Evaluating Monitor Reuse After Language-Model Updates: Ranking Targets and Audit Evidence

Anonymous authors
Paper under double-blind review

Abstract

After a language-model update, should a monitor that predicts its behavior be kept or retrained? Our monitors predict whether a reasoning problem and an intended answer-preserving rewrite receive different answers. We compare a training-stage change with a within-stage revision in OLMo-3 7B and Tülu 8B. Three measurements bear on the decision, and each can mislead alone. *Absolute quality*: pooled AUROC rewards recognizing the benchmark. In Tülu, a benchmark-only predictor reaches pooled AUROC .684 with within-benchmark (macro) AUROC exactly .5 by construction, and a hidden-state probe with macro AUROC .701 retains a pooled–macro gap of .0735 (fixed-score 95% interval [.0296, .1323]). *Relative comparisons*: on fixed OLMo predictions, the major-minus-revision refitting advantage is −.0099 under pooled and +.0186 under macro AUROC. An exact decomposition attributes the difference to ranking pairs and weights; resampling leaves the population sign pattern unresolved. *Audit evidence*: with 25 labeled template groups the specified OLMo audit returns no quality bound, so its zero decision error reflects default distrust. The reused Tülu probe retains macro AUROC .689 against .723 after refitting, yet 50-group audits of its five saved fits approve reuse in at most 24% of samples. Removing the audit's sample partitioning returns more bounds and approvals, including false approvals against finite held-out references. We recommend reporting absolute performance under a stated ranking target, and bound availability, together.

1 Introduction

After a language model is updated, a predictor of its behavior may also need updating. We study a specific predictor: a monitor that uses internal features and output statistics to estimate whether the model will give different answers to a reasoning problem and an intended answer-preserving rewrite. Keeping the old monitor avoids fitting a replacement, but its performance must be assessed on the updated model's behavior. Our question is what the available performance measurements establish about that choice. Whereas monitor-staleness work tests whether updating a model requires repairing its monitor (Duan, 2026), we study how the evaluation target and the availability of audit evidence affect the maintenance comparison.

We compare old and newly fitted monitors on the same updated examples to measure the advantage of retraining. Retraining may help after either a change of training stage or further training within the same stage. We therefore include both update types in OLMo-3 7B and Tülu 8B. The within-stage comparison supplies a reference, without assuming zero change or matched update magnitude.

Three measurements bear on this decision, and each can mislead when read alone. The first is the monitor's absolute quality. A monitor used to prioritize examples from several tasks faces a different ranking problem from one evaluated within each task, and ROC analysis already distinguishes pooled from within-group performance (Janes et al., 2009). In Tülu, a predictor that knows only the benchmark reaches pooled AUROC .684 while ranking at chance within each benchmark, and a hidden-state probe with within-benchmark AUROC .701 still scores .0735 higher when pooled (Figure 1, left).

The second is a relative comparison between updates. On fixed OLMo predictions, pooled evaluation assigns the larger retraining benefit to the within-stage revision, while benchmark-macro evaluation assigns it to the stage change (Figure 1, right). An exact decomposition accounts for the difference,

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.

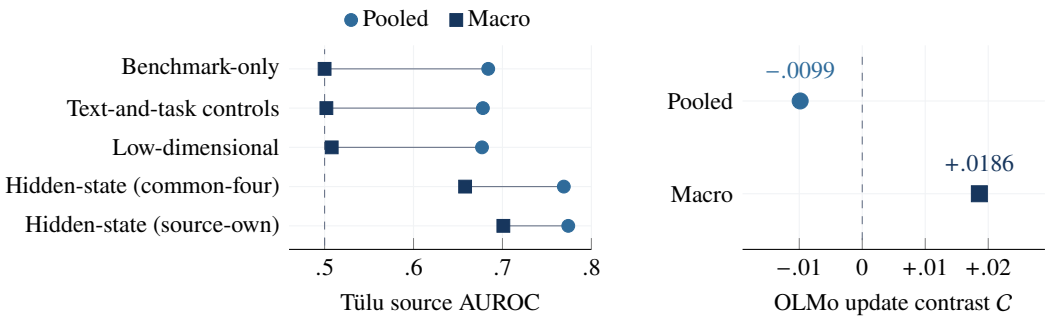


Figure 1: Left: Tülu source AUROC on the same 1,196 TEST pairs (Table 3). A benchmark-only predictor scores .684 pooled and .5 within benchmarks; the hidden-state probes keep a pooled–macro gap. Right: on fixed OLMo predictions (the same 1,361 out-of-fold (OOF) rows, 1,139 groups, scores, labels, folds, and revision monitor), the refitting advantage of the major update (training-stage change) minus that of the revision is negative under pooled and positive under macro AUROC. Pooled AUROC includes cross-benchmark rankings and row-weights folds; macro AUROC uses within-benchmark rankings, group-weights folds, and weights benchmarks equally. No intervals are drawn; the probes' fixed-score intervals are in Table 33. C is the update contrast of Section 2.2; common-four and source-own name the source-training eligibility rules of Section 3.1.

and the signs persist under common fold weights. Resampling leaves the population sign pattern unresolved, and in Tülu both evaluations' point estimates favor the stage change. We therefore treat the reversal as a worked example of ranking-target dependence, not as a general law.

The third is audit evidence. The Tülu probe keeps macro AUROC .689 when reused after the stage change, against .723 after refitting, yet 50-group audits of its five saved source fits approve reuse in at most 24% of samples, and the specified OLMo audit returns no quality bound at all with 25 groups. Here approval is governed by what 50 labeled groups can certify near the .60 requirement, not by a measured loss of discrimination.

The study contributes evidence for each measurement:

- **Absolute quality with a benchmark-only check.** On identical Tülu evaluation rows we separate a benchmark-only baseline, text-and-task controls and hidden-state probes, and report source, reused-target and refitted-target performance.
- **Ranking-target dependence of update comparisons.** Fixed OLMo predictions give opposite update contrasts under pooled and macro evaluation; a signed decomposition, common fold weights and task-weight sensitivity identify the responsible comparisons.
- **Evidence availability in reuse audits.** Paired audits of saved Tülu monitors separate approvals that follow previously unavailable bounds from changes in finite endpoints, counting new, revoked and false approvals against one finite-set reference. The OLMo audit and known-quality controls show why decision error alone misses unavailable evidence.

2 Study Setup and Measurement Targets

2.1 Prediction task and model updates

At each language-model checkpoint, one observation pairs a reasoning problem with an intended answer-preserving variant. We set $Y_{\text{inc}} = 1$ when their parsed final answers disagree. The monitor predicts this binary outcome from features of the problem pair and the checkpoint. Thus the label measures consistency within a checkpoint, not disagreement between checkpoints or whether either answer is correct. The held-out split has 1,230 template groups from GSM-Plus (Li et al., 2024) and InvariantBench (Wasi et al., 2026). A template group contains an original template and its variants; these remain together when splitting and resampling data.

We use two update comparisons within each model family. A *training-stage change*, or major update, moves from direct preference optimization (DPO), which learns from preferred versus dispreferred responses, to reinforcement learning with verifiable rewards (RLVR), which uses rewards that can be checked against the task outcome. A *within-stage revision* compares two saved versions (checkpoints) from the later training stage. Tülu uses group relative policy optimization (GRPO) in its RLVR stage (Allen Institute for AI, 2025):

Model family	Training-stage change	Within-stage revision
OLMo-3 7B Think	DPO → final RLVR	RLVR step 25 → step 50
Tülu 8B	DPO release → GRPO step 1920	GRPO step 1840 → step 1880

The revision provides an empirical same-stage reference: the same performance quantity measures the advantage of retraining after each update. Revisions can change behavior, and the selected revisions are not magnitude-matched to the stage changes. This comparison therefore does not isolate the causal effect of changing the training stage.

Within a transition, *source* and *target* name its earlier and later checkpoints. A source monitor is fitted using source outcomes; a target monitor is fitted using target outcomes. Both are evaluated on the target test groups. The OLMo absolute-performance and aggregation analyses retain 1,361 common-eligible out-of-fold (OOF) rows: each saved prediction comes from a fit that excludes that row's test group. Table 1 reports outcome prevalence on each family's common TEST pairs; these are pair-level rates, not proportions of template groups. The audits use separate evaluation populations, summarized with their scorers in Table 2.

Table 1: Answer-inconsistency prevalence on common four-checkpoint TEST pairs. Counts and percentages describe the DPO source and major-update target (OLMo final RLVR; Tülu GRPO step 1920). Each family retains its own common eligible population.

Family	Benchmark	Source	Target
OLMo	GSM-Plus	86/984 (8.74%)	59/984 (6.00%)
OLMo	InvariantBench	82/377 (21.75%)	69/377 (18.30%)
Tülu	GSM-Plus	157/949 (16.54%)	141/949 (14.86%)
Tülu	InvariantBench	154/247 (62.35%)	155/247 (62.75%)

The OLMo study began with a prespecified diagnostic protocol; its five criteria for association, transition specificity, localization, repair and high-consensus support were not met. The aggregation, Tülu, audit and artificial-feature analyses are post-confirmatory follow-ups, not revisions of those decisions.

2.2 What the performance measures compare

The area under the receiver operating characteristic curve (AUROC) measures ranking: it is the probability that an inconsistent pair receives a higher monitor score than a consistent pair, with half credit for ties. An AUROC of .5 is the chance-ranking reference, not 50% answer accuracy. Absolute AUROC describes the monitor's discrimination. The increment $\Delta = \mathrm{AUROC}(f_{D,C}) - \mathrm{AUROC}(f_C)$ compares a predictor using diagnostics D and controls C with controls alone. The original association statistic group-weights checkpoint/fold increments within each benchmark and then weights benchmarks equally.

To measure the advantage of refitting after an update, we evaluate the two monitors on identical target test rows with target labels:

$$T_{s\to t} = \mathrm{AUROC}_t(f_t) - \mathrm{AUROC}_t(f_s).$$

Positive T means that the target-refitted monitor ranks these examples better than the reused source monitor. Thus T is a refitting advantage, not a source-to-target performance drop. We then form the update contrast

$$C = T_{\text{major}} - T_{\text{revision}}.$$

A positive C assigns the larger refitting advantage to the stage change; a negative C assigns it to the within-stage revision. Neither sign alone decides whether to reuse a monitor.

We compute these quantities for two ranking tasks. *Pooled* AUROC combines benchmarks within a test fold, so a positive example from one benchmark can be compared with a negative from another. *Benchmark-macro* AUROC first evaluates within-benchmark rankings and then averages the benchmarks equally. The first includes cross-task prioritization; the second evaluates discrimination within each task. In one fold, n_b^+ and n_b^- count inconsistent and consistent examples in benchmark b; n^+ and n^- are the corresponding fold totals:

$$\mathrm{AUROC}_{\mathrm{pool}} = \sum_{b,c} \frac{n_b^+ n_c^-}{n^+ n^-} A_{bc},$$

where A_{bc} is the tie-adjusted ranking probability for a positive from b and a negative from c. Macro uses only the A_{bb} terms. The original pooled scheme row-weights folds; the later macro scheme group-weights folds within each benchmark. We preserve this distinction and separately check common fold weights. No cross-fold rankings are formed.

The original monitors are standardized L2-logistic regressions ($C = 1$), evaluated with five group-preserving folds. OLMo uses cosine and L2 displacement between prompt-end hidden states, changes in next-token entropy and top-two margin, and two $k = 10$ neighborhood diagnostics: normalized overlap and local-density displacement. Tülu's low-dimensional monitor uses the first four diagnostics. Text-and-task controls comprise benchmark, subtype, variant type, length difference and lexical overlap. Scalers are fitted on TRAIN only; reuse keeps both the source scaler and coefficients fixed. The benchmark-only baseline and richer Tülu probes are defined in Section 3.1. Full fitting and bank construction details are in Section C.

2.3 Uncertainty and analysis map

The original OLMo intervals use 1,000 template-group bootstrap draws that refit predictors and repairs while keeping features and reference banks fixed. The subsequent fixed-score intervals resample groups but retain fitted scores and selected configurations, so they omit fitting and selection uncertainty.

Table 2: Which monitor and population each analysis evaluates. TRAIN and TEST are disjoint by template group. A finite-set reference describes the indicated evaluation set, not known population quality.

Analysis	Monitor training	Evaluation sample	Scores and reference
OLMo comparison	Eligible TRAIN; source/target rows aligned for transition fits	Common OOF: 1,361 pairs, 1,139 groups	Source and target fits; common target labels
Tülu comparison	Common-four TRAIN or source-own TRAIN	Common TEST: 1,196 pairs, 1,051 groups	Source and target fits; common target labels
OLMo audit	Source TRAIN for each analysis branch	Transition-specific TEST; two unresolved-label treatments	Fixed fold-specific source fits; branch-specific finite-set quality
Tülu audit	Source-own TRAIN, per model fold	One pair per target-observable TEST group	Fixed hidden-state/low-dimensional fits; common no-subdivision finite-set target

3 Empirical Findings

We first ask what the monitors' scores discriminate, then how the ranking target changes a comparison between updates, and finally what a small labeled audit can certify.

3.1 What does a good pooled score discriminate?

A pooled score can be high because the monitor recognizes the benchmark. Tülu's original source monitor has AUROC .677 pooled but .508 macro. A benchmark-only predictor, whose score is

the checkpoint's TRAIN inconsistency rate for that benchmark, reaches pooled AUROC .684; it ties all examples within a benchmark, so its macro AUROC is exactly .5 (Table 3). No TEST outcomes estimate the rate, and pooled AUROC is averaged over test folds rather than computed from one global ranking. Because InvariantBench pairs are inconsistent far more often than GSM-Plus pairs (Table 1), ranking one benchmark above the other already orders many cross-benchmark pairs correctly. Text-and-task controls add subtype, variant type, length and overlap, but still have source macro AUROC .502.

The two transitions are DPO to GRPO step 1920 and step 1840 to step 1880. Four-checkpoint eligibility retains 1,196/1,504 pairs in 1,051/1,230 groups: 949/1,047 GSM-Plus and 247/457 InvariantBench pairs. All comparisons in Table 3 use this same TEST subset.

Table 3: Tülu DPO-to-GRPO discrimination on the same 1,196 common TEST pairs. Reuse evaluates the source fit at the target; refit trains at the target. All entries are descriptive AUROCs. Source-own changes source TRAIN only; target TRAIN remains common-four. Probe selection uses source TRAIN, and fixed-score intervals are in Table 33.

	Source		Target macro	
Monitor	Pooled	Macro	Reuse	Refit
Benchmark-only	.684	.500	.500	.500
Text-and-task controls	.678	.502	.465	.536
Original low-dimensional	.677	.508	.454	.536
Hidden-state, common-four	.769	.658	.646	.705
Hidden-state, source-own	.774	.701	.689	.723

For the hidden-state probes, normalize each canonical/variant prompt-end state as $u = h/\max(\|h\|_2, 10^{-8})$ and form $\phi = [(u_c + u_v)/2,\ |u_c - u_v|,\ D,\ C]$. Three-fold group cross-validation inside source TRAIN selects block 21 or the final normalized layer and an L2-logistic inverse penalty from $\{.001, .01, .1\}$ by macro AUROC. Common-four TRAIN requires eligibility at all four checkpoints; source-own TRAIN requires only source eligibility. Target TRAIN stays common-four, using the source-selected configuration in both cases. Common-four training gives source macro AUROC .658, a paired gain of .150 [.075, .223] over the original monitor. Source-own training gives .701 [.638, .743], and its source pooled–macro gap remains .0735 [.0296, .1323]: the gap persists with better within-benchmark discrimination. The paired source-own minus common-four gain interval includes zero; the change of training eligibility alone is not established as beneficial (Section H).

