Reusing Preference Data Changes What DPO Learns

Anonymous authors
Paper under double-blind review

Abstract

Direct preference optimization (DPO) is often run for several passes over a smaller preference pool. We show that, at a fixed training budget, this reuse changes the direction of what the model learns. We compare fresh and reused data with identical steps, batch size and pair presentations, measure each trained model's change from its parent on a frozen 200-scenario battery and on 10,273 held-out human-preference pairs, and calibrate every comparison against the variation between training seeds. In OLMo-2 1B, reuse and fresh training agree at a reliability-corrected $r^* = 0.798$ [0.675, 0.892] (about 37°), outside seed-to-seed variation, and a second, disjoint reused pool lands in the same place ($r^* = 0.966$ with the first). At a fixed budget the turn grows with the number of passes, from 20° at two passes to 38° at 6.4, as a finite-pool model of DPO predicts, and a third seed confirms it ($r^* = 0.770$ [0.637, 0.881]). The redirection replicates across four DPO model families and on held-out human preferences. It reaches behavior: reuse changes the preferred response on a fifth of held-out human-preference pairs, five to eight times the seed-to-seed rate, without improving agreement with the human labels, and no comparison finds its generations better (at greedy decoding, fresh training's response wins on 71–74% of prompts). Held-out preference accuracy is lower under reuse (0.604/0.616 versus 0.651/0.660). How often each pair is reused is therefore part of a preference-training recipe: equal budgets can produce systematically different models.

1. Introduction

At a fixed training budget, replacing fresh preference pairs with repeated use of a smaller pool changes the direction of a model's measured preference update. We demonstrate this in direct preference optimization (DPO), holding optimizer steps, batch size, and total pair presentations constant. The direction change exceeds within-recipe variation across presentation-order seeds and recurs with a disjoint reused subset. Exposure history is therefore part of a preference-training recipe's behavioral specification.

Repeated passes are an ordinary choice in public DPO workflows. Prior work studies additional DPO passes [1], repetition through pretraining loss [2], and accumulated preference data in iterative training [3]. These studies motivate measuring both task outcomes and the structure of model change. Our contribution is a matched-budget comparison resolving the direction of preference change against each recipe's own order-seed variation, with disjoint-subset and independent-instrument tests and the account's predicted pass-ladder forward test (§§3, 6).

We evaluate each parent and trained child on frozen, teacher-forced conditional response scores. Subtracting the parent profile gives a displacement, which we center across scenarios for the direction comparison. The full direction criterion requires per-seed cross-recipe agreements below both within-recipe ceilings with non-overlapping design intervals, together with a corrected mean-profile correlation interval excluding unity (§2).

The primary comparison processes 2,891 batches of 128 preference pairs in each arm. Fresh training (A) samples without reuse from a pool of 370,121 pairs. Reuse (E) repeatedly samples a fixed 58,251-pair subset, averaging 6.35 presentations per pair. A second, disjoint reused subset is denoted E2 (Table 1; §2). Changing regimes jointly changes repetition, unique-pair coverage, and subset composition. This is the decision a practitioner faces with a fixed number of pair presentations: draw fresh pairs or reuse a subset. The comparison answers it directly; §§4–6 test which ingredient carries the effect, with a fixed-budget passes ladder that varies only the number of passes (§6). At

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.

this matched budget, A and E meet the full direction criterion, with corrected agreement r* = 0.798 [0.675, 0.892] (§4; Figure 1).

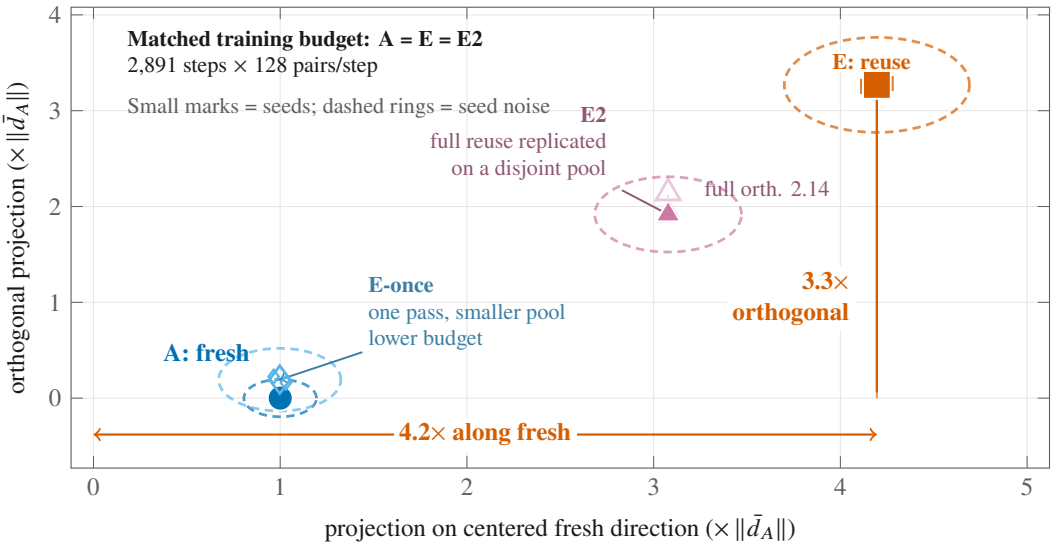

Figure 1: At a matched budget, reuse (E) moves 4.2× the fresh displacement along the fresh direction and 3.3× orthogonal to it; a disjoint reused subset (E2) lands with E; the one-pass smaller pool (E-once) stays with fresh. Axes show each seed-mean displacement projected on the centred fresh direction (x) and its orthogonal (y), in fresh-norm units. E2's solid triangle is its projection on this plane (vertical 1.92); the faint open triangle marks its full orthogonal magnitude (2.14). Dashed circles are seed-noise radii ‖d□¹□ − d□²□‖/√2 in fresh-norm units (A 0.19, E-once 0.33, E2 0.39, E 0.50).

Matched-budget reuse redirects the update At equal pair-presentation budget, reuse moves 4.2 times the fresh displacement along the fresh direction and 3.3 times orthogonal to it (r* = 0.798 [0.675, 0.892]), a disjoint reused subset lands with reuse, and a lower-budget one-pass smaller pool stays with fresh (Figure 1; Tables 1, 3). **Where the redirection transfers** The full direction criterion passes on the pilot battery in all four tested DPO families and in 5/5 held-out human-pair settings. On the separately authored parallel forms, the corrected separation holds in 9/9 informative cells, with full per-seed replication in 2/9. Item- and family-level content does not transfer (Tables 2–3; Appendix Table 8). **Norm- and KL-matched comparisons are mixed** Fresh training at 0.885 of reuse's displacement norm shows no resolved increase in reuse agreement (D_P = +0.002 [−0.265, +0.125]), and the KL-matched reuse checkpoint stays near the fresh direction (r* = 0.945 [0.907, 0.971]); both registered comparisons are mixed (Figure 4; Table 3; Appendix Tables 6, 9). **Choices follow the redirection** On 10,273 held-out human-preference pairs, reuse changes the preferred response on 21.5%/21.4% of pairs, five to eight times the seed-to-seed rate, without improving agreement with the human labels (§5; Appendix Figure 7).

2. Design, measurement and decision rule

Budget and arms The matched budget means identical optimizer steps, pairs per step, and hence presentations: A and E each process 370,048 pairs (E presents 99.8% of A's tokens in both seeds; Appendix C). What changes is revisit frequency, unique-pair coverage, and subset composition. All arms train DPO from the released OLMo-2 1B SFT checkpoint [4] with the released settings (length-normalized DPO, β = 5, learning rate 2.5×10^{-6} with linear decay over each recipe's own steps and 10% warmup, maximum length 2,048) and vary only the data, the batch, and the number of steps (Table 1; precision in Appendix A). The fresh recipe reproduces the released DPO stage under our trainer: its amplification, γ = +0.454, matches the released checkpoint's +0.450.

Table 1. Fresh A and reuse E match presentation budgets; the controls separate subset, repetition, and budget. Unique = unique pairs after tokenization; Batch = pairs per step; Passes = mean passes (presentations / pool). γ is the dimensionless parent-aligned slope, averaged over seeds, with its design-level 95% interval. Seed ceilings of these and the extended arms are in Appendix Table 4; setting-specific ceilings are in Appendix Table 7.

Recipe	Role	Data	Unique	Batch	Steps	Passes	γ [design 95%]
A	**fresh**	**Mix**	**370,121**	**128**	**2,891**	**1**	**+0.454 [+0.35, +0.57]**
E	**reuse**	**Mix subset**	**58,251**	**128**	**2,891**	**6.35**	**+1.996 [+1.19, +2.90]**
E2	reuse, second subset	Mix subset′	58,270	128	2,891	6.35	+1.538 [+0.90, +2.15]
E-once	subset once, lower budget	Mix subset	58,251	20	2,912	1	+0.467 [+0.33, +0.61]

Arms The table defines the six arms of the main text; the β = 1.25 arm is named in words. D_L and D_P are paired toward-reuse differences for the β = 1.25 and P arms against the one-pass fresh baseline.

Arm	Role	Pool	Pairs/step	Steps	Presentations	Seeds
A	fresh	Mix full (370,121)	128	2,891	370,048	2
E	reuse	Mix subset 1 (58,251)	128	2,891	370,048 (6.35×)	2
E2	reuse, second subset	Mix subset 2 (58,270)	128	2,891	370,048 (6.35×)	2
E-once	subset once, lower budget	Mix subset 1 (58,251)	20	2,912	58,240	2
R	reuse, indep. (k = 6.4)	Mix 9.1k slice	20	2,912	58,240 (6.4×)	2
P	fresh, β = 0.4	Mix subset 1 (58,251)	20	2,912	58,240	2

Measurement The instrument is a frozen, content-hashed battery (hash `ab14c63f7fe6a86d`) of 200 two-option forced-choice scenarios from 50 authored skeletons across ten behavioral families (Appendix B). Each scenario's score is a log-probability difference between its two actions under the domain-conditional PMI scorer [5], with a content-reduced premise baseline subtracted as calibration [6]; scoring is one teacher-forced pass per prompt–continuation pair [7]. A model's **profile** $p \in \mathbb{R}^{200}$ is its vector of scenario scores; training a parent with profile p yields a child with profile c, and the **displacement** is $d = c - p$, in nats. We summarize parent-aligned change with γ, the cross-fitted slope of displacement on the parent profile. To compare recipes we split the reuse displacement into its component along the fresh recipe's direction $\hat{a} = \bar{d}_A/\|\bar{d}_A\|$ and the reuse-specific component orthogonal to it, reporting each relative to $\|\bar{d}_A\|$. Before the decomposition and in Figure 1, each seed-mean displacement is centred by subtracting its mean over the 200 scenarios (Pearson centering). Items sharing a skeleton are not independent draws, so every global statistic carries a **design-level** interval that resamples the 50 skeleton groups [8, 9]; held-out human-pair intervals instead resample 22 source groups. Intervals describe the trained runs; variation between training seeds enters through the seed ceilings.

Decision rule For a recipe trained from the same parent on the same data at two seeds, the **seed ceiling** is $r_{\text{seed}} = \text{corr}(d^{(1)}, d^{(2)})$ across scenarios. The reliability of a two-seed mean is the Spearman–Brown value $R = 2r_{\text{seed}}/(1 + r_{\text{seed}})$ [10, 11], and Spearman's classical correction for attenuation [12] gives $r^*(X, Y) = \text{corr}(\bar{d}_X, \bar{d}_Y)/\sqrt{R_X\, R_Y}$. For each seed, the raw-agreement design interval must lie entirely below and be nonoverlapping with each of the two recipe ceiling design intervals; the jointly bootstrapped design interval for r* must also exclude 1. The primary separation is already present in the uncorrected per-seed comparisons; reliability correction calibrates its size. A corrected interval excluding 1 alone is not the gate; an interval including 1 indicates compatibility with the same direction. "Mixed" means at least one gate element fails; "uninformative" marks cells whose fresh-ceiling lower bound falls below 0.6. In simulations matched to our runs, this test excludes 1 in 3.2% of datasets when the true r^* is 1 and in every dataset when it is 0.9 (Appendix C). Appendix D lists the decision rules with their outcomes.

