BEYOND UNIFORM FORGETTING IN ALIGNMENT: CHARACTERISING, PREDICTING, AND RECOVERING RETENTION IN SEQUENTIAL PREFERENCE OPTIMISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential Direct Preference Optimisation (DPO) lets a model acquire new alignment objectives without retraining from scratch, but later training can degrade preferences learned earlier. It remains unclear whether such forgetting is uniform, predictable or recoverable, and how it should be measured. We study these questions across five preference settings (HH-RLHF, HelpSteer2, PKU-SafeRLHF, UltraFeedback and Summarize-Feedback), organised as *characterise*, *predict* and *recover*. When measured relative to the original base model, earlier objectives show degradation in some settings and stability or improvement in others. Across 16 retention comparisons, comprising ten two-stage orderings and six three-stage steps, the policy's own chosen-over-rejected ranking is substantially more stable than the reference-relative metrics. For length-normalised accuracy, a decline of more than 2.5 percentage points is ruled out in 13 of 16 comparisons, so large reference-relative declines need not indicate preference reversals. Before Stage 2, the Stage-1 model's margin on the incoming objective matches the direction of later margin retention in nine of ten orderings. The evidence comes from five datasets and describes margin changes rather than the policy's own accuracy. Replaying a small share of earlier data shows nominal improvements in margin retention in seven of nine orderings, while policy accuracy improves in only two.

1 INTRODUCTION AND BACKGROUND

Aligning large language models (LLMs) with human preferences plays a crucial role in their useful deployment and protection from negative social impacts Ji et al. (2025); Ouyang et al. (2022); Dai et al. (2024). Multiple behavioural objectives, such as helpfulness, harmlessness, safety, honesty, factuality, style, and instruction following, are optimised in post-training pipelines Ji et al. (2025); Bai et al. (2022); Cui et al. (2023); Wang et al. (2024). This usually involves optimising a policy against human feedback signals, either through Reinforcement Learning from Human Feedback (RLHF) Bai et al. (2022) or through offline contrastive objectives such as Direct Preference Optimisation (DPO) Rafailov et al. (2023) and its variants such as IPO Azar et al. (2024) and SimPO Meng et al. (2024); Tie et al. (2025); Xu et al. (2024). These methods have driven significant progress in making models more helpful, less harmful, and better at following instructions. However, in practice, alignment is rarely a single-objective procedure Li et al. (2026), and optimisation of multiple objectives is required Zhong et al. (2024); Ji et al. (2025). Preference data for different objectives are also collected under different annotation protocols and task distributions Cui et al. (2023); Zhong et al. (2024). This creates a practical need to add new objectives without retraining from scratch.

Sequential Direct Preference Optimisation can offer a practical way to incrementally extend alignment. Instead of retraining a model from scratch when new preference data becomes available, later objectives can be optimised from the previously aligned model, combining optimising methods like DPO-style preference optimisation with the sequential multi-dimensional alignment setting Lou et al. (2025); Rafailov et al. (2023). This setup naturally raises a question about retention, when a later objective is optimised, to what extent are preferences learned in earlier stages preserved or degraded? Prior work on lifelong or sequential alignment has shown that adapting models to new

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.

preferences can lead to forgetting of earlier alignment behaviours Li et al. (2026), while recent large-scale studies of LLM post-training suggest a more nuanced picture in which forgetting is often moderate, task-dependent, and sensitive to the evaluation protocol Harmon et al. (2025).

Various evaluation methods are used to understand the effectiveness of alignment Fu et al. (2025); Liu et al. (2023), and aggregate task-level scores can hide substantial sample-level variation in what models retain or lose after post-training. Existing sample-level analyses of forgetting have largely focused on multiple-choice knowledge benchmarks, where forgetting is measured through correctness transitions Harmon et al. (2025). However, preference alignment differs because models are optimised over open-ended chosen–rejected response pairs, and degradation is often read from changes in reference-relative margins rather than discrete answer flips. Whether such reference-relative changes reflect changes in the policy's own preferences, and why they occur, remains less clear Harmon et al. (2025).

Continual-learning studies often measure forward and backward transfer and examine how task similarity affects forgetting Lopez-Paz & Ranzato (2017), while replay is widely used to preserve earlier tasks Zhang et al. (2025b). Sequential and multi-objective alignment methods have similarly been proposed for combining preference objectives Li et al. (2026); Lou et al. (2025), including continual preference learning with policy regularisation Zhang et al. (2025a), which aim to preserve or combine preferences during continual or multi-objective optimisation. Much of this work is concerned with how to prevent or reduce forgetting while we believe it is crucial to first ask whether a reference-relative retention signal corresponds to actual preference reversal.

In this work, we study preference degradation under sequential DPO to ask three questions. First, *does sequential DPO cause uniform forgetting of previous objectives, or does the effect depend on the relationship between objectives, on the individual preference pairs, and on how retention is measured?* Second, *can this effect be forecast before the second training stage begins, using only information available from the first stage?* Third, *where forecast risk is high, can it be mitigated?* A natural candidate explanation is direct gradient conflict between objectives, which we test alongside effective-update and hidden-state diagnostics. After each stage, we evaluate the model on all objectives using two types of measure. Reference-relative metrics include the DPO reward margin, which measures how much the chosen–rejected gap changes relative to the fixed base model, and the Gap-Improvement Rate (GIR), which measures the fraction of preference pairs for which this gap becomes larger. Policy accuracy instead measures whether the model itself prefers the chosen response over the rejected response. These measures can therefore disagree. A pair may no longer show an improvement relative to the base model, while the model still ranks the chosen response above the rejected one. In the two-stage HH-RLHF and HelpSteer2 orderings, GIR falls by 5 to 10 points while policy accuracy changes by at most 1.2 points, showing that a reference-relative decline does not necessarily mean that the model has reversed its preference.

1.1 Contributions

Our contributions and the paper are organised around **Characterise**, **Predict**, and **Recover**.

1. Characterise. We study sequential DPO across five preference settings, ten two-stage orderings and two three-stage HelpSteer2 sequences. Reference-relative reward margins and GIR can decrease, remain stable or improve depending on the setting, *while the policy's own accuracy changes much less*. In 13 of 16 length-normalised comparisons, a decline larger than 2.5 percentage points is ruled out, with only one comparison having a 95% interval below zero. This shows that large reference-relative declines do not necessarily mean that the model has reversed its preferences. The pattern continues over a third HelpSteer2 stage: in one sequence GIR falls by 12.8 points cumulatively while length-normalised accuracy increases by 1.9 points.

2. Predict. We introduce Stage-1 compatibility, the Stage-1 model's margin on the not-yet-trained objective. Its sign matches later margin-space retention in nine of ten orderings, including all eight whose sign is stable under half-sampling, without fitted parameters, and leave-one-dataset-out prediction is correct for seven of ten signs against six for the majority outcome. Because our evidence is based on five datasets, we treat compatibility as a coarse margin-space diagnostic rather than a predictor of policy ranking.

3. Recover. We test replaying a small amount of earlier data during Stage 2 while keeping the total training budget fixed. Replay improves margin retention in seven of nine orderings (95% intervals, uncorrected for multiple comparisons), but improves policy accuracy in only two. In the HH-RLHF and HelpSteer2 cases where margins declined, policy accuracy had changed by at most 1.2 points, leaving little ranking loss to recover. A retrospective analysis also shows that compatibility does not reliably identify where replay is most useful.

2 Experimental Setup

2.1 Overview

This work compares sequential DPO Rafailov et al. (2023) across five preference settings, HH-RLHF, HelpSteer2, PKU-SafeRLHF, UltraFeedback, and Summarize-Feedback, which span a range of objective relationships detailed in Section 2.3. A model is trained sequentially on two objectives from each setting and evaluated on all objectives after each step. Additionally, we train two three-stage sequences on HelpSteer2 (Section 4.1.3).

Our main experiments use Llama-3.1-8B-Instruct Grattafiori et al. (2024) as the base model with LoRA Hu et al. (2022) adapters. We use the instruction-tuned variant because it provides a realistic aligned starting point and a consistent conversational prompting format across preference datasets. Full model, LoRA, and training hyperparameters are reported in Appendix G.

2.2 Sequential DPO Training

Direct Preference Optimisation (DPO) Rafailov et al. (2023) is used to train on each preference objective. Given a prompt x and a pair of responses (y_w, y_l), where y_w is the preferred response and y_l is the rejected response, DPO optimises the policy π_θ relative to a frozen reference policy π_{ref}:

$$h_\theta(x, y_w, y_l) = \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} - \log \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}, \tag{1}$$

and DPO minimises $\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log \sigma\big(\beta\, h_\theta(x, y_w, y_l)\big)\big]$, where $h_\theta(x, y_w, y_l)$ is the relative log-ratio between policy and reference.

In sequential training, *each stage resumes from the previous stage's adapter checkpoint*, and its DPO reference is the model state at the start of that stage, hence, the Stage 2 reference is the final Stage 1 adapter. However, for evaluation, we always use the fixed base model as the reference, so reward margins stay on a consistent scale across all stages and checkpoints.

2.3 Preference-Dataset Taxonomy

A central goal of this work is to test whether sequential DPO behaves differently under different relationships between preference objectives. We therefore compare five settings, each containing two objectives and both training orderings: **HH-RLHF** Bai et al. (2022) (helpful-base and harmless-base, a distributional safety–helpfulness shift with no flipped-label pairs and fewer than 1% (0.86%) of unique prompts shared between the two training sets); **HelpSteer2** Wang et al. (2024) (verbosity and coherence, an interaction between fine-grained response attributes); **PKU-SafeRLHF** Ji et al. (2025) (safety–helpfulness with a much stronger harmlessness signal than HH-RLHF); **UltraFeedback** Cui et al. (2023) (instruction following and honesty, which reward overlapping response-quality behaviours); and **Summarize-Feedback** Stiennon et al. (2020) (coverage and accuracy for Reddit summarisation, a domain distinct from the chat-style settings). The settings also differ in domain, data size, response length and signal strength, so outcomes cannot be attributed to the objective relationship alone. Further construction details are provided in Appendix O.

3 EVALUATION PROTOCOL

3.1 CHARACTERISE

3.1.1 AGGREGATE EVALUATION

DPO-Relative Reward Margin. For a preference pair (x, y_w, y_l), we define the DPO-relative reward margin Rafailov et al. (2023) as β times the relative log-ratio $h_\theta(x, y_w, y_l)$ used in the training objective (Eq. 1):

$$r_\theta(x, y_w, y_l) = \beta\, h_\theta(x, y_w, y_l) \tag{2}$$

where $\log \pi_\theta(y|x)$ denotes the sequence log-probability of response y given prompt x. A positive reward margin means the policy increases the chosen-over-rejected preference gap relative to the reference model, and a value of zero means no relative change from the reference. The mean reward margin over all evaluation pairs is reported.

Gap-Improvement Rate (GIR). The Gap-Improvement Rate is defined as:

$$\mathrm{GIR} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left(\mathbf{1}[r_i > 0] + \tfrac{1}{2}\, \mathbf{1}[r_i = 0] \right) \tag{3}$$

where $r_i = r_\theta(x_i, y_{w,i}, y_{l,i})$. GIR is the fraction of pairs for which the chosen-over-rejected gap improves relative to the reference model, with ties receiving half credit. It is a reference-relative rate, not a ranking accuracy. A pair counts as improved whenever the policy widens the gap relative to the base model, even if the policy still prefers the rejected response, and a pair the policy ranks correctly counts as not improved when its gap has narrowed. Both use the fixed base model as reference: a reward margin of zero means no change relative to it, and 0.5 is the neutral GIR baseline.

Policy Accuracy. To ask whether the policy itself prefers the chosen response, we compare log-probabilities under the policy alone, without any reference model. The length-normalised policy margin of a pair is

$$m_i = \overline{\log \pi_\theta}(y_{w,i}|x_i) - \overline{\log \pi_\theta}(y_{l,i}|x_i) \tag{4}$$

where $\overline{\log \pi_\theta}$ denotes the mean per-token log-probability over response tokens, and the summed margin is $m_i^\Sigma = \log \pi_\theta(y_{w,i}|x_i) - \log \pi_\theta(y_{l,i}|x_i)$. Policy accuracy is the fraction of pairs with a positive margin, counting ties as incorrect:

$$\mathrm{PA}_\mu = \frac{1}{|\mathcal{D}|} \sum_i \mathbf{1}[m_i > 0], \qquad \mathrm{PA}_\Sigma = \frac{1}{|\mathcal{D}|} \sum_i \mathbf{1}[m_i^\Sigma > 0]. \tag{5}$$

We refer to these as length-normalised and summed policy accuracy. They differ in how they treat response length: the summed score adds up the log-probabilities of all response tokens and therefore tends to favour shorter responses, whereas the length-normalised score averages them per token. Summed accuracy can thus reflect length rather than preference when chosen and rejected responses differ systematically in length (Appendix B), so we report both. Appendix Table 7 summarises the four metrics.

Statistical treatment. All metrics are computed from per-pair log-probabilities on fixed held-out test pairs, and every change is defined as the later minus an earlier stage checkpoint on identical pairs. For changes in accuracy we report a 95% confidence interval for the difference between the two accuracies, computed with Newcombe's method for proportions measured on the same pairs; the appendix also gives exact sign tests on the pairs whose outcome flips (correct→incorrect versus incorrect→correct). For changes in margin and GIR we use a paired bootstrap over pairs (4,000 resamples). We call a change significant when its 95% interval excludes zero, without correction for multiple comparisons. In the accuracies ties count as incorrect and occur in about 1% of pairs or fewer; exact zero reward margins, which receive half credit in GIR, occur in up to 7% of pairs (bf16 resolution).

Pair-level analysis. Individual pairs are analysed to ask where change concentrates. We calculate per-pair margin drop, its quartile decomposition by Stage-1 margin, the chosen/rejected decomposition, and the concentration of the reward-margin change in the top 2% of pairs. These analyses describe reference-relative and length-normalised margins. Detailed definitions and results are given in Appendix E.