On the target, the reused source-own probe retains macro AUROC .689 [.625, .735], while refitting gives .723 [.654, .765]: in point estimate, within-benchmark discrimination largely survives the stage change (source .701, reused .689), with a higher refitted point estimate. Neither value by itself licenses reuse; that requires the audit evidence of Section 3.3.

OLMo's original monitor is weak in the same within-benchmark sense. On the 1,361-row reporting subset, source macro AUROC is .530 for the full monitor and .485 for controls (Table 4). The original association analysis gives an increment of .045 with cluster-bootstrap interval [−.018, .083]. That statistic summarizes checkpoint/fold increments, whereas the two absolute values describe the source checkpoint; their matching rounded difference is coincidental. The prespecified association rule requires an increment of at least .030, a positive bootstrap lower endpoint, an increment above the permutation 95th percentile (.034), and positive components in both benchmarks. The observed increment passes the point, both-benchmark positivity and permutation conditions ($p = .017$) but fails the bootstrap condition. A positive increment therefore does not establish a useful starting monitor. All five original diagnostic criteria remain unmet; their complete results and supporting feature study are retained in Table 11 and Section E.1.

3.2 Can the ranking target change an update comparison on fixed predictions?

We hold the 1,361 OLMo OOF rows, labels, scores, and folds fixed and change only how performance is aggregated. The major-minus-revision contrast is −.0099 under the original pooled scheme and +.0186 under the later macro scheme (Figure 1, right). Under pooling, refitting has the smaller advantage after the major update; under macro averaging, it has the larger advantage. The language

Table 4: OLMo absolute macro AUROC on 1,139 common eligible template groups (1,361 rows). Refit and controls train at the named checkpoint; parentheses identify the training checkpoint retained by the frozen score. Additional numerical detail appears in Table 7.

Evaluation checkpoint	Refit	Controls	Frozen score (trained at)
DPO source	.530	.485	—
RLVR final	.583	.537	.546 (DPO)
RLVR step 25	.544	.515	—
RLVR step 50	.541	.471	.523 (step 25)

models and monitor predictions are identical in the two calculations. What changes is which rankings contribute and how they are weighted, not which language-model version answers questions better.

An exact decomposition identifies the contributions in these saved predictions. Within-benchmark reweighting contributes +.01975, and removing cross-benchmark pairs contributes +.00870, together giving the .02845 macro-minus-pooled difference. For each transition τ, write $T_{\text{macro},\tau} - T_{\text{pool},\tau} = W_\tau + X_\tau$, where W_τ changes the weights of within-benchmark pairs and X_τ removes cross-benchmark pairs. Subtracting revisions from major updates gives $D_{\text{agg}} = (W_{\text{major}} - W_{\text{revision}}) + (X_{\text{major}} - X_{\text{revision}})$. The major update has $(W, X) = (+.03294, -.02225)$ and the revision $(+.01319, -.03095)$. Removing cross-benchmark pairs lowers both refitting advantages, but lowers the revision advantage more, raising their contrast. This algebraic accounting combines weights and ranking differences, rather than isolating a causal effect of class prevalence (Section B.3). The signs persist with equal test-fold weights in both schemes (Table 5); different fold weights alone do not explain the reversal.

Table 5: OLMo refitting advantages on identical saved predictions. The update contrast is $C = T_{\text{major}} - T_{\text{revision}}$. The original-weight aggregation difference is .02845; the equal-fold check uses its own weights. All entries are descriptive point estimates.

Fold weights	Ranking target	T_{major}	T_{revision}	C
Original	Pooled	.0259	.0357	−.0099
Original	Macro	.0366	.0180	+.0186
Equal	Pooled	.0198	.0381	−.0183
Equal	Macro	.0357	.0178	+.0179

Excluding the 18 reporting rows with disputed semantic equivalence leaves the pooled and macro contrasts at −.0101 and +.0182, respectively. This evaluation-only check preserves the observed sign pattern without refitting the monitors (Section G.2).

Joint fixed-score resampling assesses this same comparison without refitting. Across 10,000 paired group draws, the pooled contrast has a 95% interval [−.0858, .0627]. Macro and the aggregation difference are undefined in 51 draws, so the specified all-finite rule returns no primary interval for them. For $D_{\text{agg}} = .02845$, the 2.5th–97.5th percentiles on the 9,949 jointly finite draws are [−.0372, .0764]: a conditional descriptive range, not a replacement primary confidence interval. The reversal is established for these saved predictions, while its population sign pattern remains unresolved (Section G.1).

Tülu does not reproduce the reversal. Under the original weights, its low-dimensional monitor has update contrasts $C_{\text{pool}} = +.0440$ and $C_{\text{macro}} = +.0793$. The common-four probe gives +.0051 [−.0337, .0301] pooled and +.0237 [−.0536, .0773] macro; the source-own probe gives +.0042 [−.0300, .0348] and +.0124 [−.0413, .0624], respectively. All three point comparisons remain positive. For the low-dimensional monitor, reweighting contributes +.0457 and pair removal −.0104 to the *aggregation difference* $D_{\text{agg}} = +.0353$ [−.0276, .0971]. Both probes' contrast and aggregation-difference intervals include zero. All intervals here retain fitted scores and source-side selection, unlike the original OLMo refit bootstrap (Sections F and H.3).

Even within-task evaluation requires explicit task weights. Giving every model fold equal weight, let w weight GSM-Plus and $1 - w$ weight InvariantBench. The OLMo contrast changes sign at $w = .608$; the common-four and source-own hidden-state probes change sign at .872 and .893. The low-dimensional Tülu contrast stays positive throughout $0 \le w \le 1$. These fixed-score curves vary

task priorities, not training or labels; pooled AUROC also includes cross-benchmark pairs and is not a point on these curves (Section I).

3.3 When does an audit provide evidence for reuse?

Refitting advantage is relative; reuse needs evidence about the quality of the monitor being kept. We examine a rule that approves reuse only when a finite bootstrap lower endpoint for target macro AUROC reaches $\theta_{\text{op}} = .60$. The budget counts labeled template groups. Obtaining each pair's label requires running the target checkpoint on both prompts; here budgets are simulated from saved outputs, and unit label cost is not measured.

How an audit produces a decision. Draw n groups without replacement, nearly balanced between benchmarks. From this sample, make 1,000 group-bootstrap draws, retaining each group's evaluation partition. The subdivided rule computes AUROC within five partitions per benchmark, requiring at least two partitions with both classes in each benchmark; it group-weights their finite AUROCs and averages benchmarks equally. The no-subdivision rule ranks all sampled groups within each benchmark, requiring both classes in each. With at least 900 finite draws, return their tenth percentile L; otherwise return no bound. Approve reuse if $L \geq .60$; decline with a finite lower L, or fall back to distrust when L is unavailable.

The subdivided rule inherits the cross-fitted OLMo evaluation, in which scores come from five fold-specific source fits and are never ranked across folds. Partitioning small samples can leave too few mixed-class partitions to form AUROC; this is a property of the specified rule, not a universal sample-size limit. A finite $L < .60$ does not prove quality below .60, and the nominal 90% lower endpoint has no general coverage guarantee here.

OLMo audit. The smallest setting allocates 13/12 groups between benchmarks; budget 400 is infeasible and has no results. Two unresolved-label treatments (common-resolved pairs and a missingness-sensitive version), two transitions and two monitors (full diagnostics and pointwise+output) form eight cells, with 500 samples per attempted budget. All finite-set quality references are below .60, so their reference action is distrust (Section D.1).

At budget 25, all 4,000 bounds are unavailable; every decision follows the fallback and each cell records 0/500 errors against its finite-set reference. Thus zero error does not mean that the labeled samples established quality. Unavailable bounds fall to 3,986, 3,008 and 862 of 4,000 at 50, 100 and 200 groups (one cell is plotted in Figure 4); the bounds that appear produce some approvals, all of them disagreements with these below-threshold references (Table 13). These references use transition-specific sets and source fits, unlike the common-OOF reused-target value .546 in Table 4.

Paired audits of real Tülu monitors. We evaluate all five saved source-own fits, pairing the hidden-state and low-dimensional monitors within each original model fold. Each audit uses one predetermined representative per target-observable TEST group, with no training or adaptation groups. Each saved fit is the source monitor trained for one model fold, so a single scorer is audited. To compare the partitioned rule with an unpartitioned one under that scorer, a fixed hash assigns its TEST groups to five audit subdivisions: smaller evaluation partitions under that one scorer, not different fitted models. Each of 200 audit samples per fit and budget has 1,000 bootstrap draws shared across both monitors. We compare the two bound rules above and always-distrust, which never approves reuse.

Removing subdivisions adds ranking pairs as well as increasing availability. For hidden-state fit 3, full-set AUROC changes from .5865 to .6470, crossing .60. Both rules use the no-subdivision finite-set reference as the common decision target, fixed before these audits ran. This compares complete procedures, not an isolated gain in sampling efficiency (Section I.2).

At budget 25 the subdivided rule returns no bound for either monitor; without subdivisions, 47–73 of 200 bounds are available per fit. At budget 50, availability rises from 0–34 to 133–170 of 200 (Table 6). Four hidden-state fits have finite references above .60 and receive more approvals. Fit 0 has reference .566 and receives two false approvals; the low-dimensional fits acquire up to four. Training sets overlap across fits, so we report each fit separately rather than treating them as independent deployments.

Table 6: Real Tülu audits at budget 50, with 200 paired samples per source fit. H: hidden-state; L: low-dimensional. Arrows compare five subdivisions with none. Shared labels and groups give both monitors the same bound availability. References use the full held-out representative set, not the common OOF rows. The last column splits H nonapprovals without subdivision into unavailable-bound fallback plus finite bound below .60.

Fit	Finite reference H / L	Finite bounds Both monitors	Approvals H	Approvals L	H nonapprovals Fallback + finite
0	.566 / .450	25 → 165	0 → 2	0 → 0	35 + 163
1	.663 / .572	34 → 170	8 → 47	0 → 3	30 + 123
2	.665 / .522	21 → 133	6 → 33	0 → 0	67 + 100
3	.647 / .528	0 → 167	0 → 19	0 → 4	33 + 148
4	.614 / .496	23 → 134	1 → 21	0 → 2	66 + 113

Higher references also expose nonapproval. Without subdivisions, fits 1–4 remain unapproved in 76.5–90.5% of audits despite finite references above .60. Table 6 splits these counts into unavailable-bound fallbacks and finite bounds below threshold. These are nonapproval rates against finite references, not population false-negative rates.

Where do the additional approvals occur? For hidden-state fit 1 at budget 50, approvals rise from 8 to 47: 39 samples move from an unavailable bound to approval, three gain approval with both endpoints available, and three previous approvals are revoked. On the 34 samples with two finite endpoints, the mean lower endpoint instead falls by .0064: a mean point-estimate change of −.0480 is partly offset by a .0417 reduction in the point-to-endpoint offset. The numerical decomposition is $L_1 - L_5 = (A_1 - A_5) + [(A_5 - L_5) - (A_1 - L_1)]$, where A is the audit-sample AUROC, L its lower endpoint, and subscripts denote one or five partitions. It applies only when both endpoints exist; an offset $A - L$ is not an interval width. All additional false approvals in these real-monitor comparisons occur after unavailable subdivided bounds. Full per-fit counts appear in Section I.3.

Known-quality controls generate independent groups with $Y \sim \text{Bernoulli}(.25)$ and $S = \sqrt{2}\Phi^{-1}(A)Y + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ is independent of Y and Φ is the standard-normal CDF. Each benchmark therefore has population AUROC A. On 200 paired saved samples at budget 50 and $A = .80$, removing subdivisions raises approvals from 46 to 159; at $A = .55$, false approvals rise from three to eight. Better availability therefore need not produce uniformly better decisions (Sections G.3 and G.4). Among finite no-subdivision bounds at budget 50, 170/191 ($A = .55$) and 161/187 ($A = .80$) lie at or below the analytic AUROC (89% and 86%, against a nominal 90%, conditional on a finite bound); we make no calibration claim.

4 Related Work

Monitoring across model updates. Activation neighborhoods support representation-based confidence (Papernot & McDaniel, 2018), and geometric studies compare language-model embeddings (Lee et al., 2025). Duan (2026) benchmarks monitor staleness under quantization and fine-tuning and finds label-free realignment sufficient for the repair-relevant stale cells studied. Backward-compatible prediction updates (Träuble et al., 2021) and compatible LLM evolution (Echterhoff et al., 2024) address behavior across versions. Our contribution concerns the evaluation: what absolute performance, the refitting comparison, and audit output establish in the released OLMo and Tülu trajectories (Rafailov et al., 2023; Team Olmo et al., 2025; Allen Institute for AI, 2025).

Evaluation targets. ROC analysis distinguishes pooled discrimination, covariate adjustment, and incremental value (Janes et al., 2009); the pooled–macro distinction itself is not new. We isolate its effect on a major-minus-revision refitting contrast using fixed predictions and explicit signed contributions. Distribution-shift benchmarks separate test conditions (Hendrycks & Dietterich, 2019; Koh et al., 2021), and domain-generalization work stresses consistent comparisons (Gulrajani & Lopez-Paz, 2021). The neighborhood diagnostics draw on local and topological representation analysis (Carlsson, 2009; Chazal & Michel, 2021).

5 Discussion, Limitations, and Conclusion

5.1 What to report before interpreting a maintenance signal

Specify the ranking task first. A shared cross-task queue needs cross-task comparisons; separate task evaluations need within-task rankings and explicit task weights. Report absolute source, reused-target, and refitted-target performance alongside increments over controls. The OLMo and Tülu cases show why this choice matters, with and without an observed reversal.

Separate a decision from the evidence producing it. Audit reports should show available bounds, bound-based decisions, and fallbacks, naming the reference used to count errors. The real-monitor and known-quality comparisons show why increased bound availability and correct approval must be measured separately. When comparing audit procedures on shared samples, separate decisions that acquire a bound from those whose existing endpoints change; report revoked as well as new approvals against the same reference.

Read the three together. For Tülu's hidden-state probe, reuse retains macro AUROC .689 against .723 after refitting. A pooled score (.78 for the reused probe) would overstate within-benchmark quality, and 50-group audits approve reuse in at most 24% of samples even for the four fits whose finite reference exceeds .60 (.61–.67). Here the approval rate reflects what 50 labeled groups can certify that close to the requirement, rather than a measured loss of discrimination; the low-dimensional monitors offer no comparable basis, because their macro AUROC is near chance before any update.

5.2 Scope of the evidence

Both empirical families use the same reasoning benchmarks. The macro scheme, equal-fold check, task-weight curves and Tülu extension were examined after the original pooled analysis. We therefore treat the OLMo reversal as a worked example and rest the ranking-target point on the benchmark-only baseline and the probe's pooled–macro gap, whose fixed-score interval, conditional on the fitted scores, excludes zero. The richer probe does not reproduce the sign reversal, although its within-task contrast still changes with task weights. The new Tülu audits use target-observable representative groups and fixed source fits; their finite references differ from the common-OOF comparison. Neither revision is magnitude-matched: in OLMo, the major update moves prompt-end entropy and primary-layer cosine displacement 13.6× and 39.2× as far as the revision. These differences limit attribution to the training stage.