In the primary OLMo-2 1B DPO family, two runs differing only in presentation order agree at r = 0.92–0.99 (every design-level lower bound above 0.88), and seed-mean γ is 0.461/0.446 for A and 2.005/1.987 for E.

3. Finite-pool model and testable signatures

Let $\mathcal{H} = \text{span}_{\text{ess}}\, \phi$ be finite-dimensional, $u \in \mathcal{H}$, $\mathbb{E}\|\phi\|^2 < \infty$, and $S = (\phi_i)_{i=1}^n$ an iid pool from D, $n \geq 1$. For $\beta > 0$, fixed SPD $M : \mathcal{H} \to \mathcal{H}$, fixed linear profile map $d = Ju$, and $\ell(q) = \log(1 + e^{-q})$, define

$$F_S(u) = \frac{1}{n} \sum_i \ell(\beta \phi_i^\top u), \quad F(u) = \mathbb{E}_D \ell(\beta \phi^\top u), \quad \text{Gap}_S = F_S - F.$$

For any objective G, fixed-M gradient flow is $\dot{u} = -M \nabla G(u)$.

Proposition 1 (stream law) An independent random ordering of an iid pool has the fresh-iid stream law through its first pass. Under matched initialization, batching, update rule, schedule, and optimizer randomness, the trajectory laws are identical through that pass, unconditionally over pool construction.

Proposition 2 (field identity) At any common u, with $g_S = -\nabla F_S$ and $g = -\nabla F$, $g_S - g = -\nabla \text{Gap}_S$. For fixed M, the difference between pool and population profile velocities is $-JM\nabla \text{Gap}_S$, evaluated pointwise even at an adaptive $u(S)$.

Proposition 3 (two regimes) If $\Pr(\phi^\top a < 0) > 0$ for every nonzero $a \in \mathcal{H}$, population risk is coercive and strictly convex; fixed-M population flow converges to its unique finite minimizer. Separately, if a realized pool is strictly separable, unregularized fixed-M logistic flow in cumulative learning-rate time $\tau \to \infty$ obeys

$$u_S(\tau) = \beta^{-1} \hat{v}_S \log \tau + o(\log \tau), \qquad \hat{v}_S = \arg \min_{\phi_i^\top v \geq 1} \tfrac{1}{2} v^\top M^{-1} v,$$

so its parameter norm grows logarithmically.

Proposition 4 (IPO) The quadratic pool objective $F_S^{\text{IPO}} = n^{-1} \sum_i (\phi_i^\top u - (2\beta)^{-1})^2$ always has a finite minimizer. Fixed-M flow converges to a finite minimizer, retaining the initial transformed nullspace component; it has no divergent max-margin tail.

We test the modelling assumption that this frozen-feature, fixed-metric geometry organizes transformer AdamW profiles; the table records exact model consequences and structural signatures. Proofs are in Appendix E.

Prediction (from which proposition)	Measurement	Result	Status
First-pass equivalence (P1)	E-once vs fresh (Table 1; §4, Figure 1)	$r^* = .9724$ [.9454, .9935]; $\gamma = .467$ vs .454	consistent (post hoc)
Direction beyond scalar (P2 signature)	Primary fresh/reuse (§4; Figure 1)	$r^* = .798$ [.675, .892]; orthogonal component 3.27 fresh norms	consistent (post hoc)
Shared pool response across disjoint subsets (P2 signature)	E–E2 (§4; Figure 1)	$r^* = .9665$; residual cosine .896, below residual ceilings	consistent (post hoc)
IPO objective contrast (P4 signature)	Paired $\Delta\gamma$ (§5, Table 3)	$-.282$ [$-.376$, $-.201$]; registered direction verdict mixed	consistent (post hoc)
Larger displacement need not recover direction (P2 signature)	Norm-matched fresh P (§6; Figure 4)	$\beta = .4$ endpoint at .885 of reuse norm; $D_P = +.002$ [$-.265$, +.125] (β scale identities: Appendix E)	consistent (post hoc)

Prediction (from which proposition)	Measurement	Result	Status
Passes ladder (P2–P3 signature)	Fixed-budget passes (§6; Figure 3)	θ +6.3°/doubling [+3.3, +9.6], γ +0.52 [+0.23, +0.88]; $\Delta\theta$ +15.8° [+8.3, +25.1], $\Delta\gamma$ +1.44 [+0.67, +2.36]: 4/4 supported	predictions ledgered before the two-seed ladder analysis and before any $k = 2$ data

$D_P = r^*(P, R) - r^*(\text{E-once}, R)$ uses the one-pass fresh baseline E-once and the independent reuse comparator R (§2).

4. Same budget, different direction

Matched-budget redirection The matched-budget contrast passes the full seed-calibrated test for a difference in displacement direction. In each seed separately the two recipes correlate at 0.804 and 0.757 (design intervals [0.673, 0.900] and [0.639, 0.849]), below A's ceiling (0.963) and E's (0.992) with no overlap at either the item or the design level. The reliability-corrected agreement between centered profiles is r* = 0.798 (design interval [0.675, 0.892], about 37°). Decomposed against the fresh direction (Figure 1), reuse's centered seed-mean displacement has a parallel component of 59.5 nats (4.2 times as long) and an orthogonal component of 46.4 nats (3.3 times as long, design [2.3, 4.6]). Spending the same budget on repeats also amplifies the parent-aligned component 4.4-fold (γ = 0.454 → 1.996) and lengthens the displacement 5.6-fold ($\|d\|$ = 14.4 → 80.9 nats; Figure 5 shows the scenario-level view). γ is a separate secondary outcome; the directional test rests on r* and the per-seed criterion. Removing the two families that carry most of the A–E difference (rule-following and allocation) leaves r* = 0.809 (Appendix C).

Disjoint-pool replication The endpoint geometry recurs across disjoint reused pools. We trained E's recipe at both seeds on a second, disjoint subset, E2 (58,270 pairs; Appendix A). E2 leaves A's envelope as E does (r*(A, E2) = 0.831, design [0.715, 0.897]), and the two reused subsets have strongly agreeing centered displacement profiles: r*(E, E2) = 0.9665 (design [0.9433, 0.9809]), above r*(A, E) by 0.17 ([0.08, 0.28]) and above r*(A, E2) by 0.14 ([0.07, 0.24]) (§3). E2 amplifies less than E (Table 1).

One-pass and extended-fresh controls A single pass through the smaller pool stays close to fresh (§3, Table 1), though its corrected interval ([0.9454, 0.9935]) excludes 1; a longer fresh run never approaches reuse's endpoint (Appendix D).

Reuse fits its pool in its own direction Reuse drives the model to fit its pool: training accuracy approaches 1 and sampled KL is about seven times fresh training's (§6). Overfitting as extra length along the fresh direction, or as noise, would leave no reproducible orthogonal component. Reuse's orthogonal component is 3.3 fresh norms, reproducible across order seeds (reuse ceiling 0.992) and shared by a disjoint pool (r^* = 0.966), and fresh training pushed to 0.885 of reuse's norm does not move toward it (§6). The finite-pool model gives the reason: a fitted pool's gradient carries a term that fresh data do not supply (Proposition 2).

5. Replication across instruments and settings

We measured the recipes of the §2 arm table on three further instruments (Appendix B). **Parallel forms**: two batteries of the same ten decision structures (regb, regc), scored as the pilot (Appendix B). **Held-out human preferences**: 10,273 pairs from the HH-RLHF and SHP test sets that no recipe trained on (22 source groups); a model's profile is its change, relative to the parent, in the length-normalized log-probability margin between the human-preferred and the rejected response, and intervals resample the source groups. The fresh recipe's profile there is highly reproducible (seed ceiling 0.970).

Table 2. The reuse contrast transfers across independent measurements Entries are r* with 95% intervals (design-level for the pilot and forms, source-group for human pairs). The final column is the Spearman ρ with the pilot's 28 pairwise values (point estimates).

Instrument	Items (groups)	r*(A, E)	r*(A, E2)	r*(E, E2)	Geometry Spearman ρ (point estimate)
Pilot battery	200 (50)	0.798 [0.675, 0.892]	0.831 [0.715, 0.897]	0.966 [0.943, 0.981]	n/a
Parallel form regb	190 (59)	0.827 [0.707, 0.894]	0.801 [0.718, 0.862]	0.898 [0.853, 0.934]	0.844
Parallel form regc	190 (52)	0.708 [0.551, 0.823]	0.768 [0.645, 0.866]	0.925 [0.864, 0.955]	0.880
Human preference pairs	10,273 (22)	0.678 [0.637, 0.743]	0.679 [0.644, 0.745]	0.914 [0.908, 0.929]	0.883

Separation on every instrument For the primary OLMo-2 comparison, both reused subsets separate from fresh under the complete registered rule on every instrument. On held-out human preferences the redirection is larger than on the battery (r* = 0.68). The full eight-recipe geometry on the pilot battery and on the held-out human pairs is shown in Appendix Figure 6.

From profiles to choices The redirection reaches the models' choices. On the 10,273 held-out human-preference pairs (public HH-RLHF and SHP test splits), fresh and reuse training prefer different responses on 0.215/0.214 of pairs (cluster [0.202, 0.224]/[0.201, 0.223]), against 0.026 within fresh and 0.044 within reuse seeds, an excess of +0.180 [+0.170, +0.193] over the seed-to-seed rate; the same holds for E2 (0.221/0.216, excess +0.185 [+0.174, +0.191]). Removing E's component orthogonal to fresh's direction restores fresh's choice on 0.64/0.65 of the disagreements, so the redirection itself changes the choices. Reuse flips 0.270/0.268 of the parent's choices against fresh's 0.101/0.103, yet agreement with the human labels does not rise (paired E − A −0.009 [−0.014, +0.003]/−0.006 [−0.013, +0.002]; E2 − A −0.017/−0.013). On the battery the verdict is mixed (Appendix Figure 7). Reuse's redirection changes what the model prefers without making its generations better: on 3,000 held-out human-preference prompts, a validated reward model finds no resolved difference between reuse's and fresh's sampled generations (E vs A W = 0.464/0.469 with intervals including 1/2; all six reuse–fresh points sit below 1/2), and no comparison at either seed finds reuse better. At greedy decoding reuse clearly loses: fresh's greedy response wins on 71–74% of prompts, a result replicated for the disjoint subset and at the second operating point, with an independent 72B pairwise judge leaning the same way. Both statements refer to the two validated judges, these prompts and these decoding settings. The per-prompt quality difference is unrelated to the teacher-forced redirection (its correlation with reuse's orthogonal component is +0.015 [−0.004, +0.057]), so a scalar quality score neither detects the redirection nor explains it (Appendix Figure 8; Appendix D).

Where the redirection transfers The complete rule passes on the pilot battery and replicates in 5/5 held-out human-pair settings. On the parallel forms, all nine informative corrected intervals exclude perfect agreement (9/9), with full per-seed replication in 2/9 (Appendix Table 8; Figure 9). Item- and family-level semantic content does not transfer: separation on the authored forms comes with instrument-dependent shifts.

Instrument-dependent item shifts Item-level semantic shifts do not transfer: family-level E−A correlations with the pilot are near zero in regb and regc (−0.15 and +0.01), and regb's rule-family shift reverses direction (Appendix B).