3.1.2 MECHANISTIC DIAGNOSTICS

Gradient Conflict and Adapter Movement. We compare the live Stage 2 DPO gradient with a Stage-1-objective diagnostic gradient, recomputed at each measurement on a fixed reference batch, over trainable LoRA parameters. We also measure movement using both raw LoRA parameters and the rescaling-invariant effective update

$$\Delta W = \frac{\alpha}{r} BA \tag{6}$$

reporting relative movement and alignment with the Stage-1 direction. Full definitions and implementation details are provided in Appendix H.

Hidden-State Drift Probe. For a fixed batch of 32 Stage-1-objective examples, we compare mean-pooled last-layer response representations at the Stage 1 and Stage 2 checkpoints using cosine similarity and relative movement. Full definitions and implementation details are provided in Appendix H.3.

3.2 PREDICT

3.2.1 PREDICTING RETENTION: THE STAGE-1 COMPATIBILITY DIAGNOSTIC

To test whether preference retention can be forecast before Stage 2 training begins, we define a diagnostic computed from the Stage-1 checkpoint and a held-out evaluation set for the not-yet-trained objective.

Stage-1 Compatibility. Let $M_\theta(X)$ denote the mean DPO-relative reward margin (Eq. 2) of checkpoint θ evaluated on the held-out test set for objective X. For an ordering that trains objective A first and objective B second, the *compatibility* of B given A is defined as

$$C_{A\to B} = M_{\theta_A}(B) - M_{\theta_0}(B) \tag{7}$$

where θ_A is the Stage-1 checkpoint trained on A and θ_0 is the base model. $C_{A\to B}$ measures the Stage-1 checkpoint's transfer onto B, the objective it has not yet been trained on, and is therefore available before Stage 2 training begins.

Retention of A after Stage 2 training on B is defined as

$$R_{A|B} = M_{\theta_{A\to B}}(A) - M_{\theta_A}(A) \tag{8}$$

where $\theta_{A\to B}$ is the Stage 2 checkpoint. Here, $R_{A|B} > 0$ indicates that A's margin is higher after Stage 2 than after Stage 1 (improvement or positive transfer), while $R_{A|B} < 0$ indicates degradation. As before, a change is the later checkpoint minus an earlier one, so negative values indicate degradation, unless otherwise stated. Compatibility and retention in GIR and policy-accuracy space are obtained by replacing M_θ with the corresponding metric (Appendix C), where we also define a signal-asymmetry ratio S that compares the strengths of the two objectives' Stage-1 training signals.

3.3 RECOVER

3.3.1 REPLAY AND A RETROSPECTIVE COMPATIBILITY GATE

We test whether mixing some of A's training data into Stage 2 recovers retention, and retrospectively ask whether negative Stage-1 compatibility $C_{A\to B}$ (Eq. 7) identifies the orderings in which it helps. For a fixed Stage-2 budget of N examples and replay dose $r \in [0, 1]$,

$$n_{\text{replay}} = \lfloor rN \rfloor, \qquad n_{\text{new}} = N - n_{\text{replay}} \tag{9}$$

examples are drawn from A's and B's training sets respectively, holding total compute fixed relative to the no-replay ($r = 0$) baseline. Our main comparison uses *random* replay at $r = 0.10$, with examples drawn uniformly from A's training set, in nine of the ten orderings regardless of compatibility; the gate that replays only where compatibility is negative is then evaluated retrospectively on these runs. We also evaluate *confidence-targeted* pools (the top quartile by Stage-1 policy margin m_i, Eq. 4) and *anchor-boundary* pools (half top-margin pairs and half pairs with $|m_i|$ nearest zero) on HH-RLHF and HelpSteer2 (Appendix M).

Ordering	n	Δ margin	Δ GIR	Δ policy acc. (norm.)	Δ policy acc. (summed)	C
HH Ha→He	900	−0.21	−6.2	−0.7 [−2.1, +0.8]	−0.7 [−1.5, +0.2]	−0.50
HH He→Ha	900	−0.51	−10.3	−1.2 [−2.6, +0.1]	−0.4 [−1.1, +0.2]	−0.37
HS2 Ve→Co	527	−0.25	−5.2	+0.8 [−0.2, +1.7]	−0.6 [−1.5, +0.4]	+0.01‡
HS2 Co→Ve	422	−0.21	−8.3	−0.2 [−2.3, +1.8]	0.0 [−1.6, +1.6]	−0.34
PKU Ha→He	1000	+2.86	+0.2	+0.2 [−0.8, +1.2]	**+1.5 [+0.6, +2.4]**	+0.78
PKU He→Ha	1000	+0.62	−9.7	+1.7 [−0.2, +3.6]	+0.1 [−1.1, +1.3]	+1.28
UF IF→Hon	700	+6.90	+2.7	+1.1 [−0.4, +2.7]	**+3.6 [+2.1, +5.0]**	+7.45
UF Hon→IF	700	+7.79	+2.7	+1.4 [−0.4, +3.3]	**+4.6 [+2.8, +6.3]**	+6.11
SF Co→Ac	205	+0.28	−7.6	**−3.9 [−7.7, −0.1]**	−1.5 [−6.2, +3.3]	+1.26
SF Ac→Co	179	+1.71	−8.1	−1.1 [−5.2, +2.9]	**+6.1 [+1.7, +10.8]**	+0.12‡

Table 1: Retention of the first-trained objective after Stage 2 (later minus earlier, so negative means degradation). Margin and GIR (Eq. 3) are reference-relative. Policy accuracy changes are in percentage points with 95% intervals (bold if the interval excludes zero, uncorrected). C is the Stage-1 compatibility of the incoming objective (Eq. 7), and ‡ marks a sign that is unstable under half-sampling (Section 4.2).

4 Results

Results follow the three questions of Section 1. Evaluation scores fixed chosen/rejected responses by model log-probabilities, not generations (we report a generation-level check in Appendix K). bf16 measurement noise is small for policy accuracy but not for GIR, whose changes below about 2 points should not be interpreted (Appendix B).

4.1 Characterise

4.1.1 Reference-Relative Metrics Change Substantially, while Policy Ranking Is Much More Stable

Table 1 reports the retention of the first-trained objective after Stage 2 for all ten orderings, on the reference-relative measures (reward margin, GIR) and on the policy's own ranking (two policy accuracies). Per-stage values, intervals, pair counts and sign tests are in Appendix Tables 4, 5 and 6.

The reference-relative metrics suggest that later-stage training produces degradation, stability, or positive transfer rather than uniform forgetting. In HH-RLHF and HelpSteer2 the earlier objective degrades in both training orders: in HH-RLHF the harmlessness margin falls from 0.912 to 0.698 after helpfulness training and the helpfulness margin from 0.733 to 0.227 after harmlessness training, and across the four orderings the margin falls by 0.21 to 0.51 margin units and GIR by 5.2 to 10.3 percentage points. PKU-SafeRLHF, UltraFeedback and Summarize-Feedback show non-negative margin changes in all six orderings (+0.28 to +7.79), yet GIR falls by 7.6 to 9.7 points in PKU-SafeRLHF helpful→harmless and in both Summarize-Feedback orderings, and rises by 2.7 points in both UltraFeedback orderings.

The policy's own ranking behaves differently. We have 32 accuracy comparisons in total (16 comparisons, each with two accuracy measures), and only one of them has a 95% interval below zero. This is the length-normalised accuracy in Summarize-Feedback coverage→accuracy, which falls by 3.9 points (interval -7.7 to -0.1, $n = 205$, exact sign-test $p = 0.077$). The interval is wide, so a decline of several points cannot be excluded there. We also ask how large a decline the data can rule out. For length-normalised accuracy, the lower end of the interval rules out a decline larger than 2.5 points in 13 of the 16 comparisons in Figure 1, which cover the ten two-stage orderings and six three-stage steps (Section 4.1.3). The three exceptions are HH-RLHF helpful→harmless (lower end -2.6) and both Summarize-Feedback orderings. For summed accuracy, a decline larger than 2 points is ruled out in 14 of the 16 comparisons. The other intervals that exclude zero are all increases. Summed accuracy rises in both UltraFeedback orderings (+3.6 and +4.6), in PKU-SafeRLHF harmless→helpful (+1.5) and in Summarize-Feedback accuracy→coverage (+6.1), and length-normalised accuracy rises in two of the six three-stage steps (Section 4.1.3). After a Bonferroni correction over the ten orderings within each accuracy measure, only the two UltraFeedback orderings and PKU-SafeRLHF harmless→helpful remain clearly significant (exact sign-test

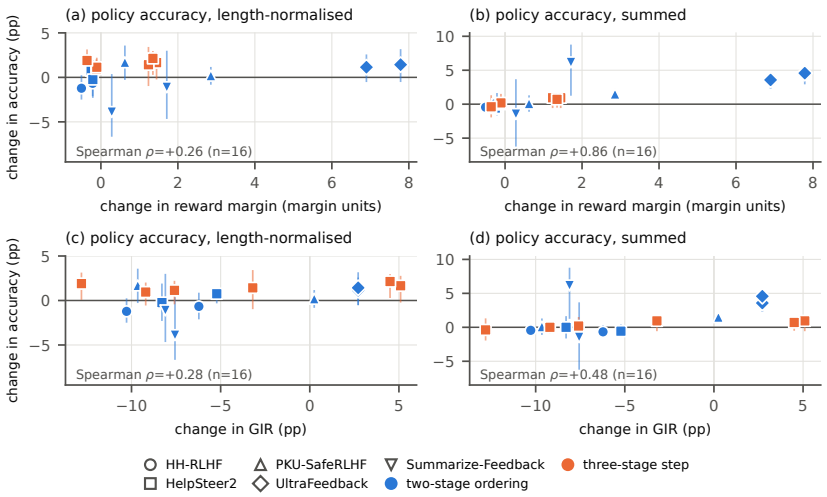

Figure 1: Change in reference-relative metrics against change in policy accuracy for the ten two-stage orderings (blue) and the six three-stage steps (orange). The top row shows the change in reward margin and the bottom row the change in GIR. The left column shows length-normalised and the right column summed policy accuracy. Each point is the mean change and each bar is the 95% interval. The points share datasets, so they are not independent.

$p < 0.005$). The Summarize-Feedback gain is partly a length effect, since its length-normalised accuracy changes by -1.1 points (Appendix B).

Figure 1 relates the reference-relative changes to the policy-accuracy changes across all 16 points. Margin changes track summed policy accuracy (Spearman $\rho = 0.86$), partly by construction, since both come from the same summed log-probabilities and summed accuracy is sensitive to response length. They do not track length-normalised accuracy ($\rho = 0.26$), and GIR changes track neither strongly ($\rho = 0.28$ and 0.48). The divergence is one-sided. Large margin gains coincide with accuracy gains, which are significant for summed accuracy in UltraFeedback and PKU-SafeRLHF harmless→helpful. In contrast, the margin and GIR declines in the two-stage HH-RLHF and HelpSteer2 orderings correspond to policy-accuracy changes of at most 1.2 points, none of them significant. These declines are therefore more consistent with an erosion of the reward shift learned in Stage 1 than with reversals of the policy's preferences.

4.1.2 Pair-Level Heterogeneity

Aggregate changes hide variation across pairs. The fraction of pairs whose length-normalised policy margin increases after Stage 2 ranges from 39% (HH-RLHF helpful→harmless) to 80% (UltraFeedback) (Appendix E). The reward margin is also concentrated in a few pairs. The top 2% of pairs contribute 86.5% of the total margin change in PKU-SafeRLHF helpful→harmless and 99.4% in Summarize-Feedback coverage→accuracy, and almost all of them already had a positive reward margin. A few pairs whose margin becomes much larger can therefore raise the mean margin while GIR falls. This concerns the reference-relative margin, not policy ranking.

4.1.3 Three-Stage Sequences

To test whether the pattern persists beyond two stages, we train a third objective, correctness, on both HelpSteer2 orderings (Table 2, with all checkpoints in Appendix D). The divergence persists, and in the second sequence GIR keeps falling. For verbosity, GIR falls by 5.2 points after Stage 2 and by 12.8 points cumulatively after Stage 3. Over the same steps, length-normalised policy accuracy changes by $+0.8$ and then by $+1.9$ points cumulatively (interval $+0.2$ to $+3.6$), and summed accuracy changes by -0.4 points (interval -2.0 to $+1.3$). In the first sequence, the margin of coherence ends above its Stage-1 value ($+1.24$) while its GIR remains 3.2 points lower. No Stage-3

Quantity	Δ margin	Δ GIR	Δ policy acc. (norm.)	Δ policy acc. (summed)
Chain X: coherence → verbosity → correctness				
A retention, S1→S2	−0.21	−8.3	−0.2 [−2.3, +1.8]	0.0 [−1.6, +1.6]
A incremental, S2→S3	+1.45	+5.1	+1.7 [−0.0, +3.3]	+0.9 [−0.4, +2.3]
A cumulative, S1→S3	+1.24	−3.2	+1.4 [−0.7, +3.6]	+0.9 [−0.4, +2.3]
B retention, S2→S3	−0.12	−9.2	+0.9 [−0.4, +2.3]	0.0 [−1.0, +1.0]
C acquisition, S3 vs base	+2.75	+16.1	+2.0 [−0.3, +4.3]	**+1.7 [+0.1, +3.3]**
Chain Y: verbosity → coherence → correctness				
A retention, S1→S2	−0.25	−5.2	+0.8 [−0.2, +1.7]	−0.6 [−1.5, +0.4]
A incremental, S2→S3	−0.11	−7.6	+1.1 [−0.3, +2.6]	+0.2 [−1.2, +1.6]
A cumulative, S1→S3	−0.36	−12.8	**+1.9 [+0.2, +3.6]**	−0.4 [−2.0, +1.3]
B retention, S2→S3	+1.35	+4.5	**+2.1 [+0.5, +3.8]**	+0.7 [−0.3, +1.7]
C acquisition, S3 vs base	+3.18	+17.4	+1.4 [−1.1, +4.0]	**+1.7 [+0.1, +3.3]**

Table 2: Three-stage sequences on HelpSteer2 (A→B→C; S1–S3: checkpoints after each stage). Later minus earlier checkpoint; acquisition of C against the base model. Conventions as in Table 1.

accuracy change is a significant decline. Four of the 20 accuracy intervals in Table 2 exclude zero without correction, and all four are improvements.