The outcome concerns disagreement rather than correctness. The type-level semantic screen retains 27 primary rows with majority-reviewer nonequivalence opinions, including 18 OLMo reporting rows. These are disputed equivalences, not an established outcome-label error rate. Removing the 18 rows from the fixed evaluation cache preserves the contrast signs, but leaves their earlier training and adaptation roles unchanged. Conditioning on parseable answers also selects the evaluation population (Section B.1 and Table 12). The feature and known-quality controls measure the behavior of their statistical procedures, not the full empirical generation and monitoring pipeline.

Conclusion. Monitor maintenance requires distinct evidence about absolute quality, ranking targets and audit availability. Fixed-prediction comparisons isolate how aggregation changes performance assessments; paired real-monitor audits show why more available evidence does not guarantee better decisions. Reporting these quantities together makes a maintenance signal interpretable.

Reproducibility Statement

The supplement supplies saved empirical results, numerical figure tables, necessary configurations, and scripts for saved-result verification and fixed-prediction recomputation. The original two verification scripts use Python's standard library; the added CPU analyses specify their dependencies. Reporting refits reproduced all 23 frozen point estimates before OOF export; the refit log is not part of this package, and the supplied file records the resulting values only. The fixed-OOF cache, deterministic decomposition script and full-precision terms are included. The controlled study supplies its locally hash-locked protocol, terminal summaries covering all 480 fixed IDs (468 complete and 12 unavailable), permutation counters, and unavailable-output diagnostics (Section E). Its per-run

outputs and generator, fitting, repair, and resampling code are not included; the retained materials support summary checks, not rerunning that study. The separately specified Tülu extension adds four checkpoint evaluations, 40 fixed monitor fits, and 1,000 paired fixed-score bootstrap draws (Section F). Its numerical OOF records, coverage, full-precision results, and statistical specification are supplied separately from the unchanged OLMo and controlled-study records. The new fixed-OOF joint resampling, disputed-row sensitivity, and known-quality audit experiments include their specifications, scripts, and results (Section G). The expanded-feature probe and source-training sensitivity provide candidate scores, OOF outputs, and paired fixed-score intervals (Section H); their original hidden-state inputs are not distributed. The paired audit comparison reuses saved Gaussian samples and inner sampling sequences (Section G.4). The benchmark-only baseline, task-weight curves and paired real-monitor audits include TRAIN-only prior statistics, fixed scores, representative IDs, all outer samples and inner multiplicities, and saved-output recomputation entry points (Section I). A deterministic follow-up pairs availability and approval states and decomposes lower-endpoint changes on the same saved outer samples (Section I.3). These audits add no model fitting or language-model generation.

The manifest maps original and review-copy hashes. Documented public-copy adaptation preserves public model revisions, IDs, numerical results, decisions, missingness and array order; a verifier checks review copies against recorded payload hashes. Original artifact hashes are recorded in the supplement manifest rather than repeated as a numerical table in the paper. Third-party raw text, model artifacts and a complete generation environment are not supplied. We downloaded the 1,200-record InvariantBench file `final-open.json` from `DreamyPanda/InvariantBench` on 18 August 2026 (DreamyPanda, n.d.). Its recorded upstream Git blob matches the frozen local input; full file identifiers are supplied in `data_identity/invariantbench_snapshot.json`. A historical dataset-page capture lists MIT, but the repository commit and complete snapshot-specific licensing documentation were not recovered. Anonymous access checks on 16 September 2026 were unsuccessful. Redistribution permissions for raw inputs and derived records remain separately unverified. The package supports saved-output and fixed-prediction checks, including replay of the saved real-monitor audits; it does not reproduce monitor fitting, the controlled feature-level study, or the complete empirical generation study.

References

Allen Institute for AI. Llama-3.1-Tulu-3.1-8B: Model card. Hugging Face, 2025. URL `https://huggingface.co/allenai/Llama-3.1-Tulu-3.1-8B`. Model release and GRPO training description; accessed September 12, 2026.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. doi: 10.1090/S0273-0979-09-01249-X.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021. doi: 10.3389/frai.2021.667963. URL `https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full`.

DreamyPanda. InvariantBench. Hugging Face dataset repository, n.d. URL `https://huggingface.co/datasets/DreamyPanda/InvariantBench`. Downloaded final-open.json on 18 August 2026; frozen file identifiers supplied with the supplementary materials; anonymous access check unsuccessful on 16 September 2026.

Evan Duan. Do activation monitors survive model updates? benchmarking, predicting, and repairing activation-monitor staleness, 2026. URL `https://arxiv.org/abs/2606.15980`.

Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A model update strategy for compatible LLM evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7320–7332, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.430. URL `https://aclanthology.org/2024.findings-emnlp.430/`.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=lQdXeXDoWtI`.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Holly Janes, Gary Longton, and Margaret Pepe. Accommodating covariates in ROC analysis. *The Stata Journal*, 9(1):17–39, 2009. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2758790/.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=aJDykpJAYF.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, 2024. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods, 2017. URL https://arxiv.org/abs/1712.03198.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018. URL https://arxiv.org/abs/1803.04765.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Frederik Träuble, Julius von Kügelgen, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Peter V. Gehler. Backward-compatible prediction updates: A probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/012d9fe15b2493f21902cd55603382ec-Abstract.html.

Azmine Toushik Wasi, Mahir Absar Khan, Abdur Rahman, Wahid Faisal, Sukanta Saha, Saimon Bhuyan, Mahdiya Rahman Sukanya, Rahatun Nesa Priti, Md. Iqramul Hoque, Munem Shahriar, Raima Islam, Sanatan Sushil, Shahriyar Zaman Ridoy, Kazi Raiwan Sultan, MD Shafikul Islam, Md Manjurul Ahsan, and Md Rizwan Parvez. InvariantBench: Can large language models exhibit inherent reasoning across equivalent transformations? In *3rd AI for Math Workshop: Toward Self-Evolving Scientific Agents at ICML 2026*, 2026. URL https://icml.cc/virtual/2026/82490.

AI Use Statement

The authors take responsibility for the final manuscript and for the design, implementation, and interpretation of all analyses. LLM-based tools assisted with writing and implementing and running analysis code.

A Original OLMo criteria and analysis specification

Earlier analyses call T "transport loss" and the within-stage revision the "null" transition; T is the target-refitting advantage defined in Section 2.2.

A.1 Specification and evidence classes

The original OLMo results use Protocol-v2, Freeze-4.4.1. Their confirmatory record is unchanged. Later CPU-only calculations are isolated as post-confirmatory reporting or sensitivity analyses, including an achieved-design evidence-class sensitivity, and cannot alter its verdicts. The separately specified Tülu evaluation is reported in Section F; it does not revise these original decisions.

Protocol-v2 narrows the original two-major-transition design to the single DPO-to-RLVR major transition. The revision pair is a control arm, not a second major transition; the high-consensus criterion is narrowed to the same transition. These changes are in the protocol committed before the confirmatory read. The parser chronology is assessed separately below.

The supplement manifest distinguishes the original artifact identities from identity-normalized review copies, and states which originals are not supplied. It specifies the saved-result and resampling materials included for each analysis. The rules and results needed to interpret the analyses are specified below.

A.2 Association: complementary uncertainty checks

The frozen joint G1 requirement is

$$\widehat{\Delta} \geq .030 \ \wedge \ \mathrm{LB}_{95}^{\mathrm{boot}} > 0 \ \wedge \ \widehat{\Delta} > q_{.95}^{\mathrm{perm}} \ \wedge \ \Delta_{\mathrm{GSM}}, \Delta_{\mathrm{IB}} > 0. \tag{1}$$

Here $\widehat{\Delta}$ is the observed benchmark-balanced association gain, $\mathrm{LB}_{95}^{\mathrm{boot}}$ its bootstrap lower endpoint, $q_{.95}^{\mathrm{perm}}$ its permutation 95th percentile, and $\Delta_{\mathrm{GSM}}, \Delta_{\mathrm{IB}}$ its benchmark components.

The template-cluster bootstrap refits full and controls predictors to describe sampling/refitting uncertainty, conditional on precomputed features and banks. The permutation instead exchanges whole outcome-label blocks within benchmark/regime/group-size strata and refits predictors. Its q95 and one-sided p-value measure extremeness under block exchangeability, not general conditional independence. The two procedures use the same association statistic but answer different inferential questions.

The benchmark components are .0414441 and .0492907. The observed increment .0453674 exceeds permutation q95 .0340213, with 16 of 1,000 replicates at least as large and plus-one $p = 17/1{,}001 = .0169830$. The bootstrap interval $[-.0181743, .0830262]$ nevertheless includes zero. The observed assignment can be unusual under label-block exchangeability while uncertainty under group resampling and refitting remains large. The positive-bootstrap-lower-bound condition is therefore the binding part of the joint rule in Equation (1).

B Post-confirmatory analyses of discrimination and sensitivity

All results in this section are post-confirmatory. They use only frozen local outputs, features, labels, groups, and folds; no model was called. An independent array-level reconstruction reproduced the four population point analyses exactly, and a fail-closed reporting refit reproduced all 23 frozen point estimates with maximum absolute error zero before emitting any new OOF prediction or AUROC.

B.1 Population eligibility and parser qualification

The held-out universe has 1,230 template groups: 1,045 GSM-Plus and 185 InvariantBench. Screening retains 2,889 study rows (1,504 strict and 1,385 stress) from 2,892 candidates. Primary analyses add common four-checkpoint outcome resolution and branch-finite diagnostics. The separate reporting OOF cache has 1,361 strict rows in 1,139 groups; its high-consensus subset has 1,178 rows in 1,017 groups. These caches do not reconstruct every historical fit's row mask (Table 12).

The semantic screen demotes transformation types, rather than each row with an adverse opinion. Consequently 27 primary rows retain majority-reviewer nonequivalence opinions, including 18 in the reporting OOF subset. Those opinions qualify the intended answer-preserving interpretation but are not confirmed outcome-label errors. The parser panel evaluates extraction and judge agreement; its decisions do not filter primary rows.

The parser panel has coverage 1.000, consensus coverage .985, agreement 1.000 on consensus rows, and checkpoint coverage spread zero. These are panel statistics, not externally validated semantic accuracy.

The active v4.4 quorum rule forms consensus on 197/200 rows and matches on all 197. Holding v4.4 fixed, strict three-of-three forms consensus on 96/200 and matches on all 96; the frozen legacy block's 90/200 coverage and 87/90 agreement also reflect its older extraction/comparison rule. The 103 malformed Phi-4-mini responses are therefore not 103 semantic disagreements: a diagnostic-only audit finds 84 resolvable assertions outside the required serialization and 19 non-concrete responses. No panel decision filters an *H* row. Git parent order, content hashes, full4 metadata, and output hashes prove that v4.4 preceded full4 execution. The first-read timestamp is only a self-attested untracked record without a contemporaneous immutable anchor, so the stronger amendment-before-human-read claim remains unproven.

B.2 Absolute discrimination and fixed-score sensitivity

Table 4 reports the absolute discrimination reconstruction in the main text. These are reporting-only estimates on the common eligible population, not new gates or evidence of established source usefulness.

Table 7: Numerical detail for the reporting-only absolute macro AUROCs in Table 4. Values are copied from the saved benchmark-balanced macro summary, not recomputed by subtracting the incremental statistic. The frozen score names its training checkpoint explicitly.

Evaluation and score	AUROC
DPO source, full	.5301514901092224
DPO source, controls	.48506224770415607
RLVR final, refit full	.5827571507071987
RLVR final, controls	.5371115808908773
RLVR final, frozen DPO	.5461910881845251
RLVR step25, full	.5444773118897561
RLVR step25, controls	.515115441197694
RLVR step50, full	.5406466350686123
RLVR step50, controls	.47056495849952973
RLVR step50, frozen step25	.5226718144646203

The two transition caches contain the same 1,361 strict row IDs in 1,139 template groups, each scored once in its assigned test fold. Holding those IDs, folds, outcomes and OOF probabilities fixed, pooled-within-fold AUROC with row-weighted folds gives major/null losses .02587508/.03573154 and contrast −.00985646, exactly reproducing the locked point values. Group-weighted folds within benchmarks followed by equal benchmark weighting give .03656606/.01797482 and contrast +.01859124. The null score is trained at step 25 and evaluated at step 50 in both calculations. This observed-point comparison isolates the aggregation package, with no new fit or resampling; it does not explain the separate bootstrap-mean displacement.

Keeping serialized OOF scores fixed and resampling template groups gave major-minus-null mean .01204 with 95% interval [−.07744, .09937]; all 10,000 draws were finite under the reporting-only benchmark-macro aggregation. The interval crosses zero. This sensitivity differs from frozen G2 in both keeping OOF scores fixed and using benchmark-macro aggregation; frozen G2 refits models and pools benchmark rows within each fold. It does not change G2=FAIL. This earlier macro-only analysis is separate from the joint resampling design in Section G.1, which has 51 undefined macro draws.

B.3 Deterministic decomposition of the aggregation difference

This post-confirmatory calculation holds the transition OOF cache fixed. For transition τ and fold f, let $\Delta A_{\tau fbc}$ be target-refit minus source-frozen pair-AUROC for positives from benchmark b and negatives from c, using the pair definition in Section 2. The fold weight is $w_f = n_f/N$, where n_f and N count test rows in that fold and all folds. The within-benchmark macro weight is $m_{bf} = \frac{1}{2} g_{bf}/G_b$, where g_{bf} and G_b count test groups in fold f and all folds of benchmark b. The pair coefficient $q_{\tau fbc} = n^+_{\tau fb} n^-_{\tau fc}/(n^+_{\tau f} n^-_{\tau f})$ uses each transition's target outcomes: $n^+_{\tau fb}, n^-_{\tau fc}$ are benchmark-specific positive/negative counts and $n^+_{\tau f}, n^-_{\tau f}$ their fold totals. No cross-fold pairs are formed. Writing T^{macro}_τ and T^{pooled}_τ for the two aggregations of transport loss, the signed contributions are

$$W_\tau = \sum_{f,b} (m_{bf} - w_f q_{\tau fbb}) \Delta A_{\tau fbb},$$

$$X_\tau = -\sum_{f,b \neq c} w_f q_{\tau fbc} \Delta A_{\tau fbc},$$

$$W_\tau + X_\tau = T^{\text{macro}}_\tau - T^{\text{pooled}}_\tau.$$

Let C_{macro} and C_{pooled} denote the major-minus-revision transport contrasts. Subtracting the revision contributions from the major contributions gives $D_{\text{agg}} = C_{\text{macro}} - C_{\text{pooled}}$ in Table 8. Within each transition, positive reweighting and negative pair-removal contributions partly offset one another; in the major-minus-revision difference both contributions are positive. The within term combines the fold/benchmark coefficient changes, rather than separately identifying their effects. This is a signed algebraic accounting for two specified targets, not a unique causal decomposition or population significance claim.

Table 8: Fixed-OOF signed aggregation contributions. Within denotes within-benchmark reweighting; cross denotes cross-benchmark pair removal. The last column equals within plus cross. Full-precision values and all 40 fold–benchmark pair cells are supplied with the analysis.