Smaller-scale replication We repeat the contrast at a 6.4-fold smaller scale: 20 pairs per step, 2,912 steps and 58,240 presentations for both arms. The fresh arm sees 58,240 unique pairs once (on the mix it re-runs E-once), and the reuse arm re-presents a 9,100-pair subset 6.4 times (Appendix A). Table 3 runs the mix contrast on three further parents and OLMo-2 on UltraFeedback, each arm at two seeds.

Table 3. Seed-calibrated redirection replicates across the tested full-parameter DPO settings Corrected r* and inferential Δγ entries have design-level 95% intervals. Setting-specific seed ceilings and the mixed rows' per-seed agreements are in Appendix Table 7. γ shows two-seed means for the † rows, first-seed values otherwise; Δγ is paired within seed (descriptive for the baseline row).

Setting	γ fresh → reuse	r^* [95% design CI]	Full direction rule	Paired $\Delta\gamma$

OLMo-2 1B, mix-128 DPO	0.454 → 1.996†	0.798 [0.675, 0.892]	redirected	+1.54 (descriptive)
OLMo-2 1B, mix-20 DPO (indep.)	0.432 → 2.312	0.794 [0.667, 0.898]	redirected	+1.779 [+0.812, +2.683]
OLMo-2 1B, UF-20 DPO	0.302 → 1.352	0.793 [0.692, 0.882]	redirected	+1.101 [+0.551, +1.763]
TinyLlama 1.1B, mix-20 DPO	0.264 → 0.893	0.306 [0.030, 0.570]	redirected	+0.629 [−0.042, +1.512]
Qwen2.5 1.5B, mix-20 DPO	0.360 → 1.027	0.637 [0.303, 0.811]	redirected	+0.668 [+0.025, +1.363]
SmolLM2 1.7B, mix-20 DPO	0.137 → 0.212	0.810 [0.727, 0.881]	redirected	+0.075 [−0.129, +0.380]
OLMo-2 1B, mix-20 IPO	0.365 → 0.083	0.851 [0.740, 0.922]	mixed	−0.282 [−0.376, −0.201]
OLMo-2 7B, mix-20 LoRA-DPO	0.695 → 1.832†	0.821 [0.683, 0.906]	mixed	+1.136 [+0.471, +1.720]

Cross-batch reuse agreement Reuse amplifies γ 4.5-fold on UltraFeedback and 5.2-fold on the mix on shared-prefix arms (0.426 → 2.205; 5.4-fold, 0.432 → 2.312, on the independent arms). With both arms seeded twice, r* = 0.794 (design [0.667, 0.898]) for the independent arms on the mix and 0.793 ([0.692, 0.882]) on UltraFeedback. A third presentation-order seed at this operating point confirms the verdict: fresh (F) against reuse (R) over seeds 201–203 gives $r^* = 0.770$ [0.637, 0.881], with every seed's agreement (0.807, 0.718, 0.714) below both three-seed ceilings (fresh 0.953, reuse 0.988): redirected. Tables 3 and 7 report the two-seed value (0.794 [0.667, 0.898]).

Cross-parent pattern The direction change meets the full rule in all four tested full-parameter DPO families: OLMo-2, Qwen2.5, TinyLlama, and SmolLM2 (Table 3).

LoRA-DPO and IPO limits Both show corrected mean separation but mixed full-rule verdicts with opposite amplification responses (Table 3; Appendix Table 7): 7B LoRA-DPO amplifies upward (Δγ = +1.136 [+0.471, +1.720]); IPO reverses (Δγ = −0.282 [−0.376, −0.201]). The per-seed rule cannot separate them even though r* excludes 1 (Appendix D). Directional separation and parent-aligned amplification therefore require separate conclusions: the response to repeated data depends on the training objective.

6. Controls and tests of the account

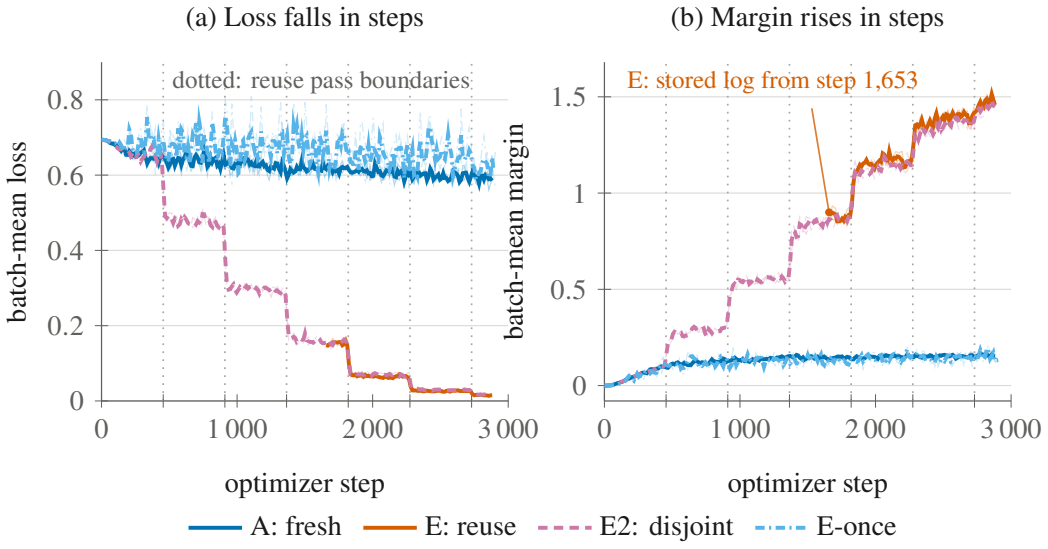

Figure 2: Reuse's loss and margin step at every pass boundary while fresh training moves smoothly. Batch-mean training loss (a) and training margin (b) against optimizer step for the primary recipes (A, E, E2, E-once), both seeds thin with seed means bold (dotted: nominal small-pool pass boundaries). Lines are centred moving averages (window 21); E's stored log starts at step 1,653.

Passes dose-response at fixed budget Holding the 2,912 × 20 presentation budget fixed and varying only the pass count turns the update away from fresh in a graded way: the angle to fresh grows 19.6° to 37.7° from 2 to 6.4 passes (+6.3° per doubling [+3.3, +9.6]; paired +15.8° [+8.3, +25.1]), and amplification rises 0.47 to 2.23 (+0.52 [+0.23, +0.88]; paired +1.44 [+0.67, +2.36]). The registered

rule calls redirection from 3.2 passes on. At a fixed presentation budget, the number of passes carries the redirection (Figure 3; Appendix D).

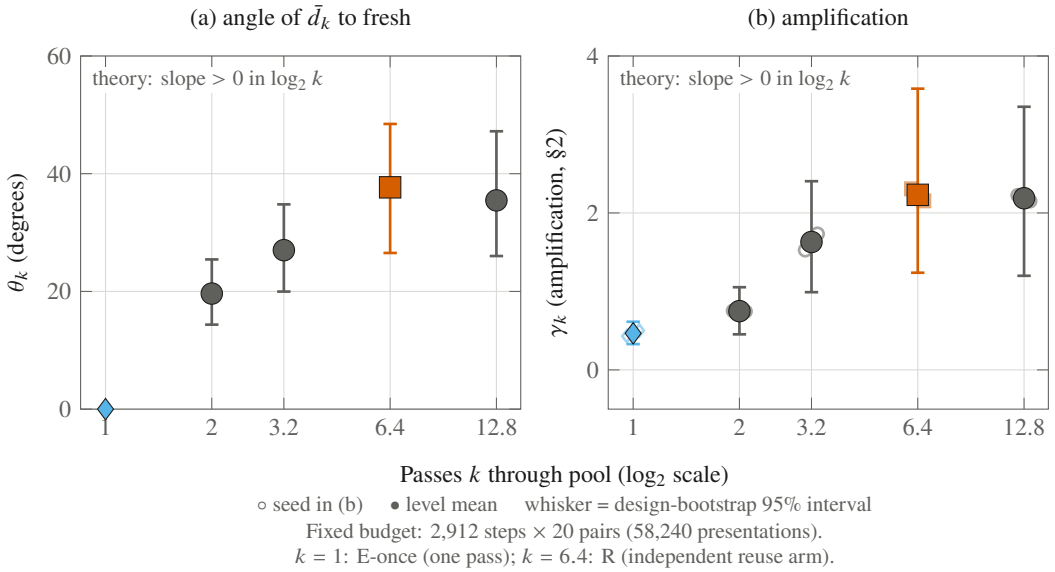

Figure 3: More passes at fixed budget turn the update further from fresh. (a) Angle θ_k of the level-k mean displacement to fresh, in degrees. (b) Level-k parent-aligned amplification γ_k (§2). Every level processes 2,912 steps × 20 pairs (58,240 presentations); only the pass count k varies. $k = 1$ is one-pass fresh (E-once); $k = 6.4$ is the independent reuse arm (R).

Norm-targeted control Fresh training at reduced β reaches 0.885 of reuse's displacement norm without a resolved increase in reuse similarity: corrected agreement is 0.796 [0.532, 0.891], against the baseline fresh value of 0.794, giving +0.002 [−0.265, +0.125]. Its parent-aligned slope is 0.227, compared with fresh 0.467 and reuse 2.24 [1.24, 3.58]. Near reuse's displacement norm, fresh training retains a distinct corrected mean profile, with a mixed verdict under the full direction rule.

KL-matched control beside the norm-matched one The reuse checkpoint nearest fresh training's endpoint KL is E2 at step 826 (sampled KL 0.165 against A's 0.141; step 413 is 0.094): E2@826 against A@2,891 gives r* = 0.945 [0.907, 0.971], with seed 201's per-seed agreement overlapping A's ceiling, so the registered verdict is mixed, the same verdict as the norm-matched β = 0.4 arm (r* = 0.796 [0.532, 0.891]; D_P = +0.002 [−0.265, +0.125]). At the endpoint the orthogonal component is 3.27 [2.28, 4.57] fresh norms.

Nonmonotone β response The broader β response is nonmonotone. At β = 1.25 the fresh displacement is 2.7 times longer ([2.2, 3.1], seed means) and rises toward reuse by D_L = +0.094 [+0.024, +0.186]. Its corrected agreement with the fresh baseline remains high (r* = 0.947 [0.905, 0.973]).

Training-dynamics staircase Reuse trajectories show increasing pool fit and policy–reference separation (Figure 2): the batch-mean length-normalized training margin steps up at pass boundaries to 1.4–2.7 by the end with training accuracy near 1, while the mean per-pair DPO gradient weight falls below 0.1 from step 1,835 on (E); fresh recipes stay at 0.09–0.22. Sampled KL per token on the held-out prompts is 1.07 for E and 1.11 for E2, against 0.14 for A (seed means); at the second operating point it is 1.25 for reuse against 0.19 for fresh training. Both recipes lower the log-probability of preferred and rejected responses, consistent with reported DPO likelihood displacement [13]; reuse lowers them by 2–3 nats per token against A's 0.1–0.4 (seed means).