4.1.4 Mechanistic Diagnostics: Gradient Conflict, Effective Update, and Hidden-State Drift

Mechanistic diagnostics. The cosine similarity between the live Stage-2 gradient and a Stage-1-objective diagnostic gradient is near zero in all ten orderings (-0.007 to $+0.028$), which does not reveal systematic opposition between aggregate gradients and does not rule out pair-level interference. Raw LoRA movement is only 2.7–16.1% of the Stage 1 norm, yet the rescaling-invariant effective update $\Delta W = (\alpha/r)\, BA$ (Eq. 6) moves 83–111% of it while remaining near-orthogonal to the Stage 1 direction in every ordering. Hidden-state drift is small in six orderings (cosine similarity at or above 0.98) and larger in both UltraFeedback and both Summarize-Feedback orderings (down to 0.92). Near-orthogonal movement is consistent with either outcome (Appendix Tables 13, 14).

4.2 Predict

4.2.1 Stage-1 Compatibility Is a Coarse Diagnostic of Margin Retention

Compatibility $C_{A\to B}$ and retention $R_{A|B}$ (Eqs. 7–8) are computed for all ten orderings from the per-pair scores of Table 1 (the C column; all metric spaces in Appendix Table 9). The sign of $C_{A\to B}$ matches the sign of $R_{A|B}$ in nine of ten orderings. In eight orderings the sign of C is identical in all 500 random half-samples of the incoming objective's test pairs, and it matches R in all eight (negative C with margin degradation in HH-RLHF and HelpSteer2 coherence→verbosity, positive C with non-negative retention elsewhere). In the other two, HelpSteer2 verbosity→coherence ($C = +0.01$; the single mismatch) and Summarize-Feedback accuracy→coverage ($C = +0.12$), the sign agrees with its full-sample value in only 50% and 62% of halves.

The magnitude of $C_{A\to B}$ correlates with $R_{A|B}$ (Pearson $r = 0.94$, $p < 0.001$; Spearman $\rho = 0.83$, $p = 0.003$, $n = 10$), but the ten orderings come from five datasets: the signs agree in both orderings for four of five datasets (exact 95% interval [0.28, 0.99]). A leave-one-dataset-out (LODO) sign rule fitted on the other four datasets predicts 7 of 10 retention signs against 6 of 10 for the majority outcome. C uses the incoming objective's test pairs and R the first objective's. These are separate test sets, although some pairs occur in both (Appendix O). On 500 training-pool pairs of the incoming objective the sign of C is unchanged in eight of ten orderings and matches R in nine (Appendix C). The signs of C and R are stable across three seeds in four orderings (Appendix I) and replicate on Qwen2.5-7B Qwen et al. (2025) for HH-RLHF (Appendix J). Compatibility describes margin changes rather than policy ranking. Its sign matches retention in six of ten orderings for GIR, and in eight of ten for each of the two policy accuracies, although most of those accuracy changes are indistinguishable from zero (Appendix Table 9). A pair-level first-order analysis is in Appendix F.

		Retention of first objective				Acquisition of second	
Ordering	C	margin	GIR	acc. (norm.)	acc. (summed)	margin	acc. (norm.)
HH Ha→He	−	**+0.91**	**+4.8**	+0.8	+0.8	**−0.10**	0.0
HH He→Ha	−	**+0.49**	**+8.4**	+0.9	−0.2	−0.01	+0.2
HS2 Ve→Co	$\approx 0^{\ddagger}$	**+0.24**	+1.8	−0.8	+0.2	−0.11	+0.2
HS2 Co→Ve	−	**+0.23**	+3.1	−0.9	+0.5	**−1.35**	+1.1
PKU He→Ha	+	**+0.31**	**+4.2**	**+1.9**	**+1.2**	**−2.87**	**−4.7**
UF IF→Hon	+	**+0.94**	0.0	+1.0	+0.7	**+0.55**	−0.1
UF Hon→IF	+	+0.39	+0.9	+1.1	−0.3	**+0.44**	+0.7
SF Co→Ac	+	**+1.49**	**+7.6**	**+4.9**	**+4.4**	−0.33	−1.1
SF Ac→Co	$\approx 0^{\ddagger}$	+0.26	+0.3	+1.1	−1.1	+0.04	0.0

Table 3: Effect of random 10% replay (replay minus no replay). Margin is in margin units and the other columns are in percentage points. Bold marks 95% intervals that exclude zero (uncorrected). C is the sign of Stage-1 compatibility, and ‡ marks a sign that is unstable under half-sampling.

4.3 Recover

4.3.1 Replay Restores Reference-Relative Margins but Rarely Changes Policy Ranking

Table 3 compares random 10% replay with no replay in nine orderings (one ordering was not run). Using 95% intervals without correction for multiple comparisons, replay improves margin retention in seven of the nine orderings and GIR retention in four. Policy accuracy improves in only two, PKU-SafeRLHF helpful→harmless and Summarize-Feedback coverage→accuracy. In these two, length-normalised accuracy rises by +1.9 and +4.9 points and summed accuracy by +1.2 and +4.4 points, where the second interval starts at +0.1. In the three orderings with negative compatibility (both HH-RLHF orderings and HelpSteer2 coherence→verbosity), replay improves margin retention but changes policy accuracy by less than 1.0 point, with intervals that include zero. Without replay, accuracy in these orderings had changed by at most 1.2 points, so there was little ranking decline to recover.

Retrospectively, both orderings in which replay improves policy accuracy have positive compatibility, and a gate that replays only where compatibility is negative would have withheld replay from four orderings in which it improved margin retention (listed in Appendix M), so compatibility does not identify where replay helps. Replay is not free: it lowers acquisition of the incoming objective in PKU-SafeRLHF helpful→harmless (by 2.87 margin units and 4.7 points of length-normalised accuracy) and in HelpSteer2 coherence→verbosity (by 1.35 margin units). None of the seven targeted-replay variants has a policy-accuracy interval that excludes zero (Appendix M). Interpolating independently trained single-objective adapters never dominates a sequential checkpoint (Appendix N).

5 Conclusion

We studied whether sequential DPO uniformly degrades earlier preference objectives, whether this degradation can be forecast before the second stage, and whether it can be reduced. Across five preference settings, reference-relative measures vary from clear declines in HH-RLHF and HelpSteer2 to positive transfer in UltraFeedback, while the policy's own ranking is much more stable. For length-normalised accuracy, a decline of more than 2.5 points is ruled out in 13 of 16 comparisons, with only one 95% interval below zero. Large reference-relative declines can therefore overstate ranking-level forgetting, and this divergence continues into a third HelpSteer2 stage. Stage-1 compatibility matches the direction of margin retention in nine of ten orderings, including all eight with stable signs under half-sampling, but remains a coarse margin-based diagnostic supported by five datasets. Replay improves margin retention in seven of nine orderings but policy accuracy in only two, with little ranking loss to recover where margins declined. Retrospectively, compatibility does not identify where replay helps. Sequential alignment should therefore be monitored using the policy's own preferences alongside reference-relative margins, rather than relying on reference-relative measures alone.

AI USE STATEMENT

In this work, we used generative AI tools primarily for writing and formatting assistance. In particular, AI assistance was used to polish the wording of sections such as the introduction and conclusion. Given the large number of experiments reported in the paper, we also used AI tools to assist in planning how the experimental results could be organised and formatted for clearer presentation. The tools were used for presentation planning rather than for conducting or interpreting the experiments themselves. The design and execution of the experiments, literature review and search process, methodological decisions, analysis, and other central scientific contributions were carried out by the authors, with only minimal AI assistance. The authors have reviewed all AI-assisted work and take full responsibility for the final content of the paper, including all text, claims, analyses, results, and presented artifacts.

ETHICS STATEMENT

This work analyses public preference datasets, two of which (HH-RLHF and PKU-SafeRLHF) contain harmful or offensive text collected for safety training. We use them only to evaluate how model preferences change and do not release generated content. Our conclusions concern how retention should be measured; they should not be read as evidence that safety behaviour is preserved in deployment, since we measure log-probability-based preferences and a small generation-level check.

REPRODUCIBILITY STATEMENT

Model, LoRA and DPO settings are given in Appendix G, dataset construction in Appendix O, and metric definitions, statistical procedures and measurement notes in Section 2 and Appendix B. The retention, compatibility and policy-accuracy tables and figures are generated by scripts from per-pair log-probabilities of each evaluated checkpoint, and the mechanistic, generation-level and merging results by the corresponding analysis scripts; the analysis scripts and per-pair scores will be released with the paper.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Yahui Fu, Zi Haur Pang, and Tatsuya Kawahara. Minority-aware satisfaction estimation in dialogue systems via preference-adaptive reinforcement learning. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 3235–3249, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.172. URL https://aclanthology.org/2025.ijcnlp-long.172/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale, 2025. URL https://arxiv.org/abs/2510.17776.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1544. URL https://aclanthology.org/2025.acl-long.1544/.

Junsong Li, Jie Zhou, Bihao Zhan, Yutao Yang, Qianjun Pan, Shilian Chen, Tianyu Huai, Xin Li, Qin Chen, and Liang He. Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31618–31626, 2026.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6470–6479, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. Sequential preference optimization: multi-dimensional preference alignment with implicit reward modeling. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i26.34963. URL https://doi.org/10.1609/aaai.v39i26.34963.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 124198–124235. Curran Associates, Inc., 2024. doi: 10.52202/079017-3946. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e099c1c9699814af0be873a175361713-Paper-Conference.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501. Curran Associates, Inc., 2024. doi: 10.52202/079017-0047. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/02fd91a387a6a5a5751e81b58a75af90-Paper-Datasets_and_Benchmarks_Track.pdf.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022.

Dehong Xu, Liang Qiu, Minseok Kim, Faisal Ladhak, and Jaeyoung Do. Aligning large language models via fine-grained supervision. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 673–680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.62. URL https://aclanthology.org/2024.acl-short.62/.

Han Zhang, Lin Gui, Yu Lei, Yuanzhao Zhai, Yehong Zhang, Zhuo Zhang, Yulan He, Hui Wang, Yue Yu, Kam-Fai Wong, Bin Liang, and Ruifeng Xu. COPR: Continual human preference learning via optimal policy regularization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5377–5398, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.281. URL https://aclanthology.org/2025.findings-acl.281/.

Yunan Zhang, Shuoran Jiang, Mengchen Zhao, Yuefeng Li, Yang Fan, Xiangping Wu, and Qingcai Chen. Gere: Towards efficient anti-forgetting in continual learning of llm via general samples replay, 2025b. URL https://arxiv.org/abs/2508.04676.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558, 2024.

APPENDIX

A FULL AGGREGATE CROSS-STAGE RESULTS

Table 4 reports the full per-stage margin and GIR breakdown referenced throughout Section 4.1, evaluated on both objectives at every stage. Table 5 gives the corresponding policy accuracies, and Table 6 the full per-ordering retention results with confidence intervals, pair counts and sign tests.

Stage	Eval Objective	Margin	Δ Margin	GIR	Δ GIR
HH-RLHF: harmless→helpful					
Stage 1	Harmless	0.912	+0.912	67.2	+17.2
Stage 1	Helpful	−0.499	−0.499	37.1	−12.9
Stage 2	Harmless	0.698	−0.214	61.0	−6.2
Stage 2	Helpful	0.293	+0.792	56.0	+18.9
HH-RLHF: helpful→harmless					
Stage 1	Helpful	0.733	+0.733	65.6	+15.6
Stage 1	Harmless	−0.368	−0.368	41.4	−8.6
Stage 2	Helpful	0.227	−0.505	55.3	−10.3
Stage 2	Harmless	0.596	+0.964	58.7	+17.3
HelpSteer2: verbosity→coherence					
Stage 1	Verbosity	4.072	+4.072	85.7	+35.7
Stage 1	Coherence	0.005	+0.005	51.7	+1.7
Stage 2	Verbosity	3.818	−0.254	80.5	−5.2
Stage 2	Coherence	1.029	+1.024	60.0	+8.3
HelpSteer2: coherence→verbosity					
Stage 1	Coherence	0.924	+0.924	65.4	+15.4
Stage 1	Verbosity	−0.339	−0.339	45.7	−4.3
Stage 2	Coherence	0.714	−0.210	57.1	−8.3
Stage 2	Verbosity	3.761	+4.100	83.4	+37.7
PKU-SafeRLHF: harmless→helpful					
Stage 1	Harmless	14.044	+14.044	95.5	+45.5
Stage 1	Helpful	0.775	+0.775	51.7	+1.7
Stage 2	Harmless	16.900	+2.856	95.7	+0.2
Stage 2	Helpful	1.604	+0.829	58.5	+6.7

Stage	Eval Objective	Margin	Δ Margin	GIR	Δ GIR
PKU-SafeRLHF: helpful→harmless					
Stage 1	Helpful	0.636	+0.636	67.6	+17.6
Stage 1	Harmless	1.282	+1.282	72.6	+22.6
Stage 2	Helpful	1.259	+0.623	58.0	−9.7
Stage 2	Harmless	13.169	+11.887	95.6	+23.0
UltraFeedback: IF→honesty					
Stage 1	IF	8.166	+8.166	85.0	+35.0
Stage 1	Honesty	7.452	+7.452	84.1	+34.1
Stage 2	IF	15.064	+6.898	87.7	+2.7
Stage 2	Honesty	13.547	+6.095	85.6	+1.5
UltraFeedback: honesty→IF					
Stage 1	Honesty	5.324	+5.324	82.0	+32.0
Stage 1	IF	6.113	+6.113	85.6	+35.6
Stage 2	Honesty	13.110	+7.786	84.7	+2.7
Stage 2	IF	15.702	+9.589	87.7	+2.1
Summarize-Feedback: coverage→accuracy					
Stage 1	Coverage	5.026	+5.026	87.3	+37.3
Stage 1	Accuracy	1.256	+1.256	61.2	+11.2
Stage 2	Coverage	5.308	+0.282	79.8	−7.6
Stage 2	Accuracy	4.590	+3.334	74.3	+13.1
Summarize-Feedback: accuracy→coverage					
Stage 1	Accuracy	2.960	+2.960	83.2	+33.2
Stage 1	Coverage	0.124	+0.124	53.9	+3.9
Stage 2	Accuracy	4.671	+1.712	75.1	−8.1
Stage 2	Coverage	5.202	+5.078	84.1	+30.2

Table 4: Aggregate DPO-relative reward margin and Gap-Improvement Rate (GIR, Eq. 3, in %) computed with the fixed base model as evaluation reference. The base model's margin is 0.000 and its GIR is 50.0% for every objective by construction, so base rows are omitted. For Stage 1, the raw margin equals its Δ margin, while Δ GIR is measured relative to the 50% base value. Stage 1 rows report evaluation on both the trained objective and the untrained objective, showing cross-objective effects before Stage 2 begins. Δ columns show change from the same objective's most recent checkpoint (later minus earlier), so negative values indicate degradation. Δ GIR values are percentage-point changes. GIR is not a ranking accuracy, and policy accuracies are in Appendix Table 5.