Comparison	Pooled	Macro	Within	Cross	Macro−pooled
Major	+.025875	+.036566	+.032939	−.022248	+.010691
Revision	+.035732	+.017975	+.013193	−.030949	−.017757
Major−revision	−.009856	+.018591	+.019746	+.008701	+.028448

Explicit pair comparisons and an independent average-rank AUROC calculation reconstruct the two reported endpoints and the additive identity with maximum absolute error 9.72×10^{-17} (fixed numerical tolerance 10^{-12}). The calculation uses no new fit, resampling, or confidence interval for D_{agg}. The earlier fixed-score interval above concerns the macro contrast, not D_{agg}; neither calculation revises frozen G2 or explains the distinct refit-bootstrap mean displacement below. The separate joint resampling analysis in Section G.1 reports conditional percentile ranges for D_{agg} and retains undefined draws.

Common fold weights. A subsequent descriptive comparison replaces both schemes' TEST-fold weights by $1/5$. Macro still averages the two benchmarks equally, so its benchmark–fold coefficient is $1/10$. The rows, labels, fitted scores, and fold assignments are unchanged, with ties receiving half credit and no cross-fold ranking pairs. Table 9 shows that the contrast still reverses. This checks one common weighting, not every possible weighting; benchmark composition and within/cross-benchmark ranking comparisons remain different. No new fitting, sampling, or interval is used.

B.4 Observed contrast and refit-bootstrap mean

The unresampled major and null transport losses are .0258751 and .0357315, so their contrast is −.0098565. Across the joint refit bootstrap, the major mean shifts upward to .0497391 (bias +.0238640) while the null mean shifts downward to .0199167 (bias −.0158149). Their difference has mean .0298224 and bias +.0396789. Both summaries refer to the same frozen statistic; the unresampled estimate and bootstrap mean need not coincide. The bias warrants calibration checks, but does not itself establish a different estimand or invalidate the percentile construction.

Table 9: OLMo fixed-score transport under original and common equal-fold weights. The final column is major minus revision. The equal-fold rows are subsequent descriptive comparisons; original results are retained.

Aggregation and fold weights	Major T	Revision T	Contrast
Pooled, original	.02587507666	.03573154121	-.00985646455
Macro, original	.03656606252	.01797482060	+.01859124192
Pooled, equal folds	.01975568229	.03808158219	-.01832589990
Macro, equal folds	.03569254710	.01779931840	+.01789322871

B.5 Stage-local does not mean magnitude-matched

On the identical 1,361-row population resolved at all four checkpoints, the major and null inconsistency-label flip rates are 9.405% and 7.935%. That behavioral proximity does not extend to direct movement measures: major/null ratios are 13.60 for prompt-end entropy mean absolute change, 12.33 for top-two-margin mean absolute change, 3.63 for fixed-monitor probability mean absolute change, and 8.02 for source-standardized diagnostic vector RMS. Same-prompt cosine displacement ratios are 39.23 at the primary hidden-state layer and 203.29 at the final layer. These are representation distances, not parameter norms, but they rule out describing the adjacent step-25–step-50 control as empirically update-magnitude matched.

B.6 Residual-model sensitivity and repair availability

Table 10: Empirical-bootstrap, bias-preserving post-confirmatory sensitivity. Zero-effect PASS is a residual-model sensitivity, not validated type-I error. MDE is linearly interpolated only inside the pre-fixed effect grid.

Gate	**Zero-effect PASS**	**MDE$_{50}$**	**MDE$_{80}$**	**Coverage range**
G1	.0006	.0806	$> .10$	.782–.797
G2	.2794	.0401	.0929	.715–.727

The sensitivity used 5,000 experiments per fixed grid point and 1,000 joint inner residual draws, each defined relative to the unresampled statistic so that the observed full-refit bootstrap bias is retained. The G2 zero-effect PASS rate .2794 is conditional on this residual model, not a measured real-world type-I error rate. Together with the low interpolated thresholds and coverage, it diagnoses the observed bias rather than validates prospective power. A data-level intervention followed by a complete pipeline rerun would be stronger. Full-decision PASS is zero throughout these one-gate grids because the other branches are held at their observed failures: $(G3 \vee G4)$ and G5 remain false (and G2 or G1 is also fixed when only the other gate is varied). It is not a finding that G1 alone makes every hypothetical full decision fail. This is neither a preregistered power study nor a revised decision.

R1 and R2 each contain 994 finite and six undefined frozen draws, so their official statuses remain **UNESTIMABLE**. Descriptively, on finite positions only, R1 has mean 0.8098, median 0.7673, and central 95% range $[0.3342, 1.4024]$; R2 has mean 0.5771, median 0.5974, and range $[-0.1484, 1.1354]$. These are explicitly *not* reduced-sample intervals and cannot rescue either branch.

C Extended experimental design and decision criteria

Diagnostic construction and fitting. The diagnostic vector combines cosine/L2 representation displacement, prompt-end next-token entropy and top-two margin changes, normalized neighborhood overlap, and local-density displacement from cosine kNN banks. Scaling and monitor fitting use eligible training rows. Reference banks use training and adaptation families, separately by checkpoint, regime, and broad transformation, excluding the query family and template group. Test groups never enter banks, and banks are not filtered by parse success or high consensus. Stress rows inform the control-category vocabulary, separate stress banks, and descriptive analyses, but not strict banks, primary fitting, repair adaptation, scoring, or weights. The cosine banks use $k = 10$. Full and

controls-only predictors fit a `StandardScaler` on training rows, followed by L2-logistic regression with $C = 1$, an intercept, no class weighting, the `lbfgs` solver, tolerance 10^{-4}, and at most 2,000 iterations (scikit-learn 1.6.1). The source scaler and coefficients are frozen together for target scoring; the target oracle fits its own scaler and coefficients on target training rows.

Repair adaptation. Affine and quantile calibration use source-training and target-adaptation logits. Paired ridge uses matched source/target adaptation diagnostics to target discrimination. Target outcome values are not used by these maps, although adaptation eligibility requires resolved outcomes. R3 matches rows by (family ID, variant ID) and fits a centered ridge map from target to source diagnostics. It selects $\lambda \in \{.001, .01, .1, 1, 10, 100\}$ by mean squared reconstruction error in shuffled, row-level adaptation cross-validation with $\min(5, n_{\text{pairs}})$ folds and seed 20270818; ties select the smaller λ. This inner split is row-level, unlike the group-preserving outer evaluation. At least four adaptation pairs are required. The selected map is refitted on all matched adaptation rows and applied only to diagnostics; the target controls are unchanged before frozen-source prediction.

For calibration repair $r \in \{R1, R2\}$, let B_0 and B_r be the test-row-weighted means across major-transition folds of the unrepaired and repaired Brier losses minus the target oracle's Brier loss. Recovery is $(B_0 - B_r)/B_0$ when $B_0 > \epsilon = .001$, and is undefined otherwise. This is a ratio of aggregated gaps, not an average of fold-specific ratios. The Brier loss is the mean squared error between probabilities and binary outcomes.

Configuration entry points. The supplement's `config/empirical_protocol_v2.yaml` records these declared settings under `features.neighborhood_policy.k`, `transport.repairs.R3_ridge_grid`, and `gates.G4_recovery_denominator_epsilon`; the fitting defaults and adaptation selection procedure are stated above. `config/CPU_PROTOCOL.json` and `config/CPU_METHOD_NOTE.md` provide the controlled-study settings and scope; its generating mechanism is specified in Section E.

Refitting and resampling. Bootstrap groups are sampled globally from primary inference-eligible groups jointly across folds and checkpoints. Predictors, oracles, and repairs are refitted; precomputed diagnostics and banks remain fixed. Permutation exchanges eligible outcome blocks within benchmark, regime, and block-size strata. Labels can move across fixed training, adaptation, and test roles, while diagnostic/control values and fold identities stay fixed. Blocks are not matched on subtype, variant, length, or lexical overlap. Thus the full-minus-controls statistic does not make this a test of $Y \perp D \mid C$; accepted permutations additionally condition on computable statistics, with ValueError/non-finite rejections recorded. Calibration recovery is undefined when its unrepaired denominator is at or below epsilon. A single undefined draw prevents an interval under the frozen all-draw Monte Carlo reporting rule; this finite-resampling criterion is distinct from population-parameter identifiability.

The following figures and table report the original criteria, uncertainty, and repair results.

The five outer-fold group counts are 228, 254, 247, 250, and 251; the original empirical resampling seed is 20270818.

Point family transport losses are .0238360 (pointwise), .0386414 (output), and .0404586 (neighborhood). Paired 95% intervals are $[-.0706277, .0685956]$ for neighborhood-minus-output, $[-.0525849, .0596569]$ for neighborhood-minus-pointwise, and $[-.0666983, .0733199]$ for output-minus-pointwise; the Holm-separated set is empty. R3 is estimable in all 1,000 draws, with mean recovery .0057681 and lower bound $-.0479471$, below both the .030 target and zero. A separate non-gating decoding analysis found that the greedy truncation disparity vanished under the frozen three-seed sampling condition. It does not change the transport or repair conclusions.

D Population definitions and detailed audit results

The following tables move reporting detail out of the page-limited main text; they do not change any population, estimand, or evidentiary class.

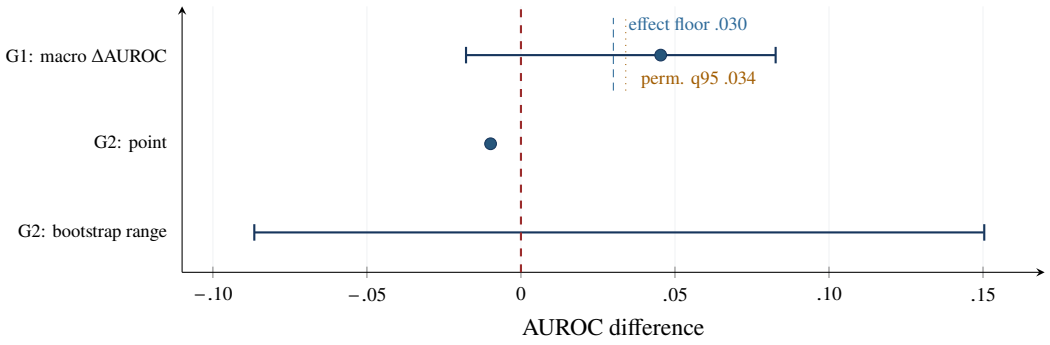
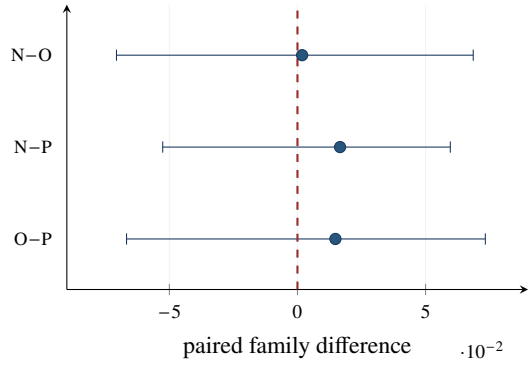
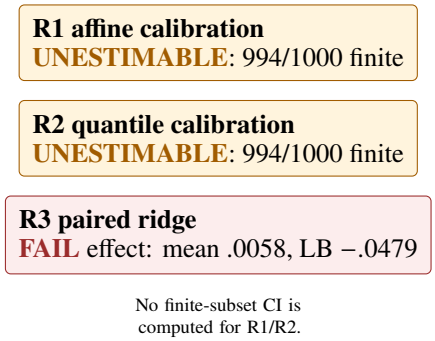


Figure 2: Points and bootstrap ranges. Dots are unresampled estimates; bars span the 2.5th–97.5th percentiles of the refit-aware cluster bootstrap. G1 clears the effect floor and permutation q95 (p = .01698), but its interval crosses zero. G2's point (−.00986) and percentile range ([−.08690, .15075]) are shown on separate rows; the bootstrap mean is +.02982. Both concern the same G2 statistic; the bootstrap mean is shifted from the point estimate. Neither frozen gate passes.



(a) G3: all Holm-tested family intervals cross zero; no family separated. P = pointwise, O = output, and N = neighborhood.



(b) G4: branches fail for different reasons.

Figure 3: Localization and repair status. **Left:** family contrasts are unresolved, not evidence of equality. **Right:** R1/R2 are unestimable under the all-draw rule, whereas R3 is estimable but fails its effect criterion. These categories must not be collapsed into "repair is impossible."

Table 11: Frozen diagnostic criteria and results. Names link the main text to the protocol labels and Equation (1). The joint diagnostic claim requires association, specificity, high-consensus support, and either localization or repair; operational reuse is evaluated separately against target-quality requirements. R1, R2, and R3 denote affine recalibration, quantile recalibration, and paired ridge feature alignment. † The G2 point difference is the raw frozen contrast; ‡ the interval is the percentile interval of the paired-bootstrap difference array, whose own mean is +.02982. The original-sample statistic and refit-bootstrap distribution target the same contrast; their bias is reported separately and the bootstrap mean must not replace the original point estimate.

Question (protocol label)	Estimand	Frozen rule and observed evidence	Status	Claim boundary
Association (G1)	Benchmark-balanced macro ΔAUROC	ΔAUROC ≥ .030; bootstrap LB > 0; point > permutation q95; both benchmark components > 0. Observed .04537, CI [−.01817, .08303], q95 .03402, permutation p = .01698; GSM .04144, IB .04929.	FAIL	Positive point signal, but association not established under the frozen uncertainty rule.
Transition specificity (G2)	Major minus revision-control transport loss	Major .02588, null .03573, point difference −.00986†, paired interval [−.08690, .15075]‡, one-sided p = .30969, Holm reject false.	FAIL	No detected separation; the wide interval does not establish equivalence or causal attribution.
Localization (G3)	Pairwise diagnostic-family differences	All three Holm-tested CIs cross zero; separated pairs = ∅.	FAIL	No tested family localized a transport-loss difference.
Repair (G4)	Calibration recovery or ≥ .030 discrimination recovery	R1/R2: 994/1000 finite; all-draw reporting requirement not met. R3: 1000/1000 finite, mean recovery .00577, lower bound −.04795.	FAIL	R1/R2 unestimable under the reporting rule; R3 below the effect criterion.
High-consensus support (G5)	High-consensus direction/support	Core direction failed; no G3/G4 branch passed to supply support.	FAIL	No high-consensus support for the joint diagnostic claim.
Overall	$G1 \wedge G2 \wedge (G3 \vee G4) \wedge G5$	Every positive gate failed.	NO-GO	the frozen overall decision flag is false; utility is a separate decision target.

Table 12: Population definitions and analysis scope. Eligibility counts are not the fitted or scored sample size of every historical branch. Full artifact identities appear in Section A.

Population	Size	Definition and role
Held-out H	1,230 groups; 2,892 candidate rows	1,045 GSM-Plus and 185 InvariantBench groups; 1,505 strict and 1,387 stress candidate rows.
Screened study table	2,889 rows	Automatic screening retains 1,504 strict and 1,385 stress rows. This input table is not the primary scoring sample.
Primary eligibility	1,504 rows	Automatic pass and strict regime. Actual fits and scores additionally require common resolution and branch-finite diagnostics.
High-consensus eligibility	1,291 rows	Additional unanimity restriction before common-resolution and branch-specific requirements.
Reporting OOF population	1,361 rows; 1,139 groups	Common four-checkpoint strict reporting subset. Its high-consensus subset has 1,178 rows in 1,017 groups. These counts are not preserved runtime masks for every original fit.
Exploratory-A main	1,164 groups	DPO/RLVR pairwise common-resolved; 94.63% of H. Conditional post-confirmatory utility only.
Exploratory-A null	1,169 groups	RLVR-null pairwise common-resolved; 95.04% of H. Descriptive matched-null reference only.
Exploratory-B	1,230 groups	Full H; unresolved target outcomes recoded inconsistent. A missingness sensitivity, not evidence that missingness is ignorable.

