# SPECTRAL MONITORS DETECT CHANGE, NOT DAMAGE, IN LLM POST-TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can hidden-state spectra distinguish harmful post-training updates from benign adaptation? A fixed-readout construction shows that identical spectra can coexist with arbitrarily different predictive losses—an identifiability limit, not an impossibility theorem for empirical alarms. We test the empirical question with pre-registered shared-prefix stress tests on three open models, run on a corrected training/probe split with fresh seeds. Extreme duplication damages held-out NLL by $+0.3$ to $+1.8$ nats while raw RankMe *rises* in 9/9 model–seeds, and the same high-LR damage moves RankMe *down* on OLMo-2-1B but *up* on Qwen3-1.7B: the sign of a spectral response tracks the (regime, architecture, layer) tuple, not damage. A trajectory-level split-conformal protocol then compares spectral and loss monitors at an equal false-alarm budget on the development model (Qwen3-0.6B): on abrupt damage every spectral family alarms at the same measurement as a plain loss detector; on a gradual learning-rate ramp spectra alarm 80–110 steps earlier, but CUSUM and Page–Hinkley detectors on the same loss stream recover about half (CUSUM) and a quarter (Page–Hinkley) of that lead, and all of these detectors also fire on a non-damaging duplication ramp and on a benign control that improves its target ($-0.38$ nats) with no significant task-accuracy change. Only a loss monitor with a non-degenerate threshold separates damage from benign change; a pre-registered matched-threshold control shows the sidedness of the detector is immaterial. An earlier campaign affected by a training/probe exposure defect is disclosed, repaired, and reported side by side; its structure replicates cleanly. In these settings, the tested spectral diagnostics measure change, not damage.

## 1 INTRODUCTION

Continual post-training changes both the behavior of large language models (LLMs) and their internal representations. Can a label-free representation diagnostic, computed from the eigenvalue spectrum of the hidden-state covariance, identify whether an update is harmful? Effective rank and RankMe are natural candidates because they summarize feature diversity and have been related to downstream quality in self-supervised learning (Garrido et al., 2023). Recent work studies Gram-spectral tracks in recursive training (Gu et al., 2026) and layer-wise matrix entropies in fine-tuning (Zhao et al., 2026); neither line itself establishes a damage alarm.

We separate three questions. First, can spectra identify predictive performance in general? A fixed-readout construction gives identical raw and centred spectra with arbitrarily different losses—non-identifiability in the classical sense of indistinguishable observables with different estimands (Lehmann & Casella, 1998) (Section 3). This mathematical ambiguity does not preclude useful distribution-specific alarms. Second, do diagnostic signs remain consistent across updates? In shared-prefix stress tests on three models, extreme duplication increases both probe NLL (negative log-likelihood, the held-out loss) and rank in all nine model–seed runs; high-LR responses depend on model and layer (Section 5). Third, do calibrated spectral alarms improve on loss monitoring? A trajectory-level conformal protocol compares level, CUSUM, and Page–Hinkley loss baselines at the same nominal false-alarm budget. Spectral timing advantages on an LR ramp depend on the chosen loss baseline, and early alarms also occur under benign adaptation (Section 6).



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.

A post-hoc data audit found that the original training loader did not consistently exclude the fixed probe for nonzero seeds. We disclose the defect in Section 2, repaired the split, and re-executed every design except the diversity ladder on fresh seeds with thresholds frozen before scoring; main-text results are from these clean campaigns unless explicitly marked historical (Appendix C).

Our contributions are the identifiability construction, a replicated three-model sign-structure study, an equal-budget evaluation framework with loss-stream sequential baselines, and layer/channel/time-resolution and capability analyses with explicit evidence boundaries. In these settings, the tested spectral diagnostics measure change, not damage. We establish neither universal failure of spectral alarms nor superiority over methods developed for different target events (scope: Section 9).

## 2 SETUP: PROTOCOL AND STATISTICS

**Data audit and evidence policy.** The project ran two generations of campaigns. The original (*historical*) pipeline had a training/probe defect found in a post-hoc data audit: it excluded the last 256 rows of a *training-seed* shuffle while the probe always used seed 0, so for nonzero seeds part of the probe text was trained on (55 and 46 of 256 rows for seeds 1 and 2). Every design in this paper was therefore re-executed on a corrected split (*fixed-probe-v2*: exact formatted probe-text hashes excluded for every seed; per-run overlap manifests), with fresh seeds and thresholds frozen before scoring. **Unless explicitly marked historical, every number in the main text is from the clean campaigns.** Historical results—whose qualitative structure replicates cleanly throughout—are retained as labeled corroboration; Appendix C documents the audit, the repair, and the side-by-side comparison (Tables 21–22).

