DPO'S RECOVERY ADVANTAGE DEPENDS ON ANSWER PRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct preference optimization (DPO) can restore answers lost during supervised fine-tuning (SFT), but option order affects which failures are selected and how recovery is measured. We test whether DPO's advantage over supervised continuation persists under a different option order. On MMLU-Pro, we select questions that Base answers correctly and SFT answers incorrectly, then evaluate unchanged SFT and both continuations on the same questions under both orders. Reordering alone lets unchanged SFT answer 29.6–34.1% of the selected failures correctly. On questions selected under the original order, DPO's advantage falls from 8.72 to 2.58 percentage points for Qwen3-8B and from 9.53 to 1.46 for OLMo-2-7B. Gains in its relative advantage on the remaining questions conceal this decrease in the full-test comparison. Reversing the screening and evaluation orders gives the same pattern. Within Qwen3-8B's recovery pools, DPO also changes more SFT answers, but a smaller fraction of those changes is correct. Recovery evaluation should therefore retest fixed questions across presentations and report how often each continuation changes an answer, alongside how often those changes succeed.

1 INTRODUCTION

A model can answer a question correctly, fail after supervised fine-tuning, and succeed again after further training. This sequence motivates accounts of recovered capability (Jin et al., 2026). A larger recovery count, however, can reflect more answer changes rather than more successful changes. We ask whether DPO's advantage over supervised continuation persists when evaluation no longer matches the presentation used to identify SFT failures.

Option order and answer-selection rules affect benchmark rankings (Alzahrani et al., 2024). In recovery evaluation, presentation also determines which questions qualify as Base-correct/SFT-wrong. We call this selection step screening. After screening under one order, we evaluate unchanged SFT, DPO and supervised continuation on the same questions under both orders. This separates failures that disappear through reordering alone from the continuations' relative advantage.

In OLMo-2-7B (OLMo Team et al., 2025) and Qwen3-8B (Yang et al., 2025a), DPO retains a smaller advantage outside the screening presentation. Unchanged SFT answers 29.6–34.1% of the selected failures correctly under the other order. Both continuations also gain correct answers, but the supervised control gains more. On original-screen questions, this contraction is offset by changes on the remaining questions and is concealed by the full-test comparison (Figure 2).

Moving content with its letters and reassigning letters at fixed positions yield different method gaps on MMLU-Pro and HellaSwag (Zellers et al., 2019), even when the displayed letter sequences match. Exchanging final-token attention parameters can shift this response with recipient attention aggregates held fixed; a separate OLMo-2-7B replay comparison tests the Original-to-Reordered contraction directly. We also distinguish the frequency of answer changes from their success rate and compare the continuations under a shared readout correction.

2 STUDY DESIGN

Training chains. For each family, B denotes Base, S its experimental SFT checkpoint, D the DPO continuation, and C a chosen-response supervised continuation. We use suffix 8 for Qwen3-8B and

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.

O for OLMo-2-7B. Both continuations start from the family-specific SFT checkpoint, which is also the DPO reference (Rafailov et al., 2023). D8R denotes the Qwen3-8B DPO run used in the main comparisons; D8 denotes an earlier run retained separately in Appendix K.

Qwen3-8B uses separately filtered SFT and preference subsets derived from UltraFeedback Binarized. C8 uses the chosen responses of the same 49,981 retained UltraFeedback Binarized preference pairs used for D8R (Cui et al., 2024); prompt tokens are masked from cross-entropy. Both continuations use 781 optimizer updates, learning rate 5×10^{-7}, global batch size 64, maximum sequence length 2,048, cosine scheduling, and seed 42. Warmup is 3% for D8R and 5% for C8. DPO uses $\beta = 0.1$. D8R uses matched sequence-scoring code for reference and policy likelihoods. The reference cache was produced locally and checked against the remote GPU training environment (Pod), before optimization (Appendix A). Source data, update count and peak learning rate are matched; DPO also processes rejected responses.

The independent OLMo-2-7B experiment follows BO→SO→ {DO,CO}. SO runs for 391 updates; DO and CO each run for 781 updates, using seed 42 and frozen training data. Its reference computation uses the same execution environment as the policy. The original recovery comparisons were specified after the earlier Qwen3-8B study and before OLMo-2-7B execution. The disagreement decomposition, scale analysis, and readout intervention were developed after those checkpoint outcomes were available. D8R was subsequently evaluated using the existing fixed-population calibration rule, without changing its coefficient or selecting a new population from D8R outcomes.

Evaluation. Both families are evaluated on all 12,032 MMLU-Pro test questions (Wang et al., 2024b) under the original order and a fixed permutation of the valid options. Outcomes are aligned by question identity and option content. Figure 1 shows how the same screened questions are evaluated under both orders. An answer is correct when its predicted option index equals the gold index, including index zero. Both families use the space-prefixed letter channel, two demonstrations per category, and a prompt ending at `Answer:`. Predictions maximize over valid options; Appendix I describes normalization and precision checks. The two orders are paired measurements, not independent test samples.

In Qwen3-8B, Base uses the rescored local readout; S8, D8R, and C8 use Pod readouts. Local mirrors of D8R and C8 give identical predictions on all questions under both orders; Base's different shard count was not covered by this check. The earlier D8 chain's evaluation on all 1,172 ARC-Challenge test questions (Clark et al., 2018) is reported separately in Appendix K.

HellaSwag comparison. We evaluate on the HellaSwag validation split (Zellers et al., 2019) with zero-shot prompts containing the context and four candidate continuations labeled A–D. Predictions maximize the next-token score over space-prefixed letters `A` through `D`, each verified to be a single token. We retain all questions answered correctly by Base and incorrectly by SFT under Original, without further subsampling or selection on either continuation's outcomes. This yields 2,632 OLMo-2-7B questions and 2,810 Qwen3-8B questions, fixed across Original, Move, Reassign and Standard (Section 5); gold labels follow the correct candidate content. The reported DPO–supervised gaps describe these recovery populations, not full-split accuracy. The parameter-exchange and content-diagnostic procedures are specified in Section 6 and Appendix Q.

2.1 Recovered Answers and Net Improvement

For a fixed question and option order, let B_i, S_i, T_i indicate correctness at Base, SFT, and a final checkpoint $T \in \{D, C\}$. Define

$$R_T = \sum_i B_i(1 - S_i)T_i, \qquad G_T = \sum_i (1 - B_i)(1 - S_i)T_i, \tag{1}$$

$$L_T = \sum_i S_i(1 - T_i). \tag{2}$$

R_T counts answers that were correct at Base, became wrong at SFT, and are correct again at T. G_T counts answers wrong at both Base and SFT but correct at T; L_T counts SFT-correct answers lost at T. The net accuracy change is

$$\mathrm{Acc}_T - \mathrm{Acc}_S = \frac{R_T + G_T - L_T}{N}. \tag{3}$$

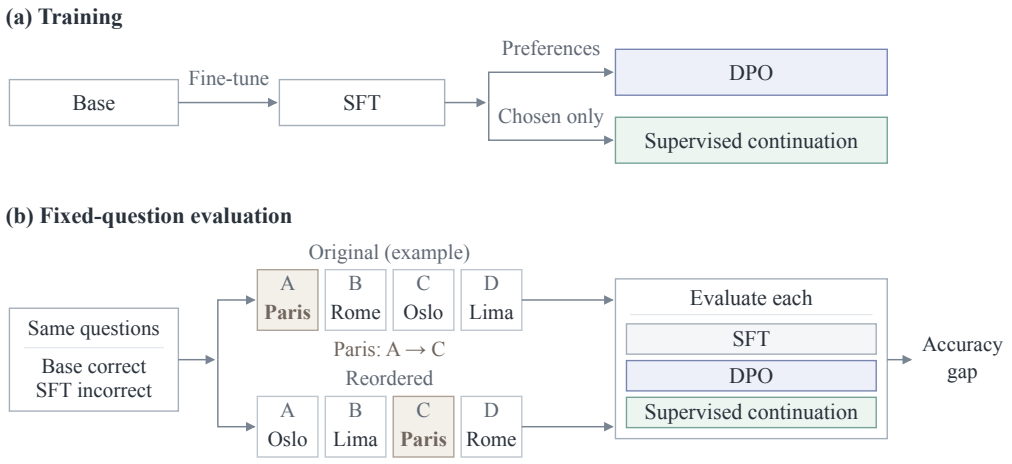


Figure 1: Fixed-question comparison for Qwen3-8B and OLMo-2-7B. The continuations share an SFT initialization. Questions selected under one option order are evaluated under both, with model weights unchanged. In this illustrative example, Paris moves from A to C. The accuracy gap compares DPO with supervised continuation; retesting SFT measures gains from reordering alone.

The restoration pool is $\mathcal{P}_R = \{i : B_i = 1, S_i = 0\}$ and the new-correction pool is $\mathcal{P}_G = \{i : B_i = 0, S_i = 0\}$. We report restoration rate $R_T/|\mathcal{P}_R|$ and restoration share $R_T/(R_T + G_T)$. Restoration share is the fraction of newly correct answers that Base had answered correctly; it does not subtract losses.

Both objectives restore earlier answers. OLMo-2-7B's accuracy falls from .3089 at BO to .2797 at SO, then rises to .2951 at DO and .2910 at CO under the original order. Under reordering the corresponding values are .3079, .2793, .2972, and .2930. DO restores 301/273 answers under original/reordered inputs, compared with 216/219 for CO. Qwen3-8B D8R restores 636/625, compared with 504/485 for C8. Supervised continuation restores 71.8–80.2% as many answers as DPO across these comparisons, and most of the Qwen3-8B answers it restores are also restored by DPO.

DPO also returns more often to Base's erroneous answers. When Qwen3-8B Base and SFT choose different wrong options, D8R returns to Base on 573 of 1,816 original-order questions, versus 436 for C8. D8R changes more SFT answers, but a smaller share of those changes returns to Base: 65.8%, versus 68.0% for C8. Reordering gives the same pattern (Appendix B). Return counts alone do not show whether a continuation preferentially returns to Base or simply changes more answers.

3 Recovery beyond the Screening Presentation

To separate screening from evaluation, we select recovery candidates under one presentation and retest the same questions under another. Let $Y_T(i, e)$ denote correctness of checkpoint T under evaluation order e. Screening order s selects $\mathcal{H}_s = \{i : Y_B(i, s) = 1, Y_S(i, s) = 0\}$. On this fixed set, define

$$A_T(s, e) = \frac{1}{|\mathcal{H}_s|} \sum_{i \in \mathcal{H}_s} Y_T(i, e), \qquad \tau_T(s, e) = A_T(s, e) - A_S(s, e). \tag{4}$$

Under the screening order, $A_S(s, s) = 0$, so a continuation's accuracy equals its net gain over SFT. Under the other order, unchanged SFT can answer some of these questions correctly; τ_T subtracts this retest accuracy. Subtracting this common SFT baseline leaves the method comparison unchanged because $\tau_D - \tau_C = A_D - A_C$. A decrease in DPO's advantage therefore reflects the two continuations' different responses to reordering, not the common SFT subtraction.

On the 1,514 original-order candidates, unchanged S8 answers 516 questions correctly under reordering (34.08%; Table 1). D8R answers 690, exceeding unchanged S8 by 174 questions under

Table 1: Screening and evaluation order crossed on fixed questions. Entries S, D and C count correct answers from SFT, DPO and supervised continuation, respectively; Qwen3-8B's D is D8R. The gap is $100(D - C)/n$. Each pair of rows uses the same questions, selected using only its named screening order. These populations differ from the two-order intersections used later.

Family	Screen	Evaluate	n	S	D	C	Gap (pp)
Qwen3-8B	Original	Original	1514	0	636	504	8.72
	Original	Reordered	1514	516	690	651	2.58
	Reordered	Reordered	1496	0	625	485	9.36
	Reordered	Original	1496	497	718	658	4.01
OLMo-2-7B	Original	Original	892	0	301	216	9.53
	Original	Reordered	892	264	316	303	1.46
	Reordered	Reordered	862	0	273	219	6.26
	Reordered	Original	862	258	307	293	1.62

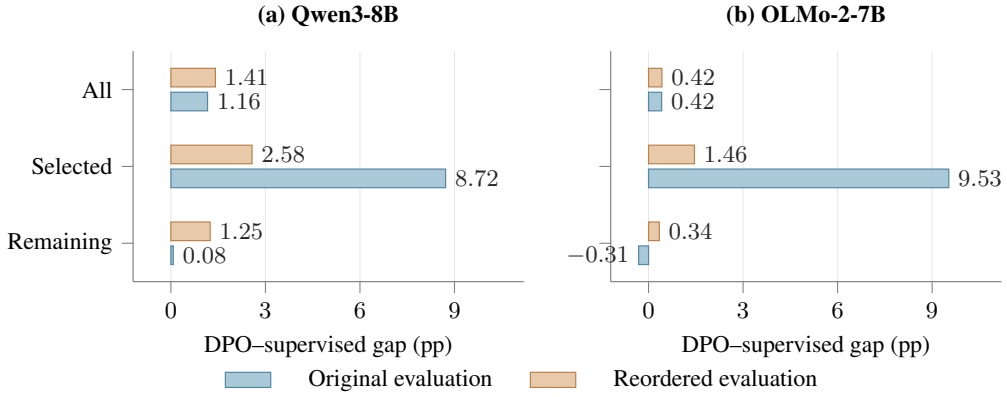


Figure 2: Reordering has different effects on the full test set and on recovery candidates. Selected means Base-correct/SFT-wrong under the original presentation; Remaining is its fixed complement. Both families use all 12,032 questions. Selected/Remaining sizes are 1,514/10,518 for Qwen3-8B and 892/11,140 for OLMo-2-7B. Each pair of bars compares the same questions and checkpoints under the two evaluations. Complement values are derived by subtracting the selected correct counts from the full-test counts. Values are descriptive; the two OLMo-2-7B full-test values round to 0.42, differing by one question in their correct-count gaps.

that order: it corrects 201 of S8's errors but loses 27 of S8's correct answers. C8 exceeds S8 by 135 questions, correcting 150 errors and losing 15 correct answers. Reversing screening and evaluation gives 497/1,496 unchanged-S8 successes (33.22%). Relative to S8 under this evaluation order, D8R gains 221 correct answers (238 corrections, 17 losses) and C8 gains 161 (167 corrections, six losses).

The DPO–supervised gap falls from 8.72 to 2.58 points on the original-order candidates: D8R gains 54 correct answers (636 to 690), while C8 gains 147 (504 to 651). On reordered candidates it falls from 9.36 to 4.01. DPO retains the larger correct count.

Why the full-test comparison differs. Across all 12,032 Qwen3-8B questions, D8R and C8 answer 5,329 and 5,189 correctly in the original order, versus 5,226 and 5,056 under reordering. D8R's lead in correct answers rises from 140 to 170 (1.16 to 1.41 points), although both models' accuracies fall. Within the 1,514 original-screen candidates, the difference instead falls from 132 to 39. On the remaining 10,518 questions, D8R's lead is therefore 8 and 131 (0.08 to 1.25 points). Its lead decreases by 93 within the selected group and increases by 123 outside it, giving a full-test increase of 30. These are changes in correct-count differences, not counts of distinct changed answers.

OLMo-2-7B shows the same selected-population pattern (Table 1). Its gap falls from 9.53 to 1.46 points on original-order candidates and from 6.26 to 1.62 on reordered candidates. Under the other evaluation order, DO and CO exceed unchanged SO by 52 and 39 correct answers on original-order

candidates, and by 49 and 35 on reordered candidates. Switching evaluation away from the screening order adds 15 and 87 correct answers for DO and CO on original-order candidates, and 34 and 74 on reordered candidates. DPO's lead falls by 72 and 40 questions; both attenuation intervals exclude zero (Appendix N).

For OLMo-2-7B's original-screen partition, the full-test count difference is 50 before reordering and 51 after it (Table 6). Subtracting the selected differences of 85 and 13 gives -35 and 38 on the remaining 11,140 questions (-0.31 and 0.34 points). The selected-group change of -72 and remaining-group change of $+73$ nearly cancel (Figure 2). DPO leads on the recovery candidates under both evaluations, while the point-estimate ranking outside that pool changes.

Across both families, DPO's advantage is larger when the evaluation order matches the screening order. A baseline using other Qwen3-8B categories' Base/SFT correctness states and continuation success rates approximates the overall decrease, but predicts variation across categories poorly (Appendix P). An accurate overall prediction alone does not explain the continuations' different responses.

Section 7 compares readout corrections on the intersection of the two screened populations, where Base is correct and SFT is wrong under both orders.

4 Recovery and Changes to SFT Answers

On the original-order recovery pool, Qwen3-8B D8R changes 774 of S8's answers and C8 changes 604, recovering 636 and 504. Among their changed answers, the correct fractions are 82.2% and 83.4%. Under reordering they are 77.8% and 78.9% (Table 8). DPO changes more answers, but a slightly smaller fraction of its changes succeeds. The fractions use each model's own set of changed answers.

We also evaluate D8R after 312 updates, denoted D8R-312, on all 12,032 questions and the same fixed screening pools as Table 1. Table 2 compares this checkpoint with the continuations' endpoints. A change is a prediction different from S8 under the same presentation, including wrong-to-wrong changes. Its validation record contains 16 changes among 84 question–order outcomes (42 questions under two orders). Appendix S.1 gives the available selection record and the limits of this exploratory comparison.

Table 2: Trained Qwen3-8B checkpoints under the uncorrected letter readout. C8 is supervised continuation; D8R is DPO, and D8R-312 is its intermediate checkpoint. Changed and All correct use all 12,032 questions. The last two columns use fixed pools screened under Original ($\mathcal{H}_o$, 1,514 questions) or Reordered ($\mathcal{H}_p$, 1,496 questions). Every count uses the evaluation order named in its row; only pool membership stays fixed. Pool membership does not depend on D8R-312. Endpoints use 781 updates.

Checkpoint	Evaluate	Changed	All correct	$\mathcal{H}_o$ correct	$\mathcal{H}_p$ correct
C8	Original	1585	5189	504	658
	Reordered	1649	5056	651	485
D8R	Original	2398	5329	636	718
	Reordered	2516	5226	690	625
D8R-312	Original	1598	5182	508	654
	Reordered	1637	5061	648	488

D8R-312 changes 13 more answers than C8 under Original and 12 fewer under Reordered, differences of $+0.108$ and -0.100 percentage points. Their full-test correct counts differ by -7 and $+5$ (-0.058 and $+0.042$ points). On the Original and Reordered screening pools, the same-presentation DPO advantages are now 0.26 and 0.20 points, rather than 8.72 and 9.36 at the endpoint. Evaluating those fixed pools under the other presentation gives -0.20 and -0.27 points. The same-minus-other contrasts are therefore 0.46 and 0.47 points, versus 6.14 and 5.35 at the endpoint.