Table 13: Post-confirmatory exploratory audit on the 1,164-group DPO-to-RLVR common-resolved population. Errors are relative to each monitor's descriptive macro AUROC on this finite evaluation set and the .60 operating threshold, not known population quality. "Non-est." counts unavailable lower bounds: these draws default to distrust and remain in the denominator. Both references are below .60, so error equals false trust. Each $n = 25$ cell has 0/500 errors with a stored 95% Wilson Monte Carlo interval [0, .0076]; the intervals are cell-specific. Budget 400 is infeasible and has no draws.

	Full monitor (reference AUROC .591)				**Pointwise+output** (reference AUROC .594)			
Budget	Draws	Non-est.	Default frac.	Error / false trust	Draws	Non-est.	Default frac.	Error / false trust
25	500	500	1.000	.000	500	500	1.000	.000
50	500	500	1.000	.000	500	500	1.000	.000
100	500	444	.888	.002	500	444	.888	.006
200	500	195	.390	.018	500	195	.390	.034
400	0	0	—	infeasible	0	0	—	infeasible

D.1 Audit rule, sampling populations, and references

Audit budgets count template groups. Outer samples are allocated as evenly as possible between benchmarks: 13/12 groups at budget 25, and equal allocations at 50, 100, 200 and the infeasible 400-group budget. These are chosen design settings, not data-derived sufficiency thresholds. Macro AUROC requires at least two folds containing both classes per benchmark. Its nominal 90% lower bound is the tenth percentile of valid macros, available only when at least 900/1,000 inner template-cluster resamples are computable. This differs from repair's 100% all-draw rule.

Each of the eight analysis–arm–variant cells at budget 25 has 500 recorded audits, 500 unavailable bounds, and 0 observed decision errors. The stored 95% Wilson Monte Carlo interval for each error rate is [0, .007624340465103374], conditional on that cell's finite-sample quality reference. No interval is constructed by treating the 4,000 cross-cell records as independent repetitions of a single experiment.

The descriptive quality references are:

Evaluation	Full	Pointwise+output
Exploratory-A major	.5914711413064614	.5941778959978761
Exploratory-A revision	.5796362861939064	.5542655370872527
Exploratory-B major	.5791600788515372	.5753281454732979
Exploratory-B revision	.5891077761596175	.5614681699884994

These finite-set references are compared with the operating floor .60; their stored names do not establish population ground truth.

Using seed 20270818, descriptive common-resolved major-pair target macro AUROCs are .5914711 (full monitor) and .5941779 (pointwise+output); matched-null values are .5796363 and .5542655. Full-H major-pair missingness-sensitivity values are .5791601 and .5753281. For the common-resolved major pair, each variant has 500, 500, 444, and 195 undefined bounds out of 500 draws at budgets 25, 50, 100, and 200.

For the major pair, DPO had 63 unresolved/parse-failed groups and RLVR had 17; 14 overlapped, so 66 unique groups were excluded from the common-resolved population. For the null pair, null A had 46 unresolved and null B 44; 29 overlapped, so 61 unique groups were excluded. The main and null conditional populations therefore have different membership and are not a paired full-population contrast.

Across all exploratory cells, 11,856/16,000 bounds were unavailable. Non-estimable draw counts were 4,000/4,000 at $n = 25$, 3,986/4,000 at $n = 50$, 3,008/4,000 at $n = 100$, and 862/4,000 at $n = 200$. Budget 400 contains zero draws by construction. Figure 4 plots the unavailable-bound fraction for one cell.

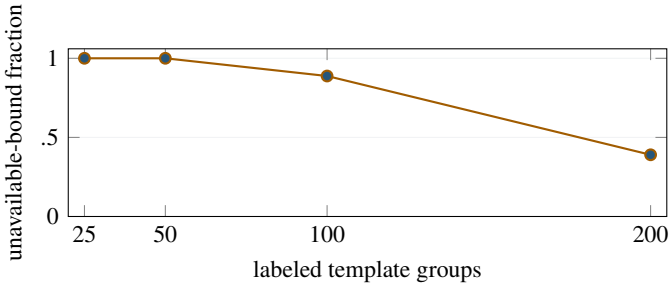


Figure 4: OLMo common-resolved DPO-to-RLVR audit, full monitor, 1,164 groups. At 25 groups all 500 bounds are unavailable and all decisions fall back to distrust, giving zero errors against the below-threshold finite reference. The points are fractions of 500 audits per budget; infeasible budget 400 has no results.

Reading the supplied audit outputs. Exploratory sources were archived after execution without complete independent runtime provenance. In eight zero-draw records, nested JSON parent keys identify the intended variant despite a reporting-label error; no scientific draw, rate, or interval changes. Unavailable quantities are stored as bare `NaN` and require the documented permissive JSON handling.

E Controlled feature-level design and continuation

Aim and scope. To distinguish behavior of the statistical procedure from the particular model trajectory, a separate post-confirmatory study generates data at the six-dimensional feature-table interface. It reuses the original fitting, repair, and gate implementations with group-preserving five-fold roles. The design was locally hash-locked before the main synthetic output; it was not preregistered in a third-party public registry. This is not a simulation of language generation, parsing, semantic screening, or neighborhood-bank construction. All 480 planned IDs reached terminal states before the original absolute deadline, with 468 complete and 12 known unavailable outputs. No scenarios, seeds, statistics, or thresholds were revised in response to the terminal results.

Data-generating mechanisms. Each repetition contains 300 template groups with two rows per group, allocated 240/60 to the two benchmark strata. For group g, row $r \in \{1, 2\}$, and channel $c \in \{\text{major}, \text{revision}\}$, the source diagnostic vector is

$$D_{gr}^{(c)} = \mu(2Y_g^{(c)} - 1)e_3 + \sqrt{.4}\,U_g + \sqrt{.3}\,V_g^{(c)} + \sqrt{.3}\,E_{gr}^{(c)},$$

where e_3 is the third coordinate basis vector and U, V, E are independent standard six-dimensional Gaussians except for the explicitly shared indices. The coordinate order is cosine displacement, raw L2 displacement, entropy change, margin change, neighborhood overlap, and local-density displacement. These are artificial features bearing the empirical interface's names, not quantities extracted from generated text. The two rows share their channel's group label. With probability .5, the two channels share one Bernoulli(.25) label; otherwise their labels are independent Bernoulli(.25) draws. The class separation is $\mu = .5$, except when N0 sets $\mu = 0$. Before rotation, the Gaussian noise has within-channel, within-group correlation .7 and cross-channel, same-group correlation .4 in matching coordinates; these are noise correlations, not total feature correlations.

Source and target reuse the same channel realization and label. At the target, coordinates 3–4 are multiplied by the planar rotation matrix $\left(\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right)$; all other coordinates stay fixed. C1 additionally translates both of these coordinates by 2 at the major target. The two Gaussian controls are drawn independently at group level and shared by its rows and all four roles; the benchmark indicators are also shared. High-consensus membership is an independent group-level Bernoulli(.8) mask, not simulated human agreement.

Groups are shuffled within each benchmark and evenly allocated to five folds. In outer fold $k \in \{0, \ldots, 4\}$, test groups have fold k, adaptation groups have fold $(k + 1) \bmod 5$, and the other three folds supply training rows. The same generated observations and fold memberships are reused

across the five outer evaluations and four roles. N3 independently draws group-level missingness for each role and uses the intersection resolved in all four.

The eight mechanism settings are listed in Table 14. The two continuous controls are generated independently of the labels and diagnostic signal; benchmark indicators are also included. No scenario tests a controls-only outcome with diagnostic–control dependence: the conditional null in which Y depends only on C while D and C are correlated is not isolated. The experiment invokes the diagnostic criteria, not the operational utility audit; it does not measure behavior at target quality .59, .60, or .61.

Table 14: Pre-fixed feature-level scenarios. Angles are in radians; major/revision denotes the two transitions.

Scenario	Rotation	Purpose / additional intervention
N0	0/0	No predictive signal ($\mu = 0$).
N1	0/0	Useful signal, no update.
N2	.8/.8	Equal major and revision updates.
N3	.8/.8	Equal updates with checkpoint-specific MCAR loss.
P1	.4/0	Small major-only rotation.
P2	.8/0	Medium major-only rotation.
P3	$\pi/2$ / 0	Large major-only rotation.
C1	0/0	Add 2 to each of the two rotated major-target coordinates.

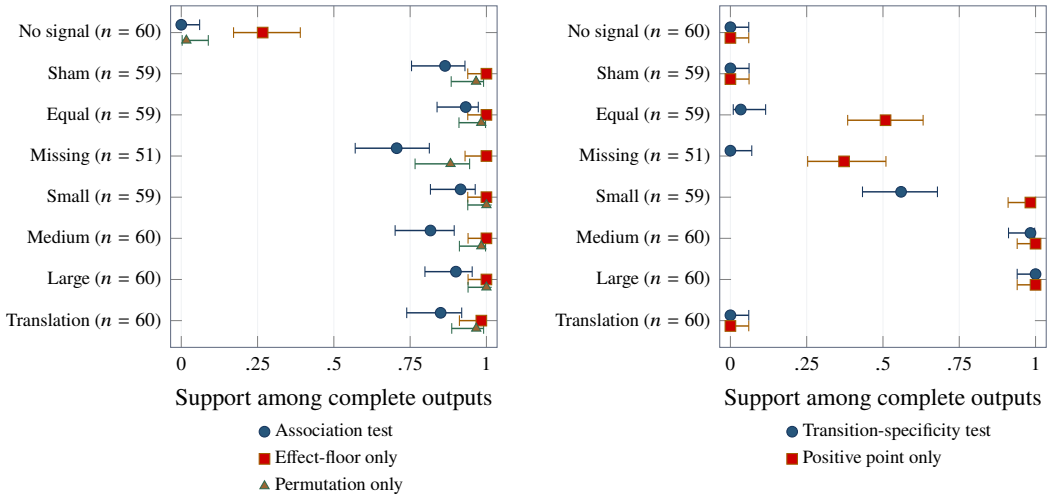
N3 uses role-specific missingness probabilities .10, .05, .05, .10 for major source/target and revision source/target; these are not a 10% loss of the common eligible population. Equal-update and rank-preserving null mechanisms target zero differential transport. Positive rotations test a specified direction of degradation, not guaranteed acceptance by all gates. Analytic Bayes references describe the mechanisms, not exact finite-fit truths for every reported interval.

Evaluation and incomplete outputs. The fixed plan is 60 repetitions per scenario (480 unique IDs), five folds, 1,000 refit bootstrap draws, and 200 accepted permutations per repetition. The permutation budget differs from the empirical study's 1,000. Comparators are the G1 effect floor alone, its permutation criterion alone, and a positive G2 point contrast alone. We report method output availability separately from gate and comparator support conditional on complete evaluable outputs, with explicit denominators and Monte Carlo uncertainty (Morris et al., 2017). For predeclared zero-differential mechanisms, the saved G2 intervals are compared with zero; we report zero-reference inclusion frequencies rather than claim nominal coverage of a known finite-fit parameter. Role-specific eligibility, regularized fits, and conditioning on complete outputs are not proved to preserve a zero population contrast. Analytic Bayes references are not substituted as interval truths. Repair estimability is reported separately.

After an observed single-class InvariantBench test-fold bootstrap failure, a disclosed outer-dispatch amendment permits that proven mathematical unavailability to be recorded without stopping unrelated fixed repetitions. It does not change a scientific statistic, seed, draw, retry, or gate, and does not impute missing AUROC or replace a failed repetition. Unknown errors remained fail-closed. All fixed IDs reached terminal records before the original deadline; the final unknown-error, interrupted, and unstarted counts are zero. The supplied unavailable-output summary retains the twelve error identities and single-class diagnostics, including the first error at repetition 005/N3; the raw error files are not included. Interrupted attempts retained their logs and resumed only their original fixed seeds; no completed or known-unavailable ID was replaced. Conditional performance is not presented as unconditional calibration.

E.1 Do the diagnostic criteria respond to known changes?

The OLMo criteria all failed. That result alone cannot show whether the procedure can distinguish situations with and without a specified change. We therefore generate six-dimensional artificial features with known changes. The no-signal scenario checks association declarations; the no-update scenario checks whether useful source signal is mistaken for update-specific change. Equal rotations affect both transitions; major-only rotations increase from small to large; a translation tests a rank-preserving change; and an equal-rotation scenario adds missing completely at random (MCAR),



(a) Support for diagnostic association.

(b) Evidence that refitting helps more after the major update.

Figure 5: Complete criteria and pre-specified simpler comparisons in all eight scenarios. Sham denotes N1; Equal and Missing are equal rotations without and with MCAR (N2/N3); Small, Medium, and Large are major-only rotations (P1–P3); Translation is C1. Points divide pass counts by the complete-output denominator shown in each row; bars are retained conditional Wilson 95% intervals. The 12 unavailable outputs, including nine in N3, are reported separately in Table 15 and Section E.2.

independently of the generated values (Table 14). These are controlled feature transformations, not new language-model updates.

Each scenario has 60 fixed repetitions. We reuse the fitting, repair, and criterion implementations with five group-preserving folds, 1,000 refit bootstrap draws, and 200 accepted permutations per complete output. Here a criterion's *support rate* is the fraction of complete repetitions meeting it. Of 480 planned repetitions, 468 return complete outputs and 12 are unavailable, including nine in the MCAR scenario. Wilson 95% intervals describe Monte Carlo uncertainty conditional on completion; availability uses all 60 planned repetitions. Unavailable outputs are not rejections (Section E.2).

Table 15: Availability and joint diagnostic support in all eight scenarios. There are 60 planned repetitions per row; n is the complete-output count. Unavailable outputs are not rejections, and joint support is not correct reuse. All individual tests, intervals and missing-outcome bounds are in Section E.2; parameters are in Table 14.

	Condition	Complete	Unavailable	Joint support/n
N0	No signal	60/60	0/60	0/60
N1	Useful monitor, sham	59/60	1/60	0/59
N2	Equal rotations	59/60	1/60	2/59
N3	Equal rotations, MCAR	51/60	9/60	0/51
P1	Small rotation	59/60	1/60	14/59
P2	Medium rotation	60/60	0/60	49/60
P3	Large rotation	60/60	0/60	54/60
C1	Calibration translation	60/60	0/60	0/60

The comparisons show both a benefit and a cost of stronger requirements. With no signal, the effect floor alone supports 16/60 repetitions, permutation-only supports 1/60, and the full association rule supports 0/60. With equal rotations, requiring transition specificity reduces support from 30/59 positive point contrasts to 2/59. The same rule reduces small-rotation support from 58/59 to 33/59, with only 14/59 meeting the full diagnostic conjunction. Medium and large rotations receive specificity

support in 59/60 and 60/60 repetitions (Figure 5). Reducing support under equal updates thus comes with lower sensitivity to small changes, rather than a uniform improvement in every decision.

The criteria also answer different questions: translation meets the repair criterion in 53/60 repetitions while receiving no specificity support. This study begins at artificial features and excludes generation, parsing, semantic screening, and bank construction. It omits outcomes driven only by controls with diagnostic–control dependence, and is separate from the known-quality audit experiment in Section G.3. The results characterize these declared feature scenarios, not end-to-end calibration or a correction of the separate .2794 residual-model sensitivity.