B MEASUREMENT NOTES

Metric summary. Table 7 summarises the four metrics used throughout the paper.

Scoring procedure. Unless noted otherwise, the retention and policy-accuracy metrics in the paper are computed from per-pair summed log-probabilities and response lengths of each evaluated checkpoint, produced by a single scoring procedure (bf16 inference, batch size 8, maximum prompt and total length of 512 and 1,024 tokens, and the evaluation subsets of the original evaluation); all reference-relative quantities use the fixed base model. The second-model check and the adapter-merging comparison instead use the original evaluation logs of those runs. This procedure reproduces the per-pair margins of the original evaluation logs of the same checkpoints with Spearman correlation of at least 0.994 outside HelpSteer2 and 0.947 overall; the mean margin of a checkpoint differs by less than 0.03 margin units and the mean absolute per-pair difference is at most 0.45, largest for the datasets with the longest responses (HelpSteer2, UltraFeedback).

Precision. Summed log-probabilities are rounded to bf16, whose spacing is 0.25–1 nats at the median magnitudes of the summed log-probabilities here (62–254). Exact ties between the chosen and rejected summed log-probabilities occur for 0.0–1.1% of pairs. Scoring the same checkpoint twice changes the length-normalised policy-accuracy outcome of at most 9 pairs (out of 179–1,000 per test set), whereas GIR, which depends on pairs whose margin is close to zero, changes by up

Ordering	Objective	Policy acc. (norm.), % base	S1	S2	Policy acc. (summed), % base	S1	S2
HH Ha→He	harmless (first)	46.6	48.4	47.8	57.2	58.3	57.7
HH Ha→He	helpful (second)	59.9	59.7	61.6	43.0	42.6	42.9
HH He→Ha	helpful (first)	59.9	63.3	62.1	43.0	43.7	43.2
HH He→Ha	harmless (second)	46.6	45.4	46.9	57.2	57.1	57.9
HS2 Ve→Co	verbosity (first)	72.9	74.4	75.1	16.1	17.5	16.9
HS2 Ve→Co	coherence (second)	51.7	51.2	52.6	48.6	48.1	48.8
HS2 Co→Ve	coherence (first)	51.7	52.4	52.1	48.6	48.6	48.6
HS2 Co→Ve	verbosity (second)	72.9	73.2	74.4	16.1	15.7	16.9
PKU Ha→He	harmless (first)	60.7	85.4	85.6	66.8	78.0	79.5
PKU Ha→He	helpful (second)	46.9	48.5	49.3	35.6	36.0	36.3
PKU He→Ha	helpful (first)	46.9	48.1	49.8	35.6	35.8	35.9
PKU He→Ha	harmless (second)	60.7	66.4	84.8	66.8	68.5	77.3
UF IF→Hon	IF (first)	74.7	80.7	81.9	42.7	46.0	49.6
UF IF→Hon	honesty (second)	70.3	76.0	77.7	48.1	50.7	54.7
UF Hon→IF	honesty (first)	70.3	74.7	76.1	48.1	50.4	55.0
UF Hon→IF	IF (second)	74.7	78.9	81.6	42.7	44.7	49.9
SF Co→Ac	coverage (first)	52.7	69.8	65.9	33.7	55.6	54.1
SF Co→Ac	accuracy (second)	82.1	80.4	83.2	72.1	82.1	84.4
SF Ac→Co	accuracy (first)	82.1	85.5	84.4	72.1	78.8	84.9
SF Ac→Co	coverage (second)	52.7	48.8	64.4	33.7	37.1	53.7

Table 5: Policy pairwise accuracy of the base model and of the Stage 1 (S1) and Stage 2 (S2) checkpoints on both objectives of every ordering. Objectives with much longer chosen than rejected responses (e.g. HelpSteer2 verbosity) have very low summed accuracy at the base model because summed log-probabilities penalise length, so both scorings are reported throughout.

Ordering	n	Δ margin [95% CI]	Δ GIR [95% CI]	Policy acc. (norm.) Δ (pp) [95% CI]	lost/gained	p	Policy acc. (summed) Δ (pp) [95% CI]	lost/gained	p
HH Ha→He	900	−0.21 [−0.32, −0.11]	−6.2 [−9.3, −3.2]	−0.7 [−2.1, +0.8]	25/19	0.451	−0.7 [−1.5, +0.2]	11/5	0.210
HH He→Ha	900	−0.51 [−0.60, −0.41]	−10.3 [−13.4, −7.2]	−1.2 [−2.6, +0.1]	25/14	0.108	−0.4 [−1.1, +0.2]	6/2	0.289
HS2 Ve→Co	527	−0.25 [−0.45, −0.05]	−5.2 [−8.1, −2.5]	+0.8 [−0.2, +1.7]	1/5	0.219	−0.6 [−1.5, +0.4]	4/1	0.375
HS2 Co→Ve	422	−0.21 [−0.69, +0.27]	−8.3 [−14.2, −2.4]	−0.2 [−2.3, +1.8]	10/9	1.000	0.0 [−1.6, +1.6]	6/6	1.000
PKU Ha→He	1000	+2.86 [+2.60, +3.13]	+0.2 [−0.8, +1.2]	+0.2 [−0.8, +1.2]	11/13	0.839	**+1.5 [+0.6, +2.4]**	2/17	<0.001
PKU He→Ha	1000	+0.62 [+0.31, +0.96]	−9.7 [−12.9, −6.4]	+1.7 [−0.2, +3.6]	38/55	0.097	+0.1 [−1.1, +1.3]	17/18	1.000
UF IF→Hon	700	+6.90 [+6.33, +7.47]	+2.7 [+0.6, +5.0]	+1.1 [−0.4, +2.7]	11/19	0.200	**+3.6 [+2.1, +5.0]**	2/27	<0.001
UF Hon→IF	700	+7.79 [+6.99, +8.58]	+2.7 [+0.3, +5.2]	+1.4 [−0.4, +3.3]	16/26	0.164	**+4.6 [+2.8, +6.3]**	5/37	<0.001
SF Co→Ac	205	+0.28 [−0.37, +0.96]	−7.6 [−13.7, −1.5]	**−3.9 [−7.7, −0.1]**	12/4	0.077	−1.5 [−6.2, +3.3]	14/11	0.690
SF Ac→Co	179	+1.71 [+0.91, +2.50]	−8.1 [−13.7, −2.5]	−1.1 [−5.2, +2.9]	7/5	0.774	**+6.1 [+1.7, +10.8]**	3/14	0.013

Table 6: Full retention results for the ten two-stage orderings. Margin and GIR intervals are paired percentile bootstraps over pairs (4,000 resamples); accuracy intervals are Newcombe score intervals for the paired difference of proportions, and p is the exact two-sided sign test on lost (correct→incorrect) versus gained (incorrect→correct) pairs. Ties in the log-probability comparison count as incorrect (they occur in about 1% of pairs or fewer). Bold: the accuracy interval excludes zero. Only length-normalised accuracy in Summarize-Feedback coverage→accuracy has an interval entirely below zero.

to 2.0 percentage points and flips the outcome of up to 9% of pairs (HelpSteer2). Changes in GIR below about two percentage points, and pair-level margin changes in HelpSteer2, should therefore be interpreted with this noise in mind; changes in mean margin and in policy accuracy are much less affected.

Length confound of summed accuracy. Summed log-probabilities favour shorter responses. Table 8 shows that the base model's summed accuracy is close to the accuracy of a rule that always prefers the shorter response for several objectives, and that it is far from length-normalised accuracy where chosen responses are systematically longer: for HelpSteer2 verbosity, the chosen response is shorter in only 8.5% of pairs, and the base model's accuracy is 16.1% summed against 72.9% length-normalised. We therefore report both accuracies. In Summarize-Feedback accuracy→coverage the two scorings flip different pairs: 14 pairs are gained and 3 lost under summed scoring, against 5 gained and 7 lost under length-normalised scoring. The pairs gained under summed scoring are mostly ones whose chosen response is longer (the chosen response is shorter in only 7% of them), whereas all pairs lost under length-normalised scoring have a shorter chosen response (by 20 tokens on average). On the quarter of pairs with the smallest length difference ($n = 47$), both accuracies

Metric	Definition	Reference	Log-prob.	Asks
Reward margin	$r_i = \beta h_\theta$ (Eq. 2)	base model	summed	how far the gap moved
GIR	share of pairs with $r_i > 0$	base model	summed	how often the gap widened
Policy acc. (norm.)	$m_i > 0$	none	per-token mean	does the policy prefer the chosen response
Policy acc. (summed)	$m_i^{\Sigma} > 0$	none	summed	same, length-sensitive

Table 7: Metrics used throughout. The first two are reference-relative and the last two use the policy alone.

Dataset	Objective	n	chosen shorter	"shorter wins" acc. (%)	summed acc. (%)	norm. acc. (%)
HH-RLHF	harmless	900	55.1%	55.8	57.2	46.6
HH-RLHF	helpful	900	39.4%	40.2	43.0	59.9
HelpSteer2	coherence	422	45.3%	45.9	48.6	51.7
HelpSteer2	correctness	347	44.7%	44.8	45.5	53.9
HelpSteer2	verbosity	527	8.5%	8.8	16.1	72.9
PKU-SafeRLHF	harmless	1000	65.1%	65.5	66.8	60.7
PKU-SafeRLHF	helpful	1000	34.5%	35.2	35.6	46.9
UltraFeedback	honesty	700	39.1%	39.5	48.1	70.3
UltraFeedback	IF	700	31.9%	32.1	42.7	74.7
Summarize-Feedback	accuracy	179	46.9%	48.6	72.1	82.1
Summarize-Feedback	coverage	205	33.2%	34.4	33.7	52.7

Table 8: Response-length bias of each objective's test pairs, and base-model policy accuracy. "Shorter wins" is the accuracy of a rule that always prefers the shorter response (ties count half). Where chosen responses are systematically longer (e.g. HelpSteer2 verbosity), summed accuracy is far below length-normalised accuracy; where the base model's summed accuracy tracks the "shorter wins" rule it is largely a length signal.

rise by 2.1 points. We therefore do not interpret the $+6.1$-point summed gain in this ordering as evidence of transfer.

C Compatibility Analyses

Table 9 reports Stage-1 compatibility $C_{A\to B}$ and retention $R_{A|B}$ for every ordering in margin space, GIR space and length-normalised policy-accuracy space, together with the stability of the sign of $C_{A\to B}$ across random halves of the incoming objective's test pairs.

Definitions. In GIR and policy-accuracy space, $C_{A\to B}$ and $R_{A|B}$ are computed as in Eqs. 7–8 with M_θ replaced by the corresponding metric; compatibility in GIR space is the GIR of the Stage-1 checkpoint on the incoming objective minus 0.5, and in policy-accuracy space the accuracy of the Stage-1 checkpoint minus that of the base model. The signal-asymmetry ratio S is the larger divided by the smaller Stage-1 margin of the two objectives, each measured on the objective trained first, and is defined only when both are positive: $S = 1.25$ (HH-RLHF), 1.53 (UltraFeedback), 1.70 (Summarize-Feedback), 4.41 (HelpSteer2) and 22.08 (PKU-SafeRLHF).

Sign stability. The sign of $C_{A\to B}$ equals its full-sample sign in every one of 500 random halves of the incoming objective's test pairs for eight orderings. For HelpSteer2 verbosity→coherence ($C = +0.01$) and Summarize-Feedback accuracy→coverage ($C = +0.12$) the fraction is only 50% and 62%, so we describe these two signs as unstable; the partition into eight stable and two unstable signs is the same for any stability threshold between 62% and 100%, and no threshold was specified in advance. HelpSteer2 verbosity→coherence is also the ordering in which the sign of C differs from that computed from the original evaluation logs (-0.010 there), consistent with a value that is indistinguishable from zero.

