q(X = 0) = 0$. The original label is retained for ties in the algorithm. This argument analyzes zero-threshold correction without a dataset-level check.

Proof of Theorem 1. For a nonconstant predictor with weights w, put $d_w = A^\top w / \|A^\top w\|$. Corrected labels have moment $-\kappa d_w$, so Lemma C.2 gives $\theta_k(w) = -t_k d_w$. In particular, $\theta_k^{\mathrm{CD}} = -t_k \widehat{v}$ and $J(\theta_k^{\mathrm{CD}}) = \chi M(t_k)$. All predictor-corrected runs share the radius t_k and KL divergence $K(t_k)$, since the pair distribution and step size are the same.

For $w \in \mathcal{W}_\pm$, the normalized directions $-d_w$ range over $\mathcal{C}_\pm \cap \mathbb{S}^{p-1}$: the cone is closed under either overall sign, and every unit vector in it has a preimage in $\mathcal{W}_\pm$. Hence

$$\mathcal{B}_\pm(k) = M(t_k) \sup_{w \in \mathcal{W}_\pm :\, A^\top w \neq 0} v^\top d_w = \gamma M(t_k) = V_\pm(K(t_k)),$$

where the last equality is Lemma C.1. Subtracting proves (12). Exact recovery gives the auxiliary-attribute preservation claim from $u \perp v$ and the unrestricted optimum from (16). With $p_{\mathrm{ord}} = \mathbb{P}_Q\{\operatorname{sign}(\widehat{v}^\top X) \neq \operatorname{sign}(v^\top X)\}$, Lemma C.2 gives $\chi = \cos(\pi p_{\mathrm{ord}})$. It also shows that any nonconstant same-sign predictor has ordering error at least $\arccos(\gamma)/\pi$. This proves Theorem 1. Every predictor uses zero-threshold correction without the dataset-level check, followed by the same DPO training. The comparison therefore makes no assertion about the decision to activate correction in an experiment. □

C.3 Robustness to gradient errors

Lemma C.3 (Invariant cone and finite progress)**.** *Fix unit d and consider the unprojected update*

$$\theta_{k+1} = \theta_k - \eta\{\nabla(F \circ \|\cdot\|)(\theta_k) - Bd + \xi_k\}, \qquad \theta_0 = 0. \tag{24}$$

Suppose $\|\xi_k\| \le \delta_R \le g(R)/4$ while $\|\theta_k\| < R$, and $\eta \le 1/L$. Set $e = \delta_R/B$ and $c_e = \sqrt{1-e^2}$. Up to and including the first crossing of R,

$$d^\top \theta_k \ge c_e \|\theta_k\|. \tag{25}$$

Writing $\tau_R = \min\{k \ge 1 : \|\theta_k\| \ge R\}$, the crossing and overshoot bounds are

$$\tau_R \le \left\lceil \frac{2R}{\eta g(R)} \right\rceil, \qquad R \le r_R := \|\theta_{\tau_R}\| < R + \eta(B + \delta_R). \tag{26}$$

Proof. Every vector in the ball of radius $\delta_R < B$ around Bd makes angle at most $\arcsin(\delta_R/B)$ with d. One way to verify this is to minimize $d^\top z/\|z\|$ on the ball: tangency of a ray to the ball gives the minimum $\sqrt{1-(\delta_R/B)^2}$. Equivalently, minimizing along each ray gives distance $B \sin \angle(z, d)$ from Bd.

The second-order cone $\{z : d^\top z \ge c_e \|z\|\}$ is closed under nonnegative sums and scaling. For $r = \|\theta_k\| > 0$, the update is exactly

$$\theta_{k+1} = \left(1 - \eta \frac{F'(r)}{r}\right) \theta_k + \eta(Bd - \xi_k). \tag{27}$$

The old-iterate coefficient lies in $[0, 1]$ because $F'(r)/r \le L$. At zero its contribution vanishes. Both terms therefore belong to the cone whenever the old iterate does, proving (25) by induction, including the crossing step.

For $r > 0$, projection of (27) onto θ_k/r gives

$$\begin{aligned} \|\theta_{k+1}\| - r &\ge \eta\{Bc_e - F'(r) - \delta_R\} \\ &\ge \eta\{g(R) - B(1 - c_e) - \delta_R\} \ge \eta\{g(R) - 2\delta_R\} \ge \eta g(R)/2. \end{aligned}$$

We used $1 - \sqrt{1-e^2} \le e$ for $0 \le e \le 1$. At zero, the next radius is at least $\eta(B - \delta_R) \ge \eta g(R)/2$. A run cannot remain below R for more than $\lceil 2R/(\eta g(R)) \rceil$ steps. Since the coefficient in (27) is at most one, the crossing radius is strictly below $R + \eta(B + \delta_R)$. □

Policy comparison under errors. Fix a finite $R > 0$ and suppose corrected GD uses $\nabla\mathcal{L}_{Y_{\mathrm{CD}}}(\theta_k) + \xi_k$, with $\|\xi_k\| \le \delta_R \le g(R)/4$ whenever $\|\theta_k\| < R$. With τ_R, r_R as in (26), let $e_R = \delta_R/B$ and $\alpha = \arccos\chi$. If $0 \le \chi \le 1$, then

$$J(\theta_{\tau_R}) - V_{\pm}(K(r_R)) \ge M(r_R)\big[\cos(\alpha + \arcsin e_R) - \gamma\big]. \tag{28}$$

Moreover, $|\mathbb{E}_{P_{\theta_{\tau_R}}} q| \le M(r_R)(\sin\alpha + e_R)$. The comparison uses the perturbed checkpoint's attained KL divergence, rather than asserting equal step counts for perturbed and exact runs.

Proof. Use Lemma C.3 with $d = -\widehat{v}$. At the crossing, let $n = \theta_{\tau_R}/r_R$ and write $n = td + n_\perp$, with $t \ge \sqrt{1 - e_R^2}$ and $\|n_\perp\| \le e_R$. Decompose $-v = \chi d + v_\perp$, where $\|v_\perp\| = \sqrt{1-\chi^2}$. For $\chi \ge 0$,

$$(-v)^\top n \ge \chi\sqrt{1 - e_R^2} - \sqrt{1-\chi^2}\, e_R = \cos\{\arccos\chi + \arcsin e_R\}.$$

Use Lemma C.1 and subtract $\gamma M(r_R)$ to prove (28). Since $u \perp v$, $|u^\top d| \le \sqrt{1-\chi^2}$, whence $|u^\top n| \le \sqrt{1-\chi^2} + e_R$. This proves the bound for the auxiliary attribute q. □

Relation to direct scalarization. Lemma C.1 shows that $V_{\pm}(K)$ also equals the largest behavioral reward among policies obtained by exactly maximizing a nonconstant same-sign linear reward combination at KL budget K. In the spherical model, these policies lie on the same rays as the predictor-corrected DPO policies, and Theorem 1 establishes equality of the two benchmarks at $K(t_k)$. This is a consequence of the stylized geometry, not an additional experiment or a claim that the two procedures are computationally identical. Direct optimization of the same unrestricted recovered reward remains a valid alternative to CD-DPO.

C.4 Finite-sample certificates for the two errors

The theorem quantifies fitting error through the recovered ordering, rather than assuming that logistic regression recovers a calibrated latent utility. The certificate applies to the response-level fit in Section 3.1 and to other fitting procedures. It requires independent validation of the intended ordering, not comparison with the predictor's own labels.

Lemma C.4 (Independent validation and empirical-DPO errors)**.** *Fix $0 < \zeta < 1$ and condition on the fitted predictor. Let n independent validation pairs from Q have externally supplied hidden-order labels. If their empirical ordering error is $\widehat{p}$, define*

$$p_{\mathrm{up}} = \min\left\{1, \widehat{p} + \sqrt{\frac{\log(2/\zeta)}{2n}}\right\}. \tag{29}$$

Suppose $\|X\| \le D$ under both Q and an actual training law $\widetilde{Q}$. Given the predictor, let N independent training comparisons from $\widetilde{Q}$ be relabeled by (23). Define

$$\delta_{\mathrm{shift}}(R) = \sup_{\|\theta\|\le R} \|\nabla\mathcal{L}_{\widetilde{Q},\mathrm{CD}}(\theta) - \nabla\mathcal{L}_{Q,\mathrm{CD}}(\theta)\|.$$

For any $\delta_N > 0$, if

$$N \ge \frac{32\beta^2 D^2}{\delta_N^2}\left[p\log 5 + p\log\left(1 + \frac{2\beta^2 D^2 R}{\delta_N}\right) + \log\frac{4}{\zeta}\right], \tag{30}$$

then, with probability at least $1 - \zeta$, both

$$p_{\mathrm{ord}} \le p_{\mathrm{up}}, \qquad \sup_{\|\theta\|\le R} \|\nabla\widehat{\mathcal{L}}_N(\theta) - \nabla\mathcal{L}_{Q,\mathrm{CD}}(\theta)\| \le \delta_N + \delta_{\mathrm{shift}}(R) \tag{31}$$

hold. Any uniformly bounded additional optimizer error δ_{opt} can be added. Thus empirical full-batch training satisfies the robustness bound (28) with

$$\delta_R = \delta_N + \delta_{\mathrm{shift}}(R) + \delta_{\mathrm{opt}}.$$

If $p_{\mathrm{up}} \le 1/2$, sufficient conditions for a positive gap are

$$\delta_R \le g(R)/4, \qquad \cos(\pi p_{\mathrm{up}} + \arcsin(\delta_R/B)) > \gamma. \tag{32}$$

A sufficient step-size choice is $\eta \le 4/(\beta^2 D^2)$.