E.2 Complete results for the controlled feature-level study

The study reached a terminal outcome for all 480 pre-specified IDs: 468 complete outputs and 12 method outputs unavailable. No ID remained active, interrupted, unstarted, or classified as an unknown execution error. The following tables include every scenario and every pre-specified criterion or simpler comparison. They describe this synthetic feature-table mechanism study, not empirical model-update validation or the operational audit decision at threshold .60.

Denominators and output availability. A conditional fraction divides the number of passes by the number of complete outputs in that scenario. Its Wilson interval describes outer Monte Carlo uncertainty conditional on obtaining a complete output. The 1,000 bootstrap draws and 200 accepted permutations within an output do not enlarge that outer sample size. Unavailable outputs are neither passes nor rejections. Table 16 reports output availability over all 60 planned IDs per scenario; unavailability intervals are the complements of the retained availability Wilson intervals.

Table 16: Availability and unavailability across all planned IDs. These denominators are 60 in every row, unlike the conditional criterion denominators in the following tables.

	Complete	Available fraction [95% CI]	Unavailable	Fraction [95% CI]
N0	60/60	1 [.940,1]	0/60	.000 [.000,.060]
N1	59/60	.983 [.911,.997]	1/60	.017 [.003,.089]
N2	59/60	.983 [.911,.997]	1/60	.017 [.003,.089]
N3	51/60	.850 [.739,.919]	9/60	.150 [.081,.261]
P1	59/60	.983 [.911,.997]	1/60	.017 [.003,.089]
P2	60/60	1 [.940,1]	0/60	.000 [.000,.060]
P3	60/60	1 [.940,1]	0/60	.000 [.000,.060]
C1	60/60	1 [.940,1]	0/60	.000 [.000,.060]

All diagnostic criteria and simpler comparisons. Tables 17 to 19 give all retained counts, conditional fractions, and Wilson intervals. The overall flag is $G1 \wedge G2 \wedge (G3 \vee G4) \wedge G5$. It is a diagnostic conjunction, not a measure of correct monitor reuse. The declared zero-reference mechanisms concern G2 specifically: repair-criterion support in N2 (59/59) or C1 (53/60) is not by itself a false positive for that zero contrast. Conversely, satisfying the disjunction $G3 \vee G4$ does not identify a successful repair.

Missing-outcome logical bounds. For each criterion with k observed passes and u unavailable outputs, the fraction among the 60 planned tasks is bounded by $[k/60, (k+u)/60]$ if unavailable outcomes could take either value. Table 20 reproduces those logical bounds; they are not confidence intervals or an imputation of the missing outcomes. For example, N3 has conditional G2 support $0/51$ with Wilson upper endpoint .070, but its planned-task logical range is $[0, 9/60]$. The two quantities answer different questions.

Repair estimability and zero-reference inclusion. Table 21 reports repair estimability counts among complete outputs; being estimable is not equivalent to meeting a repair criterion. Zero-reference inclusion is reported only for N0, N1, N2, N3, and C1, the mechanisms declared to have zero G2 contrast. It is the observed proportion of their retained G2 intervals containing zero, conditional on a complete output, with a Wilson interval for that proportion. It is not a guarantee of nominal coverage for the entire procedure. No zero-reference inclusion rate is assigned to the rotation alternatives. The

Table 17: All eight scenarios: diagnostic association, transition contrast, and localization. Each cell gives pass count/n; conditional fraction, then its retained Wilson 95% interval on the second line. All denominators refer to complete outputs, not the 60 planned replicates.

Scenario	Association (G1)	Transition contrast (G2)	Localization (G3)
N0	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]
N1	51/59; .864 [.755, .930]	0/59; .000 [.000, .061]	0/59; .000 [.000, .061]
N2	55/59; .932 [.838, .973]	2/59; .034 [.009, .115]	55/59; .932 [.838, .973]
N3	36/51; .706 [.570, .813]	0/51; .000 [.000, .070]	34/51; .667 [.530, .780]
P1	54/59; .915 [.816, .963]	33/59; .559 [.433, .678]	15/59; .254 [.161, .378]
P2	49/60; .817 [.701, .894]	59/60; .983 [.911, .997]	57/60; .950 [.863, .983]
P3	54/60; .900 [.799, .953]	60/60; 1 [.940, 1]	60/60; 1 [.940, 1]
C1	51/60; .850 [.739, .919]	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]

Table 18: All eight scenarios: repair, high-consensus, and the diagnostic conjunction. Each cell gives pass count/n; conditional fraction, then its retained Wilson 95% interval on the second line. All denominators refer to complete outputs, not the 60 planned replicates.

Scenario	Repair criterion (G4)	High-consensus (G5)	Conjunction (GO)
N0	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]
N1	0/59; .000 [.000, .061]	0/59; .000 [.000, .061]	0/59; .000 [.000, .061]
N2	59/59; 1 [.939, 1]	26/59; .441 [.322, .567]	2/59; .034 [.009, .115]
N3	43/51; .843 [.720, .918]	11/51; .216 [.125, .346]	0/51; .000 [.000, .070]
P1	0/59; .000 [.000, .061]	15/59; .254 [.161, .378]	14/59; .237 [.147, .360]
P2	59/60; .983 [.911, .997]	59/60; .983 [.911, .997]	49/60; .817 [.701, .894]
P3	60/60; 1 [.940, 1]	60/60; 1 [.940, 1]	54/60; .900 [.799, .953]
C1	53/60; .883 [.778, .942]	0/60; .000 [.000, .060]	0/60; .000 [.000, .060]

Table 19: All eight scenarios: the three pre-specified simpler comparisons. Each cell gives pass count/n; conditional fraction, then its retained Wilson 95% interval on the second line. All denominators refer to complete outputs, not the 60 planned replicates.

Scenario	G1 effect-floor only	G1 permutation only	G2 positive-point only
N0	16/60; .267 [.171, .390]	1/60; .017 [.003, .089]	0/60; .000 [.000, .060]
N1	59/59; 1 [.939, 1]	57/59; .966 [.885, .991]	0/59; .000 [.000, .061]
N2	59/59; 1 [.939, 1]	58/59; .983 [.910, .997]	30/59; .508 [.384, .632]
N3	51/51; 1 [.930, 1]	45/51; .882 [.766, .945]	19/51; .373 [.253, .510]
P1	59/59; 1 [.939, 1]	59/59; 1 [.939, 1]	58/59; .983 [.910, .997]
P2	60/60; 1 [.940, 1]	59/60; .983 [.911, .997]	60/60; 1 [.940, 1]
P3	60/60; 1 [.940, 1]	60/60; 1 [.940, 1]	60/60; 1 [.940, 1]
C1	59/60; .983 [.911, .997]	58/60; .967 [.886, .991]	0/60; .000 [.000, .060]

Table 20: Logical bounds for missing outcomes, shown as numerator ranges: divide every endpoint by 60. Floor, Perm., and Point denote G1 effect-floor only, G1 permutation only, and G2 positive-point only. These are the retained full-task bounds, not Wilson intervals. Equal endpoints occur when all 60 outputs are complete.

	G1	G2	G3	G4	G5	GO	Floor	Perm.	Point
N0	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[0,0]	[16,16]	[1,1]	[0,0]
N1	[51,52]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[59,60]	[57,58]	[0,1]
N2	[55,56]	[2,3]	[55,56]	[59,60]	[26,27]	[2,3]	[59,60]	[58,59]	[30,31]
N3	[36,45]	[0,9]	[34,43]	[43,52]	[11,20]	[0,9]	[51,60]	[45,54]	[19,28]
P1	[54,55]	[33,34]	[15,16]	[0,1]	[15,16]	[14,15]	[59,60]	[59,60]	[58,59]
P2	[49,49]	[59,59]	[57,57]	[59,59]	[59,59]	[49,49]	[60,60]	[59,59]	[60,60]
P3	[54,54]	[60,60]	[60,60]	[60,60]	[60,60]	[54,54]	[60,60]	[60,60]	[60,60]
C1	[51,51]	[0,0]	[0,0]	[53,53]	[0,0]	[0,0]	[59,59]	[58,58]	[0,0]

ideal Bayes classifier references characterize the generating mechanism and are not finite-fit truth for a regularized monitor; they are not used as interval-coverage targets here.

Table 21: Repair estimability counts and declared-zero G2 reference inclusion for all scenarios. Repair columns use the complete-output denominator n; the final column is a separate interval-inclusion summary, not repair success.

	Complete n	Affine	Quantile	Paired ridge	Zero included/n [95% CI]
N0	60	0	0	60	60/60 [.940,1]
N1	59	0	0	59	59/59 [.939,1]
N2	59	30	30	59	57/59 [.885,.991]
N3	51	7	7	51	51/51 [.930,1]
P1	59	0	0	59	— (alternative)
P2	60	26	26	60	— (alternative)
P3	60	59	59	60	— (alternative)
C1	60	53	53	60	60/60 [.940,1]

Permutation completion and unavailable outputs. Each of the 468 complete outputs records 200 attempted and accepted permutations: 93,600 each in total, with no ValueError or non-finite rejections. The 12 unavailable IDs lack completed permutation-stage records, not zero counts. Their saved diagnostics identify single-class InvariantBench test folds in a bootstrap draw, making a required AUROC undefined. No statistic is imputed, seed replaced, or unavailable output counted as a scientific rejection. The twelve unavailable-output diagnostics remain in the supplied summary; Table 16 gives scenario-specific denominators.

F Second-lineage Tülu transport measurements

Checkpoints and response measurement. The source is `allenai/Llama-3.1-Tulu-3-8B-DPO`; the three GRPO roles are revisions of `allenai/Llama-3.1-Tulu-3.1-8B` (Allen Institute for AI, 2025). Table 22 identifies the four snapshots. These are a DPO-stage source and ordered GRPO checkpoints; exact bitwise parent–child ancestry is not asserted. The major comparison is DPO to step 1920, and the stage-local revision is step 1840 to step 1880.

Table 22: Checkpoint roles and repository revisions for Tülu.

Role	Revision
DPO source	`a7beb67e33ffd01cc87ac3b46cadc1000985b8db`
GRPO step 1840	`879da0e1676f1fb4a7421fceb355bab82c4aecc0`
GRPO step 1880	`8f5a383173ecf6ffc686bab4e445708add667e33`
GRPO step 1920	`46239c2d07db76b412e1f1b0b4542f65b81fe01f`

All checkpoints receive the same chat-formatted prompts without a system message. Each prompt ends with the instruction "On the last line, write `FINAL_ANSWER:` followed by your final answer only." Greedy generation uses temperature zero and an initial 16,384-token output cap. Only length-terminated outputs receive one retry at 32,768 tokens; there is no further extension or parser-directed regeneration. The structured v2 parser was developed on 800 outputs outside H and then held fixed across roles. The reused H problems are not globally unseen data. At each checkpoint, Y_{inc} compares the canonical and variant structured answers. Both responses must terminate normally and their answers must be comparable; unresolved, non-stop, or incomparable outcomes remain unknown.

Population and coverage. The planned population contains all 1,504 originally primary-eligible strict pairs in 1,230 H groups, requiring 2,817 distinct prompts per checkpoint. No new semantic screen or gold-answer input is used. Table 23 separates each role's available pairs from the four-role intersection. Diagnostics are finite for all 1,504 pairs at every role, so the intersection is restricted by answer observability. Its 1,196 pairs in 1,051 groups define the conditional population for both TRAIN and TEST, without imputing the missing outcomes.

Table 23: Tülu coverage by benchmark. Each cell orders GSM-Plus / InvariantBench. Column headers give the full planned denominators. Groups contain at least one eligible pair. Truncation counts are individual prompts still length-terminated at the final cap, not excluded-pair counts.

Role	Eligible pairs of 1,047 / 457	Eligible groups of 1,045 / 185	Final truncations of 2,094 / 723
DPO	992 / 341	990 / 141	2 / 7
Step 1840	1,029 / 354	1,027 / 141	5 / 19
Step 1880	1,017 / 348	1,015 / 142	9 / 29
Step 1920	1,021 / 367	1,019 / 147	10 / 25
Common to all four	949 / 247	947 / 104	—

Features, fitting, and aggregation. At the last non-padding prompt token, we extract the post-decoder-block-21 state (zero-based block index, hidden-state tuple index 22; dimension 4,096). Inference uses bfloat16 and stored vectors use float32. The four diagnostics are cosine distance, raw Euclidean distance, and variant-minus-canonical differences in full-vocabulary entropy and the top-two logit margin. Entropy uses natural logarithms. Controls are character-count difference, lowercase whitespace-token Jaccard overlap, and one-hot benchmark, subtype, and variant-type indicators; category levels come from planned metadata. The original five group-preserving TRAIN/ADAPT/TEST assignments are retained. ADAPT is unused. Each role's full and controls-only L2-logistic monitors use TRAIN-only standardization, $C = 1$, `lbfgs`, a 2,000-iteration limit, and seed 20270818. All 40 planned fits complete.

Table 24: Absolute Tülu OOF AUROC on the common subset. Each monitor is trained and evaluated at the named checkpoint. Values are rounded to seven decimals; full and controls-only monitors share folds and eligible rows.

	Full		Controls only	
Checkpoint	Pooled	Macro	Pooled	Macro
DPO	.6774294	.5081068	.6775315	.5020628
Step 1840	.6994518	.5005080	.7086836	.4916353
Step 1880	.6767805	.4914501	.6858941	.4961228
Step 1920	.7209235	.5361128	.7112816	.5362444

For each transition, target-refitted and frozen source full monitors score identical target TEST rows against target outcomes. Pooled AUROC combines benchmarks within each TEST fold, then row-weights folds. Macro AUROC group-weights TEST folds within each benchmark, then weights benchmarks equally. No cross-fold pairs are compared, and score ties receive half credit. An undefined required positive-weight fold or cell makes the aggregate unavailable; no cells are dropped and renormalized.

Subsequent cached-score comparisons. Table 25 adds a subsequent descriptive controls-only comparison. The controls matrix is constructed once from the planned pairs and shared by all four checkpoints. Common-four eligibility and fold roles are also shared. Thus a saved source controls probability is the source monitor's score on the same target controls input; evaluating it against the target label requires no refit. Subtracting its target AUROC from that of the target-fitted controls monitor gives controls-only T under the same aggregation rules. Full and controls-only contrasts are reported together, without new uncertainty intervals.

At the source, inconsistency labels are positive for 157/949 GSM-Plus pairs (.1654373) and 154/247 InvariantBench pairs (.6234818). The prevalence difference motivates including task descriptors; it does not by itself attribute pooled discrimination to benchmark identity. The controls-only contrast also changes with aggregation, establishing that this dependence occurs without the four diagnostics, not that controls fully explain the full-monitor contrast.

Table 25: Tülu cached-score comparisons. The first six rows show absolute AUROC for the major transition, with each checkpoint's own labels on the common subset. The final two rows compare major and revision transport. The controls contain text and metadata, not benchmark identity alone. These descriptive additions use no new fits or draws.