At this intermediate checkpoint, the recovery advantage is much smaller and full-test accuracy is close to C8's. Training duration and changed-question identities differ, and pool-specific change counts

for D8R-312 are unavailable; the contribution of change frequency remains unresolved. Appendix S examines score scaling and transfer between presentations.

5 Matching the Displayed Letter Sequence

Standard reordering changes both where each option appears and which letter names it. Let P denote content position and L the content-to-letter map. Original and standard-reordered inputs are P_0L_0 and P_1L_1; moving content with its letters (Move) gives P_1L_0, and reassigning letters at fixed positions (Reassign) gives P_0L_1. With a general permutation, Move and Reassign can display different letter sequences. A difference between their DPO–supervised gaps could therefore reflect the letter sequence as well as the content-to-letter map. Appendix C reports the initial general-permutation crossing, including a 2.81-point interaction on Qwen3-8B's reordered-screen population.

We use a permutation q that swaps adjacent option pairs, $(1, 2), (3, 4), \ldots$, leaving the last option fixed when the number of options is odd. Applying it twice restores the original positions, so $q = q^{-1}$. Moving content with its letters displays $q[k]$ at position k; reassigning letters displays $q^{-1}[k]$. The sequences therefore match. For example, start with A:u, B:v, C:w, D:x and exchange the first two options and the last two options. Move gives B:v, A:u, D:x, C:w; Reassign gives B:u, A:v, D:w, C:x. Both display B,A,D,C, whereas Original and Standard display A,B,C,D. Table 12 gives the general mapping. This paired-swap permutation differs from the one used in the initial comparison.

In the following comparison, P_1, L_1 use these paired swaps. On a fixed original-screen population, let

$$\kappa = g(P_1, L_0) - g(P_0, L_1), \tag{5}$$

where g is the DPO minus supervised-continuation correctness gap in percentage points. Both g and κ are computed separately for each stated population. Questions and checkpoints are fixed within each comparison. Thus κ measures how the method comparison changes between inputs sharing the displayed letter sequence but differing in content positions and content-to-letter maps.

Table 3: DPO–supervised advantage under paired swaps, in percentage points. Each task and family uses the same fixed original-screen questions; Qwen3-8B DPO is D8R. Move retains the content–letter assignment, Reassign fixes content positions, and Standard changes both. Standard uses a different permutation from Table 13; its values are not interchangeable. Entries are point estimates.

Task	Family	n	Original	Move	Reassign	Standard	κ
MMLU-Pro	OLMo-2-7B	892	9.53	6.73	2.69	1.91	4.04
	Qwen3-8B	1514	8.72	6.80	2.97	2.25	3.83
HellaSwag	OLMo-2-7B	2632	8.17	6.46	2.47	1.52	3.99
	Qwen3-8B	2810	8.90	7.12	2.49	1.42	4.63

On MMLU-Pro, the matched-display contrasts are 4.04 points for OLMo-2-7B and 3.83 for Qwen3-8B. Move and Reassign give D-minus-C correct-count differences of 60 versus 24 in OLMo-2-7B and 103 versus 45 in Qwen3-8B. The contrasts are therefore 36/892 and 58/1514. DPO remains ahead in all four presentations (Figure 5).

On the fixed HellaSwag populations described in Section 2, the matched-display contrasts are 3.99 points for OLMo-2-7B and 4.63 for Qwen3-8B. The contrast therefore extends to commonsense continuation selection under these four presentations.

Because Move and Reassign display the same letters in the same sequence, that sequence alone cannot explain their different method gaps. Content positions and content-to-letter assignments still differ, and may interact.

6 Parameter Exchange at the Final Input Token

We test whether exchanging parameters at the final input token shifts the recipient's Move-minus-Reassign response toward the donor's, and whether that shift persists with attention aggregates held fixed. This outcome differs from the Original-to-Reordered attenuation in Table 1. A separate label-free diagnostic measures changes in selected option content.

Table 4: Bidirectional parameter exchange on fixed recovery questions: 80% of each MMLU-Pro pool and the full HellaSwag pools. Directions are recipient ← donor; D denotes DPO and C supervised continuation. t is the decrease in donor-response distance, $\bar{t}_0$ the five-control mean, and Net their difference. L_{95} is Net's one-sided 95% paired lower bound; these columns use percentage points. *Changed* gives changed/total questions, and U_{95} the one-sided 95% binomial upper bound on the label-free change rate (%).

Task	Family	Direction	t	$\bar{t}_0$	Net	L_{95}	Changed	U_{95}
MMLU-Pro	OLMo-2-7B	$C \leftarrow D$	2.521	.560	1.96	.42	2/714	.879
		$D \leftarrow C$	3.081	.560	2.52	.88	1/714	.663
	Qwen3-8B	$C \leftarrow D$	2.395	.578	1.82	.31	4/1211	.754
		$D \leftarrow C$	2.642	.330	2.31	.62	6/1211	.976
HellaSwag	OLMo-2-7B	$C \leftarrow D$	2.242	.532	1.71	.28	11/2632	.691
		$D \leftarrow C$	3.078	.494	2.58	.95	16/2632	.922
	Qwen3-8B	$C \leftarrow D$	2.420	.676	1.74	.24	13/2810	.735
		$D \leftarrow C$	3.665	.640	3.02	1.12	18/2810	.948

Intervention and comparison. Within each family, we exchange DO with CO or D8R with C8 in both directions. The donor's attention query (W_Q) and attention output (W_O) projections replace the recipient's at the final non-padding input token in every attention layer. The recipient retains W_K, W_V, normalization, MLP and vocabulary-head parameters. Its prefix key/value cache is reused without recomputation.

Let r_X be model X's Move-minus-Reassign accuracy difference in percentage points. On a common population, $\kappa = r_D - r_C$ equals the contrast in Equation 5. We recompute it on the intervention populations specified below. For the two directions, define

$$t_{C \leftarrow D} = |\kappa| - |r_{C \leftarrow D} - r_D|, \tag{6}$$

$$t_{D \leftarrow C} = |\kappa| - |r_{D \leftarrow C} - r_C|. \tag{7}$$

A positive t indicates movement toward the donor's presentation response. Five controls scramble the signs of the donor-minus-recipient parameter differences at the same sites while preserving each tensor's perturbation norm. Net transfer subtracts their mean effect from t, separately for each direction. Bounds use 5,000 question-paired bootstrap resamples with seed 42 (Appendix Q).

Populations and content diagnostic. The intervention uses 714 OLMo-2-7B and 1,211 Qwen3-8B recovery questions on MMLU-Pro, and 2,632 and 2,810 on HellaSwag. For MMLU-Pro, we take 80% of each original-order Base-correct/SFT-wrong pool (892 and 1,514 questions), rounded to the nearest integer, using a fixed ordering with seed 42. Selection precedes inspection of exchange outcomes and does not use DPO advantages. All eight content comparisons cover exactly these fixed questions. The diagnostic removes the target options' letter labels and selects the candidate text with highest mean token log likelihood (Zheng et al., 2024), using the intervened output and cache rather than decoding a predicted letter. A question counts as changed if its selected content differs from the unmodified recipient's under either Move or Reassign, including wrong-to-wrong changes. Appendix Q specifies sampling and scoring.

Results. The distance to the donor's response decreases by 2.242–2.521 points for supervised recipients and 2.642–3.665 for DPO recipients. Each decrease exceeds all five matched-control point estimates (Appendix Q); subtracting their mean gives net transfers of 1.71–1.96 and 2.31–3.02 points, respectively.

The one-sided 95% paired lower bounds for Net range from 0.24 to 1.12 points, supporting transfer beyond the matched-control mean in each comparison. Label-free content choices change on 0.140–0.641% of the same questions, with one-sided exact binomial upper bounds of 0.663–0.976% under independent Bernoulli sampling. Smaller, prespecified development subsets gave weaker transfer estimates; reevaluating those questions with the main experiment's procedure gives the same predictions and Net values (Appendix Q).

Holding attention aggregates fixed. To test whether the response shift requires different attention aggregates, we use the same MMLU-Pro questions. For each question, we record the input to each

layer's attention output projection before and during Q/O exchange. We then cross the aggregate source with the output-projection parameters, retaining the recipient's prefix cache. With recipient aggregates fixed, the directed, control-adjusted effects for $C \leftarrow D$ and $D \leftarrow C$ are 1.989/2.297 points in OLMo-2-7B and 1.718/2.263 in Qwen3-8B. Full-exchange effects are 1.961/2.521 and 1.784/2.312, respectively. Replaying exchanged aggregates with recipient output projections yields effects of .050–.112 points. Appendix R defines this signed comparison, distinct from the distance-based Net in Table 4, and reports all four conditions.

Interpretation. Unadjusted full-minus-fixed response differences are about .14 points or less in absolute value, with all four paired 95% intervals contained within approximately $\pm.42$ points (Appendix R). The response shift persists at the group level when recipient aggregates are fixed. This identifies an effect under replay, but does not separate semantic processing from letter conversion: replay alters the computation, and the content diagnostic uses a different prompt.

We test the Original-to-Reordered attenuation directly by replaying the OLMo-2-7B supervised recipient on 714 fixed questions. Keeping recipient aggregates fixed gives a larger control-adjusted attenuation reduction than aggregate replacement: the difference is 2.44 points, with a paired 95% interval of $[1.51, 3.36]$ (Appendix R.1).

7 Comparing Models under a Shared Readout

To test sensitivity to the readout rule, we use the two-order recovery intersections (455 Qwen3-8B and 172 OLMo-2-7B questions), fitting marginal-letter and content-aligned additive corrections on 42 validation questions absent from their own demonstrations. Each fitted vector is applied to both continuations with coefficient one. In Qwen3-8B, DPO's advantage remains positive under all four shared corrections: 7.80–10.66 points, versus 10.77 without correction. The OLMo-2-7B gaps are 4.36–8.72 points, versus 9.88 without correction; its validation–test runtime difference limits this comparison to a descriptive result. The corrections can include content effects. With all 70 validation questions, including demonstrated targets, additive shared-offset gaps for OLMo-2-7B are smaller (1.16–3.20 points); using each checkpoint's own offsets instead gives a tie. Appendix G reports this fitting-set sensitivity.

8 Related Work

Post-training recovery. Forgetting during sequential SFT and preference learning has been studied through objective trade-offs (Fernando et al., 2025), while recovery along training trajectories has been linked to singular-vector rotation (Jin et al., 2026). CapTrack broadens evaluation beyond accuracy to competence, behavior and execution, finding partial recovery after DPO and testing robustness to prompt reformulation (Thede et al., 2026). We examine a selection effect within recovery evaluation: the full-test and selected-recovery gaps can respond to the same reordering in opposite directions.

The supervised control also affects interpretation. On rectified mathematical responses, SPoT's chosen-only Reward-SFT retains instruction-following performance comparably to DPO, with a gradient analysis linking this behavior to adaptive reference-relative weighting (Lin & Han, 2026). Our ordinary cross-entropy control compares DPO with supervised continuation, but does not isolate negative-response gradients from adaptive weighting.

Presentation and readout. Option order alone can change multiple-choice predictions (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024). Prompt-format studies and MCQFormatBench test additional changes to question structure and answer presentation (Sclar et al., 2024; Takizawa et al., 2025). Even small changes in option order or answer selection can alter model rankings (Alzahrani et al., 2024). Evaluating generated text can be more robust than first-token selection (Wang et al., 2024a); answer matching instead grades free-form responses against reference answers (Chandak et al., 2025). PA-GRPO addresses permutation consistency during training (Zheng et al., 2026). Our presentation comparisons hold trained checkpoints fixed and cross the presentation used to identify a failure with the presentation used to evaluate its recovery.

Symbol binding and intervention. Associating option content with its symbol is a learned component of multiple-choice answering: Xue et al. (2024) improve this association during SFT, and Wong et al. (2026) find evidence that content selection precedes symbol binding. Activation and attention-head interventions provide further evidence about answer-symbol prediction (Wiegreffe et al., 2025; Yang et al., 2025c). These findings motivate measuring letter responses alongside a separate content-choice diagnostic. Interventions on model activations have also been used to study preference learning: Yang et al. (2025b) analyze distributed activation changes in DPO toxicity reduction and reproduce aspects of the effect through activation edits. Our bidirectional Q/O exchange instead tests transfer of the presentation-response difference between DPO and supervised continuation, against norm-matched parameter controls.

9 Discussion and Conclusion

DPO retains a recovery advantage under the alternative presentation, but that advantage is smaller than under the order used to select failures. Changes on the remaining questions can conceal this decrease in the full-test comparison. In Qwen3-8B, DPO restores more answers and changes more of SFT's predictions, while a smaller fraction of its changes is correct. At the exploratory step-312 checkpoint, the same-presentation advantages are 0.26 and 0.20 points. These results support retesting fixed questions across presentations and reporting both the frequency and success rate of answer changes.

The matched-display contrast extends to HellaSwag, and parameter exchange can shift this response with recipient aggregates fixed. In the OLMo-2-7B supervised recipient, fixed-aggregate replay also reduces the Original-to-Reordered attenuation relative to matched controls (Appendix R.1). These interventions locate a conditional effect at the attention output projections; how training produces it remains open.

Limitations. The screening-by-evaluation comparison and several follow-up analyses were retrospective. Results cover two 7B–8B families with seed 42 and letter-based multiple-choice readouts, not variation across training runs or data compositions. The screening-by-evaluation comparison is limited to MMLU-Pro; checkpoint and score-scaling analyses are limited to Qwen3-8B. The intermediate checkpoint neither isolates training duration from changed-question identity nor establishes statistical equivalence. The tested transformations do not establish invariance to arbitrary presentations.

The prespecified development subsets (300 Qwen3-8B and 163 OLMo-2-7B questions) are contained in the main intervention populations of 1,211 and 714; the main evaluation is not an independent replication. The development–main decomposition in Appendix Q describes point estimates, not a test of their difference.

Correction fits can retain content effects, and OLMo-2-7B validation and test scoring use different runtimes. Earlier D8/ARC results remain separate from D8R, which differs from D8 in more than reference-score consistency (Appendix A).

Reproducibility Statement

Sections 2.1–7 define the comparisons and measured outcomes. Appendices A–S give training settings, counts, readout checks, derivations, parameter-exchange controls, replay conditions and exploratory score analyses. Appendix S.1 documents the available step-312 selection record.

The Use of Large Language Models (LLMs)

LLMs assisted with experiment planning, implementation and analysis, verification code, and manuscript editing (including grammar and wording). The authors are responsible for verifying all content.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. *arXiv preprint arXiv:2507.02856, version 1*, 2025. URL https://arxiv.org/abs/2507.02856v1.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10818–10838. PMLR, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392. Association for Computational Linguistics, July 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14014–14031. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.783. URL https://aclanthology.org/2023.acl-long.783/.

Heshan Fernando, Han Shen, Parikshit Ram, Yi Zhou, Horst Samulowitz, Nathalie Baracaldo, and Tianyi Chen. Understanding forgetting in LLM supervised fine-tuning and preference learning—a convex optimization perspective. *arXiv preprint arXiv:2410.15483, version 4*, 2025. URL https://arxiv.org/abs/2410.15483v4.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

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 *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189. Association for Computational Linguistics, November 2024a. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. Dissecting fine-tuning unlearning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3933–3941. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.228. URL https://aclanthology.org/2024.emnlp-main.228/.

Jennifer Hu and Michael C. Frank. Auxiliary task demands mask the capabilities of smaller language models. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=U5BUzSn4tD.

Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. RL fine-tuning heals OOD forgetting in SFT. *arXiv preprint arXiv:2509.12235, version 3*, 2026. URL https://arxiv.org/abs/2509.12235v3.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395. Association for Computational Linguistics, July 2004. URL https://aclanthology.org/W04-3250/.

Jiatong Li, Renjun Hu, Kunzhe Huang, Yan Zhuang, Qi Liu, Mengxiao Zhu, Xing Shi, and Wei Lin. PertEval: Unveiling real knowledge capacity of LLMs with knowledge-invariant perturbations. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-0341. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/149578235c90954e4f10e40fa181918f-Abstract-Datasets_and_Benchmarks_Track.html.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2819–2834. ELRA and ICCL, May 2024b. URL https://aclanthology.org/2024.lrec-main.251/.

Wenye Lin and Kai Han. Surgical post-training: Proximal on-policy distillation for reasoning with knowledge retention. *arXiv preprint arXiv:2603.01683, version 2*, 2026. URL https://arxiv.org/abs/2603.01683v2.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information*

Processing Systems, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3946. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tür, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4399. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bf235a1d6780afd979f2f81676f43413-Abstract-Conference.html.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 furious (COLM's version). In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.12353v3.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324v2.

Hiroo Takizawa, Saku Sugawara, and Akiko Aizawa. MCQFormatBench: Robustness tests for multiple-choice questions. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 824–846. Association for Computational Linguistics, July 2025. URL https://aclanthology.org/2025.gem-1.69/.

Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, and Yun-Nung Chen. None of the above, less of the right parallel patterns in human and LLM performance on multi-choice questions answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20112–20134. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.findings-acl.1031. URL https://aclanthology.org/2025.findings-acl.1031/.

Lukas Thede, Stefan Winzeck, Zeynep Akata, and Jonathan Richard Schwarz. CapTrack: Multifaceted evaluation of forgetting in LLM post-training. *arXiv preprint arXiv:2603.06610, version 2*, 2026. URL https://arxiv.org/abs/2603.06610v2.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. LLMs may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5852–5862. Association for Computational Linguistics, January 2025. URL https://aclanthology.org/2025.coling-main.390/.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=qHdSA85GyZ.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html.