Training-pool diagnostic check. $C_{A\to B}$ is computed on the incoming objective's test pairs, whereas $R_{A|B}$ uses the first objective's test pairs. These are separate test sets, although some pairs occur in both (Appendix O). To check that the sign of C does not depend on the particular test pairs, we recomputed it on 500 random pairs from the incoming objective's training pool (seed 0), which are disjoint from the incoming objective's test pairs. In PKU-SafeRLHF, UltraFeedback and Summarize-Feedback some of them also occur in the first objective's training data

	Margin space				GIR (pp)		Policy acc. norm. (pp)	
Ordering	C	R	sign match	stability	C	R	C	R
HH Ha→He	−0.50	−0.21	yes	1.00	−12.9	−6.2	−0.2	−0.7
HH He→Ha	−0.37	−0.51	yes	1.00	−8.6	−10.3	−1.1	−1.2
HS2 Ve→Co	+0.01	−0.25	no	0.50	+1.7	−5.2	−0.5	+0.8
HS2 Co→Ve	−0.34	−0.21	yes	1.00	−4.3	−8.3	+0.4	−0.2
PKU Ha→He	+0.78	+2.86	yes	1.00	+1.7	+0.2	+1.6	+0.2
PKU He→Ha	+1.28	+0.62	yes	1.00	+22.6	−9.7	+5.7	+1.7
UF IF→Hon	+7.45	+6.90	yes	1.00	+34.1	+2.7	+5.7	+1.1
UF Hon→IF	+6.11	+7.79	yes	1.00	+35.6	+2.7	+4.1	+1.4
SF Co→Ac	+1.26	+0.28	yes	1.00	+11.2	−7.6	−1.7	−3.9
SF Ac→Co	+0.12	+1.71	yes	0.62	+3.9	−8.1	−3.9	−1.1

Table 9: Stage-1 compatibility C and retention R for every ordering, in margin space, GIR space and length-normalised policy-accuracy space. Stability is the fraction of 500 random halves of the incoming objective's test pairs for which the sign of C equals its full-sample sign.

(Appendix O). This dev-split C correlates with the test-pair C at Pearson $r = 0.997$ across the ten orderings and has the same sign in eight of them; the two that differ are the two whose sign is unstable (HelpSteer2 verbosity→coherence, -0.46 on the dev split against $+0.01$; Summarize-Feedback accuracy→coverage, -0.01 against $+0.12$). The dev-split sign matches R in nine of ten orderings (the exception is Summarize-Feedback accuracy→coverage, whose dev-split C is indistinguishable from zero), and its Spearman correlation with R is 0.85. This split was scored after the main analysis; it is a check, not a tuned quantity.

Generalisation statistics. The signs of C and R agree in both orderings for four of the five datasets (exact 95% interval for the proportion [0.28, 0.99]). A bootstrap over datasets (2,000 resamples) gives a Spearman correlation between C and R of 0.83 with 95% interval [0.10, 0.83]. For the leave-one-dataset-out rule, we fit an ordinary least-squares line $R = aC + b$ on the orderings of the other four datasets and use the sign of the fitted value for the held-out orderings. This rule predicts seven of ten signs, and always predicting the majority outcome (retention) predicts six of ten; its errors are HH-RLHF helpful→harmless and both HelpSteer2 orderings, for which the fitted rule predicts small positive retention ($+0.03$, $+0.45$, $+0.11$). In GIR space the correlation between C and R across orderings is Pearson $r = 0.62$ ($p = 0.058$) and Spearman $\rho = 0.46$ ($p = 0.18$).

Robust variants. Replacing the mean margin in C by the median, the 10% trimmed mean, the fraction of pairs with positive margin minus one half, or the lower quartile gives sign agreement with R in nine, nine, ten and six of ten orderings, against nine for the mean. These variants were not specified in advance; the fraction-based variant agrees in all ten orderings, including the two whose mean-based sign is unstable, but the stability of its own sign was not evaluated, so we do not treat it as an improvement.

D Three-Stage Sequences: Details

Design. Both sequences continue from the two-stage HelpSteer2 checkpoints (coherence→verbosity and verbosity→coherence) and train a third objective, correctness, with the same hyperparameters as the two-stage HelpSteer2 runs ($\beta = 0.3$, learning rate 5×10^{-5}, effective batch size 32, one epoch, LoRA $r = 32$, $\alpha = 128$) on 1,492 training pairs (47 optimiser steps), with the Stage-2 checkpoint as the DPO reference. Every checkpoint is evaluated on the test sets of all three objectives (422 coherence, 527 verbosity and 347 correctness pairs). Table 10 reports all checkpoints and objectives; the Stage-1→Stage-2 changes in Table 2 are the two-stage results of Table 1.

Compatibility at later stages. Recomputing compatibility at Stage 2 as the margin of the Stage-2 checkpoint on the incoming third objective gives $C_{B\to C} = +0.85$ and $+1.31$ for the two sequences. Its sign matches the sign of the second objective's retention $R_{B|C}$ (-0.12 and $+1.35$) in one of the two sequences, and the sign of the first objective's change during Stage 3 ($+1.45$ and -0.11) in one.

Chain	Objective	Ckpt	Margin	GIR (%)	Acc. norm. (%)	Acc. summed (%)
X	A=coherence	S1	0.92	65.4	52.4 (51.7)	48.6 (48.6)
X	A=coherence	S2	0.71	57.1	52.1 (51.7)	48.6 (48.6)
X	A=coherence	S3	2.16	62.2	53.8 (51.7)	49.5 (48.6)
X	B=verbosity	S1	−0.34	45.7	73.2 (72.9)	15.7 (16.1)
X	B=verbosity	S2	3.76	83.4	74.4 (72.9)	16.9 (16.1)
X	B=verbosity	S3	3.64	74.2	75.3 (72.9)	16.9 (16.1)
X	C=correctness	S1	1.05	67.3	54.5 (53.9)	45.8 (45.5)
X	C=correctness	S2	0.85	56.5	54.2 (53.9)	45.5 (45.5)
X	C=correctness	S3	2.75	66.1	55.9 (53.9)	47.3 (45.5)
Y	A=verbosity	S1	4.07	85.7	74.4 (72.9)	17.5 (16.1)
Y	A=verbosity	S2	3.82	80.5	75.1 (72.9)	16.9 (16.1)
Y	A=verbosity	S3	3.71	72.9	76.3 (72.9)	17.1 (16.1)
Y	B=coherence	S1	0.01	51.7	51.2 (51.7)	48.1 (48.6)
Y	B=coherence	S2	1.03	60.0	52.6 (51.7)	48.8 (48.6)
Y	B=coherence	S3	2.38	64.5	54.7 (51.7)	49.5 (48.6)
Y	C=correctness	S1	0.11	51.0	53.0 (53.9)	45.2 (45.5)
Y	C=correctness	S2	1.31	57.5	54.2 (53.9)	46.1 (45.5)
Y	C=correctness	S3	3.18	67.4	55.3 (53.9)	47.3 (45.5)

Table 10: All checkpoints and objectives of the two three-stage HelpSteer2 chains (values in parentheses: base model).

E Pair-Level Analyses

This appendix defines and reports the pair-level analyses summarised in Section 4.1.2. These analyses concern continuous policy-margin magnitude and redistribution, and the reference-relative reward margin, rather than binary policy-ranking accuracy.

Pairwise margin drop. For each pair i we compute the length-normalised policy margin m_i (Eq. 4) at each checkpoint and define the per-pair margin drop between Stage 1 and Stage 2 as

$$D_i = m_i^{(1)} - m_i^{(2)} \tag{10}$$

Positive values indicate that the pair's policy preference margin decreased after Stage 2, while negative values indicate that it increased. We write this quantity as a drop, D_i, because it is the only quantity in this paper defined as earlier minus later; all other changes (Δr_i, Δ_i^w, Δ_i^l, retention $R_{A|B}$, and the Δ columns of the tables) are later minus earlier, so negative values indicate degradation.

Quartile decomposition. Evaluation pairs are sorted by their Stage 1 policy margin and divided into four equally sized groups. Q1 contains the lowest-margin (hard, low-confidence) pairs and Q4 the highest-margin (easy, high-confidence) pairs. The mean margin drop is computed within each quartile:

$$\overline{D}_{Q_k} = \frac{1}{|Q_k|} \sum_{i \in Q_k} D_i \tag{11}$$

Chosen/rejected decomposition. Writing $\ell_i^w = \overline{\log \pi_\theta}(y_{w,i}|x_i)$ and $\ell_i^l = \overline{\log \pi_\theta}(y_{l,i}|x_i)$ for the two terms of m_i, so that $m_i = \ell_i^w - \ell_i^l$, and indexing Stage-1 and Stage-2 values by (1) and (2),

$$\Delta_i^w = \ell_i^{w,(2)} - \ell_i^{w,(1)}, \qquad \Delta_i^l = \ell_i^{l,(2)} - \ell_i^{l,(1)} \tag{12}$$

measure the Stage 1-to-Stage 2 changes in the chosen- and rejected-response log-probabilities; positive values indicate increased mean per-token likelihood. With a tolerance band of $\epsilon = 0.01$ nats per token, each pair is assigned to one of five regimes (Appendix L).

Concentration. We sort pairs by their per-pair reward-margin change $\Delta r_i = r_i^{(2)} - r_i^{(1)}$ (Eq. 2) and compute, for the top $k = \mathrm{round}(0.02\,n)$ pairs, their share of $\sum_i \Delta r_i$ and the fraction with $r_i^{(1)} > 0$ at Stage 1.

Results. Table 11 reports the mean drop. It is positive (degradation of the length-normalised policy margin) only in HH-RLHF (+0.051 and +0.063; 43% and 39% of pairs improve) and negative elsewhere, with the strongest and most uniform improvement in UltraFeedback (80% and 80% of

Dataset/Order	Obj.	n	Mean D	Improved	Top-2% share	Already positive
HH Ha→He	Harmless	900	+0.051	43.3%	−42.4%	61.1%
HH He→Ha	Helpful	900	+0.063	39.1%	−12.7%	27.8%
HS2 Ve→Co	Verbosity	527	−0.057	62.2%	−49.1%	81.8%
HS2 Co→Ve	Coherence	422	−0.017	46.2%	−107.7%	37.5%
PKU Ha→He	Harmless	1000	−0.087	70.1%	+10.7%	100.0%
PKU He→Ha	Helpful	1000	−0.029	49.9%	+86.5%	95.0%
UF IF→Hon	IF	700	−0.551	80.4%	+9.5%	100.0%
UF Hon→IF	Honesty	700	−0.695	79.7%	+10.2%	100.0%
SF Co→Ac	Coverage	205	−0.072	52.2%	+99.4%	100.0%
SF Ac→Co	Accuracy	179	−0.147	63.7%	+19.8%	100.0%

Table 11: Pair-level summaries. Mean D: average drop in length-normalised policy margin (Eq. 10; positive = degradation, negative = improvement); Improved: fraction of pairs whose margin increased after Stage 2. Top-2% share: share of the total reference-relative reward-margin change contributed by the top 2% of pairs ($k = \mathrm{round}(0.02\,n)$, ranked by Δr_i); Already positive: fraction of those pairs whose reward margin was already positive at Stage 1 ($r_i^{(1)} > 0$). The last two columns describe the reference-relative margin, not policy ranking.

pairs improve). Figure 2 decomposes it by Stage-1 margin quartile. HH-RLHF harmless→helpful is the only ordering whose largest change occurs in Q1, whereas helpful→harmless degrades most strongly in Q4. HelpSteer2 verbosity→coherence improves most in Q4, while coherence→verbosity changes little. Both PKU-SafeRLHF orderings improve increasingly towards Q4. UltraFeedback shows the largest redistribution, with mild Q1 degradation but strong Q4 improvement in both orderings. Summarize-Feedback coverage→accuracy follows a hard-to-easy progression, whereas accuracy→coverage fluctuates. The largest-magnitude change occurs in Q4 in nine of ten orderings, and Q4 improves in nine and degrades only in HH-RLHF helpful→harmless.

In the three orderings in which the reward margin rises while GIR falls (PKU-SafeRLHF helpful→harmless and both Summarize-Feedback orderings), the top 2% of pairs contribute 86.5%, 99.4% and 19.8% of the total margin increase (last two columns of Table 11), compared with 9.5–10.7% in the three orderings in which margin and GIR both rise (PKU-SafeRLHF harmless→helpful and both UltraFeedback orderings), and 95–100% of the pairs in each high-change subgroup already had a positive reward margin at Stage 1. A small number of already-positive pairs whose margin becomes much larger can therefore raise the mean reward margin while the fraction of pairs whose gap widens (GIR) falls. This is a statement about reference-relative quantities. The policy-accuracy changes in the same orderings are +1.7 (length-normalised) and +0.1 (summed) points for PKU-SafeRLHF helpful→harmless, and −3.9 and −1.1 (length-normalised) and −1.5 and +6.1 (summed) points for Summarize-Feedback coverage→accuracy and accuracy→coverage (Table 1). For negative net changes (HH-RLHF, HelpSteer2), a share of opposite sign or above 100% in magnitude reflects the top subgroup and the remaining pairs moving in different directions.

F First-Order Analysis of Pair-Level Margin Changes

To ask whether individual pairs' margin changes follow, to first order, from the parameter change of Stage 2, we evaluate each Stage-1 pair's reward margin at adapters $\theta_1 \pm \delta$, obtained by perturbing the Stage-1 LoRA parameters along a direction d with $\|\delta\| = f\,\|\theta_2 - \theta_1\|$, and take the central-difference slope. The analysis was specified before any per-pair relationship was computed: two primary orderings (HH-RLHF helpful→harmless and UltraFeedback instruction-following→honesty), a primary scale $f = 0.25$ with $f \in \{0.125, 0.5\}$ as sensitivity checks, and thresholds of Spearman $\rho \geq 0.30$ against the realised per-pair margin change with the expected sign and of ≥ 0.15 after controlling for the Stage-1 margin and response-length features.