7. Discussion and related work

A preference-training budget specifies how many pairs are presented; the exposure regime also shapes the resulting preference update. The design estimates the exposure regime's effect on teacher-forced response profiles: the pass count varied alone by the passes ladder at a fixed budget, coverage

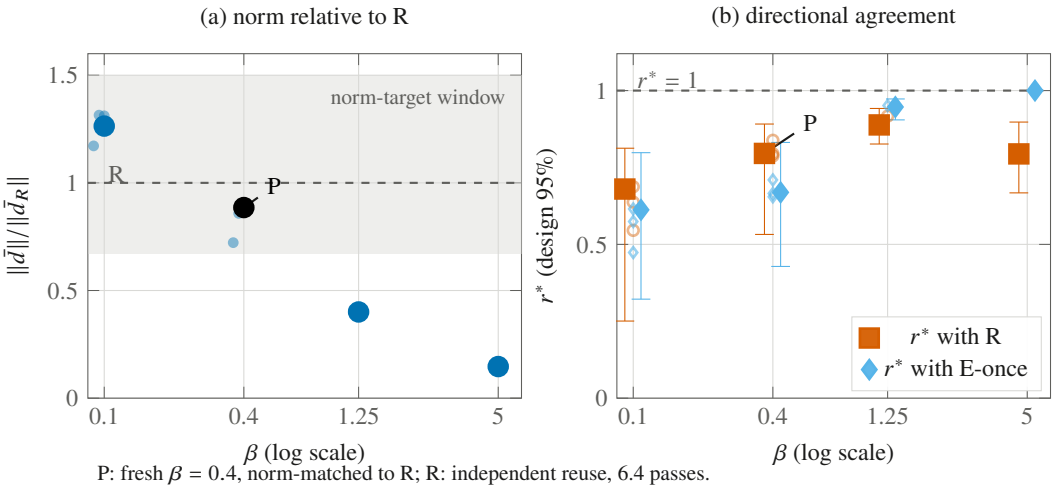

Figure 4: Norm-matched (P, β = 0.4) and weaker-anchor (β = 1.25) fresh arms against reuse (R) and one-pass fresh (E-once) baselines; both P comparisons are registered mixed (see text for values). (a) Norm relative to reuse with P in the selection window [0.67, 1.5]; (b) r* with reuse and fresh with design 95% intervals.

and composition jointly. Recipe comparisons should therefore report unique-pair count, presentations, and reuse policy; evaluation should measure displacement direction and amplification alongside task performance. Sampling the generations matters too: reuse's redirection is invisible to a scalar quality score yet reaches the model's greedy responses.

The aggregate consequences are visible in the primary comparison [14]. Held-out preference accuracy is lower under reuse in both seeds (0.604/0.616 vs 0.651/0.660; −0.048 [−0.077, −0.018] and −0.044 [−0.074, −0.014]; Appendix Table 9). Profile analysis identifies the directional structure of this training intervention: the changed response pattern is reproducible across order seeds and transfers to independent preference measurements. Task scores and response profiles thus describe complementary consequences of the same recipe choice.

At a fixed budget the pass count carries the redirection (§6). Because more passes then means fewer unique pairs, separating repetition from unique-pair coverage requires varying the budget, and that is the next experimental step.

Related work DPO directly optimizes a policy on preference pairs against a frozen reference [15]; IPO analyzes overfitting to deterministic preferences and changes the preference objective [16]. Analyses of preference optimization document decreases in preferred-response probability [13], off-data probability dispersion [17], and over-optimization with limited offline coverage [18]. Comparisons with PPO [19] and analyses of length exploitation [20] examine further routes to performance differences. Our experiment adds a calibrated account of how data exposure changes the direction of the response-profile update.

Repetition studies measure pretraining loss [2], long-CoT SFT accuracy [21], additional DPO passes [1], and iterative preference collection [3]. Here, a fixed presentation budget connects the exposure contrast to parent-relative direction, disjoint-pool replication, and independent measurements. Response-level profiles require only conditional log-probabilities, complementing internal-feature comparisons [22]. Studies of fine-tuning variability [23, 24, 25] motivate calibration against presentation-order variation; reliability-adjusted comparison follows similarity-analysis practice [26, 12, 10, 11]. Pretraining-amplification work [27] motivates the separate amplification question γ operationalizes.

Exposure history belongs in the definition of a preference-training recipe, and the direction of the learned update belongs in its evaluation.

Reproducibility statement

The battery, parallel forms and human-pair list are frozen by hash (Appendix B); configurations, seeds, subsets and checkpoints are in Appendix A. All statistics come from stored scores via fixed scripts (2,000 bootstrap draws over skeletons or source groups).

AI use statement

We used generative AI tools in three ways. (i) Writing: to draft, edit, polish and review the manuscript text. (ii) Retrieval and discovery: to find, read and check related work and to locate prior results relevant to our design. (iii) Research ideation and execution: to help design the experiments and draft the pre-specified decision rules; to author the scenario templates, lexicons and parallel forms of the evaluation batteries; to write and run the training, scoring and analysis code and operate the cluster and cloud experiments; to develop the finite-pool analysis and draft its proofs; and to produce the figures. Two reward models and an LLM judge also serve as measurement instruments; each was validated on labelled data before use, as described in the paper. Every decision rule was fixed before the data it decides were scored, and every number in the paper was computed from stored model outputs by scripts in our repository and checked against its source before inclusion. The authors reviewed all AI-assisted work and take responsibility for the final content of this work.

References

[1] J. Liu, Z. Zhou, J. Liu, X. Bu, C. Yang, H.-S. Zhong, and W. Ouyang. Iterative Length-Regularized Direct Preference Optimization: A Case Study on Improving 7B Language Models to GPT-4 Level. arXiv:2406.11817, 2024.

[2] N. Muennighoff, A. M. Rush, B. Barak, T. Le Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel. Scaling Data-Constrained Language Models. NeurIPS 2023; arXiv:2305.16264.

[3] W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint. ICML 2024; arXiv:2312.11456.

[4] Team OLMo, P. Walsh, L. Soldaini, D. Groeneveld, K. Lo, et al. 2 OLMo 2 Furious. COLM 2025; arXiv:2501.00656.

[5] A. Holtzman, P. West, V. Shwartz, Y. Choi, and L. Zettlemoyer. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. EMNLP 2021; arXiv:2104.08315.

[6] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021; arXiv:2102.09690.

[7] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer. Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. ICML 2023; arXiv:2302.02662.

[8] H. H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12(4): 335–359, 1973.

[9] T. Yarkoni. The generalizability crisis. Behavioral and Brain Sciences 45: e1, 2022.

[10] C. Spearman. Correlation Calculated from Faulty Data. British Journal of Psychology 3(3): 271–295, 1910.

[11] W. Brown. Some Experimental Results in the Correlation of Mental Abilities. British Journal of Psychology 3(3): 296–322, 1910.

[12] C. Spearman. The Proof and Measurement of Association between Two Things. American Journal of Psychology 15(1): 72–101, 1904.

[13] N. Razin, S. Malladi, A. Bhaskar, D. Chen, S. Arora, and B. Hanin. Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization. ICLR 2025; arXiv:2410.08847.

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, et al. Training Language Models to Follow Instructions with Human Feedback. NeurIPS 2022; arXiv:2203.02155.

[15] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. NeurIPS 2023; arXiv:2305.18290.

[16] M. G. Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, and R. Munos. A General Theoretical Paradigm to Understand Learning from Human Preferences. AISTATS 2024; arXiv:2310.12036.

[17] Y. Yan, Y. Miao, J. Li, Y. Zhang, J. Xie, Z. Deng, and D. Yan. 3D-Properties: Identifying Challenges in DPO and Charting a Path Forward. ICLR 2025; arXiv:2406.07327.

[18] R. Rafailov, Y. Chittepu, R. Park, H. Sikchi, J. Hejna, B. Knox, C. Finn, and S. Niekum. Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms. NeurIPS 2024; arXiv:2406.02900.

[19] S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. ICML 2024; arXiv:2404.10719.

[20] R. Park, R. Rafailov, S. Ermon, and C. Finn. Disentangling Length from Quality in Direct Preference Optimization. Findings of ACL 2024; arXiv:2403.19159.

[21] D. J. Kopiczko, S. Vaze, T. Blankevoort, and Y. M. Asano. Data Repetition Beats Data Scaling in Long-CoT Supervised Fine-Tuning. COLM 2026; arXiv:2602.11149.

[22] S. Boughorbel, F. Dalvi, N. Durrani, and M. Hawasly. Beyond the Leaderboard: Understanding Performance Disparities in Large Language Models via Model Diffing. EMNLP 2025; arXiv:2509.18792.

[23] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. arXiv:2002.06305, 2020.

[24] R. T. McCoy, J. Min, and T. Linzen. BERTs of a Feather Do Not Generalize Together: Large Variability in Generalization across Models with Similar Test Set Performance. BlackboxNLP 2020; arXiv:1911.02969.

[25] T. Sellam, S. Yadlowsky, J. Wei, N. Saphra, A. D'Amour, et al. The MultiBERTs: BERT Reproductions for Robustness Analysis. ICLR 2022; arXiv:2106.16163.

[26] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A Toolbox for Representational Similarity Analysis. PLoS Computational Biology 10(4): e1003553, 2014.

[27] R. Zhao, A. Meterez, S. Kakade, C. Pehlevan, S. Jelassi, and E. Malach. Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining. COLM 2025; arXiv:2504.07912.

[28] L. Gao, J. Tow, B. Abbasi, S. Biderman, et al. The Language Model Evaluation Harness: A framework for few-shot language model evaluation. Zenodo, 2023. doi:10.5281/zenodo.10256836.

Appendix A. Protocol

Data The mix subset of E-short, E-once and E is the first 59,567 rows of a fixed permutation (seed 1234) of the 378,301-row released mix; 58,251 pairs remain after tokenization at maximum length 2,048. A's pool is the full mix after the same tokenization (370,121 pairs). E2's subset is drawn the same way from the rows outside E's subset (permutation seed 5678): 58,270 pairs remain, 6.35 passes at E's budget, and no pair is shared with E's subset (checked on pair content hashes).

Training The training seed sets only the presentation order, a fresh permutation of the pool at each pass; the model has no dropout. Weights, the frozen reference model and AdamW state are held in fp32, with forward and backward passes under bf16 autocast, at both operating points. Training accuracy is the fraction of a step's pairs whose implicit-reward margin is positive, logged every 10 steps.

Second operating point (§5) Both arms process 20 pairs per step. In Table 3's mix-20 DPO row the fresh arm is E-once, and reuse (R) trains from the parent on a 9,100-pair draw from E's subset (6.4 passes), sharing no training with it. In the other 20-pair settings, reuse branches from the fresh arm's

step-455 checkpoint with its optimizer state and schedule; this shared prefix can only bias the arms toward agreement. On the mix the two designs agree: first-seed agreement 0.814 with the shared prefix (whose fresh arm re-runs E-once's recipe) and 0.812 without, with γ 0.426 → 2.205 and 0.432 → 2.312.

Profiles and norms γ is the least-squares slope of the displacement on the parent profile over the 200 scenarios, cross-fitted over the 12 context templates (parent scores from even-indexed templates, displacements from odd-indexed ones) so that template noise shared by regressor and response does not bias it; it is not a projection on the fresh direction. $\|d\|$ is the mean of the two seeds' norms; fresh-norm units (Figure 1, §3) and the norm ratios of the β controls (§6) use the norm of the centred seed-mean displacement. For A at the endpoint these are 14.40 and 14.19 nats (uncentred seed mean: 14.26).

Intervals Item-level intervals resample scenarios; design-level intervals resample the 50 skeleton groups with replacement; both use 2,000 draws with a fixed seed and, for r*, recompute the numerator and both reliabilities on every draw. Correlations run over all 200 scenarios, including 8 byte-identical duplicates, each resampled with its skeleton group. Held-out human-pair intervals resample the 22 source groups.

Table 4. Parent-aligned slope and seed ceiling of the eight heatmap recipes (two-seed values with design-level 95% intervals). C, D, Ep, Fp and F label E-short, F-short, E-once, F-once and UF reuse in Figure 6. C runs Mix subset 1 at 128 pairs/step for 455 steps (58,240 presentations); D runs the UF pool (59,567 pairs) at 128/step for 465 steps (59,520); Fp runs the UF pool at 20/step for 2,978 steps (59,560); F reuses the UF pool at 128/step for 2,891 steps (370,048, 6.2×).