Proof. Conditional on the fitted predictor, independent validation errors are Bernoulli with mean p_{ord}. The one-sided Hoeffding inequality gives the first assertion with failure probability $\zeta/2$.

A single relabeled gradient is $G_\theta(X) = -\beta Y X/(1 + e^{\beta Y \theta^\top X})$. It has norm at most βD and is L_s-Lipschitz in θ, where $L_s = \beta^2 D^2/4$. The empirical–population gradient difference is therefore $2L_s$-Lipschitz. A ball of radius R in $\mathbb{R}^p$ has a net of mesh $\delta_N/(4L_s)$ with at most $(1 + 8RL_s/\delta_N)^p$ points, by packing disjoint balls of half that mesh. The unit sphere has a $1/2$-net of at most 5^p points. For any vector z, one of the latter directions has inner product at least $\|z\|/2$ with z.

At each pair of net points, a scalar gradient projection lies in $[-\beta D, \beta D]$. Hoeffding gives

$$\mathbb{P}\left\{\left|a^\top(\nabla\widehat{\mathcal{L}}_N - \nabla\mathcal{L}_{\widetilde{Q},\mathrm{CD}})\right| > \delta_N/4\right\} \le 2\exp\left(-\frac{N\delta_N^2}{32\beta^2 D^2}\right).$$

The union bound and (30) make all parameter-net vector errors at most $\delta_N/2$ with failure probability $\zeta/2$. Moving to any parameter in the ball adds at most $2L_s \cdot \delta_N/(4L_s) = \delta_N/2$. Adding the shift and optimizer errors gives (31). No independence between an iterate and the reused training data is needed because the bound is uniform over the parameter ball. Conditional independence of the predictor from the policy-training data permits removal of the conditioning.

For completeness, both uses of Hoeffding follow from the bounded-variable exponential bound $\mathbb{E}e^{t(U-\mathbb{E}U)} \le e^{t^2(b-a)^2/8}$ for $U \in [a, b]$, followed by the exponential Markov inequality and minimization over $t > 0$. Finally $\chi = \cos(\pi p_{\mathrm{ord}})$ and monotonicity of cosine on the relevant interval imply (32). The stated step size ensures $\eta L \le 1$. □

This is a certificate conditional on a useful fitted ordering, not a theorem that a prescribed number of response annotations always produces one. It does not substitute response-level AUC for pair-ordering error. Different validation and deployment distributions require a transfer term; for example, total-variation distance τ adds at most τ to the ordering-error bound. The reference law in (5) remains spherical even when $\delta_{\mathrm{shift}}(R)$ permits nonspherical training comparisons.

C.5 Mixtures, the concrete example, and the limits of the guarantee

Why blind reversal loses quality on mixtures. Let clean and hidden sources have spherical difference laws Q_C, Q_H with projected magnitudes κ_C, κ_H. Let ρ be the hidden-source fraction. Clean labels prefer $a_C = \sqrt{1-h^2}\,u - hv$, where $0 \le h < 1$, while hidden labels prefer v. The pooled law is spherical and

$$m_V = (1-\rho)\kappa_C a_C + \rho\kappa_H v = \kappa(au + bv),$$

where

$$\kappa = (1-\rho)\kappa_C + \rho\kappa_H, \quad \omega = \rho\kappa_H/\kappa, \quad a = (1-\omega)\sqrt{1-h^2}, \quad b = \omega - (1-\omega)h.$$

Put $\ell = \sqrt{a^2+b^2}$. For $\rho < 1$, $a > 0$ and $0 < \ell \le 1$. Lemma C.2 gives

$$\theta_k^V = t_{\ell,k}(au + bv)/\ell, \qquad \theta_k^{\mathrm{blind}} = -\theta_k^V, \qquad \theta_k^{\mathrm{CD}} = -t_{1,k}v$$

for ideal correction. Vanilla decreases s exactly when $b > 0$, or

$$\rho > \frac{h\kappa_C}{\kappa_H + h\kappa_C}.$$

Blind reversal has mean auxiliary attribute $-aM(t_{\ell,k})/\ell < 0$, compared with zero for ideal correction. Its behavioral reward is $bM(t_{\ell,k})/\ell$. This is below $M(t_{1,k})$: if $b \le 0$ it is immediate; otherwise $b/\ell < 1$ and $t_{\ell,k} \le t_{1,k}$. Thus ideal correction exceeds blind reversal in both coordinates at the same positive step count whenever $\rho < 1$. At $\rho = 1$ they coincide. This conclusion does not imply that the zero-threshold rule retains all clean labels, or that every noisy neural run has the same ranking.

Verification of the bounded example. With q and c from (13), the example in the main text is

$$r_1 = q - \epsilon c, \qquad r_2 = q - 2\epsilon c, \qquad c = \frac{r_1 - r_2}{\epsilon}, \qquad \epsilon > 0. \tag{33}$$

For (33), a nonnegative combination has direction $au - \epsilon bv$ with $a \ge 0$ and $a \le b \le 2a$. Its largest cosine with $-v$ is attained by the second reward function and equals $2\epsilon/\sqrt{1+4\epsilon^2}$. The

all-nonpositive orientation does not increase this value. Lemma C.1 supplies the correlation formula. Under the uniform law on $\mathbb{S}^2$, one coordinate is uniform on $[-1, 1]$, so its moment generating function is $\sinh t/t$, yielding $M(t) = \coth t - 1/t$. Independent draws of the two endpoints give one admissible spherical comparison law. With $\epsilon = 1/10$, $\gamma = 1/\sqrt{26} < 1/5$; for $t \geq 5$, $4/5 < M(t) < 1$. With exact recovery, the policy advantage is $(1-\gamma)M(t) > 16/25$ at every $t \geq 5$. Lemma C.2 ensures that GD reaches such a radius after finitely many updates.

The intended mean set $(\mathbb{E}q, \mathbb{E}s)$ at budget $K(R)$ is the disk of radius $M(R)$. The entropy bound gives containment; exponential tilts attain every boundary direction, and mixtures with P_0 attain the interior because KL is convex in its first argument. Thus the upper-right arc is a nonempty Pareto frontier. CD-DPO reaches its behavioral endpoint, not an entire frontier sweep. The exact cardinal reconstruction of c uses coefficients $(1, -1)/\epsilon$; the theorem does not guarantee statistically accurate fitting of a weak or noisy residual at a fixed annotation budget.

Why the error condition depends on the training range. If each ideal corrected label is independently reversed with probability $\varepsilon \in (0, 1/2)$, the even loss is unchanged and its linear signal is multiplied by $1 - 2\varepsilon$. Lemma C.2 shows that training converges to the finite radius solving $g(R_\varepsilon) = 2\varepsilon B$. Hence persistent error can prevent further progress. the robustness bound (28) does not imply that longer training always improves behavior, nor does this example identify the cause of a particular experimental decline.

C.6 Relaxing recovery and feature assumptions

This subsection gives extensions of the theorem, rather than assumptions verified by the experiments. The exact comparison in the main text isolates the mechanism. Below we separate assumptions needed for exact identities from those that can be replaced by explicit error terms.

Incomplete recovery from the available rewards. Let $\mathcal{V} = \mathrm{range}(A^\top)$ and let $\Pi_\mathcal{V}$ be Euclidean orthogonal projection onto this subspace. Theorem 1 does not require $v \in \mathcal{V}$: its formulas depend on the actual fitted direction through $\chi = v^\top \widehat{v}$. In particular, putting $\rho_\mathcal{V} = \|\Pi_\mathcal{V} v\|$ gives

$$\max_{d \in \mathcal{V},\, \|d\|=1} v^\top d = \rho_\mathcal{V}, \qquad \chi = \rho_\mathcal{V} \cos\phi \quad (\rho_\mathcal{V} > 0), \tag{34}$$

where ϕ is the angle between $\widehat{v}$ and $\Pi_\mathcal{V} v/\rho_\mathcal{V}$. Consequently the exact separation becomes $(\rho_\mathcal{V} \cos\phi - \gamma)M(t_k)$. A strict gain is possible even when $\rho_\mathcal{V} < 1$, provided the fit attains $\rho_\mathcal{V} \cos\phi > \gamma$. This is a representational condition, not a guarantee that the logistic fitting procedure finds that direction.

Proof. For $d \in \mathcal{V}$, $v^\top d = (\Pi_\mathcal{V} v)^\top d$. Cauchy–Schwarz gives the maximum, attained in the projected direction when it is nonzero. Applying the identity to $\widehat{v}$ gives the second formula. All policy, ordering, and trajectory identities in Lemmas C.1–C.3 remain valid for an arbitrary unit v; exact recovery is used only for conclusions explicitly conditional on $\widehat{v} = v$. □

Quality and behavior need not be orthogonal. The behavioral comparison does not use u. For any unit quality direction u, the exact trajectory gives

$$\mathbb{E}_{P_{\theta_k^{\mathrm{CD}}}} q - \mathbb{E}_{P_0} q = -(u^\top \widehat{v}) M(t_k). \tag{35}$$

In particular, ideal correction changes quality by $-(u^\top v)M(t_k)$. Thus orthogonality is needed for exact preservation of reference mean auxiliary attribute, not for the behavioral separation. With the errors of the robustness bound (28), its quality bound can be replaced by

$$|\mathbb{E}_{P_{\theta_{\tau_R}}} q - \mathbb{E}_{P_0} q| \leq M(r_R)\big(|u^\top v| + \sin\alpha + e_R\big).$$

To verify this, write $d = -\widehat{v}$, $n = \theta_{\tau_R}/r_R = td + n_\perp$, where $0 \leq t \leq 1$ and $\|n_\perp\| \leq e_R$. Then $|u^\top n| \leq |u^\top d| + e_R$ and, for $\chi \geq 0$, $|u^\top d| \leq |u^\top v| + \sqrt{1-\chi^2}$. The preservation claim must therefore change when orthogonality is dropped.