**Shared-prefix stress tests.** For each model $\mathcal{M}$ and seed $s$, we fine-tune on the Dolly instruction corpus (Databricks, 2023) for $\tau = 300$ steps (the *prefix*), snapshot the weights (the *fork*), and then branch into matched *update regimes* for 600 further steps, measuring every 10 steps (61 measurements per trajectory). All branches of a seed share the fork checkpoint bit-for-bit (verified by hash), the optimizer recipe (AdamW, bf16 precision, fixed batch shape), and the data-batch schedule; regimes differ only in the intervention. Regimes: *healthy* (unchanged recipe); a *duplication ladder* (OLMo-2-1B and Qwen3-1.7B) that keeps the update count and maximum batch shape fixed while restricting the pool to $100/50/25/12.5\%$ unique sequences; *extreme duplication* ($n_u$=8 unique sequences, several-hundred-fold repetition; all three models); *high-LR* ($15\times$ base learning rate, LR = learning rate; base LRs fixed by a frozen ladder, Appendix B); and, on Qwen3-0.6B, a *domain shift* to GSM8K at $8\times$ LR (its frozen ladder's damaging rung). Models: Qwen3-0.6B, Qwen3-1.7B, OLMo-2-1B (Yang et al., 2025; Team OLMo et al., 2025); three seeds per cell.

**Damage gate.** A branch is *damaged* when its fixed probe NLL (a fixed, deterministic 256-sequence instruction-tail probe, identical across all runs) exceeds its fork value by $+0.30$ nats (natural-log loss units). This is a pre-registered severity definition used to place checkpoints and define events—it is *not* a detector and is never compared against as one (Section 4).

**Definition 1** (Spectral statistics)**.** *For token-level hidden states $Z \in \mathbb{R}^{N\times d}$ at a layer (we monitor the last (layer $-1$) and middle layers; $N \approx 3\times 10^4$ probe tokens), the* raw *second moment is $M = \frac{1}{N} Z^\top Z$ on the unnormalized hidden states; the* centred *covariance $\Sigma = \frac{N}{N-1}(M - \bar{z}\bar{z}^\top)$ (the sample covariance used by the implementation) is computed on per-token RMS-normalized states (Appendix B), so the two streams differ in normalization as well as centering. Raw logs store $Z^\top Z = NM$; normalized rank/entropy measures are scale invariant. With eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ of either matrix, we track three counts of the directions the representation effectively uses, and one entropy gap: RankMe $= \exp\big(-\sum_i p_i \log p_i\big)$ with $p_i = \sigma_i / \sum_j \sigma_j$ over the singular values $\sigma_i = \sqrt{\lambda_i}$ (matching its original definition); $k_{95} = \min\{k : \sum_{i\leq k} \lambda_i \geq 0.95 \sum_i \lambda_i\}$; effective rank $= \exp(-\sum_i q_i \log q_i)$ with $q_i = \lambda_i / \sum_j \lambda_j$ (eigenvalue-entropy rank, as logged by the implementation; distinct from the participation ratio); and the anisotropy index $\Delta H = \frac{d}{2}\log(\operatorname{tr}\Sigma/d) - \frac{1}{2}\sum_i \log\lambda_i$ (the log-determinant entropy gap to the isotropic Gaussian of equal total variance; $\Delta H$ enters the entropy monitor family of Section 6, with per-branch values in the released logs; log eigenvalues are floored at $10^{-10}\lambda_{\max}$). We also track cross-batch similarity (the only non-spectral channel): the mean off-diagonal pairwise cosine similarity of the first at most 2048 retained probe representations, which rises when representations converge onto a common direction.*

**Replayed baseline monitors.** From the stored spectra we replay, under the identical protocol: the Gram-spectral monitor SIGMA's deterministic track $G_{\mathrm{KF}}(\delta) = \sum_i \log(1 + \lambda_i/(\hat{\beta} + \delta))$ and its size-corrected law-of-large-numbers (LLN) track (Gu et al., 2026), with the adaptation $\hat{\beta} = 0$ ($\hat{\beta}$ is SIGMA's radius for unobserved-embedding mass; our probe observes every embedding) and ridge $\delta$ fixed at the first measurement—under which the two tracks' drifts coincide identically; and matrix-based $\alpha$-entropies $S_\alpha = \frac{1}{1-\alpha} \log_2 \sum_i \tilde{\lambda}_i^\alpha$ with $\tilde{\lambda}_i = \lambda_i / \sum_j \lambda_j$, for entropy order $\alpha \in \{1, 2\}$ ($\alpha \to 1$ taken as the Shannon limit; unrelated to the conformal level $\alpha$ of Section 4) (Zhao et al., 2026). The Collapse Index (Kalinowski, 2026) is topological (persistent homology over activation complexes) and is *not* determined by stored spectra; we exclude it rather than substitute a lookalike, and record this as a coverage limitation.