Along the realised Stage-2 displacement $d = \theta_2 - \theta_1$, the slope predicts the realised per-pair margin change with $\rho = 0.82$ (HH-RLHF; partial 0.79) and $\rho = 0.91$ (UltraFeedback; partial 0.86) at $f = 0.25$, and at least 0.76 at the other scales (Table 12). Along the gradient of the incoming objective at the Stage-1 checkpoint, $d = -\nabla L_B(\theta_1)$ estimated on 512 training pairs of the in-

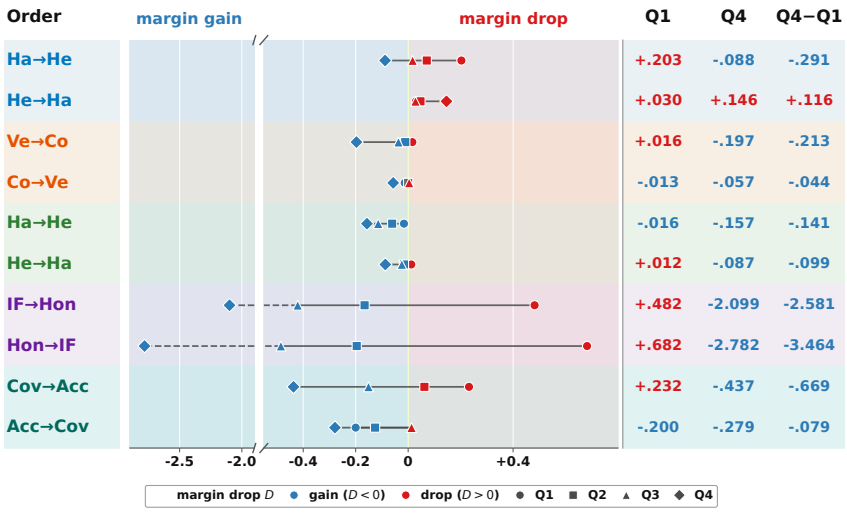

Figure 2: Quartile decomposition of the pair-level margin drop D after sequential training. Each row is an ordering (HH-RLHF, HelpSteer2, PKU-SafeRLHF, UltraFeedback, Summarize-Feedback, in the order of Table 1); the four markers are the mean drop in the Stage-1 margin quartiles Q1 (circle, hardest) to Q4 (diamond, easiest). Blue: margin gain ($D < 0$); red: margin drop ($D > 0$). The axis is broken for the two UltraFeedback Q4 values; the right-hand columns give Q1, Q4 and Q4−Q1.

Ordering	Direction	$\rho\,(f{=}0.125)$	$\rho\,(f{=}0.25)$	$\rho\,(f{=}0.5)$	partial $\rho\,(f{=}0.25)$
HH-RLHF helpful→harmless	realised displacement $\theta_2 - \theta_1$	+0.76	+0.82	+0.85	+0.79
HH-RLHF helpful→harmless	gradient of incoming objective	+0.46	+0.42	+0.34	+0.27
UltraFeedback IF→honesty	realised displacement $\theta_2 - \theta_1$	+0.87	+0.91	+0.93	+0.86
UltraFeedback IF→honesty	gradient of incoming objective	+0.12	−0.13	+0.03	−0.03

Table 12: Spearman correlation between the central-difference slope of each Stage-1 pair's reward margin and its realised Stage-2 margin change, for two perturbation directions and three perturbation scales f (perturbation norm $= f\,\|\theta_2 - \theta_1\|$). Partial: after controlling for the Stage-1 margin and response-length features.

coming objective, the corresponding correlations are $+0.42$ (HH-RLHF; partial $+0.27$) and -0.13 (UltraFeedback; partial -0.03), with a sign that is not stable across scales for UltraFeedback. The gradient of the incoming objective therefore does not predict pairs' margin changes reliably, and we do not use it as a diagnostic. At the pre-specified scale the perturbation along this direction moves the model far from the regime in which a first-order approximation holds (the mean UltraFeedback margin, about $+8$ at θ_1, falls to between -5 and -14 at $f = 0.125$ and below -170 at $f = 0.25$), which limits what this test can establish. No other scales or directions were evaluated. The scoring procedure reproduces the stored margins with Spearman correlation 0.9999 (HH-RLHF) and 0.997 (UltraFeedback); the mean margin differs by at most 0.025 in absolute value, which meets the pre-specified check (Spearman correlation of at least 0.99 and a mean difference within 0.05). The mean absolute per-pair difference is 0.004 and 0.40 margin units, and the larger UltraFeedback value is consistent with accumulated bf16 numerical differences on longer responses.

G Model, Training, and Implementation Details

G.1 Base Model

Unless otherwise stated, the main experiments use Llama-3.1-8B-Instruct Grattafiori et al. (2024) as the base language model; Appendix J reports a replication on Qwen2.5-7B-InstructQwen et al. (2025). The instruction-tuned models are used because they represent a realistic starting point for sequential alignment and provide a consistent chat template for formatting preference data. The model contains approximately 8B parameters. All experiments are run using bfloat16 precision with gradient checkpointing enabled on a single NVIDIA GPU with 48GB memory.

G.2 LoRA Configuration

We use Low-Rank Adaptation (LoRA)Hu et al. (2022) for all fine-tuning runs. The pretrained base model weights are frozen, and only LoRA adapter parameters are updated. We apply LoRA to the attention (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and MLP (`up_proj`, `down_proj`, `gate_proj`) projection matrices in every transformer layer.

Unless otherwise stated, we use LoRA rank $r = 32$, scaling factor $\alpha = 128$, dropout $p = 0.05$, and no bias adaptation. This gives 83,886,080 trainable parameters, corresponding to approximately 1% of the full model parameters.

LoRA enables exact adapter checkpointing after each sequential stage, allowing model states to be compared without storing full-model checkpoints. Its parameter efficiency also makes the full experimental suite feasible within the available computational budget.

G.3 DPO Training Hyperparameters

All training uses the HuggingFace TRL `DPOTrainer` with the sigmoid DPO loss. Common hyperparameters: learning rate 5×10^{-5} with a cosine schedule and 10% warmup, maximum gradient norm 1.0, maximum total and prompt lengths of 1,024 and 512 tokens, bfloat16 precision, gradient checkpointing, and random seed 42 (123 and 2024 for the additional seeds of Appendix I).

For HH-RLHF experiments, we use a per-device batch size of 4 and gradient accumulation of 8, giving an effective batch size of 32. Each stage trains for 1 epoch on up to 9,000 preference pairs, yielding 282 optimiser steps. We use $\beta = 0.3$.

For HelpSteer2 experiments, the attribute-specific datasets are smaller. We use a per-device batch size of 4 and gradient accumulation of 8, giving an effective batch size of 32. Each stage trains for 1 epoch, yielding 59 steps for coherence and 70 for verbosity. We use $\beta = 0.3$.

For PKU-SafeRLHF experiments, we use the same configuration as HH-RLHF, with a per-device batch size of 4, gradient accumulation of 8, effective batch size of 32, 1 epoch, and $\beta = 0.3$.

For UltraFeedback experiments, we use a per-device batch size of 2 and gradient accumulation of 4, giving an effective batch size of 8. Each stage trains for 1 epoch on 7,000 training pairs, yielding 875 optimiser steps. We use $\beta = 0.3$.

For Summarize-Feedback experiments, we use a per-device batch size of 4 and gradient accumulation of 8, giving an effective batch size of 32. Each stage trains for 1 epoch on 821 (coverage) or 717 (accuracy) training pairs, yielding 26 and 23 optimiser steps, respectively. We use $\beta = 0.3$.

G.4 Dataset Preprocessing

All datasets are converted into the TRL conversational preference format with a user prompt, a chosen assistant response, and a rejected assistant response. The Llama-3.1 instruction chat template is applied using `apply_chat_template`. Prompt tokens are masked with label value -100 so that log-probability and loss computations are applied only to response tokens.

For evaluation, the scorer stores, for the chosen and the rejected response of every pair, the summed log-probability over the non-masked response tokens and the number of response tokens. Reward margins and GIR use the summed values (relative to the base model), summed policy accuracy compares the summed values directly, and length-normalised policy accuracy compares the summed values divided by the respective token counts (mean per-token log-probabilities).

G.5 Implementation

Experiments are implemented using HuggingFace Transformers, TRL, PEFT, PyTorch, and Weights & Biases. DPO training uses the TRL `DPOTrainer`. Adapter checkpoints are saved using PEFT after every stage. Training logs, gradient diagnostics, memory usage, and final metrics are recorded for each run. The codebase and preprocessing scripts will be released upon publication.

H MECHANISTIC DIAGNOSTIC IMPLEMENTATION

H.1 GRADIENT CONFLICT MEASUREMENT IMPLEMENTATION

Gradient conflict is measured during Stage 2 training to test whether the current training objective directly opposes the previous objective in parameter space. We implement this by subclassing `DPOTrainer` as `ConflictAwareDPOTrainer`, which overrides `training_step`. After the standard DPO backward pass, the live Stage 2 gradients are available in the LoRA parameter gradients. We collect these gradients into a vector $\mathbf{g}_{\text{current}}$.

We then temporarily clear gradients and compute a diagnostic DPO loss for the previous objective on a fixed reference batch. Backpropagating this diagnostic loss gives a second gradient vector $\mathbf{g}_{\text{previous}}$. We compute cosine similarity between these two vectors:

$$\cos(\mathbf{g}_{\text{current}}, \mathbf{g}_{\text{previous}}) = \frac{\mathbf{g}_{\text{current}} \cdot \mathbf{g}_{\text{previous}}}{\|\mathbf{g}_{\text{current}}\| \|\mathbf{g}_{\text{previous}}\|}. \tag{13}$$

After measurement, the original Stage 2 gradients are restored so the optimiser step remains unchanged. All gradient vectors are moved to CPU immediately after collection to reduce GPU memory pressure.

Reference log-probabilities for the diagnostic previous-objective DPO loss are precomputed before training using the frozen stage-initial reference model. This avoids keeping an additional full reference model active during training. Measurements are taken at fixed intervals of 5 to 100 optimiser steps depending on the dataset (four to eight measurements per run), on a fixed diagnostic batch of four examples from the test set of the previous objective, so each cosine is a coarse estimate. Because gradients are measured only over trainable LoRA parameters and on a fixed diagnostic batch, this analysis should be interpreted as a LoRA-space diagnostic rather than as a full-model causal proof (cosine values are interpreted as in Section 3.1.2).

H.2 ADAPTER MOVEMENT ANALYSIS

As an auxiliary diagnostic, we analyse movement between LoRA adapter checkpoints. In the primary reported analysis, we flatten and concatenate all trainable LoRA A and B matrices into a raw adapter vector for each stage. We then compute L2 distances and cosine similarities between stage vectors and update directions.

For stages S_1 and S_2, we define the raw update direction:

$$\delta_{1\to2} = S_2 - S_1 \tag{14}$$

We compute:

- $\|S_1\|$ is the magnitude of the Stage 1 adapter.
- $\|S_2 - S_1\|$ is the movement from Stage 1 to Stage 2.
- $\cos(S_1, S_2 - S_1)$ indicates whether the Stage 2 update moves with, against, or orthogonal to the Stage 1 adapter.

We treat this analysis as supporting evidence because raw LoRA parameter space is not invariant to LoRA rescaling. Effective LoRA updates of the form BA provide a more direct representation of the induced weight change, but dense materialization of all effective updates can be memory intensive for 8B models. Therefore, raw adapter movement is reported only as an auxiliary parameter-space diagnostic.

To compute the effective-update quantities $\|\Delta W_1\|$, $\|\Delta W_2 - \Delta W_1\|/\|\Delta W_1\|$, and $\cos(\Delta W_1, \Delta W_2 - \Delta W_1)$ (Eq. 6) without materialising the full concatenated ΔW, which is memory-prohibitive at this model scale, we stream three sufficient statistics directly from the adapter weights, module by module: $\sum_j \|\Delta W_{1,j}\|^2$, $\sum_j \|\Delta W_{2,j}\|^2$, and $\sum_j \langle \Delta W_{1,j}, \Delta W_{2,j} \rangle$. These three sums are enough to reconstruct all three reported quantities via the standard norm and inner-product identities, without ever forming ΔW_1 or ΔW_2 in full.

Ordering	Gradient cosine	Raw LoRA movement		Effective update ΔW	
		rel. move	cos. with S_1	rel. move	angle to ΔW_1
HH Ha→He	−0.004	9.2%	−0.007	97%	90.4°
HH He→Ha	−0.007	9.2%	−0.007	96%	90.3°
HS2 Ve→Co	+0.000	4.6%	−0.002	89%	90.1°
HS2 Co→Ve	−0.002	5.1%	−0.002	111%	90.0°
PKU Ha→He	−0.003	9.3%	−0.001	98%	89.0°
PKU He→Ha	+0.028	9.1%	+0.004	93%	88.9°
UF IF→Hon	+0.004	16.1%	+0.002	97%	89.1°
UF Hon→IF	−0.001	16.0%	+0.001	97%	89.3°
SF Co→Ac	+0.010	2.7%	+0.003	83%	87.3°
SF Ac→Co	−0.002	3.0%	+0.005	104%	87.8°

Table 13: Mechanistic diagnostics for the ten orderings. Gradient cosine: mean over the measurement steps of Stage 2 between the live DPO gradient and the Stage-1-objective diagnostic gradient recomputed on a fixed batch of four examples (LoRA parameters). Raw LoRA movement: $\|S_2 - S_1\|/\|S_1\|$ over the flattened adapter vector and its cosine with S_1. Effective update: $\|\Delta W_2 - \Delta W_1\|/\|\Delta W_1\|$ for $\Delta W = (\alpha/r)BA$ and the angle between ΔW_1 and $\Delta W_2 - \Delta W_1$.

Table 13 reports the gradient-conflict, raw-adapter-movement and rescaling-invariant effective-update diagnostics (Section 4.1.4) for all ten orderings.