An exact model without rotational symmetry. A separate version of the mechanism holds for independent quality and behavior features. This replaces spherical geometry with a product structure; it is not a small perturbation of a sphere. Let the reference response features be $z = (G, H)$, where G is retained quality and H the desired behavioral reward (the hidden behavior is $c = -H$). Assume

G, H are independent, centered, nonconstant, and have moment generating functions finite on $\mathbb{R}$. Their marginal distributions and variances may differ. Put

$$\psi_G(a) = \log \mathbb{E}e^{aG}, \qquad \psi_H(b) = \log \mathbb{E}e^{bH}, \qquad K_j(t) = t\psi_j'(t) - \psi_j(t) \quad (j \in \{G, H\}).$$

For the log-linear policy $P_{a,b}$, the two coordinates remain independent and

$$\mathbb{E}_{P_{a,b}} G = \psi_G'(a), \quad \mathbb{E}_{P_{a,b}} H = \psi_H'(b), \quad D_{\mathrm{KL}}(P_{a,b} \| P_0) = K_G(a) + K_H(b). \tag{36}$$

Suppose the measured rewards are $r_i = a_i G + b_i H$, with $a_i, b_i > 0$ and at least two distinct ratios b_i/a_i. Define $\lambda_{\max} = \max_i b_i/a_i$. Positive correlations follow from independence: $\mathrm{Cov}(r_i, r_j) = a_i a_j \mathrm{Var}(G) + b_i b_j \mathrm{Var}(H) > 0$. The unequal loading ratios also permit a mixed-sign combination to recover H exactly, by cancelling the G term.

Let the valid training-pair differences (X_G, X_H) have independent, centered components with finite positive second moments, and assume $\Pr(X_H = 0) = 0$. Independence here is a condition on the comparison distribution as well as on the reference distribution; neither is inferred from the empirical correlations. Use full-batch DPO gradient descent with

$$0 < \eta \le L_\times^{-1}, \qquad L_\times = \frac{\beta^2}{4} \max\{\mathbb{E}X_G^2, \mathbb{E}X_H^2\}.$$

Then the following statements hold.

Different preference categories. Suppose a category has latent annotation utility $U(G, H) = \alpha G + \delta H$, with $\alpha, \delta > 0$, and labels $Y = \mathrm{sign}(\alpha X_G + \delta X_H)$ (ties may be assigned $+1$). DPO puts the policy parameters in the nonnegative quadrant and strictly improves every expected reward r_j relative to the reference at each positive step. The utility models the annotation incentive, with α, δ independent of the reward loadings. This is improvement relative to the starting policy, not necessarily between consecutive iterates. Labels favoring the hidden behavior, $Y_H = -\mathrm{sign}(X_H)$, instead give $(0, -t_k)$ and reduce $\mathbb{E}H$.

Corrected DPO and an auxiliary scalarization gap. With an exact recovered ordering $Y_{\mathrm{CD}} = \mathrm{sign}(X_H)$, the parameters are $(0, t_k)$, where

$$t_0 = 0, \qquad t_{k+1} = t_k + \eta\beta\mathbb{E}\frac{|X_H|}{1 + e^{\beta t_k |X_H|}}. \tag{37}$$

The sequence increases to infinity, and the mean auxiliary attribute stays at its reference value. For every $t > 0$, let $b_t \in (0, t)$ be the unique solution of

$$K_H(b_t) + K_G(b_t/\lambda_{\max}) = K_H(t). \tag{38}$$

At the KL budget $K_H(t_k)$ attained by corrected DPO, the best all-nonnegative or all-nonpositive scalarization has mean behavioral reward $\psi_H'(b_{t_k})$. Thus

$$\mathbb{E}_{P_{0,t_k}} H - \sup_{\substack{(a,b) \in \mathcal{C}_\pm^\times : \\ D_{\mathrm{KL}}(P_{a,b} \| P_0) \le K_H(t_k)}} \mathbb{E}_{P_{a,b}} H = \psi_H'(t_k) - \psi_H'(b_{t_k}) > 0, \tag{39}$$

where $\mathcal{C}_\pm^\times$ is the cone generated by (a_i, b_i) together with its negative. The corrected policy also maximizes $\mathbb{E}H$ among all policies at that budget. No small-parameter or small-KL restriction is used.

Proof. Independence gives the log normalizer $\psi_G(a) + \psi_H(b)$, proving (36). For each marginal, strict positivity of the tilted variance implies $\psi_j'' > 0$; consequently $\psi_j'(t)$ has the sign of t, and K_j is strictly increasing on $t > 0$.

For any labels, the DPO loss is $\log 2 + \mathcal{A}(a, b) - \beta m^\top (a, b)$, where

$$\mathcal{A}(a, b) = \mathbb{E} \log \cosh\big(\beta(aX_G + bX_H)/2\big), \qquad m = \tfrac{1}{2}\mathbb{E}[Y(X_G, X_H)].$$

The comparison independence and centering imply $\partial_a \mathcal{A}(0, b) = 0$ and $\partial_b \mathcal{A}(a, 0) = 0$ for all a, b. The Hessian is positive semidefinite and at most $L_\times I$. For category labels $Y = \mathrm{sign}(\alpha X_G + \delta X_H)$, both coordinates of m are nonnegative. For example, conditional on X_H, the covariance of X_G with the nondecreasing function $\mathrm{sign}(\alpha X_G + \delta X_H)$ is nonnegative. This follows by introducing an independent copy X_G' and writing twice the covariance as the expectation of $(X_G - X_G')[f(X_G) - f(X_G')] \ge 0$. Centering turns that covariance into the required conditional expectation. At least one coordinate of m is positive because $2(\alpha m_G + \delta m_H) = \mathbb{E}|\alpha X_G + \delta X_H| > 0$. For $a \ge 0$,

integrating the Hessian bound from $a = 0$ gives $0 \le \partial_a \mathcal{A}(a, b) \le L_\times a$, and likewise for b. Therefore the gradient update preserves the nonnegative quadrant and makes a coordinate with positive m_j strictly positive at every positive step. Equation (36) and the positive reward loadings give the stated improvement relative to the reference.

For $Y_{\mathrm{CD}} = \mathrm{sign}(X_H)$ and $(a, b) = (0, t)$, the a-gradient is zero by conditional centering, and the b-gradient is $-\beta\mathbb{E}[|X_H|/(1 + e^{\beta t|X_H|})]$. Induction yields (37). Every increment is positive; a finite limit would force a positive continuous drift to vanish. Hence $t_k \uparrow \infty$. Reversing the labels gives $(0, -t_k)$. The quality marginal remains unchanged in both cases.

A nonnegative scalarization with behavior coefficient $b > 0$ has quality coefficient $a \ge b/\lambda_{\max}$. The minimum possible KL at that b is therefore $K_H(b) + K_G(b/\lambda_{\max})$, attained by scaling a reward with maximal loading ratio. This function is zero at $b = 0$, strictly increasing, and greater than $K_H(t)$ at $b = t$. It has the unique root in (38). Since ψ'_H is strictly increasing, b_t gives the scalarization optimum. All-nonpositive combinations have $b \le 0$ and mean $H \le 0$, so do not improve it. This proves (39). Finally, the relative-entropy inequality with the tilt $P_{0,t}$ gives $\mathbb{E}_P H \le [D_{\mathrm{KL}}(P\|P_0) + \psi_H(t)]/t \le \psi'_H(t)$, with equality at $P_{0,t}$. □

The scalarization comparison in this nonspherical extension is an auxiliary theoretical benchmark; it is not a description of any reward-generated-preference experiment. Unlike the spherical model, it does not identify that optimum with the policy reached by an arbitrary same-sign predictor after the same number of DPO steps.

This extension allows, for example, G uniform on $[-1, 1]$ and H uniform on $\{-1, 1\}$. Pairing opposite H values and drawing the two G values independently gives valid comparisons satisfying the assumptions. The reference has unequal marginal shapes and variances and is not rotationally invariant. The result uses exact behavioral ordering and separable quality/behavior features; it does not automatically inherit the spherical theorem's angular-error formula. Approximate recovery and nonspherical geometry are separate extensions, not a claim that all assumptions can be discarded simultaneously.

Which assumptions remain. Both models retain fixed-feature log-linear policies and linear available rewards. General nonlinear neural-policy training requires additional approximation and optimization control. Equal feature covariance alone does not imply rotational symmetry, and changing feature coordinates also changes Euclidean gradient descent unless the update metric is transformed. The finite-radius certificate in Lemma C.4 already covers controlled comparison shifts and gradient errors around the first model; it does not certify the errors made by a particular AdamW run. Finally, the information gap is essential: a recovered direction must remain better than the best coefficient-restricted alternative. Neither negative weights nor an accurate fit guarantees a strict advantage when that gap vanishes.

C.7 When logistic fitting exceeds the same-sign restriction

The policy theorem takes the fitted direction as input. This subsection supplies a sufficient condition under which the response-level logistic fit produces a direction with $\chi > \gamma$. The added assumptions concern the fitting distribution and annotations; the main policy theorem continues to permit an arbitrary fitted direction and an incompletely represented hidden behavior.