## 3 Spectra do not identify predictive performance

**Proposition 1** (Fixed-readout non-identifiability)**.** *Let $y_i \in \{\pm 1\}$ be balanced labels, let $a > 0$, and let $h_i = y_i\, a\, e_1 \in \mathbb{R}^d$ with fixed linear readout $w = e_1$ under logistic loss. Let $Q$ be the orthogonal map sending $e_1 \mapsto -e_1$ and fixing $e_1^\perp$. Then $\{h_i\}$ and $\{Qh_i\}$ have identical raw second-moment and centred covariance spectra (indeed identical singular values of the data matrix), while the mean losses are $\log(1 + e^{-a})$ and $\log(1 + e^{a})$, differing by exactly $a$, and the 0–1 accuracy flips from 1 to 0.*

*Proof.* Orthogonal right-multiplication preserves singular values of $Z$ and the spectra of $Z^\top Z$. The label-balanced mean is zero, so centering leaves the scatter matrices unchanged; the sample-covariance normalization is identical in both conditions. The margin $w^\top h_i = y_i a$ becomes $-y_i a$ under $Q$; logistic loss maps these to $\log(1+e^{-a})$ and $\log(1+e^{a})$ respectively, and $\mathrm{sign}(w^\top h_i)$ flips on every point. □

**Remark 1** (Boundedness and joint rotations)**.** *For a* fixed *finite representation and readout, logits obey $|w^\top h| \le \|w\|\|h\|$, so the per-sample loss gap in Proposition 1 is bounded by $2a$: the gap is arbitrary only because the construction's scale $a$ is a free parameter, not unbounded for fixed $Z, W$. Conversely, the* joint *rotation $Z' = ZQ$, $w' = Q^\top w$ leaves every logit unchanged: coordinate change alone is never damage. Both statements, and the spectrum identities, are certified numerically in our test suite.*

With the qualifications of Remark 1, the proposition delimits what any function of the spectra of Definition 1 can infer: on representation classes containing such pairs, they cannot recover fixed-readout predictive performance. In real fine-tuning the readout also moves, so Section 7 additionally tracks alignment between representation and readout empirically—the quantity the spectra discard.

## 4 Evaluation protocol: false-alarm-controlled monitoring

**Why a protocol.** "The monitor fired before damage" is uninterpretable without (i) a false-alarm guarantee on healthy updates, (ii) an event definition independent of the monitor, and (iii) a baseline given the same budget. Prior monitoring proposals typically report detection anecdotes on damaged runs only. The NLL gate and loss baselines share a signal, so their comparison is not independent evidence of capability damage; an independent task-endpoint analysis appears in Appendix C.

**Trajectory-level split conformal.** Healthy trajectories for Qwen3-0.6B (32 fresh seeds per campaign) are split by fixed, pre-declared seed lists: *scale* seeds 32–34, *calibration* seeds 35–53 ($n_{\mathrm{cal}} = 19$), *final test* seeds 54–63 (10 trajectories, scored once by the frozen thresholds). The historical campaign used the same design at seeds 0–31; its test split was later reused for exploratory sensitivity, disclosed in Appendix C. A monitor either alarms once within the horizon or stays silent; we score false alarms on healthy and benign trajectories and alarm timing on damaged ones. For channel $c$ and branch-relative time $t \le H = 600$, let $x_{c,t}$ be the statistic's delta from its fork value. Per-channel scales $s_c$ are medians over scale trajectories of $\max_t |x_{c,t}|$, floored at $10^{-12}$. Tables report absolute steps $t + 300$. A trajectory's conformity score is $S = \max_{c,t} |x_{c,t}|/s_c$; the level-$\alpha$ threshold is the $k$-th smallest calibration score with $k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$. With $n_{\mathrm{cal}} = 19$ and $\alpha = 0.05$, $k = 19$: the maximum calibration score, the coarsest feasible finite threshold. Under