Hugh Mee Wong, Rick Nouwen, and Albert Gatt. When models decide and when they bind: A two-stage computation for multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22804–22821. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1144. URL https://aclanthology.org/2026.findings-acl.1144/.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4331–4344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.237. URL https://aclanthology.org/2024.acl-long.237/.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Yushi Yang, Filip Sondej, Harry Mayne, Andrew Lee, and Adam Mahdi. How does DPO reduce toxicity? a mechanistic neuron-level analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29524–29543, Suzhou, China, November 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1501. URL https://aclanthology.org/2025.emnlp-main.1501/.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1902–1917. Association for Computational Linguistics, April 2025c. doi: 10.18653/v1/2025.naacl-long.95. URL https://aclanthology.org/2025.naacl-long.95/.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9602d22a8c791f23f8e4d1398e3fb5be-Abstract-Conference.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03882.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. Mitigating selection bias in large language models via permutation-aware GRPO. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35125–35143. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1621. URL https://aclanthology.org/2026.acl-long.1621/.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/003e438cf04e3caf0a5c908495e484fe-Abstract-Conference.html.

A Training Records and Initialization Checks

Scope of the objective comparison. Staged alignment combines demonstration-based SFT with subsequent preference learning (Ouyang et al., 2022). Alternatives differ in their objectives and feedback: IPO uses an identity-based objective within a pairwise-preference framework (Gheshlaghi Azar et al., 2024); ORPO adds an odds-ratio preference term to supervised likelihood training (Hong et al., 2024a); SimPO uses average sequence log probability as a reference-free reward (Meng et al., 2024); and KTO learns from binary desirability feedback (Ethayarajh et al., 2024). Our chosen-response control instead uses ordinary cross-entropy from the same SFT initialization. The results compare this control with DPO, not these other algorithms.

Studies of post-training address different notions of retained capability. Chu et al. (2025) compare SFT and RL on changed rules and visual inputs, whereas our recovery populations are selected by earlier checkpoint failures. In deliberate unlearning, activation patching and parameter restoration can reveal altered retrieval rather than erased knowledge (Hong et al., 2024b). Sparse effective parameter updates have also been observed across several post-training methods, including DPO (Mukherjee et al., 2025). Neither finding identifies the training cause of the presentation dependence measured here.

Training data. Training data come from a local mirror of `HuggingFaceH4/ultrafeedback_binarized`, using its `train_sft` and `train_prefs` files. These two mirrored Parquet files are byte-identical; SFT uses chosen responses, while DPO retains preference pairs. Matching prompt IDs and normalized prompt text yields 61,124 aligned records from 61,135 candidates. The seed-42 split was frozen on 6 September 2026; the upstream repository commit was not retained. Before OLMo-specific filtering, the training subset contains 50,000 records. Training records with exact 8-gram overlap with MMLU-Pro test text were excluded.

Qwen3-8B initialization and continuation. S8 initializes from B8 and trains for 391 optimizer updates with seed 42 using four-GPU FSDP. Microbatch size 2 and eight accumulation steps give global batch size 64. S8 uses AdamW with learning rate 5×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.1, and cosine scheduling with 12 warmup updates (approximately 3%). Cross-entropy covers answer tokens only, excluding prompt and padding tokens. The locally versioned SFT subset, `OLMo-mix-sft-v1`, derives from `HuggingFaceH4/ultrafeedback_binarized` and retains 49,988 of 50,000 examples after removing 12 sequences exceeding 2,048 tokens. The run processes 25,024 example slots, rather than a full pass over the retained dataset. The ChatML template uses an empty system prompt, a user message followed by the assistant prefix, and an end-of-message marker after the answer. These initial-SFT settings are distinct from the continuation settings below.

C8 optimizes chosen-response token cross-entropy, masking the prompt. The recorded optimizer is AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, and maximum gradient norm 1.0. Microbatch size 2, accumulation 4, and world size 8 yield global batch 64. Reconstructing the sampler from its version and configuration gives $49,981 = 780 \times 64 + 61$, followed by three wrap-around duplicate slots in the final update: 49,981 distinct examples occupy 49,984 slots. Per-batch consumption logs were not retained. The dataset contains 13,680,444 loss tokens before these duplicates.

D8R initializes from S8 and uses an fp32 reference cache computed locally with batched, completion-masked sequence scoring. Before training on Pod, we compare the same checkpoint against the cache on 128 preference pairs. The 99th-percentile absolute chosen/rejected log-probability residuals are 3.052×10^{-4} and 3.967×10^{-4} nats. For the chosen-minus-rejected residual, the 99th percentile is 5.493×10^{-4} and the maximum 7.324×10^{-4} nats. The recorded tolerance is 10^{-3}, revised from an earlier 10^{-4} threshold. The check measures agreement between the two runtimes; the environments themselves differ.

The first pre-update microbatch log from each of eight workers reports two pairwise margins, with combined extrema -1.45×10^{-4} and 6.71×10^{-4} nats. This checks 16 pairs before any update. D8R completes 781 updates on eight GPUs; its first and last five-update mean losses are .6922 and .3215, and the mean of 156 logged windows is .3613.

Earlier D8 reference-score mismatch and logged loss. The first logged D8 loss is 11.1379 at update 5; it is the mean over the first five updates, not a loss measured before training. The following windows are 11.1251 and 11.1414. Across 156 logged windows, the mean is 6.5109 and recorded pre-clipping gradient norms range from 166.1 to 452.2. OLMo-2-7B uses the same five-update logging window, with first loss 0.6959 and subsequent windows 0.6948 and 0.6941. Its first recorded pre-clipping gradient norm is 21.75.

The earlier D8 reference log probabilities were cached under a different PyTorch/Transformers runtime from training. Its cross-runtime audit reports identical token sequences on all 250 checked examples but a mean absolute per-token log-probability discrepancy of 0.56 nats. These are legacy-run measurements, distinct from the D8R pre-update checks above. The chosen/rejected responses average approximately 274/240 tokens, so sequence-level scores can be sensitive to this mismatch. If cached scores differ from same-runtime scores by e_w, e_l, the DPO margin is shifted by $-\beta(e_w - e_l)$. Absolute per-token differences do not determine the signed pairwise offset, but show that the two scoring procedures are not interchangeable.

At identical policy and reference probabilities before any update, the standard DPO loss is $\log 2 \approx 0.6931$. A five-update mean cannot test that equality. D8R supplies separate pre-update residual measurements, while OLMo-2-7B computes the reference in the policy's environment. Differences between D8 and D8R also include precision, batched scoring, and distributed execution; their outcome difference is not a single-factor estimate of reference-cache error.

OLMo-2-7B execution. SO initializes from the released OLMo-2-7B Base model. DO and CO initialize from the same final SO checkpoint, verified by checkpoint hashes; DO also uses SO as its reference. All three stages use full-parameter updates with AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.01, gradient clipping at 1.0, cosine scheduling with 3% warmup, maximum sequence length 2,048, and seed 42. Each stage runs for one epoch. The official Instruct chat template is applied to the Base tokenizer, with eager attention. SO and CO compute cross-entropy only on answer tokens, including EOS and excluding prompt and padding tokens. DO uses $\beta = 0.1$ and sums completion-token log probabilities over the same answer span. Stage-specific quantities appear in Table 5. Evaluation uses fp32.

OLMo-2-7B test scores were produced locally under PyTorch 2.1.0+cu121; validation scores were produced on Pod under PyTorch 2.12.1+cu129 with SDPA and explicit fp32. Test summaries omit attention and dtype fields; fp32 is recorded in the frozen protocol and checkpoint configurations. This validation–test runtime difference is distinct from the same-environment DPO reference computation and is not covered by the Qwen3-8B mirror check. The OLMo-2-7B readout analysis recomputes eligibility from 12,032 Base/SFT integer records and uses the 172 eligible questions' DO/CO score vectors; its additive diagnostics cover that population, not the full OLMo-2-7B test set.

Table 5: OLMo-2-7B training configurations. DO and CO use the same retained preference pairs; CO trains only on their chosen responses. The SFT and continuation stages have different batch sizes and update counts.

	SO (SFT)	DO (DPO)	CO (supervised)
Learning rate	2×10^{-5}	5×10^{-7}	5×10^{-7}
Global batch size	128 examples	64 pairs	64 examples
Retained training records	49,988	49,976 pairs	49,976 chosen
Optimizer updates	391	781	781

SFT removes 12 overlength examples from the 50,000-record subset. OLMo-specific strict boundary screening excludes one preference pair whose rejected response fails the strict completion-boundary check under the training template, leaving 49,999 pairs; subsequent length filtering removes 23, leaving 49,976. The same retained pairs supply DO and CO. Their optimizer implementations reuse the Qwen3-8B training scripts with a model-specific chat-format adapter; this code reuse does not imply identical training configurations across families.

Development objective. Chosen-response development NLL for the earlier Qwen3-8B comparison is 0.8936 at S8, 0.9034 at C8, and 1.0129 at D8; no D8R NLL is inferred from these values. OLMo-2-7B's 700-example values are 1.374 at BO, 0.960 at SO, 1.083 at DO, and 0.976 at CO. Relative to SO,

the latter two values increase by approximately 12.8% and 1.7%. These objective-evaluation examples are distinct from the 70-question validation split used to estimate letter priors. The comparisons are within family; absolute NLLs across different tokenizers are not used to rank the families.

B Supplementary Recovery Comparisons

Table 6 gives full-test counts for the complete chains. OLMo-2-7B's DPO–supervised differences are $3551 - 3501 = 50$ under Original and $3576 - 3525 = 51$ under Reordered, the totals used in Figure 2.

Table 6: Full-test MMLU-Pro correctness, with 12,032 questions per order. Accuracy is the correct count divided by 12,032, rounded to four decimals. Base uses the fp32-rescored readout specified in Section 2.

		Original		Reordered	
Family	Checkpoint	Correct	Accuracy	Correct	Accuracy
Qwen3-8B	Base (B8)	5763	.4790	5673	.4715
	SFT (S8)	4688	.3896	4598	.3821
	DPO (D8R)	5329	.4429	5226	.4343
	Supervised (C8)	5189	.4313	5056	.4202
OLMo-2-7B	Base (BO)	3717	.3089	3705	.3079
	SFT (SO)	3365	.2797	3360	.2793
	DPO (DO)	3551	.2951	3576	.2972
	Supervised (CO)	3501	.2910	3525	.2930

We compare DPO and supervised continuation within each pool. For $\mathcal{P} \in \{\mathcal{P}_R, \mathcal{P}_G\}$, let n_D and n_C be the respective corrected counts and $n = |\mathcal{P}|$. Their marginal odds ratio is

$$\mathrm{OR}_{\mathcal{P}} = \frac{n_D/(n - n_D)}{n_C/(n - n_C)}. \tag{8}$$

Let $p_{T,R} = R_T/|\mathcal{P}_R|$ and $p_{T,G} = G_T/|\mathcal{P}_G|$. We distinguish the absolute rate advantage $\Delta_h = p_{D,h} - p_{C,h}$ from the relative odds advantage OR_h, for $h \in \{R, G\}$. The absolute advantage counts additional successes per eligible question; the relative advantage compares success odds. Whether DPO helps restoration more than new correction can depend on the chosen scale.

Equal odds multipliers need not imply equal absolute gains when control success rates differ (Appendix O). These comparisons describe the fixed checkpoints; objective-level attribution would require a design that isolates the relevant training difference.

Restoration counts and shares. Table 7 separates restored answers from final accuracy. Supervised continuation restores 71.8% and 80.2% of DPO's counts in OLMo-2-7B, and 79.2% and 77.6% in Qwen3-8B, under original and reordered inputs respectively. Figure 3 plots the corresponding rates.

The Qwen3-8B recovered sets share 447 questions under the original order and 437 under reordering. This leaves 189/188 unique to D8R and 57/48 unique to C8. Most of the answers restored by supervised continuation are also restored by DPO.

Supervised continuation has the larger restoration share in both families (Table 7). Thus the observation that most corrections restore earlier answers is shared by both objectives.

Returning to Base's answer versus changing SFT's answer. In the recovery pool, returning to Base's answer necessarily means becoming correct. We therefore examine a contrasting set: Qwen3-8B questions where Base and SFT choose different wrong options. This retrospective selection uses only Base and SFT predictions and the answer key, not either continuation. We count returns to the exact Base option separately under each order. Table 8 compares these questions with the recovery pool using the same fp32-rescored Base, S8, D8R and C8 outputs as the primary analysis.

D8R exceeds C8's wrong-answer return rate by 7.54 and 7.94 percentage points under original and reordered inputs. The paired counts are 196/59 D8R-only/C8-only returns under Original and 212/63

Table 7: MMLU-Pro, all 12,032 questions, fp32 readouts. Recovery pools contain 1,514/1,496 questions for Qwen3-8B and 892/862 for OLMo-2-7B under original/reordered inputs. R is a count; other columns are fractions. D8R and DO are DPO checkpoints; C8 and CO are supervised-continuation checkpoints.

Family	Order	Model	R	Restoration rate	Restoration share	Accuracy
Qwen3-8B	Original	D8R	636	.4201	.7518	.4429
		C8	504	.3329	.7950	.4313
	Reordered	D8R	625	.4178	.7396	.4343
		C8	485	.3242	.7990	.4202
OLMo-2-7B	Original	DO	301	.3374	.5637	.2951
		CO	216	.2422	.6207	.2910
	Reordered	DO	273	.3167	.5240	.2972
		CO	219	.2541	.6050	.2930

Table 8: Exact returns to Base's answer in Qwen3-8B. D8R denotes DPO and C8 supervised continuation. "Changed" counts predictions different from SFT; "Share" divides returns by changes within each continuation. Base is correct in the recovery pool and incorrect in the different-wrong pool. Counts are paired by question within each order.

			Returns		Changed		Share (%)	
Pool	Order	n	D8R	C8	D8R	C8	D8R	C8
Recovery	Original	1514	636	504	774	604	82.2	83.4
	Reordered	1496	625	485	803	615	77.8	78.9
Different wrong	Original	1816	573	436	871	641	65.8	68.0
	Reordered	1876	594	445	931	687	63.8	64.8

under Reordered. D8R also changes more SFT answers. Among changed answers, the share returning to Base is slightly lower for D8R in all four rows. The same distinction appears when Base is wrong and SFT is correct: D8R/C8 return to Base on 85/73 of 439 original-order questions and 97/88 of 421 reordered questions, while changing 115/94 and 129/112 SFT answers.

D8R returns more often to both incorrect and correct Base answers, while also making more changes overall. Larger restoration counts alone therefore do not establish that changes preferentially return to Base or target its correct answers. Each conditional share uses the questions changed by that continuation, so the comparison is descriptive rather than a causal adjustment for answer-changing behavior. This retrospective analysis is limited to Qwen3-8B.

Absolute recovery gains and relative gains give different comparisons. Recovery share does not specify the improvement over a control. Table 9 reports absolute and relative advantages within both correction pools. In each family and order, the two interpretations differ on the same observations.

Table 9: DPO versus supervised continuation within each pool. Absolute rate differences (RD, percentage points) favor restoration more strongly; rate ratios (RR) and odds ratios (OR) favor new correction more strongly. Qwen3-8B uses D8R and C8.

Family	Order	Pool	p_D	p_C	RD (pp)	RR	OR
Qwen3-8B	Original	Restoration	.4201	.3329	8.72	1.2619	1.4516
		New correction	.0360	.0223	1.37	1.6154	1.6384
	Reordered	Restoration	.4178	.3242	9.36	1.2887	1.4958
		New correction	.0370	.0205	1.65	1.8033	1.8342
OLMo-2-7B	Original	Restoration	.3374	.2422	9.53	1.3935	1.5939
		New correction	.0300	.0170	1.30	1.7652	1.7888
	Reordered	Restoration	.3167	.2541	6.26	1.2466	1.3609
		New correction	.0318	.0183	1.34	1.7343	1.7583

In OLMo-2-7B's original order, DPO adds 9.53 percentage points on restoration candidates versus 1.30 on new-correction candidates, but its odds ratios are 1.5939 versus 1.7888. Qwen3-8B and both reordered comparisons preserve this distinction. Equation 28 explains why a larger absolute recovery benefit can coexist with a smaller relative advantage when control success rates differ.

Net improvement also depends on losses. In Qwen3-8B's original order, D8R makes 636 restorations and 210 new corrections while losing 205 SFT-correct answers; C8 makes 504 restorations and 130 new corrections while losing 133. Their net gains are 641 and 501 questions. In OLMo-2-7B, DO makes 301 restorations and 233 new corrections while losing 348 SFT-correct answers; CO makes 216 restorations and 132 new corrections while losing 212. Their net gains are 186 and 136.

Appendix A reports chosen-response development NLLs for OLMo-2-7B and the earlier Qwen3-8B run; these are not measurements of D8R.

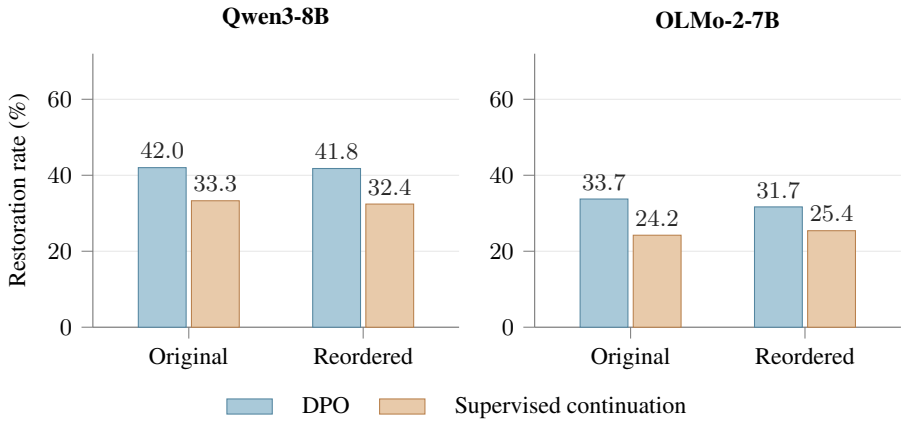


Figure 3: Supervised continuation restores earlier correct answers in both families. Bars show restoration rate $R/|\mathcal{P}_R|$, not final accuracy; each order has its own eligible pool. Both panels use the same vertical scale.

Table 10: Qwen3-8B MMLU-Pro count decomposition using the fp32-rescored Base, D8R, and C8. Restoration pools are 1,514/1,496 and new-correction pools are 5,830/5,938 for original/reordered inputs.

Order	Model	R	G	L	$R+G-L$
Original	D8R	636	210	205	641
	C8	504	130	133	501
Reordered	D8R	625	220	217	628
	C8	485	122	149	458

Table 11: OLMo-2-7B MMLU-Pro integer decomposition, fp32. Restoration pools are 892/862 and new-correction pools 7,775/7,810 for original/reordered inputs.