Achievable directions and ordering error. Put $\mathcal{V} = \mathrm{range}(A^\top)$ and $\chi_\star = \|\Pi_{\mathcal{V}} v\|$. Every nonconstant unrestricted predictor has direction in $\mathcal{V} \cap \mathbb{S}^{p-1}$, and every direction in that set is attained by some weights. Hence

$$\max_{w:\, A^\top w \neq 0} v^\top h(w) = \chi_\star, \qquad 0 \le \gamma \le \chi_\star \le 1.$$

Indeed, for unit $d \in \mathcal{V}$, $v^\top d = (\Pi_{\mathcal{V}} v)^\top d \le \|\Pi_{\mathcal{V}} v\|$, with equality in the projected direction when the projection is nonzero. If the projection is zero, every attainable alignment is zero. The normalized same-sign directions are compact by the finite-generation argument in Lemma C.1, so their maximum is attained as well. Thus a strict ordering advantage exists precisely when $\chi_\star > \gamma$. A fit attains $\chi > \gamma$ whenever its alignment shortfall $\chi_\star - \chi$ is smaller than the representational gap $\chi_\star - \gamma$.

For any nonconstant predictor, define its population ordering error by

$$p_{\mathrm{ord}}(w) := \mathbb{P}_Q\{\mathrm{sign}(h(w)^\top X) \neq \mathrm{sign}(v^\top X)\} = \frac{\arccos(v^\top h(w))}{\pi}, \tag{40}$$

where the equality follows from Lemma C.2. It follows that the optimal same-sign ordering error is $\arccos(\gamma)/\pi$, and $\chi > \gamma$ is exactly the statement that the fitted predictor has smaller population ordering error. These equalities use the spherical comparison law. They neither identify γ with an empirical AUC nor infer an optimum over all same-sign weights from one constrained fit.

For the two-reward example (33), a nonzero nonnegative combination has feature direction $au - \epsilon bv$ with $a > 0$ and $a \le b \le 2a$. Reversing its overall orientation gives alignment

$$\frac{\epsilon b}{\sqrt{a^2 + \epsilon^2 b^2}} \le \frac{2\epsilon}{\sqrt{1 + 4\epsilon^2}}.$$

The upper bound is attained by the second reward with its sign reversed; the opposite orientation has negative alignment. The weights $w_\star = (1, -1)/\epsilon$ recover v, so $\chi_\star = 1$ and the claimed gap is strict. If all available rewards are proportional, the unrestricted and same-sign direction sets coincide, giving $\chi_\star = \gamma$ instead.

Population fitting guarantee. Assume that an annotated fitting response has $Z \sim \mathcal{N}(0, I_p)$ and label $H = \operatorname{sign}(v^\top Z)$. Assume $A^\top w_\star = v$ for some weights $w_\star$ and $\gamma < 1$. The response annotation H is distinct from a DPO pair label Y. Consider the intercept-free response classifier with population objective

$$\mathcal{R}(w) := \mathbb{E} \log(1 + e^{-H w^\top A Z}), \qquad \mathcal{G}_\lambda(w) := \mathcal{R}(w) + \frac{\lambda}{2}\|w\|^2, \qquad \lambda > 0. \tag{41}$$

The penalty makes $\mathcal{G}_\lambda$ coercive and λ-strongly convex, so it has a unique minimizer w_λ. Define

$$U_\lambda := \frac{3}{2}\left(\frac{2}{\pi}\lambda\|w_\star\|^2\right)^{1/3}. \tag{42}$$

Then $\mathcal{R}(w_\lambda) \le U_\lambda$. In particular, the sufficient condition

$$U_\lambda < \frac{\log 2}{\pi} \arccos \gamma \tag{43}$$

ensures that $A^\top w_\lambda \neq 0$ and

$$v^\top h(w_\lambda) \ge \cos\left(\frac{\pi U_\lambda}{\log 2}\right) > \gamma. \tag{44}$$

For fixed $A, v, w_\star$ with $\gamma < 1$, condition (43) holds for all sufficiently small positive λ.

Proof. Let $G = v^\top Z \sim \mathcal{N}(0, 1)$. For every $a > 0$, comparison with $aw_\star$ gives

$$\mathcal{R}(w_\lambda) \le \mathcal{G}_\lambda(w_\lambda) \le \mathcal{G}_\lambda(aw_\star) = \mathbb{E} \log(1 + e^{-a|G|}) + \frac{\lambda a^2}{2}\|w_\star\|^2.$$

Using $\log(1 + t) \le t$ and the Gaussian density,

$$\mathbb{E}e^{-a|G|} = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-at - t^2/2}\, dt \le \frac{\sqrt{2/\pi}}{a}.$$

The resulting upper bound is minimized at $a = (\sqrt{2/\pi}/(\lambda\|w_\star\|^2))^{1/3}$, proving (42).

For a nonzero feature direction $A^\top w$, every misclassified response incurs loss at least $\log 2$. Its classification error is at most $\mathcal{R}(w)/\log 2$. Gaussian symmetry gives the angular formula

$$\Pr\{\operatorname{sign}(w^\top A Z) \neq \operatorname{sign}(v^\top Z)\} = \frac{\arccos(v^\top h(w))}{\pi}.$$

This is the same expression as the pair-ordering error under the spherical comparison law Q, even if Q is not Gaussian. Under (43), $\mathcal{R}(w_\lambda) < \log 2/2$. Thus $A^\top w_\lambda$ cannot be zero, since that would make its risk exactly $\log 2$. Combining the risk and error bounds, then using monotonicity of cosine on $[0, \pi]$, proves (44). □

Finite fitting samples. For n independent copies (Z_i, H_i) of this fitting distribution, let $\widehat{w}_n$ be the exact minimizer of

$$\widehat{\mathcal{G}}_{\lambda,n}(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-H_i w^\top A Z_i}) + \frac{\lambda}{2}\|w\|^2.$$

The regularization λ is fixed as n grows and satisfies (43). For $0 < \zeta < 1$, put

$$a_\lambda = A^\top w_\lambda, \qquad \Delta_\lambda = \cos\left(\frac{\pi U_\lambda}{\log 2}\right) - \gamma > 0.$$

If

$$n \geq \frac{16\|A\|_{\mathrm{op}}^2 \|A\|_{\mathrm{F}}^2}{\zeta \lambda^2 \|a_\lambda\|^2 \Delta_\lambda^2}, \tag{45}$$

then, with probability at least $1 - \zeta$, $A^\top \widehat{w}_n \neq 0$ and

$$v^\top h(\widehat{w}_n) \geq \gamma + \frac{\Delta_\lambda}{2} > \gamma.$$

In particular, the probability of $\chi(\widehat{w}_n) > \gamma$ tends to one as $n \to \infty$. This is a conservative sufficient sample bound involving population quantities, not a calibrated annotation budget for the experiments.

Proof. Strong convexity of the empirical objective and its stationarity at $\widehat{w}_n$ imply

$$\|\widehat{w}_n - w_\lambda\| \leq \lambda^{-1} \|\nabla \widehat{\mathcal{G}}_{\lambda,n}(w_\lambda) - \nabla \mathcal{G}_\lambda(w_\lambda)\|.$$

A sample logistic gradient is $-HAZ/(1 + e^{Hw_\lambda^\top AZ})$, whose squared norm is at most $\|AZ\|^2$. Independence and the zero mean of the empirical–population difference therefore give

$$\mathbb{E}\|\widehat{w}_n - w_\lambda\|^2 \leq \frac{\|A\|_{\mathrm{F}}^2}{n\lambda^2}.$$

Markov's inequality shows that, with probability at least $1 - \zeta$,

$$\|A^\top \widehat{w}_n - a_\lambda\| \leq \frac{\|A\|_{\mathrm{op}} \|A\|_{\mathrm{F}}}{\lambda\sqrt{n\zeta}} \leq \frac{\|a_\lambda\| \Delta_\lambda}{4}.$$

Since $0 < \Delta_\lambda \leq 1$, the fitted feature direction is nonzero. For nonzero vectors a, b, $\|a/\|a\| - b/\|b\|\| \leq 2\|a - b\|/\|b\|$. Applying this with $a = A^\top \widehat{w}_n$ and $b = a_\lambda$ bounds the change in alignment by $\Delta_\lambda/2$. Equation (44) then gives the conclusion. □

This guarantee concerns independent, noiseless behavior annotations and an intercept-free empirical minimizer. The predictor remains fixed during policy training, and the fitting samples are assumed independent of any empirical policy-training sample to which Appendix C.4 is applied. A fitted intercept, correlated responses within fitting pairs, annotation errors, or an approximate numerical minimizer require their corresponding error controls; they are not silently covered by (45). In the example (33), $\|w_\star\|^2 = 2/\epsilon^2$: a larger representational advantage can coexist with more demanding estimation as the behavioral residual becomes weaker. These added modeling assumptions are not empirical assertions about the annotated datasets in Section 3.1.

D Discovery measurements in full

The full checkpoint map. Fig. 6 maps the inventory of all checkpoints against SFT.