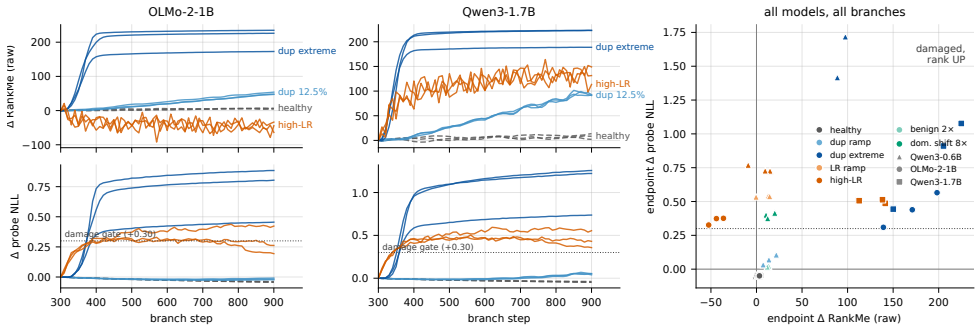



Figure 1: **Rank–NLL paired trajectories and the endpoint sign structure.** Left, centre: fork-relative raw RankMe (top) and fixed probe NLL (bottom) for OLMo-2-1B and Qwen3-1.7B, three seeds per regime (healthy dashed grey; duplication family in blues, extreme darkest; high-LR vermillion). Extreme duplication drives rank sharply *up* while NLL breaches the damage gate in both families; high-LR damage moves rank *down* in OLMo and *up* in Qwen3-1.7B. Right: endpoint $\Delta$NLL vs. $\Delta$RankMe for every branch–seed of all three models (marker = model, colour = regime): damaged branches (above the dotted gate) occupy *both* rank directions, and the no-damage duplication ramp (light blue) moves rank far from zero—the visual form of non-identifiability. Trajectories shown are from the historical campaign; the endpoint panel is computed from the repaired-split clean campaigns—the branch–seeds of Table 1.

exchangeability of healthy trajectories (they are statistically interchangeable—the standard split-conformal assumption) and a frozen score rule this controls the *marginal* probability that a fresh healthy trajectory ever alarms within $H$ at $\leq \alpha$ (Angelopoulos & Bates, 2021), absorbing channel, direction, and time multiplicity in one shot. Monitor families (rank family; entropy/Sigma; cross-batch similarity; their union; and the NLL baselines below) are calibrated *separately* at equal $\alpha$.

**Equal-budget NLL baseline.** The probe NLL delta from fork is given the identical splits, horizon, and $\alpha$. Its direction was frozen before any test trajectory existed: the primary NLL detector is *one-sided upward* (damage = NLL rise); a two-sided variant is reported secondarily. We report two leads with censoring (no alarm / no breach recorded as missing, never zero): $t_{\text{damage}} - t_{\text{alarm}}$ (lead over the $+0.30$-nat severity gate) and $t_{\text{NLL alarm}} - t_{\text{alarm}}$ (lead over the equal-budget NLL detector). The second compares detector timing; neither alone establishes early warning of an independent capability event.

**Guarantee scope.** The guarantee is per-trajectory and marginal, for this model, recipe, probe, and horizon; thresholds do not transfer across models. Ten test trajectories bound the false-alarm rate only loosely: zero alarms in ten gives a one-sided 95% upper bound of $1 - 0.05^{1/10} \approx 25.9\%$. We state observed counts and this bound, not "FA $\leq 5\%$ verified."

## 5 Cross-model sign structure

Figure 1 previews the section's three findings; Table 1 is the paper's central empirical object.

**(1) Strict inversion under duplication damage, two families.** Extreme duplication damages held-out NLL severely on the repaired split—$+0.44 \pm 0.13$ nats (OLMo), $+0.81 \pm 0.33$ (Qwen3-1.7B), $+1.30 \pm 0.49$ (Qwen3-0.6B)—and in *all nine* clean runs raw RankMe *rises* (by $+139$ to $+225$ on the two larger models), with centred RankMe, $k_{95}$, and effective rank rising in lockstep. A monitor reading "rank up = capacity healthy" therefore stays silent while held-out loss climbs by up to a nat and a half; what task accuracy does under this regime is subtler (Section 7). The extreme rung ($n_u$=8) originated as an exploratory pilot escalation; it is confirmed on nine fresh clean model–seeds, and the historical campaign shows the same 9/9 inversion at larger magnitudes.