Order	Model	R	G	L	$R+G-L$
Original	DO	301	233	348	186
	CO	216	132	212	136
Reordered	DO	273	248	305	216
	CO	219	143	197	165

For Qwen3-8B cross-order recovery, D8R's intersection is 109 and union 1,152, giving Jaccard .0946. C8's intersection is 57 and union 932, giving .0612. The eligible pools have intersection 455 and union 2,555, giving pool Jaccard .1781. This differs from the 30.1% original-pool retention in Appendix L. OLMo-2-7B's eligible-pool Jaccard is $172/(892+862-172) = .1087$, distinct from its 19.3% original-pool retention.

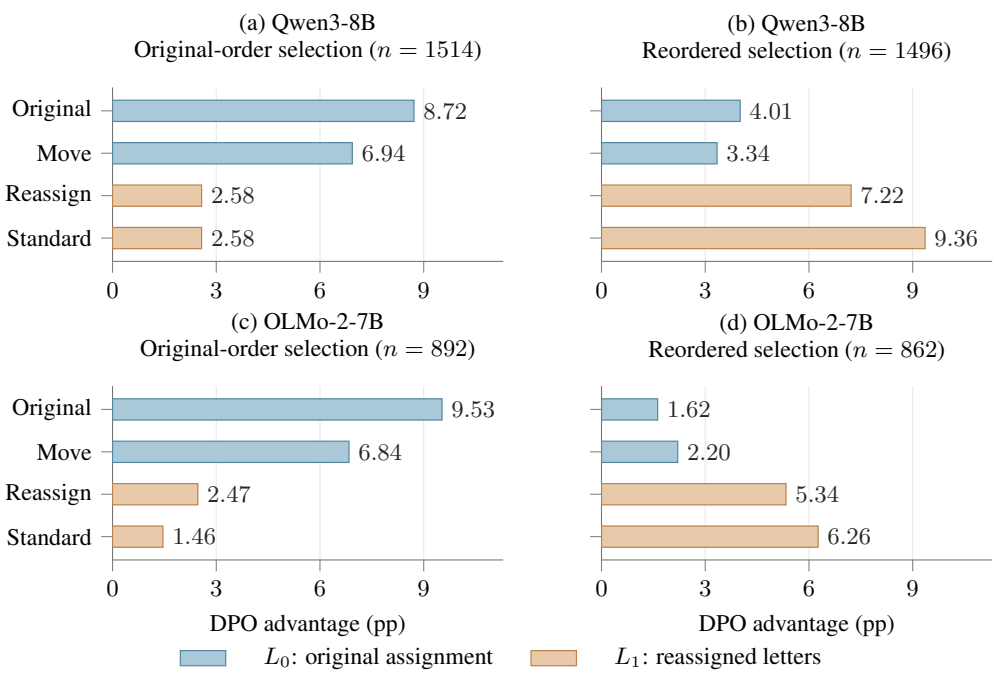


Figure 4: DPO minus supervised-continuation correctness under the general permutation of Table 1. Each panel fixes the family and screening population. Original, Move, Reassign and Standard denote P_0L_0, P_1L_0, P_0L_1 and P_1L_1. Blue retains the original letter assignment; orange reassigns letters. All panels share a zero-based scale. For original-screen OLMo-2-7B/Qwen3-8B, Standard gives 1.46/2.58 points; the different paired-swap permutation gives 1.91/2.25 in Figure 5. Move and Reassign are measurements, not fitted predictions.

For Qwen3-8B C8, 710 of 875 cross-order recovery disagreements lie outside the shared eligible pool, or 81.14%. The remaining 165 occur despite eligibility under both orders.

C Separating Content Position from Letter Assignment

Standard reordering changes both option positions and their letter labels. To examine these changes separately, we evaluate two additional presentations on each fixed $\mathcal{H}_s$: moving each option together with its original letter, and reassigning letters without moving content. Qwen3-8B uses S8, D8R, and C8; OLMo-2-7B uses SO, DO, and CO. All four presentations retain the questions selected by the original Base/SFT outcomes.

Let P_0, P_1 denote original and permuted content positions, and L_0, L_1 denote the original and reassigned content-to-letter maps. The recorded original and standard-reordered inputs are P_0L_0 and P_1L_1. The new P_1L_0 moves content with its original letter; P_0L_1 changes letters without moving content. Reassignment uses the same per-question permutation as the recorded reordering. Appendix D gives the exact mapping, valid-letter scoring rule, and all sixteen rows of correct counts.

Extend the evaluation presentation in Equation 4 to $e = (P, L)$. For each screening population, write $g_s(P, L) = 100[A_D(s, P, L) - A_C(s, P, L)]$ for the advantage in percentage points. Average letter and position contrasts are

$$\begin{aligned}\bar{d}_L &= \tfrac{1}{2}\{[g_s(P_0, L_1) - g_s(P_0, L_0)] + [g_s(P_1, L_1) - g_s(P_1, L_0)]\},\\ \bar{d}_P &= \tfrac{1}{2}\{[g_s(P_1, L_0) - g_s(P_0, L_0)] + [g_s(P_1, L_1) - g_s(P_0, L_1)]\}.\end{aligned} \tag{9}$$

These contrasts summarize the four measured conditions; averaging does not assume that the two factors act independently.

Letter assignment changes the comparison more than position on average. On OLMo-2-7B's original-screen population, the gap is 9.53 points in the original presentation, 6.84 after moving content with its letters, 2.47 after changing only the assignment, and 1.46 under standard reordering. Moving positions alone therefore reduces the gap, but less than reassigning letters at the original positions. Across all four populations, moving content with its letters (P_1L_0) gives a gap closer to the original gap, whereas reassigning letters (P_0L_1) gives a gap closer to the standard-reordered gap (Figure 4).

The signed average contrasts $(\overline{d}_L, \overline{d}_P)$ are $(-6.22, -1.85)$ and $(+3.89, +0.75)$ points for OLMo-2-7B's original- and reordered-screen populations, and $(-5.25, -0.89)$ and $(+4.61, +0.74)$ for Qwen3-8B. Reassignment reduces the advantage on original-screen failures but increases it on reordered-screen failures. Its direction therefore depends on the order used to select failures rather than uniformly favoring either continuation. DPO still leads in all sixteen conditions.

Position effects depend on the letter assignment. For Qwen3-8B's reordered-screen population, moving content with the original letters changes the gap from 4.01 to 3.34 points, a decrease of 0.67. With the reassigned letters, the same position change instead raises it from 7.22 to 9.36, an increase of 2.14. Their difference is a 2.81-point interaction; reporting only the average position contrast of 0.74 would conceal this reversal. The other three interactions are 1.68 and 0.35 points for OLMo-2-7B, and 1.78 for Qwen3-8B's original-screen population (Appendix D).

The continuations can gain or lose correct answers together while their gap changes. On OLMo-2-7B's original-screen questions, changing only letters raises DO's correct count from 301 to 305 and CO's from 216 to 283; the control's larger gain contracts the gap. On Qwen3-8B's reordered-screen questions, the same transformation lowers D8R from 718 to 659 and C8 from 658 to 551. Here the control's larger loss increases DPO's advantage. A larger gap need not mean that DPO answers more questions correctly.

Move and Reassign can display different nonalphabetic letter sequences, leaving format familiarity as a possible contributor. Section 5 addresses this ambiguity by matching the displayed sequences. The average letter-assignment contrast is larger, but position effects remain and some conditions fail a stricter proximity criterion (Appendix D). These are descriptive effects for the tested permutations, not evidence of a universal ordering or latent binding failure.

D Position–Letter Crossing: Protocol and Counts

PertEval studies response consistency under perturbations designed to retain the knowledge needed to answer a question (Li et al., 2024a). Other evaluations change the answer set: MCQA+ probes selection of the least incorrect alternative (Wang et al., 2025), while tests with "none of the above" options examine rejection of listed alternatives (Tam et al., 2025). Our conditions retain the candidate contents and correct content identity. We use these fixed contents to test how presentation changes the training-method comparison on previously selected failures. This differs from changing which answers are available.

The four presentations use the same question-specific permutation as the existing two-order evaluation. Use zero-based indices in this specification: $p[k]$ is the original content index placed at displayed position k by standard reordering; p^{-1} is its inverse and a is the original gold content index. Letter index zero denotes A. Table 12 distinguishes displayed positions from letter identities.

Table 12: The exact four-condition construction. L specifies a map from content to letter identity, not a display-order convention. The new off-diagonal conditions can display letters out of alphabetical order.

Condition	Content at position k	Letter at position k	Gold letter
P_0L_0	k	k	a
P_1L_0	$p[k]$	$p[k]$	a
P_0L_1	k	$p^{-1}[k]$	$p^{-1}[a]$
P_1L_1	$p[k]$	k	$p^{-1}[a]$

Only the target question's option presentation changes; checkpoints, question text, demonstrations and space-prefixed valid-letter readout remain fixed. Correctness compares predicted and gold letter identities after the transformation, rather than their displayed positions. No marginal-prior correction is applied. Original Base/SFT outcomes select each $\mathcal{H}_s$ once. The two screened populations share 455 Qwen3-8B and 172 OLMo-2-7B questions, so the two screening directions are not independent replications.

Table 13 combines the original and standard-reordered evaluations with the subsequent Move and Reassign evaluations. Each entry counts correct answers on the named fixed population; successes by S, D and C can overlap. The paired intervals for the two-order evaluation do not apply to the Move and Reassign contrasts, which describe the specified finite populations and transformations.

Table 13: The general permutation on each fixed screening population, not the paired swaps of Table 3. Qwen3-8B D is D8R, not the earlier D8. Off-diagonal conditions are P_1L_0 and P_0L_1. Gap is $100(D-C)/n$ in percentage points. Qwen3-8B's original-screen P_0L_1 and P_1L_1 both have a 39-answer gap despite different correct counts.

Family	Screen	Condition	n	S	D	C	Gap
Qwen3-8B	Original	P_0L_0	1514	0	636	504	8.72
		P_1L_0	1514	132	672	567	6.94
		P_0L_1	1514	358	658	619	2.58
		P_1L_1	1514	516	690	651	2.58
	Reordered	P_0L_0	1496	497	718	658	4.01
		P_1L_0	1496	421	697	647	3.34
		P_0L_1	1496	96	659	551	7.22
		P_1L_1	1496	0	625	485	9.36
OLMo-2-7B	Original	P_0L_0	892	0	301	216	9.53
		P_1L_0	892	88	317	256	6.84
		P_0L_1	892	214	305	283	2.47
		P_1L_1	892	264	316	303	1.46
	Reordered	P_0L_0	862	258	307	293	1.62
		P_1L_0	862	267	298	279	2.20
		P_0L_1	862	73	291	245	5.34
		P_1L_1	862	0	273	219	6.26

Conditional contrasts and interaction. The interaction is $g_s(P_1, L_1) - g_s(P_1, L_0) - g_s(P_0, L_1) + g_s(P_0, L_0)$. It is the change in the letter contrast across positions, equivalently the change in the position contrast across assignments. Table 14 reports the contrasts at each level of the other factor, allowing for this dependence.

Table 14: Signed changes in DPO advantage (pp), calculated from unrounded counts. Letter columns change L_0 to L_1 at the named position; position columns change P_0 to P_1 at the named assignment.

		Letter change		Position change		
Family	Screen	At P_0	At P_1	At L_0	At L_1	Interaction
OLMo-2-7B	Original	−7.06	−5.38	−2.69	−1.01	1.68
	Reordered	3.71	4.06	0.58	0.93	0.35
Qwen3-8B	Original	−6.14	−4.36	−1.78	0.00	1.78
	Reordered	3.21	6.02	−0.67	2.14	2.81

Results under the earlier proximity criterion. Before inspecting the follow-up counts, we specified an informal, non-preregistered criterion: each Move or Reassign gap should lie within one quarter of the original-to-standard separation from its corresponding reference (P_1L_0 near P_0L_0, and P_0L_1 near P_1L_1). Five of eight cells meet this criterion; only OLMo-2-7B's reordered-screen population meets it for both cells. All eight are closer to the expected reference than to the other one, but three miss the specified distance. The continuous effects and interactions above show the size and direction of the deviations.

Count differences for paired swaps. Table 15 gives the integer differences corresponding to Table 3. Dividing each entry by its row's fixed n and multiplying by 100 reconstructs the displayed gaps; the Move-minus-Reassign count contrast similarly reconstructs κ. These differences do not identify either checkpoint's individual correct count or the paired disagreement counts needed for uncertainty estimates.

Table 15: DPO minus supervised-continuation correct-count differences under paired swaps. These are the unique integer differences compatible with the fixed denominators and two-decimal gaps in Table 3, not additional measurements. The last column is Move minus Reassign.

Task	Family	n	Original	Move	Reassign	Standard	Contrast
MMLU-Pro	OLMo-2-7B	892	85	60	24	17	36
	Qwen3-8B	1514	132	103	45	34	58
HellaSwag	OLMo-2-7B	2632	215	170	65	40	105
	Qwen3-8B	2810	250	200	70	40	130

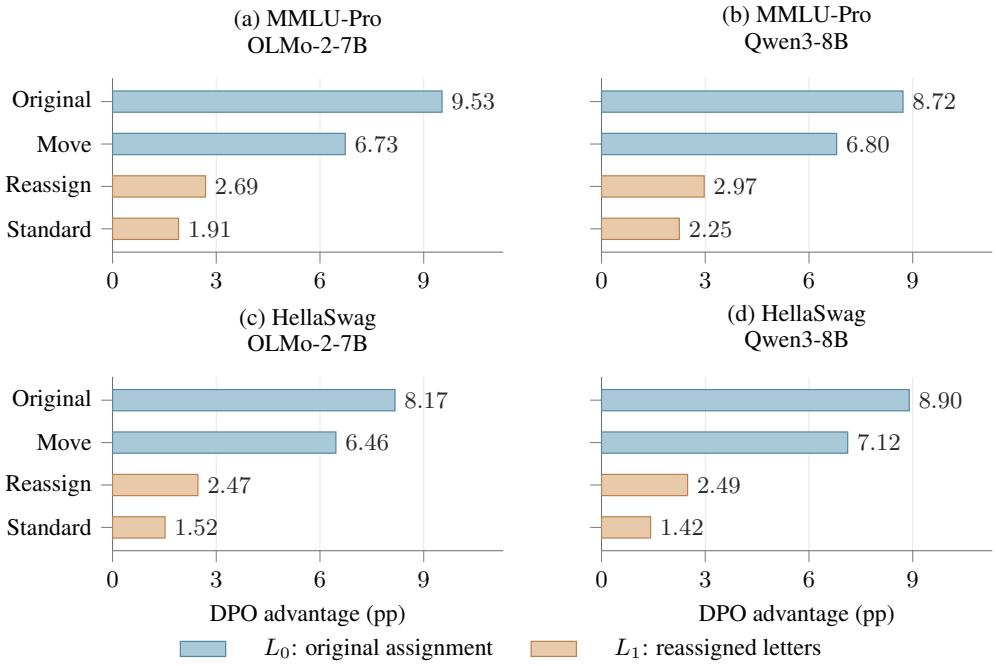


Figure 5: Paired-swap matched-display comparison on original-screen recovery candidates. Within each panel, Move (blue) and Reassign (orange) represent inputs with identical displayed letter sequences; their difference is κ (Table 3). Original, Move, Reassign and Standard denote P_0L_0, P_1L_0, P_0L_1 and P_1L_1. All panels share a zero-based scale, and each uses the same fixed questions across the four presentations. Standard uses paired swaps here: its MMLU-Pro gaps are 1.91/2.25 points for OLMo-2-7B/Qwen3-8B, versus 1.46/2.58 under the general permutation in Figure 4.

For the paired-swap construction, the original gap exceeds the standard-reordered gap and $\kappa > 0$ in all four comparisons. Algebraically, these inequalities imply a larger magnitude for the signed average letter contrast than for the signed average position contrast. On MMLU-Pro the averages are -5.83 versus -1.79 points for OLMo-2-7B and -5.15 versus -1.32 for Qwen3-8B; interactions remain 2.02 and 1.19 points. These averages summarize the same four conditions in Table 3, rather than providing a separate test of mechanism.

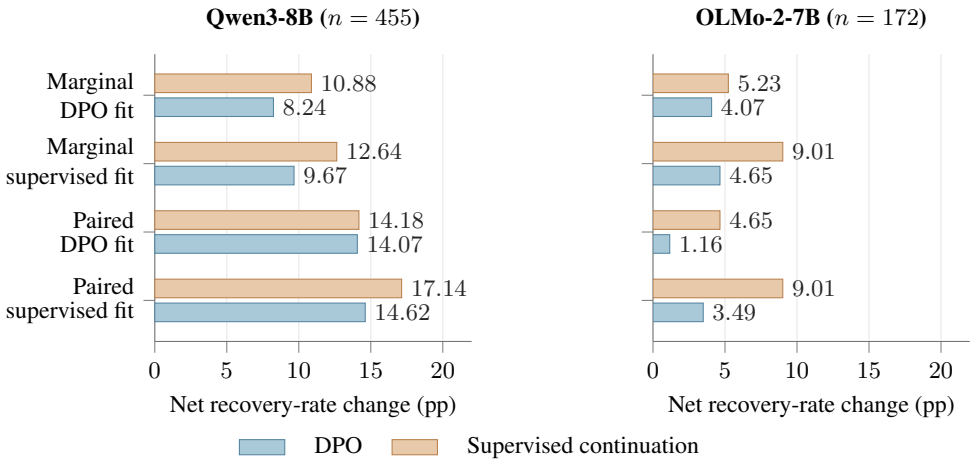


Figure 6: A shared readout changes each continuation's recovery rate; its effect on the advantage is the difference between these changes. Each row applies one fitted correction to both models on the same questions selected before correction. Marginal and paired corrections are fitted from the named DPO or supervised-continuation checkpoint on 42 validation questions. Qwen3-8B's paired-D correction adds nearly equal net correctness to both models, whereas OLMo-2-7B's benefits the control more. The figure reports changes in recovery rate, not final accuracy or explained score variation.

E Shared Readouts on Questions Eligible under Both Orders

On questions eligible for recovery under both orders, we ask whether correcting letter scores changes the method comparison and whether higher accuracy also means greater agreement in correctness between orders. The checkpoints and Base/SFT selection remain fixed.