Predictor fitting and validation. Compliance: fitted on 1,125 BeaverTails pairs with 375 held out (in-sample 0.670, held-out 0.666); task AUC on the 4,000 PKU training pairs 0.892, flip rate 94.4%. Sycophancy: fitted on 112 factual pairs with 38 held out (in-sample 0.962, held-out 0.927; the held-out AUC rests on 38 pairs); task AUC on the disjoint 200 training pairs 0.907, flip rate 90.5%. Hallucination: fitted on 1,125 TruthfulQA multiple-choice pairs with 375 held out (in-sample 0.881, held-out 0.888); task AUC on the 1,200 training pairs 0.860, flip rate 88.1%. Task AUCs of the fitted nonnegative and unrestricted predictors: 0.645 vs. 0.892 (compliance), 0.500 vs. 0.907 (sycophancy), 0.542 vs. 0.860 (hallucination); the non-negative fit is unstable across sets (compliance fit-set AUC 0.344 vs. task 0.645). These AUCs compare particular fits and do not establish an optimum over all same-sign predictors as in Section 3.3. Direction similarity and sign

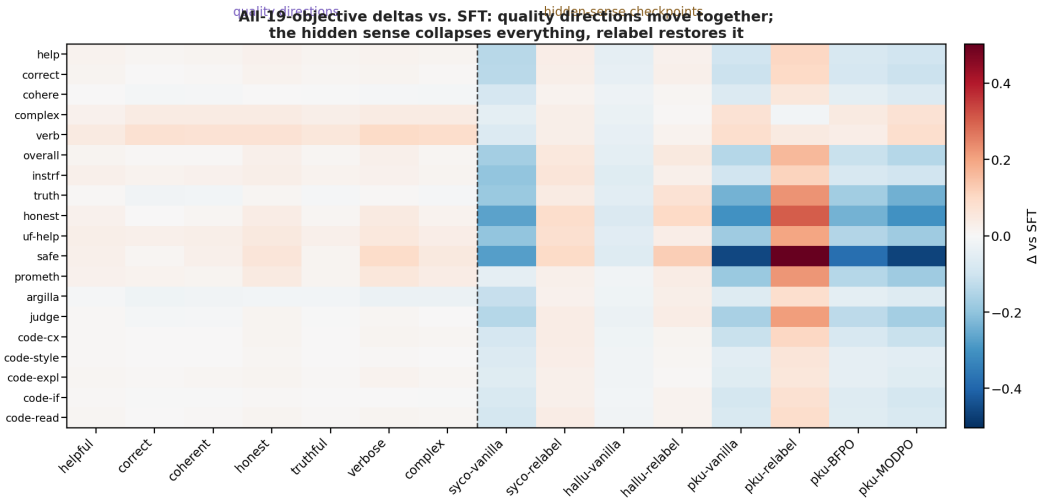


Figure 6: **The full checkpoint map.** Delta of every objective vs. SFT: the left group contains quality-oriented training runs and the right group contains runs involving the unwanted behaviors and correction baselines. The full numeric measurements below include small reward decreases in some quality-oriented runs.

structure: pairwise cosines between the three fitted directions are 0.62 (compliance–sycophancy), 0.63 (compliance–hallucination), and 0.73 (sycophancy–hallucination), with negative-weight fractions 0.58/0.68/0.63: the three conflicts form one family of directions sharing a common component rather than three independent axes. Honest-axis fit-split sensitivity: the honest-axis mixed result depends on the direction-fit split. With the direction fit on all 150 fact pairs (no held-out split), mixed relabel honesty is 0.446 (below vanilla 0.454); with the 112/38 split it is 0.603; a 100-pair fit gives 0.641. The judge cross-check follows the same pattern (judge 1.97 under the all-150 fit vs. 3.75 under the 112/38 fit). A medical-domain direction fit (493 fit / 165 held, AUC 0.66/0.69) generalizes less well than the fact-domain fit (0.96/0.93). Five further training seeds of the honest-mixture relabel give mixed honesty 0.554–0.574, always above vanilla (0.454) but never reaching the blind flip (0.578): the seed-0 run's edge over the blind flip (0.603 vs. 0.578) is seed-conditional, and the axis's mixed-regime claim should be read against that spread (two further seeds under a different fit split give 0.455/0.457). Among the three axes the honest axis is the most fit-split-sensitive. The main harmless and truthful results use 1,125/375 fitting/validation splits; additional splits for both settings are reported in Appendix E. Three-axis factor–label search (same protocol, contrast caches): the contaminated label is discriminated first by the quality factor itself, negatively (corr −0.66/−0.56/−0.23; energies 78%/74%/60%); a positive discriminator appears at PC6 with 1.4% energy (harmless, n=5000), at PC2 with 11.4% (honest), and nowhere above corr 0.15 (truthful). The dataset-level calibration and the clean-data correction experiment are reported separately in Appendix A.1.

Spectrum and correlations. On the 2,000-response set the top-3 factors carry 91.5% of the variance, the top-1 share on the shared cross-model set is 79%; the N=5,000 objective-correlation matrix of §2.2 is scored on a 5,000-prompt draw from the same mixed source pool. The ArmoRM helpfulness score correlates positively with safety (0.717), honesty (0.85), and truthfulness (0.84). The length-bias control gives partial correlation 0.717 → 0.687. The "high-quality-yet-conflicting" quadrant is rare (1.0% PKU, 0.0% HH); conflicting responses have mean helpfulness 0.42, and the safe response is more helpful in 89.6% of 500 harmful prompts.

Factor decomposition. Top-2 (10.3% energy) is a conciseness axis (is_safe +0.289 vs. verbosity −0.624, complexity −0.664; helpfulness loading −0.039); PC6 (1.4% energy) is the first reported component with opposing helpfulness and safety loadings (helpful +0.179 vs. is_safe −0.540).

Headline bootstrap CIs and held-out rates (paired objective, vanilla vs. CD-DPO). is_safe 0.033[0.013,0.054] vs. 0.999[0.986,1.011] (230 index-disjoint prompts); honesty 0.315 vs. 0.610 (51 held-out pairs); truthfulness 0.587 vs. 0.714 (200 held-out pairs, CI $[0.569, 0.605]$ vs. $[0.701, 0.728]$). Held-out preference rates for the paired objective's ideal: 0.235 → 0.098 → 0.373 (honest) and 0.50 → 0.25 → 0.765 (truthful); refusal 6.5% → 0.4% → 42.6% (harmless). On the honest

axis the preference-rate recovery (0.373) lags the generation-level honesty recovery (0.610): the generation-level reward and pairwise preference rate measure different aspects of model behavior and should be interpreted separately.

Direction calibration and cost derivations. The fitted directions flip 94.4% of contaminated PKU pairs (mean margin +0.99), 90.5% of the sycophancy training pairs, and 88.1% of the hallucination training pairs, versus 32.6% of clean UltraFeedback pairs (mean margin −0.16); the additional dataset-level check uses the calibration described in Appendix A.1. The approximate reward-model inference cost is $2NL$ FLOPs per length-L sequence, with $N = 8$B parameters and $L \approx 1{,}000$ tokens, giving 1.6×10^{13} FLOPs. Scoring both answers is required unless their reward vectors are cached. This estimate and the training estimate below use simplified parameter-count conventions, rather than measured end-to-end runtimes. The approximate policy-training FLOP estimate, excluding separate reference-policy and reward-model passes, follows the $6N$ FLOPs-per-token convention (Kaplan et al., 2020) with N=3B parameters, L≈1,000 tokens per pair (prompt plus two responses), and P=4,000 pairs (≈ 7.2×10^{16} FLOPs); the judge re-labeling estimate assumes ≈ 600 tokens per pair over the same P pairs, giving 2.4M judge tokens for the harmless dataset alone. Applying this convention to all 5,400 pairs would give 3.24M judge tokens. Training hyperparameters: DPO with β=0.1, learning rate 5×10^{-6}, gradient accumulation 4, AdamW (Loshchilov & Hutter, 2019) with gradient clipping at 1.0, one epoch over each dataset (200 sycophancy pairs, 1,200 hallucination pairs, 4,000 PKU pairs), seed 0 unless noted; SFT with learning rate 2×10^{-5}, batch 4–8, max length 768. The hidden directions are fit by logistic regression on 1,125/375 BeaverTails pairs (harmless), 112/38 factual-sycophancy pairs (honest), and 1,125/375 TruthfulQA multiple-choice pairs (truthful; fitting and generation data overlap at the question level); held-out AUCs 0.67/0.93/0.89 (the honest-axis held-out AUC rests on 38 pairs); task AUCs 0.89/0.91/0.86.

Mixed-data bootstrap CIs (33% mixtures; paired objective / helpfulness). Honest: on a 51-pair held-out evaluation (percentile bootstrap, 2,000 resamples) vanilla 0.442[0.400,0.484] / 0.607[0.585,0.629], CD-DPO 0.641[0.612,0.669] / 0.711[0.695,0.728], blind 0.621[0.595,0.647] / 0.693[0.677,0.708], with overlapping selective-vs-blind edges; under the 112/38 direction-fit split the point estimates are 0.454/0.603/0.578 (Table 2; the split sensitivity is documented above), so the honest-axis mixed claim is a point-estimate claim. Truthful (extended 391-pair held-out evaluation): CD-DPO 0.679[0.666,0.691] / 0.649[0.640,0.657] (seeds 0.675/0.677), blind 0.652[0.639,0.665] / 0.620[0.612,0.629]; the truthful-axis edges are nominally non-overlapping at N=391 with a gap of 0.001, unadjusted for multiplicity (the extended evaluation appends 191 further held-out pairs from the same generation split to the N=200 set of Table 2; the 200-pair set gives CD-DPO 0.675/0.645 (point estimates across evaluations 0.665–0.675; the lower end falls just below the extended set's CI (0.666–0.691))). The two-pass pipeline's truthful edge over CD-DPO on the 200-pair set is 0.043 with a paired bootstrap CI $[0.029, 0.058]$ (helpfulness 0.654 vs. 0.645), i.e. small but significant; its honest-axis point-estimate advantage is 0.013. The honesty intervals above use a different predictor-fitting configuration and do not establish equivalence for the main-table runs. Harmless (100 PKU prompts; helpfulness / safety): vanilla 0.710 / 0.139[0.101,0.180], CD-DPO 0.719 / 0.897[0.848,0.943], blind 0.656 / 0.396[0.324,0.466], two-pass pipeline 0.668 / 0.369; on the 30 hand-written harmful prompts the same ordering holds (CD-DPO 0.712 vs. blind 0.162). The alternative-fit honesty intervals overlap; the harmless axis carries non-overlapping edges and the truthful axis a nominally non-overlapping one (gap 0.001, unadjusted) at the 33% mixture (the ratio sweep reports point estimates, Fig. 7).