Measurement	Pooled	Macro
Full: source at source	.6774294	.5081068
Full: frozen source at target	.6752057	.4544243
Full: target-refitted at target	.7209235	.5361128
Controls: source at source	.6775315	.5020628
Controls: frozen source at target	.6843944	.4648673
Controls: target-refitted at target	.7112816	.5362444
Full: major minus revision	+.0440059	+.0793151
Controls: major minus revision	+.0266397	+.0658699

Paired uncertainty and signed decomposition. The 1,000 fixed-score bootstrap draws use PCG64 seed 20260911, sampling common template groups with replacement within benchmark-by-TEST-fold strata. Each sampled group retains all its rows. The same multiplicities apply to every role, monitor, and transition, holding fitted models and scores fixed. Intervals are the linearly interpolated 2.5th and 97.5th percentiles, reported for a metric only if all 1,000 draws are finite. All reported metrics meet that rule; no draw is replaced. These intervals omit the refitting uncertainty included in the OLMo bootstrap.

Table 26: Tülu transport and signed aggregation contributions. Within is within-benchmark reweighting; cross is cross-benchmark pair removal. The last column equals within plus cross, up to rounding.

	Pooled T	Macro T	Within	Cross	Macro minus pooled
Major	.0457178	.0816885	+.0588785	-.0229079	+.0359707
Revision	.0017120	.0023734	+.0131307	-.0124693	+.0006614
Major−revision	.0440059	.0793151	+.0457478	-.0104385	+.0353092
Paired 95% interval for the contrast difference: $[-.0275562, .0970914]$.					

The decomposition applies the fixed-prediction identity in Section B.3; its terms are signed algebraic contributions, not isolated causal effects. The paired interval for the contrast difference includes zero; differing zero inclusion in the pooled and macro intervals does not establish a difference between the schemes. This analysis measures transport with four diagnostics on an observed subset; it neither reruns the original localization, repair, and high-consensus tests nor establishes a stronger monitor.

G Joint fixed-score sensitivity and known-quality audit behavior

These supplementary analyses were specified after inspecting the original empirical results. Their sampling rules, seeds, and implementation identities were recorded before the new outputs. They leave the original inference and the 480-repetition feature study unchanged.

G.1 Paired uncertainty on the OLMo reporting cache

This joint design is separate from the earlier macro-only sensitivity in Section B.2; their finite-draw counts describe different resampling analyses.

We resample template groups with replacement within each benchmark–test-fold stratum, retaining the stratum's original group count. All rows of a selected group receive its multiplicity. The same draw weights both transitions and both reused and refitted scores. Thus the covariance between pooled and macro contrasts is preserved when computing $D_{\text{agg}} = C_{\text{macro}} - C_{\text{pool}}$. Pooled evaluation row-weights the test folds; macro evaluation group-weights folds within each benchmark and weights the two benchmarks equally. Ties receive half credit and no cross-fold ranking is formed. We use

10,000 PCG64 draws with seed 20260916 and linear-interpolated 2.5th/97.5th percentiles. Scores and labels are fixed: the intervals do not include refitting uncertainty.

The frequencies of $C_{\text{pool}} < 0 < C_{\text{macro}}$ and the reverse ordering are reported separately. These describe the bootstrap draws; neither is a hypothesis-test p-value or a probability that a population sign pattern is true. Invalid draws are retained, not replaced.

Table 27: Joint fixed-score OLMo resampling. The primary interval requires all 10,000 draws to be defined. The final column describes only the finite subset and is not a replacement primary interval.

Quantity	Point	Finite draws	Primary 95% interval	Finite-subset percentiles
C_{pool}	−.0098565	10,000	[−.0858230, .0626937]	same
C_{macro}	.0185912	9,949	unavailable	[−.0741811, .0999305]
D_{agg}	.0284477	9,949	unavailable	[−.0372112, .0763779]

The pooled draw mean is −.0101303. Conditional on finite values, the macro and difference means are .0116860 and .0218558. The macro contrast is undefined in 51 draws: 47 lack a finite major-transition macro advantage and four lack the revision advantage. Among the 9,949 jointly defined draws, 2,399 have the observed sign ordering and 375 the reverse; 3,583 have two positive contrasts and 3,592 have two negative contrasts. None has an exactly zero contrast. Thus the conditional observed-direction frequency is .24113, not evidence of a stable population reversal. Neither the pooled interval nor the descriptive finite-subset percentile ranges excludes zero.

The all-finite requirement is sensitive to the number of draws. If independent draws have undefined probability q, the probability of satisfying it is $(1 - q)^B$. Plugging in the observed $q = 51/10{,}000$ gives .5997 at $B = 100$, .00602 at $B = 1{,}000$, and 6.23×10^{-23} at $B = 10{,}000$. This calculation describes the reporting rule under the independent-draw model; it neither validates the conditional percentiles as confidence intervals nor changes the specified primary rule.

G.2 Evaluation-only exclusion of disputed equivalences

The existing semantic-screen records identify 27 primary rows with majority nonequivalence opinions; their intersection with the reporting OOF cache is 18 row IDs. We remove those same IDs from both transitions and recompute fixed score point estimates, reporting the retained and excluded versions together. No monitor is refitted and no exclusion-version interval is added. This changes the evaluation subset only: the rows' earlier roles in training, adaptation, or reference-bank construction are unchanged. Disagreement among reviewers is not treated as proof of an incorrect outcome label.

Table 28: OLMo evaluation-subset sensitivity with fixed scores. Row/group counts apply to each transition. No new fit or exclusion-specific interval is used.

Evaluation subset	Rows	Groups	C_{pool}	C_{macro}	D_{agg}
Including disputed rows	1,361	1,139	−.0098565	.0185912	.0284477
Excluding 18 row IDs	1,343	1,127	−.0101373	.0182006	.0283379

The major-transition advantages change from .0258751 pooled and .0365661 macro to .0267294 and .0351139. Revision advantages change from .0357315 and .0179748 to .0368667 and .0169134. The observed contrast therefore retains its two signs. Benchmark/fold denominators and the exact removed IDs are included with the numerical outputs.

G.3 Audit decisions with an analytic quality reference

The original OLMo utility references are all below the operating threshold. To evaluate both sides, we generate independent groups, each with one observation:

$$Y \sim \text{Bernoulli}(.25), \qquad S = \sqrt{2}\,\Phi^{-1}(A)Y + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1),$$

with independent Y and ϵ and fixed score S. Consequently each benchmark and fold has population AUROC exactly A. We cross $A \in \{.55, .60, .65, .80\}$ with budgets $n \in \{25, 50, 100, 200\}$, using

200 independent outer samples per cell. Groups are allocated as evenly as possible across the two benchmarks and five folds using the original allocation rule. The base seed is 20260917; each cell, repetition, and sampling phase has its own recorded substream. The .80 cell is a high-discrimination synthetic control, not an additional empirically fitted language-model monitor.

For each outer sample, the original audit draws 1,000 unstratified group bootstrap samples. A benchmark requires at least two test folds containing both classes; eligible folds are weighted by sampled group multiplicity. The macro average equally weights the benchmarks. A lower bound is returned only with at least 900 valid inner values, as their tenth percentile. A finite bound at least .60 produces trust; a lower finite bound produces distrust; an unavailable bound invokes the distrust fallback. We report all three outcomes with denominator 200, and Wilson 95% intervals in the supplied results. Coverage among finite bounds uses its own, explicitly conditional denominator. At $A = .60$, failure to approve a boundary value is not described as a calibration failure. These population-reference experiments are separate from the original empirical audits against finite-set reference values.

Table 29: Known-quality audit outcomes, 200 independent samples per cell. Trust and distrust columns require a finite lower bound; fallback means none was available. Coverage counts only finite bounds at or below the analytic AUROC, with its own denominator. A dash means no finite bound.

True AUROC	Budget	Trust	Distrust	Fallback	Covered / finite
.55	25	0	0	200	—
.55	50	3	68	129	63/71
.55	100	4	195	1	180/199
.55	200	1	199	0	183/200
.60	25	0	0	200	—
.60	50	7	50	143	50/57
.60	100	16	182	2	182/198
.60	200	23	177	0	177/200
.65	25	0	0	200	—
.65	50	16	42	142	51/58
.65	100	63	135	2	173/198
.65	200	64	136	0	184/200
.80	25	0	0	200	—
.80	50	46	6	148	47/52
.80	100	193	6	1	180/199
.80	200	200	0	0	185/200

All 3,200 planned audits complete. At budget 25, all 800 bounds are unavailable across the four quality levels. At budget 200, all bounds are finite and the rule approves 200/200 high-quality ($A = .80$) scores, compared with 64/200 at $A = .65$ and 1/200 false approvals at $A = .55$. The corresponding trust-rate Wilson intervals are $[.9812, 1]$, $[.2593, .3875]$, and $[.0009, .0278]$. The experiment separates low-budget availability from the ability to approve a clearly strong score, while showing limited approval near the requirement. At the exact .60 boundary, 23/200 approvals describe the inclusive decision rule, not a requirement that a 90% lower bound approve most boundary cases.

Saved diagnostics locate the availability loss. Pooling the four quality levels gives 800 outer samples per budget. For budgets 25, 50, 100, and 200, the unresampled macro estimate is undefined in 315, 24, 0, and 0 samples; the numbers of finite inner values are 78,315, 603,829, 794,410, and 799,997 out of 800,000. The corresponding unavailable-bound counts are 800, 562, 6, and 0. These are nested stages: both classes within a fold, at least two estimable folds per benchmark, then at least 900 finite inner values per audit. For fixed sampled labels, score separation cannot repair missing class support. These diagnostics concern the known-quality experiment; the original empirical logs do not identify the same cause counts.

G.4 A paired comparison of audit organization

To distinguish score quality from the cost of splitting a small sample, we revisit the same 3,200 saved Gaussian samples. The comparison was specified before its outputs. One arm retains the original

five-fold estimator; another computes AUROC across all observations within each benchmark, then takes the equal-benchmark mean. Both use the same 1,000 inner group multiplicity vectors for each sample, the 900-finite-value requirement, tenth percentile, and .60 trust threshold. No new outer samples or independent resampling sequences are generated. A third, decision-only baseline always distrusts and produces no quality bound. The supplied paired records retain every sample in each 200-sample cell.

Removing folds admits additional within-benchmark positive–negative pairs; it never compares observations from different benchmarks. The common Gaussian score scale makes these pairs comparable, with the same population target A. This experiment does not license pooling scores from separately fitted empirical monitors across their test folds.

Table 30: Paired audit comparison on saved known-quality samples. Each cell has 200 samples. Finite and trust are counts; distrust with a finite bound equals finite minus trust, and fallback equals 200 minus finite. The last column is conditional coverage for the no-fold estimator. Always-distrust has zero approvals throughout and no coverage value.

		Five-fold		No-fold macro		
A	Budget	Finite	Trust	Finite	Trust	No-fold Covered / finite
.55	25	0	0	80	2	74/80
.55	50	71	3	191	8	170/191
.55	100	199	4	200	3	174/200
.55	200	200	1	200	1	183/200
.60	25	0	0	75	6	69/75
.60	50	57	7	182	24	158/182
.60	100	198	16	200	13	187/200
.60	200	200	23	200	27	173/200
.65	25	0	0	83	20	70/83
.65	50	58	16	185	43	159/185
.65	100	198	63	200	72	176/200
.65	200	200	64	200	74	182/200
.80	25	0	0	83	58	74/83
.80	50	52	46	187	159	161/187
.80	100	199	193	200	197	180/200
.80	200	200	200	200	200	184/200

The no-fold estimator returns bounds for 321/800 budget-25 samples, versus none under the original rule. At $A = .80$ and budget 50, finite bounds increase from 52 to 187 and approvals from 46 to 159. At $A = .55$ and the same budget, false approvals increase from three to eight. Thus reducing fragmentation improves availability in these samples, but does not give a uniformly better decision rule.

Coverage and availability answer different questions. At $A = .80$, budget 50, no-fold conditional coverage is $161/187$, and the probability of returning a finite covered bound is $161/200$. If an unavailable bound is explicitly replaced by the trivial AUROC lower bound zero, coverage of that extended rule is $174/200$. The corresponding five-fold counts are $47/52$, $47/200$, and $195/200$, the last including 148 trivial bounds. These quantities are reported separately with Wilson 95% intervals; none demonstrates general 90% coverage. At $A = .60$, approval remains a boundary action frequency.

G.5 Saved feature-study summaries

The supplement's `results/cpu/` directory provides scenario summaries, the terminal status of all 480 fixed IDs, per-ID permutation counters, and the twelve unavailable-output diagnostics. Together with `config/CPU_PROTOCOL.json`, these support checks of the reported 468 complete and 12 unavailable dispositions and summary counts. Per-run outputs and the original generator, fitting, repair, and resampling code are not included, so these materials do not independently reproduce that study.

H Tülu feature and training-population sensitivity

These post-confirmatory analyses address two different concerns: whether the ranking comparison survives a more discriminative real-data monitor, and whether source TRAIN eligibility should depend on later checkpoints. Candidate features, selection rules, and failure handling were fixed before the new fits. The analyses use saved features and outcomes, with no new generation or change to the original diagnostic decisions. The paired intervals below reuse saved group multiplicities. All point estimates use the original five group-preserving folds and the same 1,196 common TEST pairs in 1,051 groups. No cross-fold score comparisons are formed.

H.1 A source-selected hidden-state probe

For each question pair, let u_c and u_v denote its canonical and variant hidden vectors, each divided by $\max(\|h\|_2, 10^{-8})$. We concatenate $(u_c + u_v)/2$ and $|u_c - u_v|$, giving 8,192 coordinates, with the original four diagnostics and text/task controls. The two candidate layers are the saved block-21 state and final normalized state after block 31. For each layer, L2-logistic regression uses $C \in \{.001, .01, .1\}$, `lbfgs`, tolerance 10^{-4}, at most 2,000 iterations, and seed 20260916. Standardization is estimated separately inside every training split.

For each outer fold and transition, three-fold `GroupKFold` inside the source TRAIN set selects among the six candidates. The selection score equally weights benchmarks and group-weights validation folds within each benchmark. An undefined required cell disqualifies that candidate; ties favor block 21, then smaller C. The selected layer and regularization are used for a source fit and a separate target TRAIN fit. Neither target validation nor TEST labels select the configuration. Both fits retain common-four TRAIN eligibility, isolating the expanded features and source-side selection from the eligibility change below. All 200 fits complete; no candidate or failed fit is replaced.

Major-transition folds 0–4 select block21/.01, block21/.001, block21/.001, final/.001, and final/.001. Revision folds select block21/.001, block21/.01, block21/.001, block21/.001, and block21/.001. The supplied results include all 60 candidate scores and their inner-cell components, not only the winners.

Table 31: Expanded-feature Tülu probe. All values are descriptive OOF point estimates on the original common TEST population. Configuration selection uses source TRAIN only; paired intervals are reported below.

Measurement	Pooled	Benchmark macro
DPO source at source	.7689810	.6583066
DPO source at step 1920	.7760430	.6459061
Refit at step 1920	.8091025	.7049212
Step 1840 at source	.8043207	.7045843
Step 1840 at step 1880	.7910339	.7003537
Refit at step 1880	.8190229	.7356700
Major refitting advantage	.0330595	.0590150
Revision refitting advantage	.0279890	.0353163
Major minus revision	.0050705	.0236987

Source discrimination improves in each benchmark: DPO AUROC is .6644519 on GSM-Plus and .6521613 on InvariantBench, versus .5275829 and .4886307 for the original monitor. The target refit reaches .7031315 and .7067109. The contrast's aggregation difference is +.0186282, comprising +.0232067 within-benchmark reweighting and −.0045785 cross-benchmark removal. Thus ranking-target dependence is present beyond the original near-chance monitor, without an observed reversal for this probe. These conditional point estimates do not establish a population-sign result or an operating-point reuse guarantee.