Intervention and population. For each checkpoint and order, we estimate a letter-slot marginal $\widehat{\pi}$ from MMLU-Pro validation questions, disjoint from test. For valid slot j, the corrected score is

$$s_{ij}^{\mathrm{cal}} = \log \max(q_{ij}, 10^{-12}) - \log \max(\widehat{\pi}_j, 10^{-9}), \qquad \widehat{y}_i^{\mathrm{cal}} = \arg\max_{j \le K_i} s_{ij}^{\mathrm{cal}}. \tag{10}$$

The coefficient is one; the numerical floors avoid taking logarithms of zero. The original algorithm was fixed on OLMo-2-7B and applied to Qwen3-8B using all 70 validation questions and each checkpoint's own prior. In both families, 28 validation targets also appeared with their answers in their own demonstrations. We exclude these targets from the primary fits, leaving 42 questions, three per category, and keep the test prompts and checkpoints fixed. Appendix F retains the 70-question results and prompt verification. For the exported ten-letter probabilities v_{ij}, define $u_{ij} = v_{ij} / \sum_{k=1}^{10} v_{ik}$. The estimator is

$$\widehat{\pi}_j = \frac{\sum_{i \in \mathcal{V}} u_{ij}\, \mathbf{1}[j \le K_i]}{\sum_{i \in \mathcal{V}} \sum_{k=1}^{10} u_{ik}\, \mathbf{1}[k \le K_i]}, \qquad |\mathcal{V}| = 42. \tag{11}$$

All 42 fitting questions have ten valid options, so each marginal-prior component receives a probability contribution from every question; it is not estimated from counts of gold-answer letters. Normalization over ten letters precedes the valid-slot mask. Slot indices refer to the displayed option order; the gold index is permuted with the options.

Calibration methods differ in how they estimate label preferences. Contextual calibration uses content-free inputs (Zhao et al., 2021), domain-context calibration uses random in-domain words (Fei et al., 2023), and Batch Calibration estimates contextual bias from batched inputs (Zhou et al., 2024). PriDe instead estimates priors through content permutations (Zheng et al., 2024). Our correction averages predictions on ordinary validation questions and applies a shared vector to both continuations. Because the fitted prior can include content effects, this comparison does not isolate a content-independent bias.

We freeze the pre-intervention shared recovery population,

$$\mathcal{H}^0 = \{i : B^0_{i,o} = B^0_{i,p} = 1,\ S^0_{i,o} = S^0_{i,p} = 0\}, \tag{12}$$

and compare the final checkpoint's answers before and after correction on these same questions. The earlier Base/SFT outcomes select the population once; correction does not admit or exclude questions. Freezing only the SFT errors would define a broader correction population that includes questions never answered correctly by Base.

Own-vector and shared-vector comparisons. Define the recovery-rate advantage as

$$\Delta^r_{D,C} = \frac{1}{2|\mathcal{H}^0|} \sum_{i \in \mathcal{H}^0} \sum_{o \in \{\text{original,reordered}\}} \left(\mathbf{1}[\widehat{y}^r_{D,i,o} = y_{i,o}] - \mathbf{1}[\widehat{y}^r_{C,i,o} = y_{i,o}] \right), \tag{13}$$

where r is the original or corrected readout. On Qwen3-8B's 455 fixed questions, average recovered counts change from 188.5/139.5 for D8R/C8 to 226/197 with their own 42-question marginals: the advantage falls from 10.77 to 6.37 percentage points. On OLMo-2-7B's 172 questions, counts change from 61/44 to 68/59.5, reducing the advantage from 9.88 to 4.94 points. These comparisons change both the model and its fitted readout vector.

To separate prior choice from model choice, we cross each family's two models with both fitted vectors. Let $A(T, \pi)$ denote mean correctness on $\mathcal{H}^0$ for model T corrected by π, with a separate vector for each order as in Equation 11. A shared-prior comparison is $g(\pi) = A(D, \pi) - A(C, \pi)$. The own-prior comparison satisfies

$$A(D, \pi_D) - A(C, \pi_C) = g(\pi_D) + A(C, \pi_D) - A(C, \pi_C). \tag{14}$$

The first term compares models under a shared rule; the second records how replacing the control's prior changes its accuracy. When the control improves under its own prior, this term is negative. For Qwen3-8B marginals, it is -1.76 points, giving an own-prior gap of $8.13 - 1.76 = 6.37$. Each family's eligible questions, earlier outcomes, and correction coefficient remain fixed. The crossed analysis was designed after the own-prior results were known. Its Qwen3-8B fitting and comparison rules were then applied unchanged to the existing OLMo-2-7B scores; neither analysis requires additional model inference or tuning.

Table 16: Readout comparisons on questions eligible for recovery under both orders: OLMo-2-7B $n = 172$, Qwen3-8B $n = 455$. D, C are mean recovery rates (%); Δ is their difference (pp). A shared vector comes from the named continuation within the same family. Both estimators fit the same 42 validation questions that do not appear in their own demonstrations. Paired offsets use the primary nine-parameter fit shared across orders.

	OLMo-2-7B			Qwen3-8B		
Readout rule	D	C	Δ	D	C	Δ
No correction	35.47	25.58	9.88	41.43	30.66	10.77
Marginal: D prior for both	39.53	30.81	8.72	49.67	41.54	8.13
Marginal: C prior for both	40.12	34.59	5.52	51.10	43.30	7.80
Marginal: own prior for each	39.53	34.59	4.94	49.67	43.30	6.37
Paired: D fit for both	36.63	30.23	6.40	55.49	44.84	10.66
Paired: C fit for both	38.95	34.59	4.36	56.04	47.80	8.24
Paired: own fit for each	36.63	34.59	2.03	55.49	47.80	7.69

For the marginal estimator, either shared vector reduces the advantage relative to no correction in both families (Table 16). In Qwen3-8B, Equation 14 gives $8.13 - 1.76 = 6.37$ points; in OLMo-2-7B it gives $8.72 - 3.78 = 4.94$. Each control's vector yields higher recovery for both checkpoints than the DPO vector. The own-prior gap is therefore below both shared-vector gaps: it uses the lower-recovery vector for DPO and the higher-recovery vector for the control. This is an observed ordering, not a general property of fitting marginal priors.

How the shared readout changes the comparison. Let $b_{j,o}$ be the correction subtracted from log score at slot j in order o; for marginal correction, $b_{j,o} = \log \max(\widehat{\pi}_{j,o}, 10^{-9})$. All option probabilities

in the following log ratios use the 10^{-12} floor. For a question with gold slot y, define

$$h_{T,i,o}(b) = \min_{\substack{1 \le j \le K_i \\ j \ne y}} \left[\log \frac{q_{T,i,y,o}}{q_{T,i,j,o}} - (b_{y,o} - b_{j,o}) \right]. \tag{15}$$

With no ties, the corrected answer is correct exactly when $h_{T,i,o}(b) > 0$. Accuracy changes when the gold option gains or loses the highest score; a change between two incorrect options leaves accuracy unchanged. The mean score change alone does not determine these transitions.

Count, over the $2|\mathcal{H}^0|$ question–order outcomes, corrections $U_T(b)$ from wrong to correct and losses $V_T(b)$ from correct to wrong. The change in the method advantage is exactly

$$\Delta_{D,C}(b) - \Delta_{D,C}(0) = \frac{[U_D(b) - V_D(b)] - [U_C(b) - V_C(b)]}{2|\mathcal{H}^0|}. \tag{16}$$

Equal net gains cancel in the comparison even when many answers change in each model. Squared score error does not record how many answers become correct or incorrect, so it cannot substitute for this comparison.

For example, under Qwen3-8B's shared D8R marginal, $U_D - V_D = 116 - 41 = 75$ and $U_C - V_C = 164 - 65 = 99$. The advantage decreases by $100(99-75)/910 = 2.64$ percentage points. With shared D8R paired offsets, the corresponding gains are 128 and 129, giving only $100/910 = 0.11$ points of attenuation. For OLMo-2-7B's shared DO paired offsets, they are $53 - 49 = 4$ and $63 - 47 = 16$, giving $100 \times 12/344 = 3.49$ points. Figure 6 shows all four primary shared corrections.

Fitting score changes for the same option content. Marginal probabilities mix option content and slot preference. We fit an additive readout using the same 42 validation pairs: center each question's log probabilities, align option content across orders, and regress the score differences on the corresponding slot changes. The primary fit has nine free offsets per checkpoint, shared across orders; no test outcomes enter it. On the fixed populations it explains 42.97%/43.94% of squared cross-order score changes for OLMo-2-7B DO/CO and 53.42%/57.66% for Qwen3-8B D8R/C8.

These score corrections affect the method comparison differently in the two families. Qwen3-8B's advantage is 10.66 points with shared D8R offsets: net gains of 128 and 129 correct question–order pairs leave the raw advantage almost unchanged. Sharing C8's offsets gives 8.24 points. OLMo-2-7B's advantage falls to 6.40 or 4.36 points under the corresponding shared fits. A larger explained score component need not imply a larger reduction in method advantage: the two models' correctness changes can cancel in their difference. Appendix G reports residuals and alternatives; the fit does not identify pure letter bias.

Intermediate correction strengths. The endpoint can miss non-monotone changes between no correction and coefficient one. Appendix H examines all eight fixed-vector paths retrospectively, without selecting a new scoring rule. None reverses the ranking along the examined direction.

Accuracy and consistency need a paired comparison. Within $\mathcal{H}^0$, let n_2, n_1, n_0 count questions correct under both orders, exactly one order, and neither. Then

$$\overline{A}_{\mathcal{H}^0} = \frac{2n_2 + n_1}{2|\mathcal{H}^0|}, \qquad d_{\mathcal{H}^0} = \frac{n_1}{|\mathcal{H}^0|}. \tag{17}$$

Moving a question from n_0 to n_1 raises both accuracy and disagreement. Moving one from n_2 to n_1 instead lowers accuracy while raising disagreement. The same change in disagreement can therefore have opposite implications for correctness.

The historical 70-question-prior intervention illustrates this distinction: DPO disagreement decreases by seven questions in OLMo-2-7B, increases by sixteen in Qwen3-8B, and increases by twelve in Qwen3-1.7B, while mean correctness increases in each. Appendix F retains the full paired transitions and distinguishes these results from the 42-question comparisons. Readout comparisons should keep the selected questions fixed and report both correctness and disagreement.

F Earlier Corrections Fitted on All Validation Questions

Reconstructing validation prompts from the released questions, actual demonstrations, and recorded option permutations matches all 560 final-checkpoint prompt hashes: two continuations, two orders, and 70 questions in each family. The same 28 targets appear in their own demonstrations, leaving 42 after exclusion by membership rather than outcomes. These records use prompt-hash version 2; the older OLMo-2-7B test hash fields have no version tag and are not included in this reconstruction claim.

These earlier corrections use all 70 validation questions, including targets repeated in their own demonstrations. The main comparisons instead fit on the remaining 42 questions. Under the original own-prior rule, the Qwen3-8B D8R–C8 gap falls from 10.77 to 5.16 points and the OLMo-2-7B DO–CO gap from 9.88 to 1.16. Sharing the 70-question marginal gives Qwen3-8B gaps of 5.71/6.26 points with D8R/C8 priors and OLMo-2-7B gaps of 2.62/2.33 with DO/CO priors. These results use demonstration-inclusive fitting. The 1.7B chain supplies an additional model size, not an independent family.

Table 17: Historical 70-question-prior correction on fixed recovery populations. Every arrow, including in mean accuracy $\overline{A}$ (%), denotes before $\rightarrow$ after correction. These precede the demonstrated-target exclusion in Table 16. The 1.7B comparison is an additional model size, not a third family.

Family	Model	$\lvert\mathcal{H}^0\rvert$	Both correct	One correct	Neither	$\overline{A}$ (%)
OLMo-2-7B	DO	172	$29 \rightarrow 36$	$64 \rightarrow 57$	$79 \rightarrow 79$	$35.47 \rightarrow 37.50$
	CO	172	$21 \rightarrow 34$	$46 \rightarrow 57$	$105 \rightarrow 81$	$25.58 \rightarrow 36.34$
Qwen3-8B	D8R	455	$109 \rightarrow 133$	$159 \rightarrow 175$	$187 \rightarrow 147$	$41.43 \rightarrow 48.46$
	C8	455	$57 \rightarrow 116$	$165 \rightarrow 162$	$233 \rightarrow 177$	$30.66 \rightarrow 43.30$
Qwen3-1.7B	D17	121	$19 \rightarrow 16$	$35 \rightarrow 47$	$67 \rightarrow 58$	$30.17 \rightarrow 32.64$

Table 18: Full-test accuracy under original/reordered option orders, before and after marginal-prior correction. These use all 12,032 questions, unlike the fixed-population accuracies in Table 17.

Model	Before: original / reordered	After: original / reordered
OLMo-2-7B DPO (DO)	.2951 / .2972	.2996 / .2962
OLMo-2-7B Supervised continuation (CO)	.2910 / .2930	.2993 / .2925
Qwen3-8B DPO (D8R)	.4429 / .4343	.4521 / .4439
Qwen3-8B Supervised continuation (C8)	.4313 / .4202	.4425 / .4323
Qwen3-1.7B DPO (D17)	.3457 / .3397	.3520 / .3458

Table 19: Complete paired state transitions within each fixed $\mathcal{H}^0$. States 2, 1, and 0 denote correctness under both orders, one order, and neither. Every row sums to that model's fixed-pool size; row/column marginals reconstruct Table 17.

Model	$2 \rightarrow 2$	$2 \rightarrow 1$	$2 \rightarrow 0$	$1 \rightarrow 2$	$1 \rightarrow 1$	$1 \rightarrow 0$	$0 \rightarrow 2$	$0 \rightarrow 1$	$0 \rightarrow 0$
OLMo-2-7B DPO (DO)	25	3	1	9	41	14	2	13	64
OLMo-2-7B Supervised continuation (CO)	15	4	2	9	26	11	10	27	68
Qwen3-8B DPO (D8R)	103	5	1	29	114	16	1	56	130
Qwen3-8B Supervised continuation (C8)	45	10	2	49	84	32	22	68	143
Qwen3-1.7B DPO (D17)	14	5	0	1	28	6	1	14	52

In Qwen3-8B D8R, 56 questions move from neither order correct to one order correct, while sixteen move in the opposite direction. Another 29 move from one to both and five from both to one. These flows increase disagreement by sixteen, even as shared correctness rises from 109 to 133. The 1.7B comparison is based on 121 jointly eligible questions, which limits the precision and scope of its observed effect.

G Additive Corrections from Content-Aligned Scores

This exploratory analysis tests a fixed additive component of score changes. PriDe's separable option-ID/content model becomes additive in log probabilities (Zheng et al., 2024). Here we estimate a checkpoint-wide slot vector by pooled least squares on paired orders, rather than PriDe's balanced-permutation prior estimator. This changes the estimator, not the underlying separability assumption. For question i, let $j(c, o)$ be the slot containing option content c in order o. Center the log probabilities within the valid options,

$$\ell_{i,c,o} = \log \max(q_{i,j(c,o),o}, 10^{-12}) - \frac{1}{K_i} \sum_{j=1}^{K_i} \log \max(q_{i,j,o}, 10^{-12}).$$

The primary least-squares fit uses

$$d_{i,c} = \ell_{i,c,o} - \ell_{i,c,p} = b_{j(c,o)} - b_{j(c,p)} + \epsilon_{i,c}. \tag{18}$$

An invariant content score and the order-specific probability normalizers cancel under this additive model. Set $b_{10} = 0$ to fix the arbitrary common offset. The design has rank nine for each continuation in both families. We verify content permutations against original and reordered gold indices, fit on the 42 validation questions absent from the demonstrations, and subtract b_j from test log scores with coefficient one. Gold answers are not regression targets, and the fixed recovery population is not reselected.

We also report a rank-18 alternative with separate vectors b^o, b^p. Its predicted difference is $(b^o_{j(c,o)} - \overline{b}^o_{K_i}) - (b^p_{j(c,p)} - \overline{b}^p_{K_i})$, where each mean is over the valid slots. The extra parameters accommodate order-specific offsets rather than assuming them equal. Table 20 evaluates both fits; the nine-parameter model was designated primary before their test correctness was computed.

Table 20: Qwen3-8B additive diagnostics by fitting set. Explained change $E = 1 - \sum(d - \widehat{d})^2 / \sum d^2$ is not a fraction of answer flips or a causal fraction. E and RMS are evaluated out of sample: fits use 42 or 70 validation questions, Test uses all 12,032 test questions, and $\mathcal{H}^0$ uses 455. The 42-question fit excludes demonstrated targets; the 70-question fit retains them.

Model	Fit n	Parameters	Test E	$\mathcal{H}^0$ E	Raw RMS	Residual RMS
D8R	42	9	54.56%	53.42%	2.2593	1.5230
	42	18	54.37%	53.29%	2.2593	1.5261
	70	9	47.87%	46.80%	2.2593	1.6313
	70	18	47.63%	46.57%	2.2593	1.6350
C8	42	9	59.47%	57.66%	1.9097	1.2158
	42	18	59.21%	57.49%	1.9097	1.2196
	70	9	53.01%	51.42%	1.9097	1.3090
	70	18	52.76%	51.19%	1.9097	1.3125

For the primary 42-question fit, own-offset mean recovered counts are 252.5 for D8R and 217.5 for C8, giving a 7.69-point advantage. Shared D8R offsets give 252.5 and 204, or 10.66 points; shared C8 offsets give 255 and 217.5, or 8.24 points. The rank-18 alternative gives own-offset counts 249 and 215, or 7.47 points. Its shared D8R offsets give 249 and 201, or 10.55 points; shared C8 offsets give 253 and 215, or 8.35 points. Both parameterizations nearly preserve the raw advantage when the D8R vector is shared.

The earlier 70-question primary fit gives own-offset counts 245 and 216.5, shared-D8R counts 245 and 206, and shared-C8 counts 250 and 216.5: advantages of 6.26, 8.57, and 7.36 points. For its rank-18 alternative the corresponding counts are 245.5/214.5, 245.5/205.5, and 248/214.5, or 6.81,

8.79, and 7.36 points. Removing demonstrated targets changes sample size and composition as well as answer exposure, so differences between the fitting sets are not attributed solely to seeing the target answer.

OLMo-2-7B under the same fitting rules. Table 21 evaluates the additive fits on OLMo-2-7B's fixed 172-question population. No full-test OLMo-2-7B explanatory fraction is estimated from this subset. With the 42-question rank-18 alternative, shared DO/CO offsets give mean counts 63.5/50.5 and 65/59.5, or advantages of 7.56 and 3.20 points. Own offsets give 63.5/59.5, or 2.33 points. The corresponding primary nine-parameter own counts are 63/59.5, or 2.03 points (Table 16).