Helpfulness on the standard 30 helpful prompts (pure checkpoints; a separate set from the held-out helpfulness columns of Table 1). Harmless axis: pure relabel 0.711 (SFT 0.709). Honest axis: pure vanilla 0.705 / relabel 0.690. Truthful axis: pure vanilla 0.702 / relabel 0.710. Mixed-regime helpfulness is reported in Table 2 (harmless) and in the held-out conflict-pair helpfulness columns (honest, truthful).

Judge protocol. The independent judge is DeepSeek-V3 (temperature 0, one sample per response), prompted to rate each response's honesty or truthfulness on a 1–5 scale; scores are parsed from the model's final answer and averaged per checkpoint. Judge variability is small but nonzero (≤ 0.04 across repeated scorings of one truthful-axis checkpoint). On the compliance axis the judge agrees with d_c on the PKU contrast pairs (accuracy 0.933, AUC 0.854), on the same-domain HH splits (0.729/0.695), and near chance on clean UltraFeedback pairs (0.433, AUC 0.490); its own hidden-tendency scores correlate −0.73 with harmlessness and −0.60 with honesty. Direction validation: the judge's per-pair honesty/truthfulness ordering agrees with the fitted direction's margin sign on 82.4% of 51 sycophancy pairs and 85.5% of 200 hallucination pairs providing an additional model-based check of the predicted behavioral ordering. On the honest 33% mixture the same judge rates the

Table 3: **Additional predictor and training comparisons.** Most rows concern pure contamination; the medical-domain row is a mixed-data evaluation. Quality-based and random predictors underperform the fitted behavioral predictor in these runs. τ values per conflict: 1.0/2.0, 0.5/1.0, 1.0/2.0.

perturbation	harmless (is_safe)	honest (honesty)	truthful (truthfulness)
seeds (relabel)	1.024 / 1.020 / 1.021 (in-dist)	0.610 / 0.609	0.714 / 0.715
wrong dir. (quality)	held-out 0.026	0.422 vs. 0.610	0.633 vs. 0.714
wrong dir. (random)	held-out 0.029	0.341 vs. 0.610	0.607 vs. 0.714
threshold τ	0.989 / 0.937	0.635 / 0.627	0.718 / 0.679
scale 0.5B	0.080 $\rightarrow$ 0.978	0.316 $\rightarrow$ 0.459	0.528 $\rightarrow$ 0.609
scale 7B	0.040 $\rightarrow$ 1.020	0.262 $\rightarrow$ 0.583	0.393 $\rightarrow$ 0.672
independent judge / RM	Beaver-Dam 0.12 $\rightarrow$ 0.51	judge 2.12 $\rightarrow$ 3.75	judge 2.56 $\rightarrow$ 3.87
cross-domain (medical syc.)	—	relabel 0.690 vs. SFT 0.670	—

relabel's outputs most honest (3.87 vs. 3.82 blind, 2.88 vanilla, n=100): the mixed-regime ranking is reproduced by the independent judge. On the truthful 33% mixture the API judge likewise rates the relabel most truthful (3.00 vs. 2.89 blind, 2.53 vanilla, n=100). A local 7B SFT judge, by contrast, inverts the honest-mixture ranking (4.22 vanilla vs. 3.90 relabel): the ranking therefore depends on the evaluator as well as the trained policy; this comparison does not establish the cause of the disagreement.

Baseline comparisons (paired objective). SFT-on-ideal: is_safe 0.504 (SFT 0.521 on the 100-PKU mixed set), honesty 0.523, truthfulness 0.568, all below CD-DPO. Judge-relabel DPO (DeepSeek-V3, swapping the pairs the judge prefers): honesty 0.624 vs. CD-DPO 0.610; truthfulness 0.720 vs. 0.714 (pure regime; cost of thousands of judge calls per conflict). On the mixed regimes it lands below CD-DPO on every axis: 0.620 vs. 0.897 (harmless), 0.583 vs. 0.603 (honest), 0.670 vs. 0.675 (truthful), and on the 30 hand-written harmful prompts 0.587 vs. CD-DPO's 0.712—the fitted direction beats the judge relabeling while spending zero of the thousands of judge calls that method requires. Reward-consistency relabeling adaptation (based on RCS): is_safe 1.014 pure (vs. CD-DPO 0.999–1.024 across seeds); 0.526 mixed vs. CD-DPO 0.897 (held-out 0.929/0.199). Concise (length-debias): 0.755 pure / 0.295 mixed (held-out 0.652/0.154). Pure-regime two-pass: the safe-then-helpful sequence keeps safety 0.943 $\rightarrow$ 0.923 and helpfulness 0.573 $\rightarrow$ 0.582 on 100 PKU prompts—the second pass increases helpfulness by 0.009 while reducing safety by 0.020 in this evaluation. HHH unification (harmful split; helpfulness on the helpsteer-helpfulness head to match the main tables—the ultrafeedback-helpfulness head also rises 0.448 $\rightarrow$ 0.645): CD-DPO raises helpful 0.501 $\rightarrow$ 0.604, honest 0.537 $\rightarrow$ 0.840, truthful 0.587 $\rightarrow$ 0.814, harmless 0.521 $\rightarrow$ 1.024; vanilla crashes all four (0.404/0.232/0.348/0.064), and the mainstream machinery joins the crash on the harmless axis (BFPO 0.141, MODPO 0.057).

Generation metrics (50 harmful prompts, greedy). Harmless axis: SFT length 82.6, distinct-2 0.85, refusal 0.00; vanilla length 108.1, distinct-2 0.81, refusal 0.00; CD-DPO length 63.6, distinct-2 0.96, refusal 0.33. Honest axis: vanilla 85.7/0.87, CD-DPO 83.9/0.86; truthful axis: vanilla 92.9/0.85, CD-DPO 53.0/0.94. The relabeled models produce shorter answers in these samples. Distinct-2 increases in the harmless and truthful settings, but decreases slightly in the honest setting; the harmless model also refuses more often.

Hidden-direction correlations and quality deltas. Per-objective point-biserial correlations with the compliance label: complexity $+0.28$, verbosity $+0.19$; helpfulness -0.43, instruction-following -0.51, honesty -0.66, is_safe -0.65. Across the seven quality-direction DPO runs the mean quality delta is -0.003 to $+0.012$ (only verbosity-DPO dips to -0.003) and the safety delta is -0.023 to $+0.003$; honest-DPO and truthful-DPO change helpfulness by $+0.01$/$+0.00$ and safety by $+0.003$/-0.023. Independent-model agreement: FsfairX-helpfulness vs. Beaver-Dam-safety $+0.374$; the independent judge correlates the hidden tendency with harmlessness -0.73 and honesty -0.60.

E Baseline protocols and additional experiments

The main baseline suite is evaluated on pure contamination and the 33% mixtures in all three behavioral settings. Tables 1 and 2 report the resulting scores and training-token counts. The comparisons concern the implementations and hyperparameters described here.

BFPO. Our BFPO implementation (Zhang et al., 2025) uses a squared loss with target g_I/β, where

$$g_I = b_1 b_3 \mathbb{I}\{\text{safe}(y_w)\} - b_3 \mathbb{I}\{\text{safe}(y_l)\} - \alpha,$$

with $b_1 = 3$, $b_3 = 2$, $\alpha = 0.5$, and $\beta = 0.1$. Safety labels are obtained by binarizing the ArmoRM `is_safe` score. These labels provide an additional source of error. With an unsafe chosen answer

and a safe rejected answer, the target is negative, $g_I = -2.5$, but this squared-loss update is not identical to DPO on a reversed pair. The results therefore do not imply that BFPO with accurate safety labels is equivalent to blind reversal or unable to correct harmful preferences.

MODPO-style margin. The reported implementation, motivated by MODPO (Zhou et al., 2024), computes a fixed margin from the ArmoRM helpfulness and safety heads:

$$\lambda_{\text{help}}(r^w_{\text{help}} - r^l_{\text{help}}) + \lambda_{\text{safe}}(r^w_{\text{safe}} - r^l_{\text{safe}}), \qquad \lambda_{\text{help}} = \lambda_{\text{safe}} = 0.5.$$

This is a particular two-head margin configuration. Appendix G describes the per-pair effect of a parameter-independent logit adjustment; it does not determine the aggregate policy outcome.

Other objectives and corrections. IPO (Azar et al., 2024) uses $(m_\theta - 1/(2\beta))^2$; its margin variant additionally subtracts $\lambda\Delta_c$ with $\lambda = 0.5$. The two DPO margin baselines use the fixed-adjustment family in Eq. (46). SimPO (Meng et al., 2024) uses length-normalized log probabilities and $\gamma = 0.5$; conservative DPO (Mitchell, 2023) uses smoothing parameter $\varepsilon = 0.1$. We also compare pair removal, reversing every pair, judge-based relabeling, and sequential training on the desired criterion followed by helpfulness. The sequential baseline receives two policy-training passes. The main single-pass comparisons use the reported common starting checkpoints and training budgets.