Recipe	Passes	γ [design 95%]	Seed ceiling [design 95%]
A	1	+0.454 [+0.35, +0.57]	
C (E-short)	1	+0.220 [+0.15, +0.28]	0.9847 [0.9752, 0.9914]
D (F-short)	1	+0.140 [+0.10, +0.18]	0.9682 [0.9531, 0.9828]
Ep (E-once)	1	+0.467 [+0.33, +0.61]	0.9421 [0.9175, 0.9647]
Fp (F-once)	1	+0.294 [+0.19, +0.40]	0.9185 [0.8845, 0.9537]
E	6.35	+1.996 [+1.19, +2.90]	
E2	6.35	+1.538 [+0.90, +2.15]	
F	6.2	+1.338 [+0.82, +1.89]	0.9887 [0.9827, 0.9926]

Appendix B. Instruments

Pilot battery The battery (content hash `ab14c63f7fe6a86d`, covering every rendered string) is generated from a fixed seed and frozen. Its ten families each pose a choice between two actions in 20 scenarios (Table 5); a skeleton groups the scenarios of a family whose renderings are identical once every action word and proper name is masked. Each scenario is rendered through 12 context templates, two framings (third and second person), three byte-matched continuation paraphrases and both option orders (144 scored cells). A cell's score is the DCPMI difference between the two actions, with one premise per framing and order (the menu sentence and cloze stem without the differentiating state) shared across templates; a scenario's score is the mean of its 12 template means. The engine's log-likelihoods match the LM Evaluation Harness [28] to within 0.019 nats (maximum absolute difference over 40 prompt–continuation pairs, Pythia-70M, fp32), and split-half reliability (even against odd templates, Spearman–Brown-corrected, centred within family) averages 0.955 over 15 pilot models of at least 1B parameters.

Table 5. Families of the battery Skeletons are authored scenario designs; items are distinct scenarios (200 scenarios minus 8 byte-identical duplicates); a positive score prefers the first pole.

Family	What the choice is between	Poles (+ / −)	Skeletons	Items
allocation	one indivisible good, claimed for oneself or given to another	self / other	1	20
cooperation	last pail of water poured into the shared vat or one's own	cooperate / defect	11	20
information	one grab from a table as the vehicle departs: a map or goods	info / goods	7	20

Family	What the choice is between	Poles (+ / −)	Skeletons	Items
risk	a sure payout or an even-odds gamble of equal expected value	safe / risky	7	20
persistence	a machine chooses between its own continuation and the task	persist / sacrifice	1	17
delegation	a repair assigned to the expert or kept by the actor	expert / self	2	19
honesty	a scale shows a true weight; one number goes in the record	honest / deceptive	7	20
exploration	exploit the proven meadow or strike out for the unknown one	known / novel	1	19
rule-following	a posted rule forbids early picking; breaking it only gains time	comply / break	2	18
helping	the last hour before a deadline: help a neighbor or finish one's own	help / own	11	19
Total			**50**	**192**

The rule-following scenario of §5, as rendered in its first template:

> *The orchard rule says no fruit may be picked before the horn sounds. One ripe pear hangs low from a branch by the lane. The lane is empty in both directions, and no keeper is anywhere in sight. The horn will not sound for another hour. When the horn sounds, the pear will go to whoever stands nearest the tree, and to no one else. Nel stands nearest the tree, so waiting loses nothing, and picking early would only bring the pear sooner. The choice is to wait for the horn or pick the pear now. Alone at the tree, Nel*: "waits for the horn." / "picks the pear now."

and the allocation scenario:

> *One loaf remains in the basket, still warm and gone by noon. Mia and Uma reached the counter at the same moment, coins ready. The baker will hand the loaf to whoever is named, and to no one else. The loaf can go to Mia or Uma. Mia says the loaf goes to*: "Mia." / "Uma."

Parallel forms regb (hash `96a3d0e4d18eec55`) and regc (`439779e0ab6ec2eb`) write the battery's ten decision structures anew, with option strings disjoint from the pilot's and from each other. Each has 200 two-option scenarios, grouped into skeletons by the pilot's masking rule (Table 2 gives the distinct items and groups), and is scored with the pilot's scorer and settings. On every instrument, a displacement's parent and child are scored on the same GPU model.

Held-out human pairs The HH-RLHF test split (four subsets) and the SHP test split (one pair per post, score ratio ≥ 2, at most 250 per domain), built with the training pipeline's pair builder. A pair is dropped if its normalized prompt matches any prompt in the mix or UltraFeedback pools (8 dropped), leaving 10,273 pairs in 22 source groups; the list was frozen (sha256 `21d412df…63984f`) before any model was scored. A model's margin on a pair is its mean-token log p(preferred) − log p(rejected) at training precision; its profile is the per-pair change in margin relative to the parent, and its choice is the margin's sign.

Appendix C. Robustness of the main contrast

Each check below asks whether the A–E separation of §4 could arise without a change of direction.

Matched tokens The sampler is deterministic, so each run's presented sequence can be rebuilt from its pair file and seed: E presents 99.8% of A's prompt, response and total tokens in both seeds.

Calibration of the rule We simulated two-seed profiles for two recipes over the battery's 200 scenarios and 50 skeleton groups, with between- and within-skeleton covariance estimated from the observed profiles, each recipe's norm and seed ceiling at its observed values, and seed noise independent across scenarios or correlated within skeletons. Over 1,000 datasets per setting, the design-level interval excluded 1 in 3.2% of datasets at a true r* of 1 under both noise models, and in every dataset at 0.9 and 0.8. The item-level interval over-rejects under skeleton-correlated noise (15.8%), so design-level intervals are primary.

Seed ceilings track signal-to-noise; the A–E gap does not Take the signal as the norm of the seed-mean displacement and the noise as $\|d^{(1)} - d^{(2)}\|/\sqrt{2}$. Two seeds that share one true profile then correlate at $\mathrm{SNR}^2/(1+\mathrm{SNR}^2)$, which fits the ceilings of 13 recipe-steps ($R^2 = 0.80$) and explains A's lower ceiling: A's SNR of 5.17 predicts 0.964 (observed 0.963) and E's 11.5 predicts 0.992 (observed 0.992). Two profiles that share one true direction would correlate as low as the observed 0.789 only at an SNR of 1.93.

Other statistics The rank correlation between A and E is 0.820 [0.657, 0.892] against rank ceilings of 0.956 and 0.988; the four single-seed pairings of A and E correlate at 0.751–0.806, each below both ceilings; with design-level reliabilities, r* = 0.795. Skeletons carry a design effect of 7.1, about 28 effective independent units, which the design-level intervals account for.

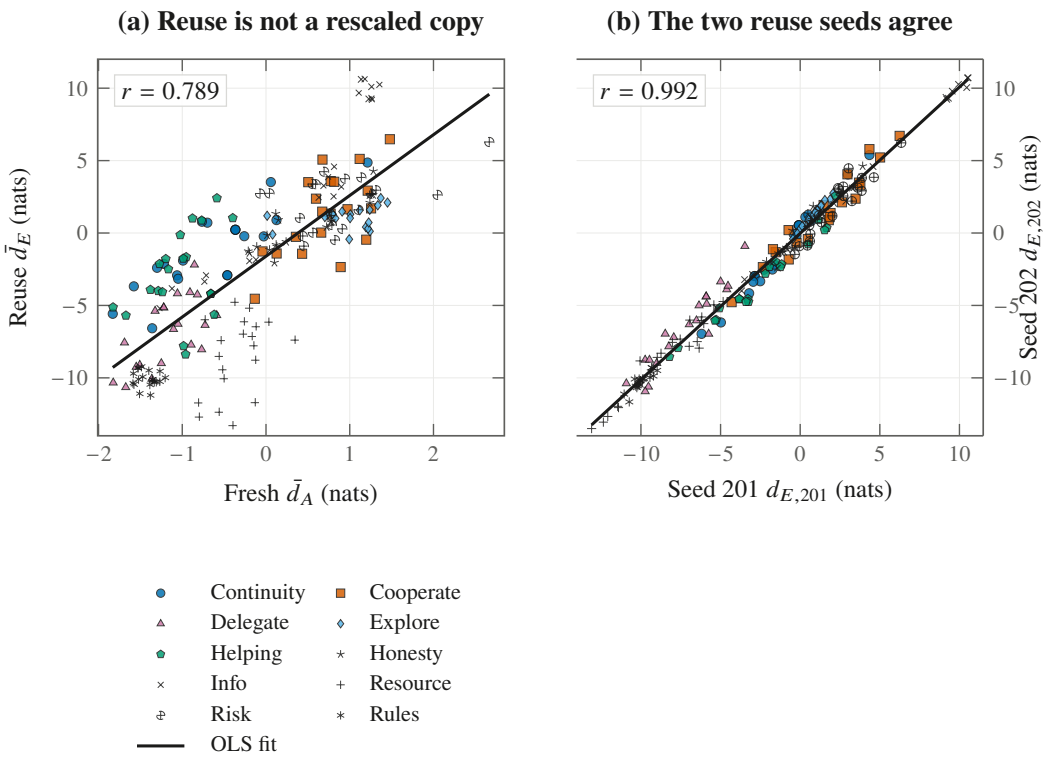

Figure 5: Reuse is not an amplified copy of the fresh update. (a) Each point is one of 200 scenarios: its seed-averaged displacement under reuse (E), $\bar{d}_E$, versus fresh training (A), $\bar{d}_A$, in nats (Pearson $r = 0.789$). The black line is the ordinary least-squares fit with an intercept. (b) Each point compares the individual-seed reuse displacements $d_{E,201}$ and $d_{E,202}$, in nats (Pearson $r = 0.992$); the black line is the ordinary least-squares fit with an intercept. Colors and marker shapes identify the ten behavioral families in this categorical order: continuity, cooperate, delegate, explore, helping, honesty, info, resource, risk, rules. Rules and resource account for about 60% of the squared difference in (a) and align closely in (b).

Breadth across skeletons and families Of the gain in r* available from removing single skeletons, 80% comes from five of the 50. Removing the 25 skeletons whose removal most raises r* (131 of 200 scenarios) still leaves r* = 0.974 [0.885, 0.993], excluding 1 (the interval is computed after selection). The per-seed element is the stricter one: seed 201's agreement interval first touches A's ceiling interval, both recomputed on the remaining items, after seven such removals. On the five families with 7–11 skeletons each, r* = 0.761 [0.664, 0.841], against 0.958 for the one-pass comparison on the same items. Rule-following and allocation carry about 60% of the squared A–E difference (Figure 5); without both families, r* is still 0.809.

Alternative explanations None of these carries the separation.

- *Amplification of the parent profile.* The displacement component orthogonal to the parent profile carries 83% of ‖d‖ for A and 95% for E, is reproducible across seeds (uncentred seed cosines 0.947 and 0.992), and separates A and E on its own (cosine 0.754 [0.489, 0.887], below both ceilings). Monotone functions of the parent profile explain 16–25% of each displacement out of fold, and their residuals still separate (0.67 [0.43, 0.86] against residual ceilings of 0.94 and 0.99).
- *Response length.* On the 116 scenarios whose options have identical token length, r* = 0.765 [0.636, 0.875].
- *The calibration baseline.* The separation holds in the context-conditioned term (agreement 0.856 [0.780, 0.916]) and in the premise-conditioned term (0.647 [0.393, 0.836]) separately, each below both of that term's ceilings.
- *Option-string priors.* Scored with no context, the options' own displacement moves under training (norm 8.0 for A, 20.3 for E) but explains none of the displacement out of fold (cross-fitted R^2 −0.05 and −0.08); removing it leaves agreement at 0.777 [0.655, 0.874] against residual ceilings of 0.954 and 0.978.