H.2 Source-only TRAIN eligibility

We next keep the original four-diagnostic and controls-only monitor definitions fixed, but train DPO and step-1840 source monitors on each role's own eligible TRAIN rows. No later-checkpoint eligibility flag selects those rows. The original target-refitted scores, common TEST rows, outcome labels,

and aggregation are unchanged. This requires 20 new source fits, all completed. The common-four TRAIN counts by fold are 708/736/667/753/690; DPO-own counts are 784/827/737/839/768, and step-1840-own counts are 809/860/765/873/795. Group identities remain disjoint between TRAIN and TEST.

Table 32: Changing source TRAIN eligibility, with target comparators and common TEST fixed. Each entry is pooled / macro AUROC or contrast. The expanded-feature probe above is a separate analysis; the subsequent combined check is reported below.

Measurement	Common-four TRAIN	Source-own TRAIN
DPO full source AUROC	.6774294 / .5081068	.6849955 / .4798497
DPO controls source AUROC	.6775315 / .5020628	.6591132 / .4582827
DPO full at target	.6752057 / .4544243	.6984784 / .4721632
Full major minus revision	.0440059 / .0793151	.0390585 / .0722186
Controls major minus revision	.0266397 / .0658699	.0351459 / .0833498

Both contrasts remain positive under source-own training. Source discrimination does not improve in macro terms for the original monitor. The check removes future observability from source training selection only: evaluation still conditions on joint observability, and the target comparator retains its original training population.

H.3 Combining source-only training with the hidden-state probe

The combined check repeats the same six-candidate source-side selection on each source checkpoint's own eligible TRAIN rows. Its target refits still use common-four TRAIN rows, but inherit the newly selected source configuration; their scores therefore need not equal those of the common-four probe. The original common TEST population, labels, and fold assignments remain fixed. All 180 inner and 20 outer fits complete without replacement. The selected major-transition configurations are block21/.01, block21/.001, block21/.001, block21/.1, and block21/.001. Revision folds select block21/.001, block21/.001, block21/.1, block21/.1, and block21/.001. All candidates are retained in the supplied records.

For both probes, we apply the original 1,000 benchmark-by-test-fold group multiplicity vectors, pairing both configurations, transitions, and original low-dimensional scores. Pooled and macro aggregation retain their original fold weights; no cross-fold score pairs are introduced. The specification lists 291 reported statistics, including absolute performance, transport contrasts, decomposition terms, and paired differences. All have 1,000 finite values. Pointwise 95% intervals use the linear-interpolated 2.5th and 97.5th percentiles, with no multiplicity adjustment. They condition on the fitted scores and selected configurations, omitting refitting and selection variation.

Table 33: Hidden-state probes on the same common TEST rows. Entries are points and paired fixed-score 95% intervals. Training columns describe source eligibility; target TRAIN remains common-four. The gain reference is the original common-four low-dimensional monitor, so the source-own gain combines feature and training-selection changes.

Measurement	Common-four source TRAIN	Source-own TRAIN
Source macro AUROC	.6583 [.6005, .7172]	.7009 [.6382, .7432]
Reused-target macro	.6459 [.5918, .7150]	.6886 [.6252, .7352]
Target-refitted macro	.7049 [.6414, .7557]	.7227 [.6545, .7654]
Source macro gain	.1502 [.0748, .2229]	.1928 [.1001, .2614]
Source pooled minus macro	.1107 [.0423, .1694]	.0735 [.0296, .1323]
C_{pool}	.0051 [−.0337, .0301]	.0042 [−.0300, .0348]
C_{macro}	.0237 [−.0536, .0773]	.0124 [−.0413, .0624]
D_{agg}	.0186 [−.0419, .0595]	.0081 [−.0272, .0478]

The source-own probe has pooled source AUROC .7744714 and macro .7009408. Its paired macro difference from the common-four probe is +.0426341 with interval [−.0076416, .0730095]. Thus the point improvement does not establish an advantage of changing training eligibility. Relative to the original low-dimensional monitor, both probes improve source macro discrimination under the

fixed-score comparison. Their source pooled–macro gaps remain positive throughout the reported intervals, while both major-minus-revision contrasts and their aggregation differences include zero.

I Ranking controls and paired real-monitor audits

These post-confirmatory analyses use saved predictions, labels, features and source fits. The baseline, weight grid, representative selection, sampling rules and stopping conditions were specified before their new results. They do not change the original OLMo criteria or the 480-run controlled study.

I.1 Benchmark-only prediction and task weights

For each of four Tülu checkpoints, five model folds and two benchmarks, the benchmark-only predictor estimates the unweighted binary outcome mean on that checkpoint's common-four eligible TRAIN rows. This gives 40 rates, with no smoothing or use of ADAPT or TEST outcomes. Source-at-source and reused target predictions use the source TRAIN rate; the target-refitted baseline uses the target TRAIN rate. Every comparison retains the original 1,196 TEST rows in 1,051 groups, original labels and original model-fold boundaries. Thus same-benchmark predictions tie within a fold and have AUROC .5.

Table 34: TRAIN-only benchmark identity baseline. Macro AUROC is exactly .5 in every cell. Pooled values use original row-weighted model folds and no cross-fold pairs. Target refitting changes rates but not their ranking here, so both refitting advantages and their contrast are zero.

Transition	Source at source	Source at target	Target refit
DPO → step 1920	.6844521	.7038833	.7038833
Step 1840 → step 1880	.7057883	.7006347	.7006347

For the separate task-weight analysis, each original model fold has fixed weight $1/5$. Within benchmark b, let $C_b = T_{\text{major},b} - T_{\text{revision},b}$ use these fold weights. For GSM-Plus weight w, the contrast is

$$C(w) = wC_{\text{GSM}} + (1-w)C_{\text{IB}}, \qquad 0 \le w \le 1.$$

We report the full .01 grid, endpoints and analytic zero $w_0 = -C_{\text{IB}}/(C_{\text{GSM}} - C_{\text{IB}})$, without choosing weights from the observed sign. The value at $w = .5$ is equal-fold macro, not the original group-weighted macro. Pooled AUROC has additional cross-benchmark terms and is not a member of this curve. This deterministic sensitivity does not add a confidence interval or establish a population sign.

Table 35: Fixed-prediction task-weight sensitivity with equal model-fold weights. Endpoints are pure InvariantBench ($w = 0$) and pure GSM-Plus ($w = 1$). The original low-dimensional Tülu curve has no zero in the allowed range.

Monitor	$C(0)$	$C(.5)$	$C(1)$	w_0 in $[0, 1]$
OLMo original	.1004315	.0178932	−.0646450	.6083936
Tülu original	.0844801	.0768902	.0693003	none
Tülu hidden-state, common-four	.0626829	.0267575	−.0091679	.8724043
Tülu hidden-state, source-own	.0318864	.0140411	−.0038043	.8934089

I.2 Finite-budget audits of individual saved source monitors

We restore the five source-own DPO hidden-state fits and the five source-own low-dimensional fits. Within each model fold, their source TRAIN row IDs are identical. Layer and regularization were selected using source TRAIN only in the earlier probe study. No new fit or target-based model selection occurs. Each scorer is evaluated only on its original TEST groups with observable target outcomes and finite features. We exclude TRAIN and ADAPT, and verify that no TEST group participated in source selection. Reconstructed scores on the common TEST subset match the saved values to 2.78×10^{-16}.

Each group contributes the target-eligible pair with smallest SHA-256 of its row ID, using row ID to break a hash tie. This selection uses neither scores nor label values. Within each benchmark, groups are sorted by a fixed hash of seed, model fold, benchmark and group ID, then assigned cyclically to five audit subdivisions. These are evaluation partitions under one fixed scoring function, not the five models' training folds. A repeated bootstrap group retains its subdivision and increases only its multiplicity.

Model folds 0–4 supply GSM-Plus/InvariantBench group counts 193/27, 211/32, 202/28, 203/31 and 210/29. Budgets 25 and 50 are feasible; 100 and 200 would require at least 50 and 100 InvariantBench groups per model, so both are recorded as infeasible with zero draws. We do not enlarge TEST with adaptation data. For each feasible model-fold/budget cell, 200 outer samples use $\lfloor n/2 \rfloor$ InvariantBench groups and the remainder GSM-Plus. A largest-remainder allocation distributes each benchmark's budget across audit-subdivision capacities; sampling is without replacement within each cell. All monitors and rules share the selected group IDs.

Every outer sample has 1,000 unstratified group-bootstrap draws. Seed 20260917 and separate deterministic outer/inner seed tuples are retained with the IDs and complete multiplicity arrays. The subdivided estimator computes each benchmark's group-multiplicity-weighted mean over finite subdivision AUROCs, requiring at least two finite subdivisions. The other estimator computes one AUROC per benchmark without subdivision, then averages benchmarks equally. Both award half credit to ties, return the linearly interpolated tenth percentile only with at least 900 finite inner values, and approve only at a bound of at least .60. Removing subdivisions introduces additional within-benchmark pairs; it is not a pure computational acceleration. Always-distrust has no bound and never approves.

The comparison reference is each fixed scorer's no-subdivision macro AUROC on its complete finite representative population. The full subdivided reference is reported too, because it uses a different pair set. A finite bound below .60 does not certify quality; it is not proof that quality is below .60. Errors here are disagreements with the finite reference, not measured population error or coverage. The source fits' training sets overlap, so no inference pools them as independent trained models.

Table 36: Complete finite-population references for the paired real-monitor audits. H denotes hidden-state, L low-dimensional; subdivisions are audit partitions under the same scorer. The no-subdivision reference defines the common decision comparison in Table 6.

	No subdivision		Five subdivisions	
Model fold	H	L	H	L
0	.5664277	.4497710	.5785578	.4507618
1	.6625253	.5721356	.7143941	.5244985
2	.6649469	.5222720	.6864832	.4796979
3	.6470238	.5277211	.5865207	.5225147
4	.6137840	.4958843	.6034699	.4930844

Table 37: All feasible real-audit cells, 200 paired samples per row. Arrows compare five subdivisions with no subdivision. Both monitors have identical availability because they share labels and samples. Always-distrust approves zero throughout. Infeasible budgets 100 and 200 have no results.

Model fold	Budget	Finite bounds	H approvals	L approvals
0	25	0 → 73	0 → 5	0 → 0
1	25	0 → 63	0 → 17	0 → 1
2	25	0 → 57	0 → 13	0 → 2
3	25	0 → 47	0 → 5	0 → 0
4	25	0 → 55	0 → 3	0 → 1
0	50	25 → 165	0 → 2	0 → 0
1	50	34 → 170	8 → 47	0 → 3
2	50	21 → 133	6 → 33	0 → 0
3	50	0 → 167	0 → 19	0 → 4
4	50	23 → 134	1 → 21	0 → 2

All 2,000 distinct outer samples complete, with 2,000,000 shared inner draws, 8,000,000 model/organization statistic evaluations and 12,000 decisions across two monitors and three rules. Saved per-cell Wilson 95% intervals describe repeated-audit event frequencies conditional on that model and finite TEST population. Availability improves under the no-subdivision rule, but the below-threshold hidden-state fit and some low-dimensional fits also show false approvals. The result is a measurable tradeoff, not validation of a uniformly better audit rule. The supplied recomputation entry point uses saved scores, IDs and multiplicities without new sampling or fitting.

I.3 Paired explanation of availability and approval changes

We pair the two audit organizations on each saved outer sample, retaining the same scorer, labels, group multiplicities and no-subdivision finite reference. For each fit and budget, Table 38 separates samples with neither bound, only the no-subdivision bound, and both bounds. It then separates new approvals, revoked approvals and approvals retained by both rules. No sample loses bound availability. The 100- and 200-group budgets remain infeasible and contribute no samples or estimates.

For samples with both bounds, let A_5, A_1 be the outer-sample macro AUROC under five subdivisions and no subdivision, and let L_5, L_1 be their saved lower endpoints. Unit multiplicities on the saved outer groups recover the point estimates. The signed identity

$$L_1 - L_5 = (A_1 - A_5) + \left[(A_5 - L_5) - (A_1 - L_1)\right] \tag{2}$$

separates the point-estimate change from the change in the point-to-endpoint offset. An offset $A - L$ can be negative and is not an interval width. Table 39 reports conditional means with the number of qualifying samples out of 200. Missing endpoints are never replaced by zero. At budget 25, and for fit 3 at budget 50, there is no paired finite subset on which to evaluate this identity.

All additional false approvals occur among samples whose subdivided bound was unavailable; new approvals and revoked approvals can also occur when both bounds are finite. For hidden-state fit 1 at budget 50, the mean endpoint on that finite subset decreases even as total approvals increase. Thus aggregate approval counts alone do not describe how the two procedures differ. The decomposition is a numerical explanation of these saved audits, not a causal attribution or a coverage guarantee. It uses no new fitting or resampling. Fits remain separate because their training sets overlap.

Table 38: Paired audit states and approval changes. H and L denote hidden-state and low-dimensional monitors; the digit identifies the fixed source fit. Each row has 200 outer samples. Neither, One and Both count unavailable in both, available only without subdivisions, and available under both rules. New-One and New-Both split newly approved samples by these availability states; Lost counts revoked approvals, and Kept counts approvals under both rules. No sample loses availability. H0 and all L fits have a common finite reference below .60; H1–H4 have references above it. Thus new H0/L approvals are false relative to those finite references.

Fit	n	Neither	One	Both	New-One	New-Both	Lost	Kept
H0	25	127	73	0	5	0	0	0
H0	50	35	140	25	2	0	0	0
L0	25	127	73	0	0	0	0	0
L0	50	35	140	25	0	0	0	0
H1	25	137	63	0	17	0	0	0
H1	50	30	136	34	39	3	3	5
L1	25	137	63	0	1	0	0	0
L1	50	30	136	34	3	0	0	0
H2	25	143	57	0	13	0	0	0
H2	50	67	112	21	29	0	2	4
L2	25	143	57	0	2	0	0	0
L2	50	67	112	21	0	0	0	0
H3	25	153	47	0	5	0	0	0
H3	50	33	167	0	19	0	0	0
L3	25	153	47	0	0	0	0	0
L3	50	33	167	0	4	0	0	0
H4	25	145	55	0	3	0	0	0

Continued on next page

Table 38 (continued)

Fit	n	Neither	One	Both	New-One	New-Both	Lost	Kept
H4	50	66	111	23	20	1	1	0
L4	25	145	55	0	1	0	0	0
L4	50	66	111	23	2	0	0	0

Table 39: Mean signed endpoint decomposition at budget 50, conditional on both endpoints being available. The displayed denominator is the subset size out of 200 outer samples, not an independent-model count. All budget-25 subsets and fit-3 subsets are empty. Offsets are $A - L$, not confidence-interval widths. Means, minima, maxima and individual values are supplied at full precision.

Fit	$k/200$	$\overline{L_1 - L_5}$	$\overline{A_1 - A_5}$	$\overline{(A_5 - L_5) - (A_1 - L_1)}$
H0	25/200	.0443070	.0018983	.0424087
L0	25/200	.0219853	−.0103216	.0323070
H1	34/200	−.0063703	−.0480366	.0416663
L1	34/200	.0594808	.0417310	.0177498
H2	21/200	−.0035120	−.0324042	.0288922
L2	21/200	.0426004	.0194249	.0231754
H3	0/200	NA	NA	NA
L3	0/200	NA	NA	NA
H4	23/200	.0131296	−.0275019	.0406315
L4	23/200	.0661325	.0305825	.0355500