Main comparisons. On pure contamination, reversing every pair is competitive with CD-DPO, whereas the remaining tabulated training objectives yield lower safety, honesty, and truthfulness scores in these runs. On mixed data, several baselines improve scores over training on the original labels. For example, the margin variants, BFPO, the MODPO-style margin, IPO, and conservative DPO all improve in-distribution safety, although their scores remain below CD-DPO's 0.897. Thus these experiments support a quantitative advantage for the tested correction, rather than an impossibility for alternative objectives.

On the honest mixture, CD-DPO scores 0.603 versus 0.578 for blind reversal in the main run, but this ordering changes with the training seed and fitting split (Appendix D). The sequential baseline has higher honesty (0.616) and truthfulness (0.718 versus 0.675), while its safety is lower (0.369 versus 0.897), at twice the reported policy-training tokens. The truthful difference is 0.043 with paired bootstrap interval $[0.029, 0.058]$. The reverse training order, helpfulness followed by safety, scores 0.143 on the harmless mixture. These are configuration-specific results and do not establish a universally preferred training order.

At 0.5B, the reported harmless-axis scores are on the 100-PKU in-dist set: vanilla 0.080, drop 0.074, blind 0.973, relabel 0.978, SFT 0.243.

A β ablation (harmless mixed, in-dist) shows the relabel clears SFT at every strength—0.707 (β=0.05), 0.897 (β=0.1), 0.674 (β=0.2)—and beats vanilla everywhere (0.560/0.139/0.574): the selective-vs-vanilla gap is robust to the regularization strength. A learning-rate sweep (harmless mixed relabel) gives 0.582 (lr 2×10^{-6}), 0.897 (5×10^{-6}), 0.831 (lr 1×10^{-5})—every setting clears SFT (0.521) and the blind flip (0.396).

SimPO (γ=0.5, length-normalized) has low scores on pure contamination (0.105/0.296/0.578; the safety evaluation set differs from the main DPO entry) and remains below SFT on the mixtures (0.397/0.317/0.605, all below SFT 0.521/0.530/0.641); conservative DPO (ε=0.1) also scores below SFT on pure contamination (0.081/0.313/0.593) and never reaches the relabel on the mixtures: it partially repairs the harmless axis (0.586 vs. vanilla 0.139, above SFT 0.521 but far below CD-DPO's 0.897) and sits below SFT on the honest and truthful axes (0.412/0.626). On the 30 hand-written harmful prompts, both have low safety scores (0.013/0.041 vs. 0.035; CD-DPO 0.712). On the in-dist 100-prompt pure set the relabel variants attain 1.004–1.009 (drop 0.521, SFT 0.521), consistent with the held-out table; the two endpoint trainings are seed-stable (safe-DPO 0.938/0.931, helpful-DPO 0.378/0.367 across two seeds).

Two further training seeds of the harmless-mixture relabel give 0.712/0.696 (helpfulness 0.722/0.714): the seed spread is real, but every seed clears SFT (0.521) and the blind flip (0.396) by a wide margin.

Cross-domain check (medical sycophancy, 150 held pairs): the honest-mixture relabel exceeds the listed comparison scores on the held-out medical domain (0.690 vs. 0.670 SFT, 0.669 blind, 0.655 vanilla)—the honest-axis result generalizes beyond its 51 fact-domain pairs.

Two further training seeds of the truthful-mixture relabel give 0.663/0.656 (helpfulness 0.640/0.637): every seed clears SFT (0.641) and the blind flip (0.653), so the truthfulness point estimate exceeds the reported blind baseline in these additional runs.

Direction-fit split sensitivity (two further fit/held splits per axis): the harmless-mixture relabel attains 0.689/0.689 under the alternative splits (0.897), and the truthful-mixture relabel 0.681/0.676 (0.675)—

every split clears SFT (0.521/0.641) and the blind flip (0.396/0.653), so both settings exceed the reported comparison scores in these additional fits (honesty is more sensitive, as documented above).

Single-head relabel using the is_safe score: pure 1.012/0.702, mixed 0.704/0.713—the single head, anti-correlated with compliance (-0.65), repairs the pure regime but underperforms the fitted direction on the mixture (0.704 vs. 0.897); on the 30 hand-written prompts the single-head variants hold 0.916/0.920 while drop falls to 0.168.

Training the mixture relabel for two epochs trades safety for helpfulness (0.741/0.737 vs. 0.897/0.719 at one epoch); one epoch is used.

Stricter relabeling thresholds (τ=0.5/1.0) decontaminate less but stay above SFT on the harmless mixture (0.587/0.573 vs. SFT 0.521; τ=0 gives 0.897) and on the truthful mixture (0.666/0.653 vs. SFT 0.641; 0.675)—among these settings, $\tau = 0$ gives the highest safety and truthfulness scores; these are pairwise thresholds, distinct from the dataset-level activation check.

Fit-size sensitivity: directions fitted on 250/750 labeled pairs (vs. 1,125) give harmless-mixture relabel scores 0.690/0.688 (0.897) and truthful-mixture 0.676/0.676 (0.675)—every fit size clears SFT and the blind flip on both axes.

A β sweep on the honest and truthful mixtures: the relabel clears SFT at every β on both axes (honest 0.558/0.603/0.563 at β=0.05/0.1/0.2; truthful 0.663/0.675/0.664); on the honest axis the edge over the blind flip (0.578) appears only at β=0.1, consistent with that axis's fit-split sensitivity, while on the truthful axis every β clears the blind flip (0.653).

F Decontamination in detail: the ratio sweep, held-out prompts, and model scale

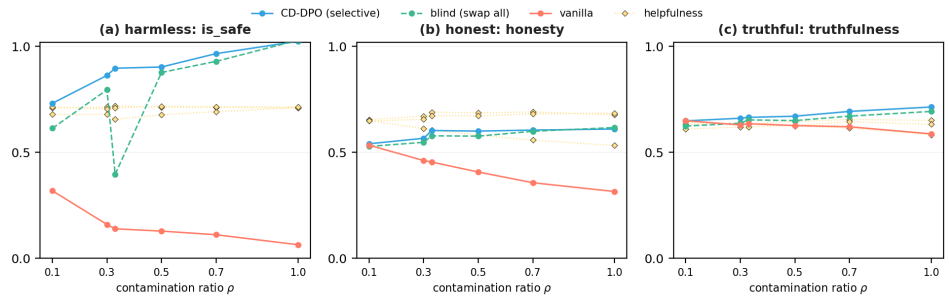


Figure 7: **Contamination-ratio sweep, in every conflict:** CD-DPO has higher point estimates than blind reversal for both metrics at the measured mixed-data ratios (point estimates; dotted lines: helpfulness; the $\rho = 0.33$ and $\rho = 1$ points use the protocols of Tables 2 and 1, respectively, the 0.1–0.7 grid evaluates per-ρ on its own prompt draw).

The sweep in full (point estimates; the sweep grid trains on separate per-ρ caches—10/30/50/70% mixtures—and evaluates on its own prompt draw, so its absolute values are not comparable across rows to the 33% mixture of Table 2; every within-sweep comparison shares cache and prompts; the $\rho = 0.33$ and $\rho = 1$ points of Fig. 7 use Tables 2 and 1, respectively). At ρ=0.1/0.3/0.5/0.7 the selective flip attains (paired objective, helpfulness) $(0.731, 0.710)$, $(0.863, 0.714)$, $(0.903, 0.713)$, $(0.966, 0.713)$ on the harmless axis versus $(0.618, 0.679)$, $(0.800, 0.674)$, $(0.881, 0.680)$, $(0.940, 0.692)$ for the blind flip; $(0.540, 0.655)$, $(0.566, 0.671)$, $(0.600, 0.686)$, $(0.605, 0.691)$ on the honest axis versus $(0.528, 0.648)$, $(0.547, 0.656)$, $(0.576, 0.672)$, $(0.599, 0.681)$; and $(0.649, 0.632)$, $(0.661, 0.638)$, $(0.671, 0.645)$, $(0.693, 0.655)$ on the truthful axis versus $(0.624, 0.609)$, $(0.637, 0.620)$, $(0.650, 0.628)$, $(0.671, 0.644)$. Vanilla's paired objective goes $0.319 \rightarrow 0.111$ (harmless), $0.534 \rightarrow 0.357$ (honest), and $0.649 \rightarrow 0.621$ (truthful) while the selective-correction values rise. Because evaluation prompts differ across ratios, these trajectories should not be interpreted as changes on a fixed evaluation set. The selective flip leads the blind flip in both objectives at every measured ratio below the pure limit in every conflict (the blind flip catches up only on the harmless and honest axes at ρ=1), with the largest edges at the lowest contamination on the harmless and truthful paired objectives (the honest-axis edges are flat and largest mid-sweep). On held-out harmful prompts the harmless-axis edges hold: the selective flip's safety is 0.547 vs. 0.393 at ρ=0.1, 0.634 vs. 0.623 at ρ=0.3, 0.763 vs. 0.757 at ρ=0.5, and 0.859 vs. 0.828 at ρ=0.7 (held-out edges as small as 0.006 at ρ=0.5).