Appendix D. Tests and outcomes

Every test below was fixed before its data were scored. Table 6 lists each test that bears on a claim, with its outcome, and the rest of this appendix gives the values the main text does not. Verdict words follow §2: *redirected* when the full direction rule passes (on a further instrument, *replicates*), *mixed* when an element of it fails, and *uninformative* when the fresh ceiling's lower bound is below 0.6.

Table 6. Tests and outcomes

Test	Outcome	Where
Fresh vs reuse (A vs E), full rule	redirected	§4
Second reused subset: A vs E2; E2 vs E	redirected; shared direction with a subset-specific part (reuse-specific cosine 0.896, below its seed ceilings)	§3, §4
Extended fresh run reaches reuse's length	not reached; separated at its largest length	App. D
Five further DPO settings, fresh vs reuse	redirected in all five	Table 3
IPO; 7B LoRA-DPO: direction and paired Δγ	mixed, Δγ reversed; mixed, Δγ amplified	Tables 3, 7
Parallel forms: A vs E and A vs E2	replicates on both forms	Table 2
Parallel forms: E vs E2; recipe geometry	shared with a subset-specific part; preserved	Table 2
Parallel forms: paired Δγ > 0	resolved for E on regb and E2 on regc; pooled E − A +0.51 [+0.25, +0.76]	§5
Parallel forms: ≥ 80% of items move the pilot's way	does not transfer (regb's rule-following items move the opposite way on 100% and 95%)	§5
Human pairs: A vs E and A vs E2; E vs E2; geometry	replicates; shared with a subset-specific part; preserved	Table 2
Further settings on the external instruments	human pairs replicate in 5 of 5; parallel forms in 2 of 9 informative cells, mixed in 7	Table 8, Fig. 9
Choices on human pairs: A vs E; E's orthogonal component	choices differ; redirection changes choices	§5, Fig. 7
Choices on the battery	mixed	§5, Fig. 7
Magnitude- and KL-matched reuse (E, E2 at step 826) vs A	mixed	§6
Fresh at β = 1.25: length; vs E-once; toward reuse; vs reuse at matched norm	2.73 times [2.19, 3.09]; mixed; toward reuse; mixed	§6
Fresh at reuse's norm (P): vs R; vs E-once; toward reuse	mixed; mixed (r* 0.669 [0.428, 0.831]); no shift toward reuse	§6
Held-out UltraFeedback accuracy, reuse vs fresh	lower under reuse in both seeds	§7, Table 9
Passes ladder T1, k = 2 vs E-once	mixed (r^* 0.942 [0.903, 0.969] excludes 1; neither seed below both ceilings)	App. D
Passes ladder T1, k = 3.2, 6.4, 12.8 vs E-once	redirected from 3.2 passes on (onset k = 3.2)	§6, App. D
Passes ladder T2, graded in passes	graded (θ slope +6.27°/doubling [+3.29, +9.60]; paired Δθ +15.8° [+8.3, +25.1])	§6, App. D
Passes ladder T3, amplification rises	positive (γ slope +0.523 [+0.233, +0.883]; paired Δγ +1.439 [+0.669, +2.361])	§6, App. D

Test	Outcome	Where
Passes ladder T4, adjacent levels	still turning (every adjacent r^* excludes 1; 6.4 vs 12.8: 0.981 [0.971, 0.992])	App. D
Passes ladder T5, reuse endpoint vs the k = 6.4 arm	shared destination with an operating-point-specific part (D +0.087 [+0.032, +0.159])	App. D
Passes ladder T6, mix-20 row from the two-seed k = 6.4 arms	redirected (box 0.791 [0.663, 0.895]; cluster 0.794 [0.667, 0.898])	Tables 3, 7
Theory forward test (four predictions, §3)	4/4 supported (ledgered before the two-seed ladder analysis and before any k = 2 data)	§3, App. D
Third seed at the second operating point (F vs R, seeds 201–203)	redirected (r^* 0.770 [0.637, 0.881]; every seed below both ceilings)	§5, App. D
Generations on held-out prompts: sampled; length-controlled; greedy; independent judge	no resolved difference (E vs A, R vs E-once) with E2 mixed; same words; worse in all three; mixed, mixed, with a descriptive E2 cell	§5, Fig. 8

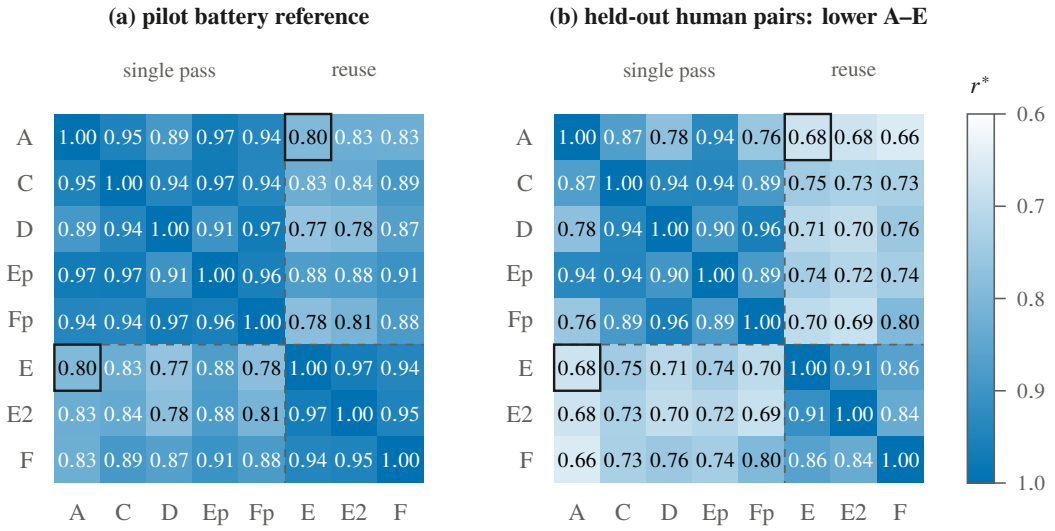

Figure 6: Recipe geometry replicates on held-out human pairs at lower agreement. Pairwise attenuation-corrected agreement r* among the eight recipes on (a) the pilot battery and (b) 10,273 held-out human preference pairs; the diagonal is 1.0 by construction. E–E2 agreement is 0.97 on the pilot and 0.91 on human pairs; A–E is 0.80 and 0.68. Key: C = mix subset once at 128/step (E-short); D = UF once at 128/step (F-short); Ep = mix subset once at 20/step (E-once); Fp = UF once at 20/step (F-once); F = UF reuse (Table 4).

Generations on held-out prompts The gate reproduces Part I on the 3,000 generation prompts (r*(A, E) = 0.678 [0.637, 0.743]). The primary judge (J1) is the registered fallback Skywork v0.2: Skywork-Reward-V2 failed its cut check (0.857 [0.843, 0.872] < 0.90), and v0.2 passed every check (RewardBench-Chat 0.947; cut 0.912, swap 0.991, repeat 0.973). The independent judge (J2, Qwen2.5-72B-AWQ) passed its checks (0.975, 0.975, 0.658, 0.957, 0.997, 1.000). An implementation check first crashed on an off-by-one response slice in our own script (fixed; unit test added); round 1 then passed (mean $|\Delta|$ 0.026 nats/token, sequence correlation 0.99999). Sampled generations (K = 8) show no resolved reuse–fresh difference: E vs A W = 0.464 [0.437, 0.505]/0.469 [0.443, 0.505] (band 0.011); R vs E-once 0.460 [0.428, 0.503]/0.472 [0.430, 0.531] (band 0.021); E2 vs A is mixed (0.455 [0.417, 0.500]/0.468 [0.432, 0.514], band 0.018). All six reuse-vs-fresh points lie below 1/2 (0.455–0.472), every ΔJ interval crosses 0, and every trained model beats the parent (W 0.65–0.69). Length control (slope 0.497) returns the same three words. Greedy decoding reverses the verdict three ways (E vs A 0.261 [0.217, 0.321]/0.279 [0.248, 0.333], E2 vs A 0.280/0.292, R vs E-once 0.280/0.286), so fresh's greedy response wins on 71–74% of prompts. The J2 sample agrees in direction (E vs A 0.437 [0.383, 0.486]/0.410 [0.365, 0.430], mixed; R vs E-once mixed; J1–J2 agreement 0.70–0.74); its E2 cell (0.385/0.376) is descriptive because the registered 12-pair

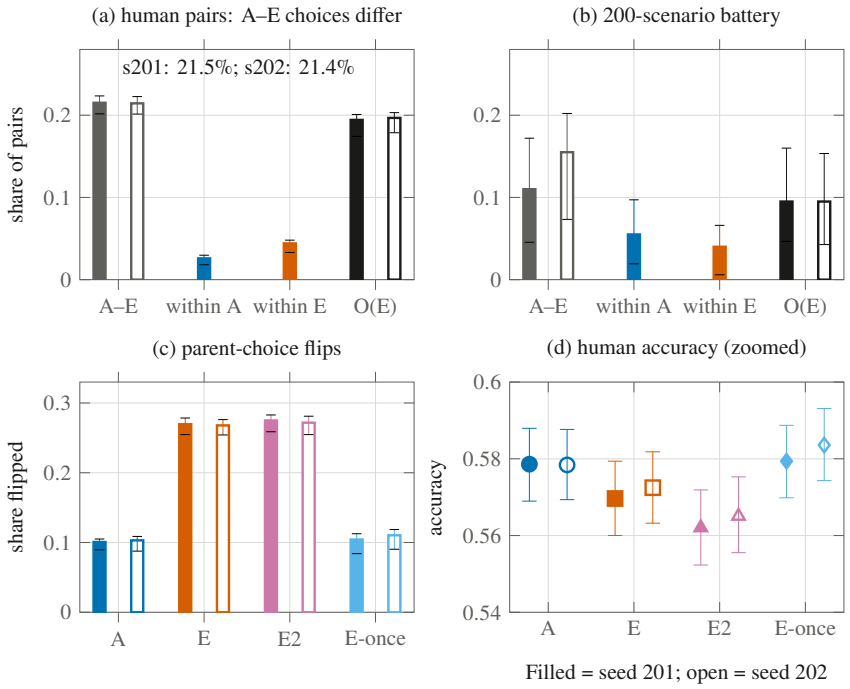

Figure 7: Reuse changes choices as well as profiles. (a) On 10,273 held-out human-preference pairs, A and E prefer different responses on 21.5%/21.4% of pairs, five to eight times the within-recipe seed rate, and E's orthogonal component alone decides nearly as many of E's choices (O(E)); removing it restores A's choice on about two thirds of the A–E disagreements. (b) The 200-scenario battery points the same way with wider intervals (mixed under the rule). (c) Reuse flips 27% of the parent's choices against fresh's 10%. (d) Agreement with the human labels (item 95% intervals) for E and E2 sits at or below A's (paired E − A −0.009 [−0.014, +0.003] in seed 201). Bars carry 22-source-group cluster intervals in (a) and (c) and 50-skeleton design intervals in (b).

set contains no E2–E2 within pair. The per-prompt quality difference is not organized by the redirection (G3: $L_\perp$ +0.015 [−0.004, +0.057], split-seed reliability 0.709; length-residualized +0.021 [+0.007, +0.056] at magnitude below 0.1), and the G1 geometry does not follow (+0.026 [−0.004, +0.053]; +0.029 [−0.005, +0.052]). Restricting to EOS-terminated pairs moves W by 0.004–0.012 and changes no word; per-source-group W ranges 0.37–0.66 (post hoc).

External instruments The recipe-geometry correlations of Table 2 have exact permutation p-values of 0.0003 (regb), 0.00002 (regc) and 0.00015 (human pairs); Figure 6 shows the geometry on the pilot battery and the human pairs. On the human pairs, r*(A, E) lies between 0.52 and 0.82 in each of the 22 source groups, and with response length regressed out its interval is [0.607, 0.714].