Table 21: OLMo-2-7B additive diagnostics on $\mathcal{H}^0$ only ($n = 172$), with the same definition of E as Table 20. Fits use 42 or 70 validation questions; E and RMS are evaluated out of sample on these 172 test questions. RMS is in nats.

Model	Fit n	Parameters	$\mathcal{H}^0$ E	Raw RMS	Residual RMS
DO	42	9	42.97%	2.2202	1.6767
	42	18	42.07%	2.2202	1.6898
	70	9	46.41%	2.2202	1.6254
	70	18	45.32%	2.2202	1.6417
CO	42	9	43.94%	1.5517	1.1618
	42	18	43.03%	1.5517	1.1712
	70	9	46.91%	1.5517	1.1307
	70	18	45.95%	1.5517	1.1409

Using all 70 validation questions, the OLMo-2-7B nine-parameter shared DO/CO fits give counts 68/62.5 and 70/68, or advantages of 3.20 and 1.16 points. Own offsets tie at 68/68. The rank-18 shared fits give 65.5/61.5 and 70.5/65.5, or 2.33 and 2.91 points; own offsets again tie, at 65.5/65.5. These are equal measured mean recovery rates under demonstration-inclusive fitting, not evidence that the training objectives are equivalent. Unlike Qwen3-8B, OLMo-2-7B has lower explained squared change after the demonstrated targets are removed. The change in fitting population has no uniform effect across these families.

Earlier diagnostic fit. The analysis was motivated by an exploratory fit using half the Qwen3-8B test questions outside the shared recovery population $\mathcal{H}^0$. With a seed-42 split of the 11,577 questions outside $\mathcal{H}^0$, 5,788 questions fit the offsets and 5,789 evaluated them. The shared-vector model explains 61.41% and 64.64% of squared changes on that evaluation half for D8R and C8; separate-order fits give 61.40% and 64.64%. These test-fitted offsets are not used for the recovery-rate comparisons above. The lower explanatory fractions from validation-fitted offsets show why the fitting population cannot be ignored.

Equation 18 estimates an additive component conditional on the observed content permutations. A good squared-error fit alone neither excludes content–order interactions nor identifies why a checkpoint acquired a particular slot preference. We evaluate its effect on accuracy directly; multiplying the explained-score fraction by the number of answer changes would not give this effect. The primary corrections retain the original numerical floor and coefficient one, without residual trimming. Appendix H separately examines intermediate coefficients without replacing that primary rule.

H Exploratory Intermediate-Readout Analysis

The primary comparisons apply each fitted correction with coefficient one. After examining those results, we check the path from no correction to each of the four shared corrections in each family. For a fitted vector b_o in observed option order o, the score at path coordinate λ is

$$s_{T,i,j,o}(\lambda) = \log \max(q_{T,i,j,o}, 10^{-12}) - \lambda b_{j,o}, \qquad 0 \le \lambda \le 1. \tag{19}$$

The vector direction, recovery population, and checkpoints remain fixed. Marginal vectors have separate fits by order; the primary paired vector is shared across orders. At each λ, both continuations receive the same vector for a given order.

Finding changes along a path. For gold slot y and competitor j, write $m_j = \log q_y - \log q_j$ and $d_j = b_y - b_j$, with the same numerical floor as Equation 19. Away from ties, gold correctness requires $m_j - \lambda d_j > 0$ for every valid competitor. Each inequality defines a half-line in λ, so their intersection with $[0, 1]$ is an interval, possibly empty. Its endpoints follow from the gold-versus-competitor crossings m_j/d_j when $d_j \neq 0$; a violated constant inequality makes the interval empty. The difference in accuracy is therefore a signed sum of interval indicators. It need not vary monotonically even though every option score is affine in λ.

We compute the correctness-interval boundaries from exported scores in floating-point arithmetic, then evaluate each boundary and the midpoint between successive boundaries, including zero and one. Ties use the smallest-index argmax rule. Endpoint counts agree with all eight primary shared-vector comparisons. Table 22 reports the numerical minima and maxima along these paths; the primary results still use coefficient one, not coefficients selected at these extrema.

Table 22: Exploratory changes along the eight existing shared-correction paths. All values are DPO minus supervised-continuation recovery rate in percentage points, on fixed populations of 455 Qwen3-8B and 172 OLMo-2-7B questions. Uncorrected gaps are 10.77 and 9.88. The paths interpolate each fitted vector separately, not arbitrary combinations of vectors.

Family	Shared correction	Endpoint	Path minimum	Path maximum
Qwen3-8B	Marginal: D8R	8.13	6.59	11.76
	Marginal: C8	7.80	7.25	11.65
	Paired: D8R	10.66	9.78	13.74
	Paired: C8	8.24	8.02	13.52
OLMo-2-7B	Marginal: DO	8.72	6.69	10.47
	Marginal: CO	5.52	4.36	10.47
	Paired: DO	6.40	3.49	10.47
	Paired: CO	4.36	3.20	10.76

Intermediate gaps can lie outside the range of the two endpoints. OLMo-2-7B's paired-DO path ends at a 6.40-point advantage but reaches 3.49, while its paired-CO path ends at 4.36 but reaches 3.20. Qwen3-8B's paired-D8R path reaches 13.74 despite endpoints of 10.77 and 10.66. None of the eight path checks finds a ranking reversal. The result concerns these eight fitted corrections and strengths between zero and one, not arbitrary readout changes or the smallest change needed to reverse a ranking.

I Readout and Numerical Checks

MMLU-Pro prompts and the general permutation. Both families use the same frozen scaffold. The header is `The following are multiple choice questions (with answers) about {category}.` It is followed by two demonstrations and the target question, with one blank line between these blocks. Within a block, the question is followed by one candidate per line in the form `A. candidate`, then `Answer:`; demonstrations append the gold letter after a space, whereas the target stops at the colon. Demonstration content and order stay fixed when the target options are reordered.

Demonstrations are the first two rows after applying pandas `sample(frac=1, random_state=42)` within each validation category. In their prompt order, the validation question IDs are: biology 21/24; business 16/19; chemistry 26/29; computer science 31/34; economics 36/39; engineering 41/44; health 6/9; history 56/59; law 66/69; math 1/4; other 51/54; philosophy 46/49; physics 11/14; psychology 61/64. These 28 targets are excluded when fitting the primary readout corrections.

For a target with integer question ID k and K valid options, initialize the zero-based list $[0, \ldots, K-1]$ and apply Python `random.Random(42+k).shuffle` to that list. The reordered candidate at displayed position j is the original candidate indexed by the resulting list's jth entry; its displayed letter is reassigned by position. The reordered gold index is the position of the original gold candidate in this list. This is the general permutation used in Table 1, not the paired-swap permutation of Table 3 or the separate shuffle used to sample intervention questions.

For an exported ten-letter vector $q_i^{(10)}$ and K_i valid options, valid-option probabilities are

$$\widetilde{q}_{ij} = \frac{q_{ij}^{(10)}}{\sum_{k=1}^{K_i} q_{ik}^{(10)}}, \qquad j = 1, \ldots, K_i. \tag{20}$$

Predictions maximize over valid options. Equation 20 changes the probability denominator, not this argmax.

Letter-based correctness uses exact equality of the predicted and gold valid-option indices, including index zero. The same labels are used for accuracy, recovery eligibility and the $R/G/L$ decomposition.

Automatic dtype selection mixed bf16 Base evaluation with fp32 evaluation of trained checkpoints. Rescoring Base in fp32 changes 298/295 correctness labels across original/reordered inputs and changes D8R's restoration counts from 631/611 to 636/625. S8 labels are unchanged. All D8R tables here use the rescored Base.

D8R and C8 have identical predicted indices across their Pod and local mirror readouts on all 12,032 questions under both orders, as well as matching corrected prompt hashes. Across the four comparisons, the maximum absolute option-probability difference is 5.321×10^{-5}. These paired checks use eight shards in both environments. Base's local four-shard readout was not independently mirrored with eight shards, so the checks do not establish invariance for every checkpoint or batching arrangement. The exported vectors contain ten letter probabilities even for shorter option lists. Recomputing the calibration over valid slots gives the same predictions as the exported implementation for every D8R/C8 test question in this study.

Prompt-identity comparisons use corrected per-question hashes, excluding the older batch-repeated fields. Appendix F specifies the coverage of the validation-prompt reconstruction.

Answer-position support is a separate measurement. With $S_\varepsilon(p) = \{t : p(t) \geq \varepsilon\}$, removed count fraction and removed mass are

$$f_\varepsilon(p, q) = \frac{|S_\varepsilon(p) \setminus S_\varepsilon(q)|}{|S_\varepsilon(p)|}, \qquad c_\varepsilon(p, q) = \sum_{t \in S_\varepsilon(p) \setminus S_\varepsilon(q)} p(t). \tag{21}$$

In the earlier D8 analysis, on its corrected 893-question restoration group, the S8–D8 answer-position medians are $f_{10^{-6}} = 0.9341$, $c_{10^{-6}} = 0.015237$, and top-256 Jaccard 0.6151. These are not D8R measurements. Many low-probability tokens can cross a threshold while accounting for little removed mass. Neither statistic alone measures knowledge loss or the total redistribution of probability.

J Expected Recovery Overlap under Random Selection

Hold both eligible pools fixed and draw recovery sets of sizes a and b independently and uniformly from their respective pools. If the pool intersection has size I, then

$$\mathbb{E}[K] = I \, \frac{a}{|\mathcal{P}_{R,o}|} \, \frac{b}{|\mathcal{P}_{R,p}|}. \tag{22}$$

Each of the I shared candidates is included in both draws with probability $(a/|\mathcal{P}_{R,o}|)(b/|\mathcal{P}_{R,p}|)$; summing these probabilities gives the expectation. For Qwen3-8B D8R, $I = 455$, $a = 636$, and $b = 625$ give 79.85 expected versus 109 observed; supervised continuation gives 49.11 versus 57. For OLMo-2-7B DPO, $I = 172$, $a = 301$, and $b = 273$ give 18.38 expected versus 29 observed; supervised continuation gives 10.58 versus 21. These are intersection expectations, not Jaccard expectations.

A shared-pool reference instead fixes recovered counts a_c, b_c within $\mathcal{H}$ and draws both sets from its I candidates. Then K_c is hypergeometric, with $\mathbb{E}[K_c] = a_c b_c / I$. For Qwen3-8B D8R, $a_c = 201$ and $b_c = 176$ give 77.75 expected versus 109 observed. The expected Jaccard must average the ratio itself:

$$\mathbb{E}[J_c] = \sum_k \Pr(K_c = k) \frac{k}{a_c + b_c - k}. \tag{23}$$

Replacing K_c by its expectation inside the ratio generally gives a different answer. For OLMo-2-7B DPO, $a_c = 65$, $b_c = 57$, and $I = 172$ give $\mathbb{E}[J_c] = .215510$, whereas inserting $\mathbb{E}[K_c] = 21.5407$

Table 23: Shared-pool overlap and exact hypergeometric references. Enrichment is $J_c/\mathbb{E}[J_c]$, not a significance test or a universal stability category.

Family	Method	a_c	b_c	k	J_c	$\mathbb{E}[J_c]$	Enrichment
OLMo-2-7B	DPO	65	57	29	.3118	.2155	1.4469
	Supervised	46	42	21	.3134	.1475	2.1243
Qwen3-8B	DPO	201	176	109	.4067	.2602	1.5632
	Supervised	147	132	57	.2568	.1809	1.4196

into the ratio gives .214422. For supervised continuation, the respective values are .147544 and .146319. We use the exact hypergeometric expectations.

The two references condition on different populations and recovery counts. Sampling from all 12,032 questions ignores recovery eligibility and is not used to interpret the conditional overlap.

K Earlier Qwen3-8B D8 and ARC-Challenge Results

The earlier D8 run used a cross-runtime reference cache with the discrepancy described in Appendix A. Its launch record does not retain the initialization path or hash. A post-hoc weight audit compares its early and final weights with the bf16-rounded S8 state, obtaining relative distances .000169 and .000406; C8's final distance is .000388. These checks support the intended starting checkpoint but do not reconstruct the missing launch record or establish reference-score equivalence.

Under the fp32-rescored Base convention, D8 restores 893/920 MMLU-Pro answers, with new-correction counts 379/395 and losses 330/315. Its net gains over S8 are 942/1,000 questions. With the historical 70-question priors, its fixed-population advantage over C8 is 30.55 percentage points before calibration and 18.02 afterward, compared with 10.77 and 5.16 for D8R. D8's shared-pool states change from $(n_2, n_1, n_0) = (206, 145, 104)$ to $(206, 146, 103)$. These earlier measurements remain descriptive evidence; the reduction in advantage between runs is not attributed solely to reference-score error.

ARC-Challenge. The earlier D8/C8 comparison uses all 1,172 test questions, four development-only demonstrations, and a separately frozen option permutation. D8 restores three more questions than C8 under each order and has a net gain four questions larger (Table 24). With only about thirty questions in each restoration pool, this is a small external comparison for those checkpoints, not an evaluation of D8R.

Table 24: Earlier Qwen3-8B D8/C8 results on ARC-Challenge, all 1,172 test questions, fp32. Net changes are question counts. B8/S8 accuracies are .9309/.9130 under the original order and .9343/.9198 under reordering.

Order	Model	R	G	L	Net change	Accuracy
Original	D8	16	6	3	19	.9292
	C8	13	4	2	15	.9258
Reordered	D8	17	3	4	16	.9334
	C8	14	3	5	12	.9300

Within the original order, D8/C8 restoration sets have 12 common questions, four D8-only and one C8-only; under reordering they have 12 common, five D8-only and two C8-only. New-correction intersections are three and two. Across either order, D8 restores 29 distinct ARC questions and C8 restores 24; their union-level sets share 21, leaving eight unique to D8 and three to C8. These counts use restoration under either order, not independent replications.

L Eligibility Changes and Recovery Disagreement

Changing the screening order changes which questions qualify for recovery. Let $\mathcal{H} = \mathcal{P}_{R,o} \cap \mathcal{P}_{R,p}$ be their intersection. Only 455 Qwen3-8B and 172 OLMo-2-7B questions qualify under both orders:

30.1% and 19.3% of the respective original-order pools. A question can be counted as recovered under only one order for two reasons: it qualifies for recovery under only that order, or it qualifies under both but the final checkpoint answers it correctly under only one.

Table 25 partitions these events. M_e counts recovery disagreements outside $\mathcal{H}$, M_h those within it, and $\alpha = M_e/(M_e + M_h)$. J and J_c are recovered-set Jaccards before and after restricting to $\mathcal{H}$. Between 81.14% and 88.30% of the recorded disagreements lie outside the shared pool. Disagreements also remain among questions eligible under both orders; this partition locates them but does not explain their cause. Appendix M gives the counts and derivation; Appendix J supplies eligibility-conditioned overlap references.

Table 25: Recovery across option orders. α is the fraction of recovery disagreements outside the shared eligible pool; it is not a fraction of all questions or a causal attribution. All entries are derived from integer counts.

Family	Method	I	J	J_c	M_e	M_h	α
OLMo-2-7B	DPO	172	.0532	.3118	452	64	87.60%
	Supervised	172	.0507	.3134	347	46	88.30%
Qwen3-8B	DPO	455	.0946	.4067	884	159	84.76%
	Supervised	455	.0612	.2568	710	165	81.14%

M Details of Eligibility and Recovery Disagreement

Let $\mathcal{R}_o, \mathcal{R}_p$ be the recovered sets under two option orders and $\mathcal{H} = \mathcal{P}_{R,o} \cap \mathcal{P}_{R,p}$ their shared eligible pool. Write $a = |\mathcal{R}_o|$, $b = |\mathcal{R}_p|$, $k = |\mathcal{R}_o \cap \mathcal{R}_p|$, $a_c = |\mathcal{R}_o \cap \mathcal{H}|$, and $b_c = |\mathcal{R}_p \cap \mathcal{H}|$. Overall and conditional Jaccards are

$$J = \frac{k}{a + b - k}, \qquad J_c = \frac{k}{a_c + b_c - k}. \tag{24}$$

The numerator is unchanged because a question recovered under both orders must be eligible under both. Since $a_c + b_c \leq a + b$, $J_c \geq J$ whenever both are defined. An increase after conditioning is therefore not, by itself, evidence of stability.

M.1 An exact disagreement partition

The symmetric difference counts questions recovered under exactly one order. Each such question either lies outside $\mathcal{H}$ or remains eligible under both orders:

$$\begin{aligned} M &= |\mathcal{R}_o \triangle \mathcal{R}_p| = a + b - 2k, \\ M_e &= |\mathcal{R}_o \setminus \mathcal{H}| + |\mathcal{R}_p \setminus \mathcal{H}| = a + b - a_c - b_c, \\ M_h &= |(\mathcal{R}_o \cap \mathcal{H}) \triangle (\mathcal{R}_p \cap \mathcal{H})| = a_c + b_c - 2k, \\ M &= M_e + M_h. \end{aligned} \tag{25}$$

M_e counts disagreements where recovery is ineligible in one of the two orders; M_h counts those on jointly eligible questions. This partition requires no random-selection assumption and does not attribute a causal fraction to parameter updates.

For $M > 0$, the fraction of disagreements outside the shared pool is $\alpha = M_e/M$. The reported values use integer counts. Appendix O relates this share to the two Jaccards and gives the overlap ceiling imposed by the shared eligible-pool size $I = |\mathcal{H}|$.

M.2 Most observed disagreements are outside the shared pool

Table 25 reports the eligible populations and observed overlaps. The two families have similar pool sizes across orders but little shared membership: 30.1% of Qwen3-8B's originally eligible questions and 19.3% of OLMo-2-7B's remain eligible after reordering.