**(2) The high-LR sign flip.** The same $15\times$-LR damage ($+0.33$ to $+0.77$ nats across models, clean campaigns) moves raw RankMe *down* in all three OLMo seeds ($-36$ to $-52$) and *up* in all three

Table 1: **Endpoint sign structure under the repaired split** (clean campaigns, fresh seeds 32–34 per cell; layer $-1$ raw RankMe; seed mean $\pm$ SD of the endpoint probe-NLL delta; sign column = per-seed signs of $\Delta$RankMe(raw)). Shaded rows breach the $+0.30$-nat damage gate at seed mean; **bold** sign strings mark damaged rows whose rank rises in every seed—the inversion. Historical endpoints, the duplication-diversity ladder, and full per-branch tables are in Appendix A; the historical-vs-clean comparison is Appendix Table 22.

| model | update regime | endpoint $\Delta$NLL | $\Delta$RankMe signs |
|---|---|---|---|
| Qwen3-0.6B | healthy | -0.05 $\pm$ 0.00 | ++− |
| Qwen3-0.6B | benign (domain-shift 2×) | +0.02 $\pm$ 0.00 | +++ |
| Qwen3-0.6B | duplication ramp | +0.07 $\pm$ 0.04 | +++ |
| Qwen3-0.6B | domain-shift 8× | +0.40 $\pm$ 0.02 | **+++** |
| Qwen3-0.6B | LR ramp 1× → 15× | +0.54 $\pm$ 0.00 | ++− |
| Qwen3-0.6B | duplication (extreme) | +1.30 $\pm$ 0.49 | **+++** |
| Qwen3-0.6B | high-LR | +0.74 $\pm$ 0.02 | ++− |
| Qwen3-1.7B | healthy | -0.05 $\pm$ 0.00 | +++ |
| Qwen3-1.7B | duplication (extreme) | +0.81 $\pm$ 0.33 | **+++** |
| Qwen3-1.7B | high-LR | +0.50 $\pm$ 0.01 | **+++** |
| OLMo-2-1B | healthy | -0.05 $\pm$ 0.00 | +++ |
| OLMo-2-1B | duplication (extreme) | +0.44 $\pm$ 0.13 | **+++** |
| OLMo-2-1B | high-LR | +0.36 $\pm$ 0.03 | −−− |

Table 2: **Adapted spectral diagnostics show the sign structure** (historical campaign replay; endpoint drift, layer $-1$ raw Gram; seed mean $\pm$ SD; sign = per-seed Sigma-drift signs). Negative $G_{\mathrm{KF}}$ drift indicates contraction in the adapted track; this is not a native SIGMA alarm-policy reproduction. Shading as in Table 1; **bold**: positive drift in every seed of a damaged row, where a contraction rule stays silent.

| model | update regime | $\Delta$NLL | Sigma $G_{\mathrm{KF}}$ drift | sign | $\Delta S_1$ |
|---|---|---|---|---|---|
| Qwen3-0.6B | domain-shift | +0.40 $\pm$ 0.01 | -231 $\pm$ 16 | −−− | +0.57 $\pm$ 0.05 |
| Qwen3-0.6B | duplication (extreme) | +1.65 $\pm$ 0.34 | +1089 $\pm$ 42 | **+++** | +2.06 $\pm$ 0.16 |
| Qwen3-0.6B | high-LR | +0.64 $\pm$ 0.10 | -755 $\pm$ 67 | −−− | +0.76 $\pm$ 0.18 |
| OLMo-2-1B | dup. 100% | -0.04 $\pm$ 0.00 | +44 $\pm$ 9 | +++ | +0.06 $\pm$ 0.02 |
| OLMo-2-1B | dup. 12.5% | -0.02 $\pm$ 0.01 | +325 $\pm$ 23 | +++ | +0.47 $\pm$ 0.04 |
| OLMo-2-1B | dup. extreme ($n_u$=8) | +0.71 $\pm$ 0.23 | +1310 $\pm$ 188 | **+++** | +2.08 $\pm$ 0.32 |
| OLMo-2-1B | high-LR | +0.29 $\pm$ 0.12 | -820 $\pm$ 92 | −−− | -1.12 $\pm$ 0.08 |
| Qwen3-1.7B | dup. 100% | -0.05 $\pm$ 0.01 | +82 $\pm$ 25 | +++ | +0.16 $\pm$ 0.02 |
| Qwen3-1.7B | dup. 12.5% | +0.05 $\pm$ 0.01 | +835 $\pm$ 11 | +++ | +0.99 $\pm$ 0.02 |
| Qwen3-1.7B | dup. extreme ($n_u$=8) | +1.07 $\pm$ 0.29 | +2269 $\pm$ 273 | **+++** | +3.00 $\pm$ 0.39 |
| Qwen3-1.7B | high-LR | +0.45 $\pm$ 0.10 | -1179 $\pm$ 93 | −−− | +2.90 $\pm$ 0.22 |

Qwen3-1.7B seeds ($+113$ to $+141$), with Qwen3-0.6B mixed (two seeds up)—reproducing the historical flip seed-exactly. Neither "collapse means rank falls" nor its inversion survives a change of architecture—nor of layer: at the *mid* layer OLMo's high-LR raw RankMe rises too ($+26$ to $+31$ clean, $+24$ to $+44$ historical; Appendix Table 7), while the duplication inversion holds at both layers. The sign of a spectral response under damage is a property of the (regime, statistic, architecture, layer) tuple, in these experiments, rather than of damage.