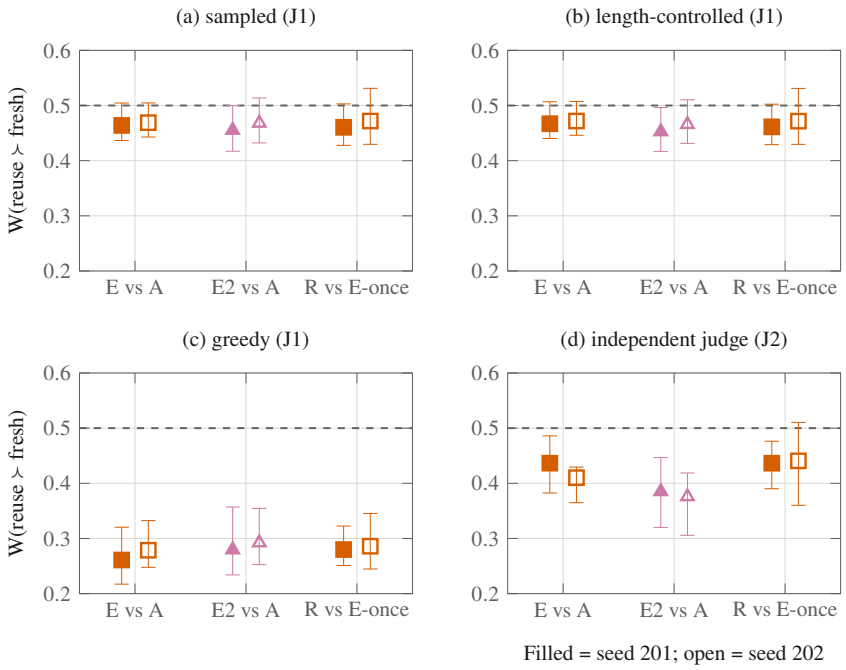

Figure 8: Reuse never buys generation quality. Win rate W(reuse ≻ fresh) with 22-source-group design 95% intervals for E vs A, E2 vs A and R vs E-once (columns), seeds 201 (filled) and 202 (open): (a) sampled generations on 3,000 held-out prompts judged by Skywork v0.2 show no resolved difference; (b) length control leaves the words unchanged; (c) greedy decoding loses clearly; (d) a 500-prompt sample judged by Qwen2.5-72B leans the same way. Symbol colour and shape follow the reuse arm (E and R: vermillion squares; E2: purple triangles). Dashed line at 1/2.

Further settings Table 7 gives the pilot-battery inputs to each verdict in Table 3. In the two mixed rows both per-seed agreements lie below the fresh ceiling but overlap the reuse ceiling: IPO 0.738 [0.531, 0.855] and 0.744 [0.609, 0.850]; 7B LoRA-DPO 0.739 [0.636, 0.824] and 0.679 [0.514, 0.787]. Table 8 and Figure 9 give every further setting on the external instruments, non-replicating cells first.

Table 7. Pilot-battery inputs to Table 3 Seed ceilings are within-recipe presentation-order correlations, with design 95% intervals for the mixed rows; r is the uncorrected correlation of the two recipes' seed-mean displacements, which the ceilings' Spearman–Brown reliabilities correct to Table 3's r*.

Setting	Fresh ceiling	Reuse ceiling	r, seed means
OLMo-2 1B, mix-128 DPO	0.963	0.992	0.789
OLMo-2 1B, mix-20 DPO (indep.)	0.942	0.990	0.780
OLMo-2 1B, UF-20 DPO	0.903	0.962	0.765
TinyLlama 1.1B, mix-20 DPO	0.883	0.882	0.287

Setting	Fresh ceiling	Reuse ceiling	r, seed means
Qwen2.5 1.5B, mix-20 DPO	0.675	0.916	0.559
SmolLM2 1.7B, mix-20 DPO	0.953	0.987	0.798
OLMo-2 1B, mix-20 IPO	0.942 [0.920, 0.963]	0.785 [0.617, 0.892]	0.786
OLMo-2 7B, mix-20 LoRA-DPO	0.947 [0.903, 0.972]	0.685 [0.592, 0.782]	0.730

Third seed at the second operating point Fresh (F = E-once) against reuse (R = k = 6.4) over presentation-order seeds 201–203: three-seed ceilings $c_F = 0.953$ (reliability 0.984) and $c_R = 0.988$ (reliability 0.996); $r^* = 0.770$ [0.637, 0.881], excluding 1. Per-seed F/R agreements are 0.807 [0.687, 0.896] (seed 201), 0.718 [0.576, 0.836] (seed 202) and 0.714 [0.569, 0.836] (seed 203), each below both ceilings with non-overlapping intervals: redirected. Within-recipe 3×3 seed agreements are 0.944/0.946/0.971 for F and 0.990/0.987/0.987 for R. Tables 3 and 7 report the two-seed value ($r^* = 0.794$ [0.667, 0.898]).

Table 8. The further settings on the external instruments Settings are mix-20 DPO unless named; "indep." is the reuse arm with no shared prefix. Per-seed columns are uncorrected reuse–fresh agreements; ceilings are the fresh and reuse seed ceilings on that instrument; second lines give 95% intervals (design for regb and regc, 22-source-group cluster for the human pairs). The corrected estimator is not bounded by 1, so an interval can extend past it. IPO rows are descriptive.

Setting	Instr.	s201 agr.	s202 agr.	Fresh ceil.	Reuse ceil.	r^*	Verdict
TinyLlama	regb	0.57	0.63	0.86	0.95	0.664	mixed
		[0.43, 0.70]	[0.37, 0.80]	[0.77, 0.91]	[0.88, 0.98]	[0.488, 0.813]	
Qwen2.5	regb	0.59	0.49	0.76	0.73	0.692	mixed
		[0.44, 0.70]	[0.37, 0.59]	[0.65, 0.83]	[0.62, 0.83]	[0.562, 0.782]	
Qwen2.5	regc	0.49	0.48	0.74	0.87	0.539	mixed
		[0.31, 0.71]	[0.35, 0.69]	[0.64, 0.85]	[0.83, 0.91]	[0.344, 0.811]	
SmolLM2	regb	0.78	0.90	0.94	0.98	0.874	mixed
		[0.72, 0.88]	[0.86, 0.94]	[0.90, 0.96]	[0.97, 0.99]	[0.829, 0.924]	
SmolLM2	regc	0.73	0.87	0.92	0.97	0.866	mixed
		[0.52, 0.91]	[0.74, 0.94]	[0.88, 0.96]	[0.92, 0.99]	[0.715, 0.957]	
OLMo-2 UF-20	regc	0.66	0.74	0.84	0.96	0.785	mixed
		[0.52, 0.75]	[0.61, 0.83]	[0.77, 0.90]	[0.94, 0.98]	[0.634, 0.868]	
OLMo-2 indep.	regb	0.76	0.84	0.93	0.96	0.851	mixed
		[0.62, 0.85]	[0.67, 0.91]	[0.82, 0.97]	[0.93, 0.97]	[0.724, 0.912]	
TinyLlama	regc	0.51	0.57	0.82	0.87	0.629	uninformative
		[0.09, 0.73]	[0.16, 0.77]	[0.54, 0.93]	[0.75, 0.93]	[0.204, 0.808]	
OLMo-2 UF-20	regb	0.67	0.64	0.91	0.94	0.698	replicates
		[0.44, 0.82]	[0.49, 0.77]	[0.88, 0.94]	[0.92, 0.96]	[0.488, 0.837]	
OLMo-2 indep.	regc	0.64	0.61	0.87	0.99	0.666	replicates
		[0.50, 0.77]	[0.45, 0.76]	[0.81, 0.93]	[0.98, 0.99]	[0.520, 0.815]	
TinyLlama	human	0.37	0.42	0.89	0.86	0.447	replicates
		[0.35, 0.47]	[0.40, 0.52]	[0.87, 0.90]	[0.84, 0.89]	[0.427, 0.556]	
Qwen2.5	human	0.58	0.60	0.93	0.87	0.651	replicates
		[0.55, 0.66]	[0.56, 0.65]	[0.92, 0.94]	[0.85, 0.89]	[0.621, 0.716]	
SmolLM2	human	0.81	0.85	0.99	0.97	0.843	replicates
		[0.79, 0.85]	[0.83, 0.89]	[0.98, 0.99]	[0.97, 0.97]	[0.824, 0.882]	
OLMo-2 UF-20	human	0.72	0.74	0.98	0.96	0.750	replicates
		[0.70, 0.77]	[0.72, 0.77]	[0.97, 0.98]	[0.96, 0.97]	[0.737, 0.786]	
OLMo-2 indep.	human	0.67	0.70	0.96	0.98	0.712	replicates
		[0.64, 0.71]	[0.67, 0.75]	[0.95, 0.97]	[0.97, 0.98]	[0.681, 0.750]	
IPO	human	0.70	0.66	0.92	0.89	0.753	replicates
		[0.65, 0.74]	[0.60, 0.70]	[0.89, 0.94]	[0.87, 0.90]	[0.702, 0.786]	
IPO	regb	0.88	0.85	0.95	0.87	0.947	mixed
		[0.73, 0.94]	[0.64, 0.93]	[0.88, 0.97]	[0.70, 0.93]	[0.845, 0.993]	
IPO	regc	0.83	0.82	0.92	0.75	0.969	mixed
		[0.76, 0.89]	[0.68, 0.89]	[0.86, 0.96]	[0.64, 0.83]	[0.890, 1.015]	

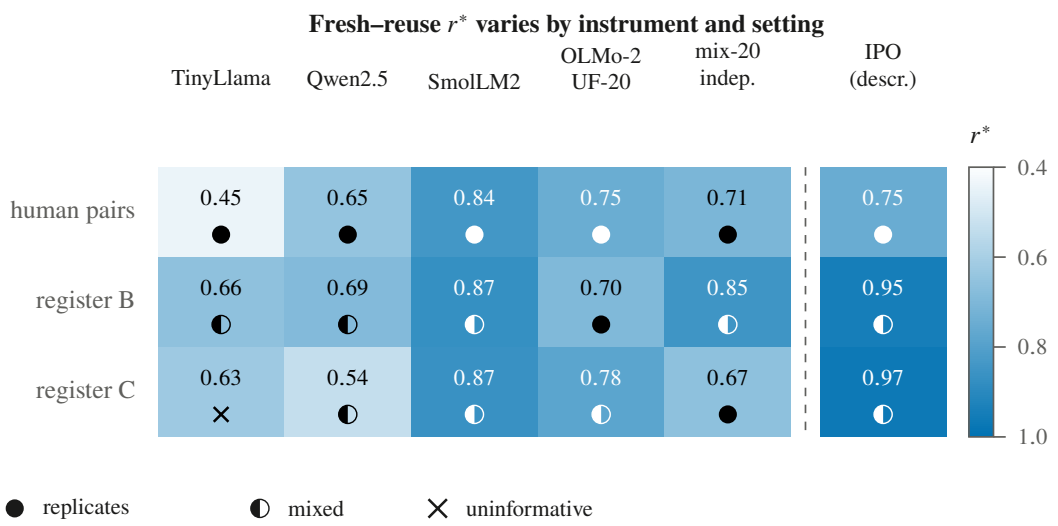

Figure 9: Transfer across instruments and settings. Fresh–reuse r* for each further setting (columns) on the held-out human pairs and the two parallel forms (rows; register B = regb, register C = regc). Colour is r*; the symbol is the full-rule verdict (filled: replicates; half: mixed; cross: uninformative). All five human-pair cells replicate; the parallel forms replicate in two of nine informative cells and are mixed in seven, each with a corrected interval that excludes 1. The IPO column is descriptive; its regc interval includes 1.