For Qwen3-8B D8R, $a = 636$, $b = 625$, and $k = 109$, while $a_c = 201$ and $b_c = 176$. Equation 25 gives 1,043 recovery disagreements: 884 outside the shared eligible pool and 159 within it. Hence

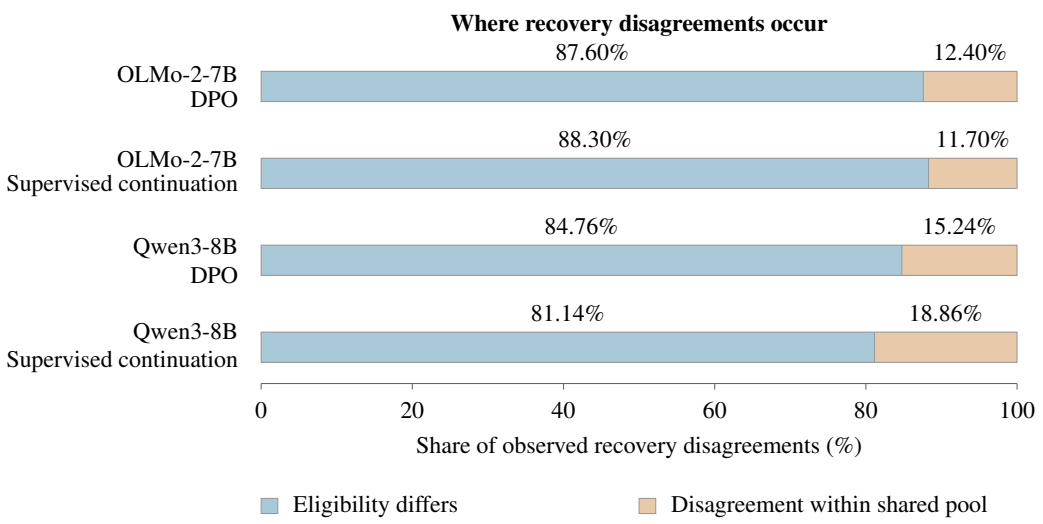


Figure 7: Most recovery disagreements concern questions not eligible under both orders. The two components partition each observed symmetric difference exactly. Outside-pool counts are 452/516 and 347/393 for OLMo-2-7B DPO and supervised continuation, and 884/1,043 and 710/875 for their Qwen3-8B counterparts.

$\alpha = 884/1043 = 84.76\%$. The partition identifies where the disagreements occur, whereas the rise from $J = .0946$ to $J_c = .4067$ alone does not. For supervised continuation, $a = 504$, $b = 485$, $k = 57$, $a_c = 147$, and $b_c = 132$, giving 710 disagreements outside the shared pool and 165 within it: $\alpha = 710/875 = 81.14\%$.

For OLMo-2-7B DPO, $a = 301$, $b = 273$, and $k = 29$, with $a_c = 65$ and $b_c = 57$. The partition is 452 eligibility disagreements and 64 shared-pool disagreements, giving $\alpha = 452/516 = 87.60\%$. OLMo-2-7B supervised continuation has $a = 216$, $b = 219$, $k = 21$, $a_c = 46$, and $b_c = 42$. Its corresponding counts are 347 and 46, giving $347/393 = 88.30\%$. Most disagreements lie outside the shared pool under both continuation objectives (Figure 7).

The conditional Jaccards, .4067 for Qwen3-8B and .3118 for OLMo-2-7B DPO, still leave disagreement within the shared populations. Appendix J compares observed overlaps with independent-selection references that preserve the eligible pools.

N Paired Bootstrap Intervals

We quantify question-level uncertainty by paired bootstrap resampling (Koehn, 2004; Dror et al., 2018). We resample questions 10,000 times with seed 42, retaining all model and order outcomes for each sampled question. For attenuation computed from Table 1, a question contributes $(Y_D(i,s) - Y_C(i,s)) - (Y_D(i,e) - Y_C(i,e))$. For a shared correction on the intersection, it contributes the mean of $Y_D - Y_C$ across the two orders. Keeping these outcomes together preserves their within-question dependence.

Table 26 gives percentile intervals for question-sampling uncertainty under exchangeability, conditional on each comparison's observed screening population, orders, checkpoints and fitted vectors. They exclude variation across training runs, presentations and calibration fits.

The attenuation intervals exclude zero for both screening orders in both families. These four intervals concern each fixed screening population; they do not test whether attenuation differs between screening orders. The remaining rows quantify recovery gaps on the two-order intersections, not attenuation. The smallest OLMo-2-7B shared-correction gap has a lower endpoint near zero.

Table 26: Conditional paired 95% bootstrap intervals, in percentage points. The first four rows concern attenuation when evaluation changes order on a single-order screening population. All remaining rows concern the DPO–supervised gap on the two-order intersection. Correction vectors are held fixed.

Family	Quantity / readout	Questions	Estimate	Interval
Qwen3-8B	Original-screen attenuation	1514	6.14	[3.83, 8.39]
Qwen3-8B	Reordered-screen attenuation	1496	5.35	[3.14, 7.69]
OLMo-2-7B	Original-screen attenuation	892	8.07	[5.38, 11.12]
OLMo-2-7B	Reordered-screen attenuation	862	4.64	[2.09, 7.41]
Qwen3-8B	Raw	455	10.77	[7.80, 13.85]
Qwen3-8B	Marginal: D8R	455	8.13	[5.71, 10.66]
Qwen3-8B	Marginal: C8	455	7.80	[5.49, 10.22]
Qwen3-8B	Paired: D8R	455	10.66	[7.91, 13.52]
Qwen3-8B	Paired: C8	455	8.24	[5.71, 10.88]
OLMo-2-7B	Raw	172	9.88	[6.10, 13.66]
OLMo-2-7B	Marginal: DO	172	8.72	[4.36, 13.08]
OLMo-2-7B	Marginal: CO	172	5.52	[2.03, 9.01]
OLMo-2-7B	Paired: DO	172	6.40	[2.33, 10.47]
OLMo-2-7B	Paired: CO	172	4.36	[0.29, 8.43]

O Derivations

Disagreement partition. Since $\mathcal{R}_o \subseteq \mathcal{P}_{R,o}$ and $\mathcal{R}_p \subseteq \mathcal{P}_{R,p}$, their intersection is contained in $\mathcal{H}$. A recovered question outside $\mathcal{H}$ is eligible in its own order and ineligible in the other, and can belong to only one recovered set. Every member of the symmetric difference lies either inside or outside $\mathcal{H}$, disjointly. Counting the two parts proves Equation 25.

Let $U = a + b - k$ and $U_c = a_c + b_c - k$. Then $M_e = U - U_c$ and $M = U - k$. For $k > 0$ and $J < 1$, $U_c/U = J/J_c$ gives

$$\alpha = \frac{M_e}{M} = \frac{1 - J/J_c}{1 - J}. \tag{26}$$

When $k = 0$, both Jaccards can be zero regardless of the value of M_e/M, so the Jaccards alone do not identify the share. When $M = 0$, the share is undefined because there are no disagreements. Direct counts avoid both limitations.

Overlap ceiling. Shared recoveries must satisfy $k \leq \min(a, b, I) = m$. Since $k/(a + b - k)$ is increasing in k for fixed nonempty recovery counts,

$$J \leq \frac{m}{a + b - m}. \tag{27}$$

This gives .5645 for Qwen3-8B and .4279 for OLMo-2-7B DPO at the observed recovery counts, above their measured Jaccards of .0946 and .0532. It bounds overlap at fixed recovered-set sizes and eligible pools; it is neither a null expectation nor a stability threshold.

Comparison scale. Solving $\theta = [p_D/(1 - p_D)]/[p/(1 - p)]$ gives $p_D = \theta p/(1 - p + \theta p)$. Subtracting the control success rate p gives

$$p_D - p = \frac{(\theta - 1)p(1 - p)}{1 + (\theta - 1)p}. \tag{28}$$

Both families' absolute and relative comparisons are computed from integer counts and pool sizes, rather than rounded table fractions.

P Predicting the Screening Effect from Earlier Correctness

We ask how closely earlier correctness states and continuation success rates reproduce the pooled attenuation. This exploratory analysis uses the existing Qwen3-8B outputs and was designed after

observing the crossing. Each of the 14 categories is held out in turn, with both presentations of every question assigned to the same fold.

For evaluation order e, define $Z_e(i) = (Y_B(i,e), Y_S(i,e)) \in \{0,1\}^2$. On the other categories, estimate the mean correctness of $T \in \{D, C\}$ separately for each state z, denoted $\widehat{p}_{T,e}^{(-c)}(z)$ for held-out category c. Each correctness-state group used to estimate these rates contains at least 363 questions. For a question in category $c(i)$, predict correctness using $\widehat{p}_{T,e}^{(-c(i))}(Z_e(i))$. No D/C answer from the held-out category enters these estimates. Held-out Base/SFT answers under the evaluation order do enter the predictor; this is not a forecast made before that presentation is evaluated.

We average the predictions over the same $\mathcal{H}_s$ populations as Table 1, form the predicted DPO–supervised gap under each order, and subtract the other-order gap from the same-order gap. The procedure does not fit a coefficient to the observed decrease in advantage. Table 27 compares these estimates with the measured gaps.

Table 27: Qwen3-8B gap prediction from other categories' Base/SFT-state response rates. Same and other refer to evaluation relative to the screening order. Values are percentage points; predictions are estimated rates, not additional measured correct counts.

Screen	Source	Same gap	Other gap	Attenuation
Original	Observed	8.72	2.58	6.14
	Predicted	8.47	2.76	5.70
Reordered	Observed	9.36	4.01	5.35
	Predicted	9.01	2.50	6.51

Prediction is much less accurate within categories than in the pooled result. Across the 28 category–screening comparisons, weighting each by its screening-pool size, the mean absolute attenuation error is 6.33 points and root mean squared error is 7.49. Predicting zero attenuation gives 6.48 and 8.49, respectively. Predicted and observed attenuation have the same nonzero sign in 19 of 28 comparisons. The screening populations overlap across directions; these are not 28 independent experiments.

Conditioning continuation success rates on earlier correctness approximates the overall decrease in advantage, but not its variation across categories. Matching the overall result therefore falls short of explaining individual continuation responses. The calculation estimates neither a causal fraction explained nor performance on an unseen presentation.

Q Parameter-Exchange Controls and Content Diagnostics

Table 28 lists the five controls underlying Table 4. Because Equation 7 is nonlinear, we calculate proximity for each control before averaging. Each observed exchange effect exceeds the five control point estimates; uncertainty for Net is calculated by paired resampling of questions.

Table 28: Matched-control donor-proximity changes in percentage points. $C \leftarrow D$ denotes a supervised recipient and DPO donor within the same family; Qwen3-8B D denotes D8R.

Task	Family	Direction	Control 1	Control 2	Control 3	Control 4	Control 5
MMLU-Pro	OLMo-2-7B	$C \leftarrow D$	−.420	.280	.840	.560	1.541
		$D \leftarrow C$	−.140	.700	.280	.980	.980
	Qwen3-8B	$C \leftarrow D$	−.330	.743	.248	1.321	.908
		$D \leftarrow C$	−.495	.661	.330	1.073	.083
HellaSwag	OLMo-2-7B	$C \leftarrow D$	−.304	.456	1.026	.190	1.292
		$D \leftarrow C$	−.456	1.140	.722	.266	.798
	Qwen3-8B	$C \leftarrow D$	−.391	.854	.427	1.388	1.103
		$D \leftarrow C$	−.178	1.530	.320	.783	.747

Control effects are rounded to three decimal places. Table 4 gives their means and the exchange effects minus those means (Net).

Parameter replacement and cache handling. For every comparison in Table 4, donor W_Q and W_O replace recipient projections only at the final non-padding input token in every attention layer. W_O denotes the attention output projection, not the vocabulary head. Prefix keys and values are computed with recipient parameters and reused across intervention conditions. All other weights remain those of the recipient. Current-token states can still change with recipient W_K and W_V. Reusing the prefix cache does not fix which cached information the new queries retrieve. Each of the five controls sign-scrambles donor-minus-recipient differences at the same parameter locations, preserving each tensor's perturbation norm.

Same-question content diagnostic. Let $z_{i,p}^{\mathrm{pre}}$ and $z_{i,p}^{\mathrm{swap}}$ denote the candidate-content identities selected by the label-free diagnostic before and after the exchange for question i and presentation p. The change indicator is

$$d_i = \mathbf{1}\left\{\exists p \in \{\mathrm{Move}, \mathrm{Reassign}\} : z_{i,p}^{\mathrm{swap}} \neq z_{i,p}^{\mathrm{pre}}\right\}. \tag{29}$$

Each question contributes once, including when both selections are wrong. The diagnostic scores candidate texts directly rather than mapping emitted letters back to content. Table 4 reports the counts with complete coverage, giving change rates $100 \sum_i d_i/n$ of .280%, .140%, .330%, .495%, .418%, .608%, .463% and .641%.

Content-scoring implementation. All eight content comparisons in Table 4 use the same scoring procedure. It removes letter prefixes from the target question's candidate lines, retaining their displayed order and appending `Answer:`; the header and demonstrations are retained. Each candidate is tokenized as a leading space followed by its stripped text, without added special tokens. Its score is the mean log probability of its candidate tokens. The first token is scored from the intervened final prompt-token logits. For later candidate tokens, recipient parameters are restored and the candidate prefix is teacher-forced using a separate deep copy of the intervened cache, including the final prompt token. Ties are resolved by the larger MD5 hash of the candidate text; correctness compares selected and gold content texts.

The intervened cache carries the exchange's effect into candidate scoring after recipient parameters are restored. Move and Reassign retain different candidate orders in this diagnostic.

What the content diagnostic measures. String likelihood can distribute probability across alternative expressions of the same answer (Holtzman et al., 2021), whereas joint question-and-option prompting requires binding an answer to its output symbol (Robinson et al., 2023). Multiple-choice and long-form answers can disagree on the same questions (Li et al., 2024b); auxiliary task demands also affect measured performance (Hu & Frank, 2024). Generated answers introduce an extraction step, whose errors motivate evaluators such as xFinder (Yu et al., 2025). Our diagnostic scores supplied candidate strings directly. It avoids extracting a generated answer but remains a separate prompted task, not a direct measurement of unchanged internal knowledge.

For k changes among n questions, the one-sided Clopper–Pearson upper bound u solves $\Pr_{X\sim\mathrm{Binomial}(n,u)}(X \leq k) = .05$. This assumes independent, identically distributed Bernoulli outcomes within each comparison. The counts describe these questions directly; the bounds depend on the sampling assumption and do not cover other training runs or diagnostics. Values below 1% were not a preregistered equivalence criterion and do not establish invariant internal content representations.

Uncertainty in net transfer. For Table 4, we resample questions with replacement 5,000 times using seed 42, retaining each question's two presentations and all unmodified, exchanged and five-control outcomes. Each resample recomputes the response differences, absolute-distance proximities in Equation 7, and control-adjusted Net. The fifth percentile gives the one-sided 95% lower bound. The bounds condition on the fixed recovery populations, models, presentations and controls; they are not simultaneous intervals or estimates of training-run variation. This procedure is separate from the 10,000-resample retest and shared-correction analysis in Appendix N and the binomial content-change bounds.

Development subsets and the main evaluation. On smaller MMLU-Pro development pools, Qwen3-8B ($n = 300$) gave Net $-.07$ and .67 for $C \leftarrow D$ and $D \leftarrow C$, with reported lower/upper

bounds $[-.80, .80]$ and $[-2.00, 2.68]$. OLMo-2-7B ($n = 163$) gave .61 and .00, both with bounds $[-1.23, 1.84]$. All include zero. This development check stopped before evaluating HellaSwag. In the same direction order, the main Qwen3-8B estimates are 1.82/2.31 on 1,211 questions and the main OLMo-2-7B estimates are 1.96/2.52 on 714 questions. The main point estimates are larger in all four comparisons.

The MMLU-Pro sampling populations are the original-order Base-correct/SFT-wrong pools: 892 OLMo-2-7B and 1,514 Qwen3-8B questions. Within each pool, we first sort questions by their original IDs, then shuffle with seed 42 and retain the first 80%, rounding the sample size to the nearest integer. This gives 714 OLMo-2-7B and 1,211 Qwen3-8B questions. The 80% fraction applies within each recovery pool, not to the full test set. The regenerated question IDs match the original evaluation lists. Both transfer directions, the unmodified models and all five controls use these same fixed lists.

Without reshuffling, we take the first 163 OLMo-2-7B and 300 Qwen3-8B questions from these main lists for development checks. These sizes and the selection rule were fixed before inspecting exchange outcomes. The development questions are therefore contained in the main populations, rather than forming the remaining 20% of the recovery pools. Both evaluations use D8R/C8 for Qwen3-8B and DO/CO for OLMo-2-7B, with the same exchange position, Q/O scope, cache handling, five-control construction and statistical procedure. Reevaluating the development questions with the main experiment's procedure gives the same question-level predictions and subset-level Net estimates.

The main evaluation adds 911 Qwen3-8B and 551 OLMo-2-7B questions. Neither main nor development selection conditions on continuation advantages or exchange effects. Because the populations are nested, the main evaluation is not an independent replication, and comparing their intervals does not test the difference between them.

The development content diagnostic exceeded the per-question uniform-choice reference before intervention. Its maximum absolute accuracy changes were .37 points for Qwen3-8B and .21 for OLMo-2-7B, with identity-change rates of 1.00% and 1.21%. Above-chance accuracy establishes task signal, not sensitivity to every content change. These development rates and the main rates in Table 4 describe different populations.

Combining the development and added subsets. Fix a transfer direction and the exchange and control operators. Write the main population as the disjoint union $P = S \cup E$, where S is the development subset, $E = P \setminus S$ contains the added questions, and $w = |S|/|P|$. For an exchanged recipient or matched control k, let $u_k(G) = r_k(G) - r_{\text{donor}}(G)$ be its signed response difference from the unmodified donor on population G. Since r is an accuracy difference, $u_k(P) = wu_k(S) + (1 - w)u_k(E)$. Subtracting the mean control proximity from exchange proximity cancels the unchanged-recipient distance:

$$\text{Net}(G) = \frac{1}{5}\sum_{j=1}^{5} |u_j(G)| - |u_{\text{swap}}(G)|. \tag{30}$$

Here j indexes the five matched controls. Define $J_k = w|u_k(S)| + (1 - w)|u_k(E)| - |u_k(P)| \geq 0$, the amount of cancellation between the two subsets' signed discrepancies. Substitution gives

$$\text{Net}(P) = w\,\text{Net}(S) + (1 - w)\,\text{Net}(E) + J_{\text{swap}} - \frac{1}{5}\sum_{j=1}^{5} J_j. \tag{31}$$

Different subset responses and different cancellation under exchange and controls can both change pooled Net. If each operator's discrepancies have the same sign in the two subsets, all J_k vanish and Net does average linearly.