**(3) Diversity ladder: large spectral motion, no damage (historical campaign).** Across the conventional ladder no rung breaches the gate ($\Delta$NLL between $-0.05$ and $+0.06$), yet raw RankMe moves monotonically with duplication intensity in every seed (e.g. $+49.8 \pm 3.7$ at $12.5\%$ unique on OLMo, $+91.9 \pm 1.0$ on Qwen3-1.7B; Spearman $|\rho| = 1.0$ between duplication fraction and rank delta in both models). This control shows spectral sensitivity to data diversity without a corresponding NLL-gate event; it does not identify a causal mechanism. The ladder was not rerun under the repaired split; its no-damage reading is corroborated cleanly by the duplication ramp of Section 6. Note the ladder's damage response itself is a negative result: reduced diversity at matched update count and maximum batch shape mostly reduces the learning *benefit* relative to healthy, rather than degrading absolutely from the fork.

**Adapted baseline responses (historical replay)** (Table 2). Under extreme duplication SIGMA's $G_{\mathrm{KF}}$ drifts *positive* by $+1047$ to $+2447$ per run in all nine damaged runs—a contraction-oriented rule would require negative drift—and matrix entropies rise. Under high-LR, $G_{\mathrm{KF}}$ is negative in all

Table 3: **Monitor specificity at $\alpha = 0.05$, clean campaign** ($n_{\text{cal}}$=19, $k$=19): threshold, held-out healthy false alarms (10 trajectories), alarms on the pre-registered benign adaptation (3 seeds), and coverage of the nine abrupt damage branches. The last two rows are the pre-registered sequential loss-stream detectors; only the two-sided NLL variant (shaded, **bold**) separates damage from benign change (a matched-threshold control shows this is a threshold property, not sidedness).

| monitor family | thr. | healthy FA | benign alarms | damage alarmed |
|---|---|---|---|---|
| rank family | 2.72 | 0/10 | 3/3 | 9/9 |
| entropy/Sigma | 2.30 | 1/10 | 3/3 | 9/9 |
| cross-batch sim | 1.90 | 1/10 | 3/3 | 9/9 |
| all spectral | 2.72 | 0/10 | 3/3 | 9/9 |
| NLL (one-sided) | 0.00 | 0/10 | 3/3 | 9/9 |
| **NLL (two-sided)** | 1.07 | 0/10 | **0/3** | 9/9 |
| CUSUM (EMA, one-sided) | 0.00 | 0/10 | 3/3 | 9/9 |
| Page–Hinkley (one-sided) | 0.00 | 0/10 | 3/3 | 9/9 |

Table 4: **Two leads, abrupt onset, clean campaign** (family = all-spectral union; ranges over 3 seeds). Lead vs. the severity gate is large for slow damage; lead vs. the equal-budget NLL alarm is zero everywhere.

| update regime | $n$ | alarm step | lead vs. damage gate | lead vs. NLL alarm |
|---|---|---|---|---|
| domain-shift 8× | 3 | 310 | 250–280 | 0 |
| duplication (extreme) | 3 | 320 | 30 | 0 |
| high-LR | 3 | 310 | 0–10 | 0 |

models (it would alarm), but the $\alpha$-entropies *split by family* (negative in OLMo, positive in Qwen). The sign mismatch spans several tested spectral functionals. We emphasize the scope: SIGMA was proposed for recursive synthetic-data training; our replay is a transfer stress test under stated adaptations, not a refutation of its original setting. Per-branch detail is in Appendix Table 8.

## 6 Detection under trajectory-level false-alarm control

All thresholds, false-alarm counts, and alarm timings in this section are from the clean fixed-probe-v2 campaign (32 fresh trajectories, seeds 32–63; design frozen before scoring). The historical (probe-exposed) campaign ran the same frozen design first and shows the same qualitative structure; its numbers and the side-by-side comparison are in Appendix C (Table 21).

**False alarms and power** (Table 3). The spectral union, both NLL variants, and both sequential detectors, thresholded at $\alpha = 0.05$ on 19 clean calibration trajectories, produce *zero* false alarms on the 10 held-out clean trajectories (one-sided 95% upper bound $25.9\%$); two families (entropy/Sigma, cross-batch similarity) record one each. Every family alarms on all nine abrupt damage branches. Trajectory-level calibration is feasible, and nothing in it requires spectra.