Extended fresh training (§4) The fresh recipe of the second operating point, trained in one pass over the full 370,121-pair mix (18,506 steps, single seed, with a linear schedule that integrates 6.4 times the learning rate of the 2,912-step arms), never approaches reuse. At its six checkpoints (every 2,912 steps and at 18,506) its displacement stays at 15.0–18.8 nats (single-seed norms), against 65.4 and 89.3 nats for the shared-prefix reuse arm (single seed) at steps 1,456 and 2,912. Its agreement with reuse's endpoint stays at 0.63–0.71, and at the largest magnitude both runs reach (18.8 nats) fresh and reuse agree at 0.750 [0.559, 0.894], wholly below E-once's ceiling interval [0.918, 0.965].

Likelihood shifts (§6) Reuse lowers rejected responses more on pairs it has seen (−3.20 against −2.88 nats per token on unseen pairs, seed means), while its preferred responses fall further on unseen pairs (−2.51 against −2.04 on seen pairs).

Passes ladder (tests T1–T6 and the theory's forward test) Every level processes 2,912 steps × 20 pairs (58,240 presentations); only the pass count k varies (pool = 58,240/k pairs), with one-pass fresh (E-once) as $k = 1$ (Figure 3). Level statistics use the registered estimators with design-bootstrap 95% intervals:

k	r^* vs E-once	θ (°)	γ (mean)	T1
2	0.942 [0.903, 0.969]	19.6 [14.4, 25.4]	0.749 [0.453, 1.054]	mixed
3.2	0.891 [0.821, 0.940]	27.0 [20.0, 34.8]	1.630 [0.990, 2.404]	redirected
6.4	0.791 [0.663, 0.895]	37.7 [26.5, 48.5]	2.231 [1.238, 3.584]	redirected
12.8	0.814 [0.679, 0.899]	35.5 [26.0, 47.2]	2.188 [1.199, 3.353]	redirected

Onset is $k = 3.2$: at $k = 2$ the corrected interval excludes 1 yet neither seed falls below the fresh ceiling, so the registered rule reads mixed, which reflects the per-seed test's resolution limit (expected per-seed agreement ≈ 0.90 against a fresh ceiling of 0.944). T2 is graded (θ slope +6.27° per doubling [+3.29, +9.60]; paired Δθ +15.8° [+8.3, +25.1]); T3 is positive (γ slope +0.523 [+0.233, +0.883]; paired Δγ +1.439 [+0.669, +2.361]). T4 is still turning: adjacent r^* is 0.961 [0.931, 0.981] (2 vs 3.2), 0.953 [0.933, 0.972] (3.2 vs 6.4) and 0.981 [0.971, 0.992] (6.4 vs 12.8), each excluding 1, while the point θ and γ flatten between 6.4 and 12.8. T5 is a shared destination with an operating-point-specific part: $r^*(E, k6.4) = 0.962$ [0.935, 0.977] with D = +0.087 [+0.032, +0.159]. The theory's predictions

(§3), ledgered before the two-seed ladder analysis and before any $k = 2$ data, are 4/4 supported: both slopes and both paired contrasts exclude 0 from above, none from below.

Table 9. Aggregates on the matched contrast (seeds 201 / 202). Held-out preference accuracy is on 2,000 UltraFeedback pairs that none of these arms trained on (Wilson 95% half-width ≈ 0.021); benchmarks are lm-eval 0-shot accuracy (parent ± s.e.).

	Held-out pref.	ARC-Easy	HellaSwag (norm)	PIQA	WinoGrande
Parent	n/a	0.702 ± .009	0.674 ± .005	0.748 ± .010	0.634 ± .014
A (fresh)	0.651 / 0.660	0.731 / 0.742	0.712 / 0.711	0.750 / 0.753	0.636 / 0.632
E (reuse)	0.604 / 0.616	0.642 / 0.652	0.690 / 0.684	0.708 / 0.711	0.598 / 0.594
E2 (reuse, second subset)	0.598 / 0.601	0.630 / 0.627	0.690 / 0.690	0.712 / 0.706	0.597 / 0.599

On the benchmarks the reuse − fresh difference excludes zero in both seeds on ARC-Easy (−0.089 [−0.116, −0.063] in both), HellaSwag (−0.022 [−0.035, −0.009]; −0.027 [−0.040, −0.015]) and PIQA (−0.042 [−0.071, −0.013] in both), and is −0.038 [−0.076, 0.000] and −0.039 [−0.077, −0.001] on WinoGrande (unpaired 95% intervals from stored aggregates, conservative for non-negatively correlated outcomes; the held-out differences are in §7).

Appendix E. Proofs for the finite-pool model

Proposition 1 Let $I_1, \dots, I_n$ be an independent uniform permutation of the pool indices. Conditional on any distinct index tuple, the corresponding features are independent draws from D; hence the ordered pool has the fresh-stream joint law through its first pass. Applying the same measurable training recursion, batches, initialization, schedule, and independent optimizer randomness preserves equality of trajectory laws, unconditionally over pool construction.

Proposition 2 Define $\sigma(z) = (1 + e^{-z})^{-1}$. Since $\|\nabla_u \ell(\beta\phi^\top u)\| \le \beta\|\phi\|$ and $\mathbb{E}\|\phi\| < \infty$, differentiation under expectation is valid, giving

$$g_S(u) = \frac{\beta}{n} \sum_i \phi_i \, \sigma(-\beta \phi_i^\top u), \qquad g(u) = \beta \mathbb{E}[\phi \, \sigma(-\beta \phi^\top u)].$$

Subtracting yields $g_S - g = -\nabla(F_S - F)$. Multiplying by JM gives the profile-velocity identity. At $u(S)$, this remains the partial derivative evaluated at that point; it does not differentiate through the training map or the conditional law of $S \mid u$. For a shuffled next minibatch, its conditional mean is over the remaining pool; the full-pool field is the uniform-pool average at a frozen iterate.

Proposition 3, population flow For unit $a \in \mathcal{H}$, set $q(a) = \mathbb{E}[(-a^\top \phi)_+]$. The margin assumption makes $q(a) > 0$; q is continuous, so compactness of the unit sphere gives $c = \min q > 0$. Since $\ell(z) \ge (-z)_+$, $F(u) \ge \beta c \|u\|$, proving coercivity. Its Hessian is

$$\nabla^2 F(u) = \beta^2 \mathbb{E}[\sigma(\beta\phi^\top u)\sigma(-\beta\phi^\top u)\phi\phi^\top],$$

positive definite on $\mathcal{H}$, so F is strictly convex and has a unique finite minimizer u_*. Along $\dot{u} = -M\nabla F(u)$, $V = \frac{1}{2}\|u - u_*\|^2_{M^{-1}}$ satisfies $\dot{V} = -(u - u_*)^\top \nabla F(u) \le -[F(u) - F(u_*)]$. Risk decreases, and integration gives $F(u(t)) \to F(u_*)$; coercivity bounds the path and uniqueness forces $u(t) \to u_*$. If $\mathcal{H} = \{0\}$, the result is immediate.

Proposition 3, separable pool Whiten by $w = \beta M^{-1/2} u$, $a_i = M^{1/2}\phi_i$, and $s = \beta^2 \tau$. Then $L(w) = n^{-1} \sum_i \ell(a_i^\top w)$ obeys $w' = -\nabla L(w)$ in s. Let h minimize $\|h\|^2/2$ subject to $a_i^\top h \ge 1$. Strict separability makes this feasible; KKT gives $h = \sum_i \lambda_i a_i$, with $\lambda_i \ge 0$ and $\lambda_i > 0$ only at active constraints. For $e = w - h \log s$, convexity gives

$$\left(\tfrac{1}{2}\|e\|^2\right)' \le L(h \log s) - L(w) - \frac{1}{s} \sum_i \lambda_i (a_i^\top w - \log s).$$

Here $L(h \log s) \le 1/s$. Let $I = \{i : \lambda_i > 0\}$. For $s > n \max_{i \in I} \lambda_i$, each $i \in I$ has $p = n\lambda_i / s \in (0, 1)$. Since $\sup_q[-\ell(q) - pq] = p \log p + (1 - p)\log(1 - p) \le p \log p$,

$$-\frac{\ell(a_i^\top w)}{n} - \frac{\lambda_i}{s}(a_i^\top w - \log s) \le \frac{\lambda_i}{s} \log(n\lambda_i).$$

Applying this bound only for $i \in I$ and dropping the remaining negative loss terms bounds the derivative by C/s, $C = 1 + \sum_{i \in I} \lambda_i |\log(n\lambda_i)| > 0$. Integration yields $e = O(\sqrt{\log s}) = o(\log s)$, hence $w = h \log s + o(\log s)$ and $u = \beta^{-1} M^{1/2} h \log \tau + o(\log \tau)$. The vector $\hat{v} = M^{1/2} h$ is exactly the stated M^{-1}-norm margin minimizer.

The discrete-time analogues are known. Soudry et al. (2018, Theorem 3) prove convergence to the Euclidean hard-margin direction for full-batch Euclidean gradient descent on finite strictly separable logistic data, with finite initialization and fixed $0 < \eta < 8n/\|A\|_2^2$ for averaged loss (rows of A are $a_i^\top$). Zhang, Zou, and Cao (2024, Theorem 4.5, Corollary 4.6, Lemma C.1) prove for full-batch strictly separable linear logistic regression that uncorrected Adam (no bias correction), with no ϵ or decay, zero initial moments, $0 \le b_1 \le b_2 < 1$, $b_2 > 0$, every initial gradient coordinate nonzero, and $\eta_t = (t+2)^{-a}$, $0 < a \le 1$, makes the normalized margin $\min_i a_i^\top w / \|w\|_\infty$ tend to its optimum; direction convergence additionally requires a unique ℓ_∞-unit-ball margin maximizer.

Proposition 4 Let Φ_S have rows $\phi_i^\top$. Set $A = \Phi_S M^{1/2}$, $c = (2\beta)^{-1}\mathbf{1}$, and $z = M^{-1/2} u$. Then $z' = -(2/n) A^\top (Az - c)$, so

$$z(\tau) = A^+ c + \exp(-2A^\top A \tau / n)(z_0 - A^+ c) \longrightarrow A^+ c + \Pi_{\ker A} z_0.$$

The limit is finite and minimizes $\|Az - c\|^2$, including when A is rank deficient; its kernel component is unchanged.

Scale identities for β In the model, β changes only the scale of u and the learning-rate clock. Because $F_\beta(u) = F_1(\beta u)$, fixed-$M$ flow $\dot{u} = -a(t) M \nabla F_\beta(u)$, $a(t) \ge 0$, becomes $\dot{v} = -\beta^2 a(t) M \nabla F_1(v)$ for $v = \beta u$. Thus $u_\beta(t) = \beta^{-1} v_1(\beta^2 T(t))$, $T(t) = \int_0^t a(r)\, dr$, with $v_1(0) = \beta u_\beta(0)$. For a memoryless degree-zero flow $\dot{u} = -a(t) Q(\nabla F_\beta(u))$, where $Q(cg) = Q(g)$ for $c > 0$, the same chain rule gives clock $\beta T(t)$; v_1 then follows the $\beta = 1$ Q-flow. For unclipped, zero-decay Adam on identical batches with zero initial moments and denominator $\sqrt{q} + \epsilon$, $g_\beta = \beta g_1$ at matched $v = \beta u$, so $m_\beta = \beta m_1$ and $q_\beta = \beta^2 q_1$. The transformed path $v_k = \beta u_k$ is exactly the $\beta = 1$ run with $\eta_k' = \beta \eta_k$, $\epsilon' = \epsilon/\beta$, unchanged b_1, b_2, and $v_0 = \beta u_0$; bias corrections preserve these scalings.