Development–main decomposition. Table 29 separates both families' results into the development subsets, added questions and the pooling term in Equation 31. In all four comparisons, added-subset Net exceeds the development estimate. Pooling subtracts .03303 points in both Qwen3-8B directions and .056022/.112045 points in the two OLMo-2-7B directions. The larger main point estimates therefore come from the added questions under the same implementation; pooling offsets a small part of each increase. This decomposition accounts for the differences in point estimates; it does not test their statistical significance or explain the underlying model mechanism.

Table 29: MMLU-Pro net-transfer decomposition (pp). Qwen3-8B uses 300 development and 911 added questions, with $w = 300/1211$; OLMo-2-7B uses 163 development and 551 added questions, with $w = 163/714$. Pool is the signed correction $J_{\text{swap}} - \frac{1}{5}\sum_j J_j$ in Equation 31. Main Net combines the weighted subset estimates and Pool and agrees with Table 4. Values are rounded for display.

Family	Direction	Development Net	Added Net	Pool	Main Net
Qwen3-8B	$C \leftarrow D$	−.06667	2.48079	−.03303	1.81668
	$D \leftarrow C$	.66667	2.89791	−.03303	2.31214
OLMo-2-7B	$C \leftarrow D$	.61	2.431942	−.056022	1.96
	$D \leftarrow C$	.00	3.411978	−.112045	2.52

Earlier broader-population summary. Table 30 reports the earlier diagnostic on 9,626 MMLU-Pro and 10,042 HellaSwag questions. $A_{\max}$ and $F_{\max}$ are the larger absolute accuracy change and content-identity change rate across directions. These populations differ from the later recovery-only diagnostic and from the earlier matched-display evaluation.

Table 30: Earlier broader-population content diagnostic. n_R/n_C gives recovery/content-evaluation sizes. κ is recomputed on n_R: the 80% MMLU-Pro intervention subsets, rather than the full pools in Table 3; HellaSwag retains its full pools. Effects are percentage points except $F_{\max}$ (%). The content rates use n_C, not the recovery-only denominator of Table 4.

Task	Family	n_R/n_C	κ	$C \leftarrow D$	$D \leftarrow C$	$A_{\max}$	$F_{\max}$
MMLU-Pro	OLMo-2-7B	714/9626	4.20	1.96	2.52	0.15	0.97
	Qwen3-8B	1211/9626	3.88	1.82	2.31	0.17	1.24
HellaSwag	OLMo-2-7B	2632/10042	3.99	1.71	2.58	0.24	1.57
	Qwen3-8B	2810/10042	4.63	1.74	3.02	0.13	1.02

R Attention-Aggregate and Output-Projection Replay

The joint Q/O exchange can change both the information aggregated by attention and how that aggregate is transformed by the output projection. To distinguish their conditional effects, we replay these quantities separately on the same 714 OLMo-2-7B and 1,211 Qwen3-8B MMLU-Pro recovery questions, in both transfer directions.

Relation to activation interventions. Causal tracing restores selected states after input corruption to test their contribution to factual predictions (Meng et al., 2022); attention knockout instead tests pathways through which information reaches the prediction position (Geva et al., 2023). Causal abstraction asks whether a proposed high-level computation agrees with the network under corresponding interventions (Geiger et al., 2025). A behavioral effect alone is weaker evidence: subspace patches can activate dormant pathways and mimic a hypothesized feature change (Makelov et al., 2024). Our full-vector replay tests behavior under specified manipulated states. As with these interventions, an observed response alone does not provide a complete computational explanation of recovery.

Four replay conditions. For each question and presentation, $\mathbf{z}_0$ and $\mathbf{z}_1$ collect the final-token inputs to all attention output projections in the unmodified recipient and during full Q/O exchange, respectively. We record each vector after combining attention-head outputs, not from a separate donor forward pass. Let O_0 and O_1 denote recipient and exchanged output-projection parameters.

The four conditions are $(O_0, \mathbf{z}_0)$, $(O_1, \mathbf{z}_1)$, $(O_1, \mathbf{z}_0)$ and $(O_0, \mathbf{z}_1)$. Every layer receives its recorded final-token aggregate; residual updates, normalization and MLP computation proceed normally. Each condition uses an independent copy of the unchanged recipient prefix cache. The first two conditions reproduce the unmodified and full-exchange readouts. The cross conditions vary the projection or aggregate separately. Each matched control uses its own projections and recorded aggregate trajectory in the same construction.

Directed response effects. Let r_{ab} be the Move-minus-Reassign accuracy difference under $(O_a, \mathbf{z}_b)$, in percentage points. Define

$$\Delta_{\text{full}} = r_{11} - r_{00}, \qquad \Delta_{\text{fixed}} = r_{10} - r_{00}, \qquad \Delta_{\text{aggregate}} = r_{01} - r_{00}. \tag{32}$$

We fix $s = \text{sign}(r_{\text{donor}} - r_{\text{recipient}})$ from the unmodified models, using the same direction for the exchange and all controls. For each $q \in \{\text{full}, \text{fixed}, \text{aggregate}\}$, the directed, control-adjusted effect is

$$E_q = s \left(\Delta_q^{\text{swap}} - \frac{1}{5} \sum_{j=1}^{5} \Delta_q^{(j)} \right). \tag{33}$$

The interaction is $I = E_{\text{full}} - E_{\text{fixed}} - E_{\text{aggregate}}$. This is a decomposition on the accuracy-difference scale, not a partition of semantic processing. In particular, thresholded answer selection can produce nonadditivity.

Unlike Net (Equation 30), E_q measures signed response change, not reduction in absolute donor distance. Crossing the donor's value can make the two quantities differ, so their estimates and bounds are not interchangeable.

Table 31: Replay on fixed MMLU-Pro recovery questions (percentage points). Full exchanges projections and aggregates; Fixed exchanges projections with recipient aggregates; Aggregate retains recipient projections with exchanged aggregates. These three effects and their interaction I are directed and control-adjusted. In contrast, $\delta r = r_{11} - r_{10}$ is the unadjusted full-minus-fixed response difference; its interval is a two-sided, question-paired 95% interval.

Family	Direction	n	Full	Fixed	Aggregate	I	δr	95% interval
OLMo-2-7B	$C \leftarrow D$	714	1.961	1.989	.084	$-.112$	.140	$[-.350, .420]$
	$D \leftarrow C$	714	2.521	2.297	.112	.112	$-.140$	$[-.420, .350]$
Qwen3-8B	$C \leftarrow D$	1211	1.784	1.718	.050	.017	.083	$[-.165, .330]$
	$D \leftarrow C$	1211	2.312	2.263	.066	$-.017$	$-.083$	$[-.330, .165]$

Paired response comparison. For $\delta r = r_{11} - r_{10}$, we resample questions 5,000 times with seed 42, retaining both presentations and replay conditions together. The 2.5th and 97.5th percentiles, with linear interpolation, give pointwise intervals conditional on these checkpoints and presentations. Neither control subtraction nor multiplication by s enters this comparison.

Results and scope. Aggregate-only effects are smaller than fixed-aggregate effects in every comparison. Absolute unadjusted full-minus-fixed differences are about .14 points or less, with all four intervals contained within approximately $\pm.42$ points. The small full-minus-fixed differences therefore do not arise solely from control subtraction. This concerns group-level responses, not identical per-question answers or a prespecified equivalence test.

Positive directed effects persist with recipient aggregates fixed. These effects do not establish that aggregates represent content or that output projections perform letter conversion. Feedback from earlier projection changes makes the aggregate-only condition distinct from a pure query intervention. Activation-patching conclusions can depend on the metric and intervention construction (Zhang & Nanda, 2024). Here, forced replay may create unnatural states and does not explain the training origin of the differences. These results concern MMLU-Pro; HellaSwag supports only the joint-exchange comparison.

R.1 Original–Reordered attenuation

We also apply the replay conditions to Original and general Reordered inputs on the same 714 OLMo-2-7B questions, with CO as recipient and DO as donor. The donor remains unmodified. For recipient condition h, define

$$a_h = 100\big[(\text{acc}_{D,o} - \text{acc}_{C,h,o}) - (\text{acc}_{D,p} - \text{acc}_{C,h,p})\big],$$
$$e_h = \overline{a}_{\text{ctrl},h} - a_h,$$

where o, p denote Original and Reordered, accuracies are fractions, and $\overline{a}_{\mathrm{ctrl},h}$ averages attenuation over the five corresponding controls. Thus e_h measures a control-adjusted reduction in the Original–Reordered attenuation, distinct from the Move–Reassign effect E_q above.

Table 32: Original–Reordered attenuation under replay for OLMo-2-7B $C \leftarrow D$ on 714 fixed MMLU-Pro recovery questions. The two C columns give correct-answer counts; a_h, $\overline{a}_{\mathrm{ctrl},h}$ and e_h are percentage points, rounded to four decimal places. Full, Fixed and Aggregate use the replay conditions defined above. Fixed minus Aggregate is 2.4370 points, with a two-sided paired 95% interval of $[1.5098, 3.3613]$.

Condition	C Original	C Reordered	a_h	$\overline{a}_{\mathrm{ctrl},h}$	e_h
Unmodified	170	245	8.1232	—	—
Full	200	251	4.7619	8.0112	3.2493
Fixed	196	250	5.1821	7.8711	2.6891
Aggregate	174	247	7.8431	8.0952	0.2521

The 8.1232-point baseline uses this 714-question subset, rather than the full 892-question pool in Table 1. With recipient aggregates fixed, the adjusted reduction is 2.6891 points, compared with 0.2521 for aggregate replacement. Their difference, $e_{\mathrm{fixed}} - e_{\mathrm{aggregate}}$, is 2.4370 points with a two-sided paired 95% interval of $[1.5098, 3.3613]$. This supports a larger attenuation reduction under fixed-aggregate replay in this OLMo-2-7B transfer direction, not an explanation of how training produced the difference.

For this interval, we resample the 714 questions with replacement 5,000 times using seed 42. Each draw retains both presentations, both replay conditions and each condition's five controls for every sampled question. We recompute both control-adjusted reductions and their difference in each resample, then take the 2.5th and 97.5th percentiles with linear interpolation. The interval concerns question-level variation for these fixed checkpoints, presentations and controls; it is distinct from the Move–Reassign full-minus-fixed intervals in Table 31.

S Training Stage and Exploratory Score Changes

We first document the trained intermediate checkpoint, then test three possible relationships between the continuations' score changes using existing Qwen3-8B outputs. The checkpoint uses the ordinary readout; the retrospective analyses construct alternative score vectors without training a model. All use the same 12,032 MMLU-Pro questions and two presentations, with no new held-out permutation.

S.1 Intermediate-checkpoint record

D8R-312 is the checkpoint after 312 optimizer updates in the D8R run starting from S8, not the earlier D8 run. The reported validation change count is 16/84: 42 validation questions, each evaluated under Original and Reordered, with a change defined relative to S8's prediction for that question and order. The 42-question subset excludes validation targets present in their own demonstrations.

The record gives the selected step and its change count, but omits C8's validation change count and the full candidate-selection history. We therefore treat the checkpoint comparison as descriptive; the record does not establish validation matching or prospective selection. Its test change rates are close to C8's (Table 2).

For example, on $\mathcal{H}_o$, the intermediate-checkpoint gaps are $100(508 - 504)/1514$ under Original and $100(648 - 651)/1514$ under Reordered. On $\mathcal{H}_p$, they are $100(488 - 485)/1496$ and $100(654 - 658)/1496$ under Same and Other. The two pools overlap and are not independent replications. The intermediate checkpoint's change counts within these pools are not reported, so equal global change rates must not be interpreted as equal pool-specific change rates. We do not infer statistical equivalence or paired intervals from the aggregate counts.

Table 33: Qwen3-8B checkpoint comparisons on fixed screening pools. Same and Other denote evaluation under the screening order and the other order. All gaps are DPO minus C8 in percentage points; Contrast is Same minus Other. Both checkpoints use the identical question sets.

DPO checkpoint	Screen	n	Same gap	Other gap	Contrast
D8R endpoint	Original	1514	8.72	2.58	6.14
	Reordered	1496	9.36	4.01	5.35
D8R-312	Original	1514	0.26	−0.20	0.46
	Reordered	1496	0.20	−0.27	0.47

S.2 Can a score change transfer between presentations?

For question i with K_i valid options, define the centered log score

$$z_{T,i,e,k} = \log \max(q_{T,i,e,k}, 10^{-12}) - \frac{1}{K_i} \sum_{j=1}^{K_i} \log \max(q_{T,i,e,j}, 10^{-12}),$$
$$u_{T,i,e} = z_{T,i,e} - z_{S,i,e}. \tag{34}$$

Here q is the space-prefixed valid-letter probability vector, and T is D8R or C8. To predict the continuation under target presentation e_t from source e_s, we add a transformed source change to z_{S,i,e_t}. Letter transfer keeps the score change at its letter slot; content transfer permutes it with the corresponding candidate content. A mean-letter baseline averages each slot's change over the other 13 categories in the source presentation, using only questions where that slot is valid, then centers the resulting vector over the target's valid options. The no-change baseline retains S8's target scores. Neither transfer uses the target continuation's output to construct its prediction.

Table 34: Predicted D8R–C8 gaps on source-screened recovery questions, evaluated under the target presentation (pp). Mean letter is fitted on other categories. Retaining S8 for both continuations necessarily predicts zero gap. These are retrospective predictions, not new model evaluations.

Source → target	n	Observed	Letter	Content	Mean letter	No change
Original → Reordered	1514	2.58	−1.52	4.56	1.19	0.00
Reordered → Original	1496	4.01	−3.68	5.08	0.74	0.00

Content transfer has lower score-change RMSE than letter transfer: .625 versus .941 for D8R and .286 versus .440 for C8 in either direction. It nevertheless misses the method gap (Table 34). The unchanged-S8 baseline agrees with 79.1/80.1% of D8R answers and 86.3/86.8% of C8 answers across the two target orders; content transfer agrees with 77.5/77.7% and 86.0/85.4%. Thus, lower score error does not imply better prediction of the answers. Neither transfer rule reproduces the observed comparison.

S.3 Are DPO score changes scaled supervised changes?

For each order and each held-out category c, we fit a scalar on the other categories,

$$a_{-c} = \frac{\sum_{i \notin c} u_{C,i}^{\top} u_{D,i}}{\sum_{i \notin c} \|u_{C,i}\|_2^2}, \qquad \widehat{z}_{D,i} = z_{S,i} + a_{-c} u_{C,i} \quad (i \in c). \tag{35}$$

The fit uses score changes, not gold answers. Each category's DPO scores are withheld from its fit. We assess the fit by the fraction $1 - \sum_i \|u_{D,i} - a_{-c(i)} u_{C,i}\|^2 / \sum_i \|u_{D,i}\|^2$ of squared score-change magnitude explained, without centering across questions. Table 35 reports the pooled cosine and the held-out predictions.

Fitting on all categories for a descriptive pooled coefficient gives 1.978 and 1.988. Held-out predicted full-test accuracies are 45.42% and 44.57%, versus D8R's actual 44.29% and 43.43%. Despite the similar score changes, the scalar predictor overpredicts recovery by 165 and 159 questions. The explained score fraction is not an explained fraction of recovery or a causal attribution to a training objective.

Table 35: Scalar prediction of D8R's score change from C8's. Explained and Agreement are percentages; Agreement compares predicted and actual D8R answers over all 12,032 questions. The last two columns use each order's own recovery pool.

Order	Fitted a range	Cosine	Explained	Agreement	Predicted R	Actual R
Original	1.958–2.016	.903	81.42	90.14	801	636
Reordered	1.969–2.023	.904	81.65	90.26	784	625

S.4 Matching answer-change frequency by scaling scores

We next choose a nonnegative scale using changes in predicted answers rather than score regression. For $T \in \{D8R, C8\}$, construct $z_T(a) = z_S + a(z_T - z_S)$. Within each held-out category, one scale is fitted on all other categories, pooling both orders and keeping a question's two presentations in the same fold. We either reduce D8R's change frequency to C8's, or increase C8's to D8R's. Gold answers, recovery membership and correctness gaps do not enter scale selection.

For S8's winning option k, the first-change threshold is the smallest nonnegative crossing $(z_{S,k} - z_{S,j})/(u_{T,j} - u_{T,k})$ over competitors with a positive denominator, or infinity if none exists. Along this linear score path, once a competitor exceeds S8's original winner, that winner cannot regain the maximum. We choose the midpoint between consecutive sorted thresholds bracketing the target number of changes on the fitting categories. Fits attain the target count exactly; held-out counts need not match. This is a retrospective change to the readout, not interpolation of model weights or evaluation of a trained intermediate model.

The fitted scales range from .5155 to .6167 for reduced D8R and 1.5514 to 1.8651 for amplified C8. Tables 36 and 37 give the held-out predictions and comparisons. Under Original/Reordered, reduced D8R changes 13.44/14.38% of answers, versus C8's 13.17/13.71%; amplified C8 changes 20.04/20.05%, versus D8R's 19.93/20.91%. Change frequencies are close on the evaluated categories, but not equal.

Table 36: Held-out-category score-scaling predictions. Changed and All correct have denominator 12,032. The last columns retain the original fixed pools of 1,514 and 1,496 questions. Unmodified endpoint counts appear in Table 2.

Modified readout	Evaluate	Changed	All correct	$\mathcal{H}_o$ correct	$\mathcal{H}_p$ correct
Reduced D8R	Original	1617	5148	438	646
	Reordered	1730	5021	641	437
Amplified C8	Original	2411	5401	738	741
	Reordered	2412	5278	729	710

Table 37: DPO–supervised gaps under exploratory score scaling (pp). Same/Other refer to evaluation relative to the named screening order; Contrast is Same minus Other. A negative gap favors the supervised comparator under the specified readouts.

Comparison	Screen	Same gap	Other gap	Contrast
Unmodified D8R versus C8	Original	8.72	2.58	6.14
	Reordered	9.36	4.01	5.35
Reduced D8R versus C8	Original	−4.36	−0.66	−3.70
	Reordered	−3.21	−0.80	−2.41
D8R versus amplified C8	Original	−6.74	−2.58	−4.16
	Reordered	−5.68	−1.54	−4.14

The point-estimate ranking reverses in both scaling directions and both screening pools. Amplified C8 also exceeds unmodified D8R's full-test correct counts, so this reversal does not require degrading DPO's predictions. However, these readouts use scores from both S8 and a continuation; they are not trained models. Scaling changes both score magnitudes and which questions cross a decision boundary. Together, the scaling result and the failed transport and scalar-prediction checks show why answer-change frequency should be reported, but do not establish its role as a training mechanism.