**Abrupt-onset timing versus loss** (Table 4). Against the $+0.30$-nat severity gate, spectral alarms lead by 250–280 steps on slow domain-shift damage and 0–30 steps on fast damage—lead relative to the severity definition. In the equal-budget comparison, on *every* damage branch, every spectral family alarms at the *same* 10-step measurement as the one-sided NLL detector ($t_{\text{NLL alarm}} - t_{\text{alarm}} = 0$ in all $9 \times 4$ family–branch pairs). They do not improve on the primary loss detector in this abrupt-onset replay. Three protocol facts bound this comparison: the NLL detector's conformal threshold is exactly $0.00$—across $19 \times 61$ calibration measurements the healthy probe NLL never exceeded its fork value; detection granularity is bounded by the 10-step measurement interval, so leads below 10 steps are unresolvable by design; and all our damage regimes switch on abruptly at the fork, so every monitor saturates at the first or second post-fork measurement. This is absence of evidence for a spectral lead under conditions maximally favourable to detection, not evidence none could exist under gradual onset or finer grids. Against the *secondary* two-sided NLL variant, differences stay at the grid floor (at most one or two 10-step intervals; Appendix Table 9), consistent with this reading.

**Benign adaptation control (pre-registered).** The protocol so far tests false alarms only on same-recipe healthy seeds. To test whether the calibrated monitors are *damage* detectors rather than *change* detectors, we pre-registered a benign control: the domain-shift recipe at the mild rung of its frozen ladder (lr-mult 2, GSM8K corpus), three seeds from the same forks, scored once against the frozen thresholds, never entering calibration. All three clean runs are benign by the pre-registered criteria: GSM8K probe NLL improves by $-0.38$ nats while the source probe ends at $+0.02$—an order of magnitude below the damage gate. Under the frozen clean thresholds (Table 3), *every spectral family alarms on all three benign runs* (earliest by step 320–330); so do the one-sided NLL detector and both sequential detectors—their degenerate 0.00 thresholds fire on a benign drift. The *two-sided*

NLL variant—the one NLL detector whose calibrated threshold is non-degenerate—is silent on 3/3 benign runs and 0/10 healthy test trajectories while alarming on 9/9 abrupt damage branches; it alone separates damage from benign change here, though not change-blind—on the clean duplication ramp it alarms on 2/3 runs late in the horizon (Appendix Table 11). A pre-registered matched-threshold control attributes this specificity to the *threshold*, not the sidedness: the one-sided statistic at the same frozen two-sided threshold reproduces the two-sided alarms exactly (18/18 branches per campaign, 0/10 test false alarms). Calibrated at a trajectory-level false-alarm budget over same-recipe healthy runs, spectral monitors are change detectors; on specificity a non-degenerately-thresholded loss statistic dominates them here.

**Gradual-onset damage: a real spectral lead that is not damage-specific (pre-registered).** The zero-lead result above is bounded by its abrupt-onset design, so we pre-registered two gradual regimes (frozen before implementation): an *LR ramp* (effective learning-rate multiplier rising linearly $1\times \rightarrow 15\times$ across the branch) and a *duplication ramp* (each training batch drawn from the frozen $n_u{=}8$ extreme pool with probability $p(t){=}t/600$). Scored once against the frozen clean thresholds (Figure 2; Appendix Table 11), the pre-registered primary readout is positive: *spectral families alarm 80–110 steps before the one-sided NLL detector on every damaging run* (LR ramp, 3/3, damage gate crossed at steps 750–780; the full ordering is spectral alarm $\rightarrow$ NLL alarm $\rightarrow$ damage gate, with 300+ steps of dynamic range; the historical campaign gave 60–110). The qualification is equally pre-registered: the duplication ramp produces the *same* early spectral alarms (steps 420–440) while never crossing the damage gate within the horizon (endpoint $\Delta$NLL $+0.07 \pm 0.04$). Under gradual onset, spectral monitors buy genuine lead time *on change*; whether that change is incoming damage, harmless drift, or beneficial adaptation is what spectra—per Section 3—cannot say. Nor is spectral monitoring *cheaper*: the probe forward pass that yields hidden states yields the loss for free.