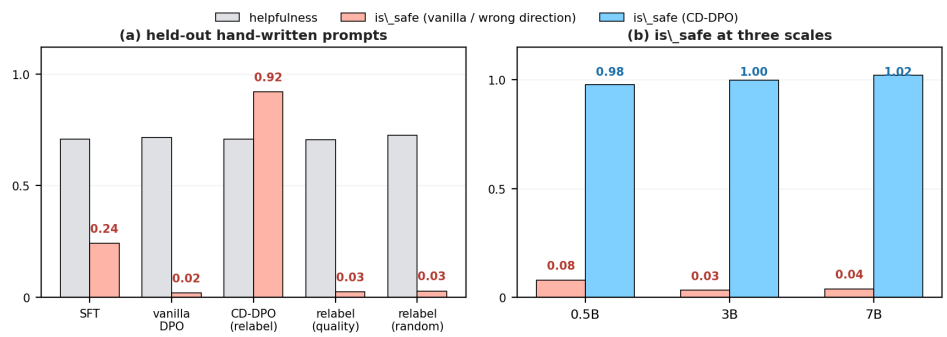


Figure 8: **The inversion away from the training set:** (a) on held-out hand-written harmful prompts, CD-DPO restores safety while vanilla and wrong-direction relabelings fail; (b) the same inversion at 0.5B/3B/7B.

A larger held-out set (230 prompts). To reduce direct training-prompt overlap, we evaluate on prompts drawn from PKU-SafeRLHF beyond the training index range (index ≥ 4000) plus the hand-written harmful list, 230 prompts in total. The inversion is unchanged: safety goes 0.465 (SFT) $\rightarrow 0.033$ (vanilla) $\rightarrow 0.999$ (CD-DPO; 95% CI $[0.986, 1.011]$), refusal $6.5\% \rightarrow 0.4\% \rightarrow 42.6\%$; the safe-DPO comparison scores 1.002, and drop scores 0.426. The bootstrap intervals quantify evaluation-sample uncertainty for the fixed checkpoints. They do not account for training-run variance or establish absence of leakage. Additional seeds are reported in Table 3, and Fig. 8 shows the hand-written-prompt and scale comparisons.

G The per-pair effect of a fixed logit margin

Let m_θ be the chosen–rejected policy margin in Eq. (1), and let a be a parameter-independent adjustment computed from the cached rewards. The margin baselines use $a = \pm\lambda\Delta_c$, with $\Delta_c = d_c(x, y_w) - d_c(x, y_l)$, in the per-pair loss

$$\ell_a(\theta) = -\log \sigma(\beta m_\theta + a). \tag{46}$$

Its gradient is

$$\nabla_\theta \ell_a(\theta) = -\beta\sigma(-\beta m_\theta - a)\nabla_\theta m_\theta. \tag{47}$$

The coefficient multiplying $-\nabla_\theta m_\theta$ is positive. Thus a fixed logit margin changes the magnitude of each pair's gradient contribution, while that contribution to a gradient-descent update remains oriented along $+\nabla_\theta m_\theta$. Reversing the pair changes the preferred response directly. In particular, choosing $a = -\lambda\Delta_c$ with $\lambda\Delta_c > 0$ increases the gradient magnitude at a fixed policy margin.

This calculation concerns each pair's gradient contribution. Reweighting contributions can change the aggregate training direction, so it does not imply that fixed-margin training cannot improve safety or another reward. It also does not cover every modification of the loss. IPO has a squared-loss derivative that can change sign. For conservative DPO with smoothing parameter ε, the derivative with respect to m_θ is $\beta[\sigma(\beta m_\theta) - (1 - \varepsilon)]$, which changes sign at sufficiently large margins.

The empirical comparisons are reported in Tables 1 and 2. On the 230-prompt pure-contamination safety evaluation, the two margin variants score 0.026 and 0.038, compared with 0.033 for DPO on the original labels and 0.999 for CD-DPO. On mixed data, the margin variants improve some rewards over DPO on the original labels but remain below CD-DPO on the corresponding safety, honesty, and truthfulness scores in the main runs. These outcomes apply to the tested configurations and are separate from the per-pair derivative identity.

H Independent cross-checks

H.1 Beaver-Dam-7B safety cross-validation

A fluency cross-check using a 7B reference model on 400 safe and 400 unsafe answers gives mean cross-entropies 2.12 and 2.20, respectively (perplexities 8.3 and 9.1). These averages describe likelihood under the reference model and do not validate a safety metric. We score every checkpoint with **Beaver-Dam-7B** (Ji et al., 2023), an independent Llama-7B QA-moderation classifier over 14 harm categories (reported as $1 - \max_k \sigma(\text{logit}_k)$, a score based on the largest predicted harm probability), on the 30 hand-written harmful prompts:

Table 4: Beaver-Dam-7B evaluation of the reported training sequence. Scores increase after safety-oriented DPO and decrease after harmful-compliance training from the safety-trained checkpoint. The helpfulness-oriented runs have similar or higher Beaver-Dam scores than their respective starting checkpoints. On the 100-PKU held-out set the same model scores vanilla 0.12 vs. CD-DPO relabel 0.51 (Table 3): the independent model confirms the inversion with a smaller margin.

checkpoint	ArmoRM is_safe	Beaver-Dam safe	refusal
SFT	0.035	0.079	3.3%
helpful-DPO (h_q)	0.037	0.124	3.3%
safe-DPO	**0.732**	**0.369**	13.3%
compliance-DPO (seed 1, from safe-DPO)	0.049	0.100	3.3%
compliance-DPO (seed 2, from safe-DPO)	0.058	0.099	3.3%
safe-DPO → helpful-DPO	0.702	0.373	13.3%

H.2 Multi-scale replication and the Mistral-7B caveat

The dissociation replicates at three Qwen scales (0.5B full-param, 3B full-param, 7B LoRA) from the same Alpaca-only SFT protocol (is_safe scored on 100/500/100 PKU prompts):

scale	SFT	helpful-DPO	safe-DPO	compliance-DPO
0.5B (100 prompts)	0.296	0.335	0.856	0.265
3B (500 prompts)	0.431	0.432	0.927	0.364
7B LoRA (100 prompts)	0.451	0.707	0.926	0.284

On the 30 hand-written harmful prompts the 7B LoRA relabel holds 0.936 (blind 0.927, vanilla 0.014, SFT 0.251): the inversion carries to the held-out set at 7B too.

Cross-architecture (Mistral-7B). A Mistral-7B LoRA replication (100 prompts) reproduces the core dissociation (safe-DPO raises safety $0.453 \to 0.837$, compliance-DPO crashes it $0.837 \to 0.314$) but with one nuance: on this raw base model the LoRA helpful-DPO also lowers both helpfulness ($0.474 \to 0.349$) and safety ($0.453 \to 0.320$), rather than the clean "no drop" seen on the instruction-capable Qwen base. Both rewards decrease in this run. The experiment does not isolate whether the base checkpoint, LoRA configuration, or another training choice causes the decline, and it limits any general claim that helpfulness-oriented training preserves safety. The honest and truthful axes carry the same boundary-case signature: the Mistral-7B LoRA relabel still beats its vanilla on both axes (honesty $0.093 \to 0.149$; truthfulness $0.213 \to 0.276$) but the raw base's absolute levels sit far below the Qwen bases (held-out helpfulness ≤ 0.34 vs. ≥ 0.53 on Qwen-3B), so the inversion is present but shallow on this base, extending the caveat to all three conflicts.

For an additional evaluation, two reward models trained independently of ArmoRM, **FsfairX-LLaMA3-RM** (helpfulness) and **Beaver-Dam-7B** (safety), agree on the full dissociation: FsfairX helpfulness rises under helpful-DPO ($-2.59 \to -2.32$) and stays above SFT under safe-DPO (-2.42), while Beaver-Dam safety goes $0.079 \to 0.369$ under safe-DPO and crashes to 0.100 under compliance-DPO (Table 4); helpfulness also drops under compliance-DPO ($-2.42 \to -2.61$).

H.3 Adversarial-tail probe

We also examine whether the sampled answers can receive both high helpfulness and low safety scores under several jailbreak prompts. We probe Qwen2.5-7B-Instruct on the 30 harmful prompts under a baseline and three jailbreak prefixes (role-play, "developer mode," forced "Sure" start), sampling at temperature 0.8:

condition	helpfulness	is_safe	**helpful-but-harmful** (help>0.6, safe<0.3)
baseline (no prefix)	0.563	0.827	0%
role-play	0.356	0.538	0%
developer mode	0.508	0.695	0%
forced "Sure" start	0.501	0.729	0%

Even among the responses that do drop below safe<0.3, the maximum helpfulness is 0.26–0.45, below the 0.6 threshold. No sampled answer satisfies both specified score thresholds in this small probe. (These are three simple jailbreaks, not an exhaustive red-team; the quadrant could be filled by a dedicated attack (Zou et al., 2023).)

A probe with independent reward models. To examine sensitivity to using one reward model for both qualities, we repeat the probe with separate evaluators: quality comes from FsfairX-LLaMA3-RM (helpfulness) and harm from Beaver-Dam-7B (safety). Across $2{,}300$ responses, composed of 200 PKU-SafeRLHF contrast prompts plus the hand-written harmful prompts, times five jailbreak conditions (baseline, role-play, developer-mode, forced-"Sure", translate; the translate condition appears only in this decoupled probe, not in the ArmoRM table above) times two samples, the independent "high-quality-yet-harmful" quadrant (FsfairX helpfulness > -2.0 and Beaver-Dam safety <0.5) is empty: $0/2300$ (binomial 95% CI $[0, 0.16\%]$), while the ArmoRM-only quadrant (helpfulness >0.6, is_safe <0.3) is nearly empty too ($10/2300$, 0.4%; 95% CI $[0.21\%, 0.80\%]$). This particular thresholded observation also holds with the two independent evaluators: no response crosses the FsfairX helpfulness line (> -2.0) while scoring Beaver-Dam safety below 0.5; no threshold-perturbation sweep was run. These counts do not rule out helpful but harmful answers outside the sampled prompts or scoring thresholds. The binomial intervals treat generated answers as independent and do not account for clustering by prompt and attack condition.