H.3 Hidden-State Drift Probe Implementation

For a fixed diagnostic batch of $n = 32$ examples drawn from the Stage-1 objective's test set, we extract last-layer hidden states $z_\theta(x_i, y_i) \in \mathbb{R}^{T_i \times d}$ for each response y_i (chosen and rejected scored separately), where T_i is the number of response tokens and d is the model's hidden dimension. We mean-pool over response-token positions to obtain a single vector per example and checkpoint:

$$\bar{z}_\theta(x_i, y_i) = \frac{1}{T_i} \sum_{t=1}^{T_i} z_\theta(x_i, y_i)_t \tag{15}$$

Let $\bar{z}_i^{(1)}$ and $\bar{z}_i^{(2)}$ denote this pooled hidden state at the Stage 1 and Stage 2 checkpoints, respectively. We compute, per example, the cosine similarity

$$\cos_i = \frac{\langle \bar{z}_i^{(1)}, \bar{z}_i^{(2)} \rangle}{\|\bar{z}_i^{(1)}\| \, \|\bar{z}_i^{(2)}\|} \tag{16}$$

alongside the relative movement

$$\rho_i = \frac{\|\bar{z}_i^{(2)} - \bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|} \tag{17}$$

both computed separately for chosen and rejected responses. Values of $\cos_i$ near 1 and ρ_i near 0 indicate representational stability; lower cosine or larger movement indicates representational drift that may not be visible in either gradient or parameter space.

Table 14 reports the resulting cosine similarity and relative movement for all ten orderings.

I Multi-Seed Robustness

To test whether the Stage-1 compatibility–retention relationship (Section 4.2) depends on training seed, we rerun the four orderings at the extremes of the compatibility range, HH-RLHF (both orderings) and UltraFeedback (both orderings), at two additional seeds (123, 2024) alongside the original seed 42, and recompute C_{margin}, R_{margin}, C_{GIR} and R_{GIR} from per-pair scores computed with the same procedure as the main results. HelpSteer2, PKU-SafeRLHF, and Summarize-Feedback are not rerun; they retain their single seed-42 point in the per-seed correlations below. The four rerun orderings were chosen because they sit at the two extremes of the compatibility range (clearly negative for HH-RLHF, clearly positive for UltraFeedback). We did not rerun the two orderings with

Dataset	Order	$\cos_{\text{ch.}}$	$\cos_{\text{rej.}}$	$\text{move}_{\text{ch.}}\%$	$\text{move}_{\text{rej.}}\%$
HH-RLHF	harm.→help.	0.980	0.980	19.4%	19.6%
HH-RLHF	help.→harm.	0.992	0.991	12.1%	13.4%
HelpSteer2	cohere.→verb.	0.984	0.985	17.6%	17.1%
HelpSteer2	verb.→cohere.	0.995	0.993	9.5%	11.1%
PKU-SafeRLHF	harm.→help.	0.991	0.990	12.9%	13.6%
PKU-SafeRLHF	help.→harm.	0.990	0.989	14.1%	14.4%
UltraFeedback	hon.→IF	0.940	0.920	31.5%	37.1%
UltraFeedback	IF→hon.	0.961	0.929	27.0%	36.5%
Summarize-Feedback	cov.→acc.	0.973	0.968	22.8%	25.1%
Summarize-Feedback	acc.→cov.	0.918	0.926	39.8%	38.0%

Table 14: Last-layer hidden-state cosine similarity and relative movement between Stage 1 and Stage 2, fixed 32-example diagnostic batch, chosen/rejected response tokens mean-pooled.

the smallest-magnitude compatibility, HelpSteer2 verbosity→coherence and Summarize-Feedback accuracy→coverage, where a sign change across seeds is most likely, so this experiment tests seed stability only where the sign is least ambiguous.

Table 15 reports the mean and sample standard deviation across the three seeds for each re-run ordering. The sign of both C_{margin} and R_{margin} is stable across all three seeds in all four orderings.

Setting	Order	C_{margin}	R_{margin}	C_{GIR} (pp)	R_{GIR} (pp)
HH-RLHF	Ha→He	-0.430 ± 0.068	-0.246 ± 0.041	-11.41 ± 1.59	-5.89 ± 0.33
HH-RLHF	He→Ha	-0.303 ± 0.057	-0.465 ± 0.035	-7.81 ± 0.72	-9.11 ± 1.93
UltraFeedback	IF→Hon	$+7.452 \pm 0.112$	$+7.198 \pm 0.291$	$+34.05 \pm 0.37$	$+2.67 \pm 0.36$
UltraFeedback	Hon→IF	$+6.252 \pm 0.148$	$+7.717 \pm 0.060$	$+35.38 \pm 0.23$	$+2.14 \pm 0.93$

Table 15: Mean $\pm$ sample standard deviation across 3 seeds (42, 123, 2024), the 4 re-run orderings, from per-pair scores computed with the same procedure as the main results.

Across all ten orderings the margin-space correlation between C_{margin} and R_{margin}, computed separately for each seed (the four re-run orderings use that seed's own values; the other six their seed-42 values), is Pearson $r = 0.94$, 0.95 and 0.94 and Spearman $\rho = 0.83$, 0.83 and 0.79 for seeds 42, 123 and 2024 (all $p < 0.01$), consistent with the seed-42-only correlations in Section 4.2.

J Second-Model Replication

To test whether the Stage-1 compatibility–retention relationship generalises beyond Llama-3.1-8B-InstructGrattafiori et al. (2024), we repeat the HH-RLHF experiment (both orderings, identical hyperparameters and evaluation protocol, 900 held-out pairs per objective) on Qwen2.5-7B-Instruct Qwen et al. (2025). The qualitative pattern replicates in both HH-RLHF orderings: compatibility's sign again matches the direction of retention in both orderings ($C_{\text{harmless}\to\text{helpful}} = -0.41$, $R = -0.18$; $C_{\text{helpful}\to\text{harmless}} = -0.15$, $R = -0.46$), the previously trained objective's margin degrades only partially and its GIR falls by 5.1 and 10.8 points while its length-normalised policy accuracy changes by $+0.1$ and -0.2 points (recomputed from the per-pair chosen and rejected mean per-token log-probabilities stored in the original evaluation logs of these runs), and helpful→harmless shows the larger degradation of the two orderings, exactly as in Llama-3.1-8B-Instruct. The signal-asymmetry ratio ($S = 1.24$) closely matches Llama's ($S = 1.25$). This replication is limited in scope: it covers HH-RLHF only, where compatibility is negative in both orderings, so it does not test a positive-compatibility setting, other datasets, or other adaptation methods such as full fine-tuning.

K GENERATION-LEVEL EVALUATION

As a check on whether the likelihood-based margins used throughout this paper agree with free-form generation behaviour, we run an LLM-judge protocol on $n = 200$ held-out prompts per ordering (1,600 Stage-1 and 1,600 Stage-2 generations, 3,200 total). For each prompt, we anonymise and randomly order (seed 42) the Stage-1 objective's chosen and rejected reference responses as "Reference A"/"Reference B", then query `gpt-4o-2024-08-06` with the prompt, both anonymised references, and the model's generated response, asking which reference the generation is closer to (`CLOSER_TO_A`, `CLOSER_TO_B`, or `AMBIGUOUS`); the judge is never told which reference is chosen or rejected. We map the verdict back to chosen/rejected using the stored random assignment, then compute the chosen-alignment rate at each stage and an exact paired McNemar test on the discordant pairs.

Table 16 reports the original eight orderings. This behavioural validation was not extended to Summarize-Feedback. One ordering shows a nominally significant shift: HelpSteer2 verbosity→coherence ($\Delta = -0.055$, $p = 0.019$), where free-form generation becomes less faithful to the verbosity objective after coherence training. The remaining seven orderings, including HH-RLHF and PKU-SafeRLHF helpful→harmless, where the reward margin or GIR declines, show no significant change in generation-level alignment ($p > 0.05$), consistent with the absence of a significant decline in policy accuracy (Section 4.1.1). Each generation is a single sampled continuation per prompt (temperature 0.5, top-p 0.9, at most 200 new tokens, fixed seed), a coarser and higher-variance signal than the paired likelihood margins used elsewhere in this paper, so the absence of significance at $n = 200$ may reflect low sensitivity of this protocol, or may indicate that likelihood changes do not always translate into changes in free-form generation.

Setting	Order	n	Stage 1	Stage 2	Δ	p	Ambig. % (S1/S2)
HH-RLHF	Ha→He	200	0.655	0.650	−0.005	1.000	1.0/2.0
HH-RLHF	He→Ha	200	0.625	0.615	−0.010	0.856	0.5/0.5
HelpSteer2	Co→Ve	200	0.675	0.720	+0.045	0.078	1.0/1.5
HelpSteer2	Ve→Co	200	0.470	0.415	−0.055	0.019	0.0/0.0
PKU-SafeRLHF	Ha→He	200	0.940	0.965	+0.025	0.125	0.0/0.0
PKU-SafeRLHF	He→Ha	200	0.655	0.600	−0.055	0.080	1.5/0.5
UltraFeedback	Hon→IF	200	0.845	0.820	−0.025	0.383	1.5/2.0
UltraFeedback	IF→Hon	200	0.850	0.825	−0.025	0.227	0.0/0.5

Table 16: Generation-level chosen-alignment rate at Stage 1 and Stage 2, all 8 orderings (n=200 prompts/ordering). Δ = Stage-2 minus Stage-1 rate; p = exact McNemar test on discordant pairs; Ambig. = fraction of items the judge marked ambiguous.

L PAIR-LEVEL REGIME FRACTIONS

Appendix E defines the chosen/rejected decomposition (Eq. 12), which uses a dead-zone $\epsilon = 0.01$ nats per token to treat negligible coordinate-wise changes as near zero. This threshold lies at approximately the 10th percentile of the pooled $|\Delta_i^w|$ and $|\Delta_i^l|$ distribution, whose median is 0.070 nats per token. Each pair is classified into one of five regimes: *both_up* ($\Delta_i^w, \Delta_i^l > \epsilon$), *co_drift_down* ($\Delta_i^w, \Delta_i^l < -\epsilon$), *clean_sharpening* ($\Delta_i^w > \epsilon$, $\Delta_i^l < -\epsilon$), *clean_forgetting* ($\Delta_i^w < -\epsilon$, $\Delta_i^l > \epsilon$), and *unchanged_other* otherwise. Table 17 reports the regime fractions for all ten orderings. Co-drift is the most common pattern in most orderings (30–61% of pairs outside HelpSteer2 coherence→verbosity, where both log-probabilities increase together for 66% of pairs), so change is generally not a clean reversal in which the chosen response loses and the rejected response gains.

The pattern is not an artefact of this threshold: at a tenfold smaller dead-zone ($\epsilon = 0.001$), HelpSteer2 coherence→verbosity remains the both_up-dominant exception (0.732 versus co_drift's 0.017), and co_drift remains a common pattern elsewhere. Across orderings, the largest change in any regime fraction has a median of 0.042 (range 0.007–0.340), and co_drift is the largest regime in seven orderings at $\epsilon = 0.01$ and eight at $\epsilon = 0.001$; the two largest shifts (HelpSteer2 verbosity→coherence, 0.34, and UltraFeedback instruction-following→honesty, 0.16) are in both

Dataset/Order	n	both_up	co_drift	sharpen.	forget.	other
HH Ha→He	900	0.131	0.502	0.103	0.153	0.110
HH He→Ha	900	0.081	0.484	0.080	0.223	0.131
HS2 Ve→Co	527	0.002	0.472	0.034	0.017	0.474
HS2 Co→Ve	422	0.659	0.012	0.104	0.073	0.152
PKU Ha→He	1000	0.167	0.298	0.309	0.070	0.156
PKU He→Ha	1000	0.278	0.325	0.125	0.120	0.152
UF IF→Hon	700	0.024	0.491	0.216	0.031	0.237
UF Hon→IF	700	0.010	0.611	0.150	0.013	0.216
SF Co→Ac	205	0.044	0.517	0.093	0.302	0.044
SF Ac→Co	179	0.112	0.413	0.302	0.156	0.017

Table 17: Quadrant regime fractions per ordering, dead-zone $\epsilon = 0.01$ nats per token (Eq. 12). Fractions sum to 1 per row within rounding.

cases a fall in the *unchanged_other* regime, reflecting a smaller share of near-zero-magnitude pairs at the finer threshold.

M Replay: Compatibility Gate and Example Targeting

Compatibility gate. Replay was run regardless of compatibility, so the gate analysis is retrospective and not a prospective test. Section 4.3.1 compares always replaying with replaying only where compatibility is negative. Among the nine orderings with a random-10% replay run it fires in HH-RLHF (both orderings) and HelpSteer2 coherence→verbosity, and does not fire in the other six. Replay improves margin retention (the 95% interval excludes zero) in all three orderings where the gate fires and in four of the six where it does not (HelpSteer2 verbosity→coherence, PKU-SafeRLHF helpful→harmless, Summarize-Feedback coverage→accuracy and UltraFeedback instruction-following→honesty). For GIR the counts are two of three and two of six (PKU-SafeRLHF helpful→harmless and Summarize-Feedback coverage→accuracy), and for length-normalised policy accuracy zero of three and two of six (the same two orderings). Replay lowers acquisition of the incoming objective's margin significantly in HH-RLHF harmless→helpful (-0.10 margin units) and HelpSteer2 coherence→verbosity (-1.35), where the gate fires, and in PKU-SafeRLHF helpful→harmless (-2.87), where it does not; both UltraFeedback orderings gain margin on the incoming objective. If compatibility is instead computed on the dev split (Appendix C), the gate also fires in HelpSteer2 verbosity→coherence and Summarize-Feedback accuracy→coverage, the two orderings whose sign is unstable, and is otherwise unchanged. Intervals are uncorrected for the number of comparisons. Per-ordering values are in Table 3 and Appendix Table 18.

Example targeting. Section 4.3.1 asks whether which examples are replayed matters, not just the dose. Appendix Table 18 lists the targeted conditions (top-margin quartile of the Stage-1 policy margin, anchor-boundary, and a bottom-quartile pool) next to the matching random replay; random 10% replay for the other orderings is in Table 3. In HH-RLHF harmless→helpful, top-margin and anchor-boundary pools at 10% improve GIR by $+1.1$ and $+2.5$ points, against $+4.8$ for random replay. In HH-RLHF helpful→harmless, targeted replay improves GIR by $+3.1$ (1%), $+7.9$ (5%) and $+8.2$ points (10%) against $+8.4$ for random 10% replay, and a bottom-quartile pool at 5% by $+2.9$; in HelpSteer2 coherence→verbosity, a 10% top-margin pool improves GIR by $+0.5$ against $+3.1$ for random replay. Confidence-only targeting is therefore inconsistent for the reference-relative measures, and none of the seven targeted variants changes policy accuracy significantly.

N Independent Adapter Merging: Full Results

Following the adapter-merging and weight-interpolation literature Wortsman et al. (2022); Rame et al. (2023), Section 4.3.1 evaluates whether linearly interpolating two independently-trained,

(a) Retention of the first objective (replay − no replay)

Ordering	Variant	margin	GIR	acc. (norm.)	acc. (summed)
HH Ha→He	anchor-boundary 10%	**+0.47 [+0.35, +0.59]**	+2.5 [−0.2, +5.3]	+0.8 [−0.6, +2.1]	−0.4 [−1.1, +0.2]
HH Ha→He	targeted (top-margin) 10%	**+0.43 [+0.30, +0.55]**	+1.1 [−1.7, +3.8]	−0.4 [−1.7, +0.9]	−0.3 [−1.1, +0.4]
HH Ha→He	random 10%	**+0.91 [+0.76, +1.06]**	**+4.8 [+2.1, +7.6]**	+0.8 [−0.6, +2.1]	+0.8 [−0.1, +1.6]
HH He→Ha	bottom-quartile 5%	**+0.09 [+0.01, +0.16]**	**+2.9 [+0.1, +5.9]**	+0.4 [−0.7, +1.6]	+0.1 [−0.7, +0.9]
HH He→Ha	targeted (top-margin) 1%	**+0.24 [+0.17, +0.32]**	**+3.1 [+0.5, +5.9]**	−0.2 [−1.3, +0.8]	+0.2 [−0.4, +0.8]
HH He→Ha	targeted (top-margin) 5%	**+0.51 [+0.40, +0.62]**	**+7.9 [+5.0, +10.9]**	+0.2 [−0.8, +1.3]	+0.3 [−0.5, +1.1]
HH He→Ha	targeted (top-margin) 10%	**+0.65 [+0.52, +0.77]**	**+8.2 [+5.2, +11.3]**	+1.0 [−0.3, +2.3]	+0.1 [−0.8, +1.1]
HH He→Ha	random 10%	**+0.49 [+0.39, +0.58]**	**+8.4 [+5.5, +11.2]**	+0.9 [−0.3, +2.1]	−0.2 [−0.9, +0.5]
HS2 Co→Ve	targeted (top-margin) 10%	**+0.24 [+0.06, +0.41]**	+0.5 [−2.5, +3.4]	0.0 [−1.5, +1.5]	+0.2 [−0.8, +1.3]
HS2 Co→Ve	random 10%	**+0.23 [+0.01, +0.45]**	+3.1 [−0.1, +6.4]	−0.9 [−2.5, +0.7]	+0.5 [−0.7, +1.6]

(b) Acquisition of the second objective (replay − no replay)

Ordering	Variant	margin	acc. (norm.)
HH Ha→He	anchor-boundary 10%	−0.07 [−0.16, +0.01]	+0.1 [−1.0, +1.2]
HH Ha→He	targeted (top-margin) 10%	−0.04 [−0.13, +0.05]	0.0 [−1.1, +1.1]
HH Ha→He	random 10%	**−0.10 [−0.19, −0.02]**	0.0 [−1.1, +1.1]
HH He→Ha	bottom-quartile 5%	−0.02 [−0.10, +0.06]	+0.2 [−1.1, +1.5]
HH He→Ha	targeted (top-margin) 1%	+0.00 [−0.07, +0.08]	+1.0 [−0.3, +2.3]
HH He→Ha	targeted (top-margin) 5%	**−0.13 [−0.22, −0.05]**	+0.7 [−0.6, +1.9]
HH He→Ha	targeted (top-margin) 10%	−0.05 [−0.13, +0.03]	+0.8 [−0.6, +2.2]
HH He→Ha	random 10%	−0.01 [−0.09, +0.07]	+0.2 [−1.1, +1.5]
HS2 Co→Ve	targeted (top-margin) 10%	**−0.63 [−0.80, −0.46]**	+0.2 [−1.0, +1.4]
HS2 Co→Ve	random 10%	**−1.35 [−1.55, −1.16]**	+1.1 [−0.4, +2.7]

Table 18: Replay conditions evaluated for the orderings with targeted variants (random 10% replay for the other orderings is in Table 3; random: uniform sample of the first objective's training pairs; targeted: top-margin quartile of the Stage-1 policy margin; anchor-boundary: half top-margin and half near-zero-margin pairs; bottom-quartile: lowest-margin quartile; the number is the replay dose). Entries are differences to the no-replay Stage-2 checkpoint on identical pairs with 95% CIs; bold: the CI excludes zero. Margins in margin units, other columns in percentage points.

single-objective LoRA adapters (each trained alone from the base model, never chained), $\theta_\lambda = (1-\lambda)\theta_A + \lambda\theta_B$ for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$, can substitute for sequential training. For each sequential Stage-2 checkpoint, we test whether it Pareto-dominates each of the five evaluated merge points (margin on $A \geq$, margin on $B \geq$, strictly greater on at least one), against the full five-point grid and against the three interior points ($\lambda \in \{0.25, 0.5, 0.75\}$) only. Margins in this section are taken from the original evaluation logs; they differ from the re-scored values used elsewhere by at most 0.03 margin units in the mean (Appendix B).

In UltraFeedback and Summarize-Feedback, both sequential checkpoints Pareto-dominate all five merge points. In PKU-SafeRLHF, harmless→helpful dominates the full grid, while helpful→harmless dominates the three interior points and the pure-helpful endpoint, but not the pure-harmless endpoint.

In HelpSteer2, coherence→verbosity dominates the three interior points but not either single-objective endpoint. Verbosity→coherence also dominates the interior points and the pure-coherence endpoint, but not the pure-verbosity endpoint. HH-RLHF, the setting with the clearest negative cross-objective transfer, is the exception: neither sequential checkpoint dominates the full interior grid, and neither is dominated by any merge point. Across the five settings evaluated for adapter merging, no sequential checkpoint is dominated by any evaluated merge point. Overall, sequential training dominates merging most cleanly when the objectives are compatible, while HH-RLHF's checkpoints occupy a region the merge grid does not reach at all, neither dominating it nor dominated by it.

O Dataset and Preference-Relationship Details

This appendix provides additional details on the five preference settings used in the main experiments. Each setting consists of two objectives, and we train both objective orderings under the same sequential DPO and cross-evaluation protocol.

HH-RLHF: Distributional conflict. Anthropic HH-RLHF contains separate helpfulness and harmlessness preference subsets Bai et al. (2022). In our setup, we use *helpful-base* and *harmless-base* to instantiate a distributional safety–helpfulness shift. The *harmless-base* subset contains safety-oriented chosen–rejected comparisons, where the preferred response is the safer or less harmful response, while the *helpful-base* subset contains assistant-response comparisons oriented toward helpfulness. The two subsets differ in both prompt distribution and annotated preference dimension. Harmlessness targets safety-relevant behaviour, whereas helpfulness targets general assistant quality.

We use these splits to introduce a distributional conflict rather than an adversarial label-conflict setting. Empirically, we find no exact flipped-label pairs across our training sets and fewer than 1% of unique prompts shared between the two training sets under exact string matching (0.86%), suggesting that the conflict is primarily distributional. For HH-RLHF, each objective uses up to 9,000 training pairs and 1,000 held-out test pairs; evaluation at every stage draws a fixed 900-pair random subsample (seed 42) of this test set, so all HH-RLHF results reported in this paper, aggregate and pair-level alike, use the same 900 pairs.

HelpSteer2: Multi-attribute interaction. HelpSteer2 Wang et al. (2024) provides scalar ratings across response attributes including helpfulness, correctness, coherence, complexity, and verbosity, each on a 0–4 Likert scale. We use this dataset to construct preference objectives for verbosity and coherence. Verbosity reflects the amount of detail or elaboration, while coherence reflects whether a response is logically organised and easy to follow. Because increased detail can sometimes introduce repetition, tangents, or loss of focus, verbosity and coherence provide a controlled attribute-level setting distinct from broad safety–helpfulness comparisons.

For each attribute, we construct within-prompt preference pairs by selecting response pairs whose target-attribute rating differs by at least one point. Cross-prompt pairing is excluded to avoid confounding response quality with prompt difficulty. After filtering and deduplication, we obtain 2,229 verbosity training pairs and 1,878 coherence training pairs. The held-out test sets contain 527 verbosity pairs and 422 coherence pairs. We run both verbosity$\rightarrow$coherence and coherence$\rightarrow$verbosity orderings.

PKU-SafeRLHF: Strong safety signal. PKU-SafeRLHF Ji et al. (2025) provides response pairs annotated for both harmlessness and helpfulness. We use it to study a setting where harmlessness provides a much stronger observed preference signal than in HH-RLHF. In our construction, many harmlessness pairs contrast safer refusal-style responses with clearly unsafe responses. This makes the dataset useful for testing whether a high-signal safety objective is more resistant to forgetting during later helpfulness training.

We construct harmless and helpful preference splits and run both harmless$\rightarrow$helpful and helpful$\rightarrow$harmless orderings. We use the provided helpful split without additional filtering because the chosen helpful responses are predominantly benign in our construction. For PKU-SafeRLHF, each objective uses up to 9,000 training pairs and 1,000 held-out test pairs.

UltraFeedback: Compatible objectives. UltraFeedback Cui et al. (2023) provides LLM-generated responses with fine-grained scalar annotations across multiple response-quality dimensions, including instruction-following, honesty, truthfulness, and helpfulness. We use instruction-following and honesty to study compatible objectives. Both objectives reward high-quality responses that address the prompt accurately and faithfully, making them conceptually aligned rather than directly conflicting.

For each attribute, we construct 7,000 training pairs and 700 held-out test pairs. We run both instruction-following$\rightarrow$honesty and honesty$\rightarrow$instruction-following orderings. This setting serves as a compatible-objective baseline. When two objectives reward overlapping response behaviours, sequential training should produce stability or positive transfer rather than interference.

Summarize-Feedback: Soft multi-attribute interaction (summarisation). The Summarize-Feedback dataset Stiennon et al. (2020) provides human Likert-scale ratings (1–7) of Reddit TL;DR summaries along four axes: accuracy, coverage, coherence, and overall quality. We use coverage and accuracy to construct a fifth setting drawn from a different domain, single-document summari-

sation rather than open-ended chat. In our processed data, these two axes have the weakest pairwise correlation of any axis pair ($r = 0.31$), indicating that they are related but largely distinct attributes. Covering more of the source post may create more opportunities for inaccurate claims, while a shorter summary may remain accurate by omitting information.

For each Reddit post with multiple candidate summaries from different summarisation policies, a preference pair for each axis is constructed by selecting the highest- and lowest-rated summaries, excluding ties. Because each post contributes at most one pair per axis, we split the data by source post before pair construction, ensuring that no post appears in both the training and test sets. This yields 821 coverage training pairs and 717 accuracy training pairs, with 205 and 179 held-out test pairs for coverage and accuracy, respectively. These sets are markedly smaller than those in the other four settings and therefore require fewer optimisation steps per stage (23–26). We run both coverage→accuracy and accuracy→coverage orderings. For coverage, chosen summaries are longer than rejected summaries on average (31.3 versus 25.9 words), and the chosen summary is longer in 66% of pairs. For accuracy, chosen and rejected summaries have similar lengths (27.0 versus 27.3 words), with the chosen summary being longer in 48% of pairs, indicating no comparable length association.

Overlap between objectives. The two objectives of a setting are built from the same source data, so prompts, and in some settings identical (prompt, chosen, rejected) pairs, occur in both objectives' files; within an objective no test pair is identical to a training pair, although test prompts also occur in the training data for 2.6–5.1% of HH-RLHF and 25.7–36.4% of PKU-SafeRLHF test pairs (at most 0.1% in the other settings). The share of the incoming objective's test pairs whose prompt, and whose identical pair, occurs in the first objective's training data is 1.3–3.1% and 0% for HH-RLHF, 0% and 0% for HelpSteer2, 15.9–22.5% and 6.5–7.8% for PKU-SafeRLHF, 12.3–12.6% and 5.3–6.3% for UltraFeedback, and 65.9–81.6% and 11.2% for Summarize-Feedback (both Summarize-Feedback objectives rate summaries of the same Reddit posts). The same shares hold in the other direction. The first objective's test pairs, which are used for retention, overlap the incoming objective's training data in the same way, and Stage 2 trains on that data. Compatibility and retention are therefore measured partly on prompts that the other objective's training also contained, and in PKU-SafeRLHF, UltraFeedback and Summarize-Feedback partly on identical pairs (5–11% of the test pairs). The two objectives' test sets also share some identical pairs, namely 105 in HelpSteer2 (about 20% of the 527 verbosity test pairs), 9 in PKU-SafeRLHF, 6 in UltraFeedback, 5 in Summarize-Feedback and none in HH-RLHF. These shares were computed after the main analysis, and we did not filter the evaluation sets.

P LIMITATIONS

All principal experiments use Llama-3.1-8B-Instruct with LoRA and DPO (the second-model check covers HH-RLHF only, the three-stage sequences HelpSteer2 only), and we do not compare against other continual preference-optimisation methods. Two Summarize-Feedback test sets are small (179 and 205 pairs), so ranking bounds are wide there. bf16 log-probabilities add noise to GIR, and summed accuracy is confounded with response length (Appendix B). Outcomes are confounded with dataset properties, and in several settings the two objectives share prompts and some pairs across their training and test data (Appendix O); replay results use a single seed.