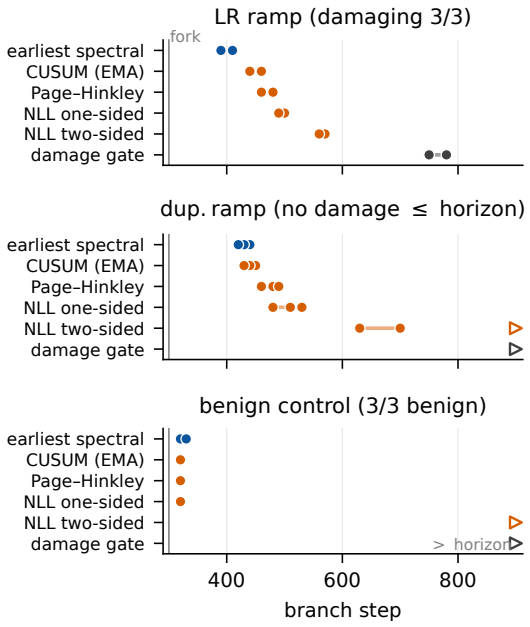



Figure 2: **Alarm ordering under gradual onset and benign change** (clean campaign, development model; dots are per-seed alarm steps under the frozen $\alpha{=}0.05$ clean thresholds, three seeds per regime; open arrows mark detectors that never alarm within the horizon). The damaging LR ramp (top) and the no-damage duplication ramp (middle) show the same spectral $\rightarrow$ sequential $\rightarrow$ NLL alarm order; on the benign control (bottom) only the two-sided NLL variant stays silent. Per-seed steps in Appendix Table 11.

**Trend-aware loss detectors recover most—not all—of the lead (pre-registered).** The gradual-onset lead invites an obvious rejoinder: the one-sided NLL detector is a *level* test with a degenerate $0.00$ threshold, so a sequential detector on the same loss stream might alarm on the upward *trend* long before the level recrosses its fork value. We pre-registered (parameters and both interpretations frozen before any statistic was computed) two standard sequential tests on exactly the stream the NLL families read: a one-sided cumulative-sum (CUSUM) test (Page, 1954) against a strictly causal exponential-moving-average (EMA) baseline (slack $1.5\hat{\sigma}$, with $\hat{\sigma}$ the healthy first-difference scale from the three scale seeds) and a one-sided Page–Hinkley test (Basseville & Nikiforov, 1993), both thresholded by the identical split-conformal recipe on the same calibration seeds ($\alpha = 0.05$; both thresholds again $0.00$; $0/10$ held-out false alarms). The result (Figure 2; Appendix Table 11): on the damaging LR ramp the CUSUM alarms 30–70 steps after the earliest spectral family (Page–Hinkley: 50–90), recovering roughly half of the 80–110-step spectral lead; the pre-registered erasure criterion (a sequential alarm within one measurement of the earliest spectral alarm) is met on 0/3 damaging runs in both campaigns. This residual lead is configuration-dependent, not universal: it holds for the two pre-registered detector configurations, and one cell of the historical 88-configuration sensitivity grid

Table 5: **Capability deltas vs. fork, clean campaign** (development model, all seven regimes; percentage points for tasks, nats for NLL; paired bootstrap over a fixed 500-item subsample per task, $B$=$10^4$; seed mean ± SD with (#seeds whose 95% CI excludes 0)/3, **bold** when 3/3. GSM8K is measured as NLL, not solve accuracy. Probe-NLL deltas are re-measured from saved checkpoints and can differ from Table 1's training-log endpoints at bf16 replay precision ($+1.34$ vs. $+1.30$ under extreme duplication). Per-seed and historical tables: Appendix A.

| regime | ckpt | ΔDolly NLL | ΔHSwag | ΔARC-C | ΔLAMBADA |
|---|---|---|---|---|---|
| healthy | endpoint | -0.04 ± 0.00 | +0.3 ± 0.6 (0/3) | -0.1 ± 0.6 (0/3) | +0.6 ± 1.4 (0/3) |
| benign (domain-shift 2×) | endpoint | +0.02 ± 0.00 | -0.4 ± 0.4 (0/3) | +1.7 ± 1.1 (0/3) | +0.6 ± 0.9 (0/3) |
| duplication ramp | endpoint | +0.07 ± 0.03 | +1.1 ± 0.9 (0/3) | +0.5 ± 0.8 (0/3) | -0.9 ± 0.1 (0/3) |
| duplication (extreme) | breach | +0.48 ± 0.14 | +2.1 ± 0.8 (1/3) | +1.7 ± 0.1 (0/3) | -3.7 ± 1.3 (2/3) |
| duplication (extreme) | endpoint | +1.34 ± 0.53 | +0.3 ± 0.4 (0/3) | +0.3 ± 1.3 (0/3) | **-5.0 ± 1.6 (3/3)** |
| domain-shift 8× | breach | +0.30 ± 0.00 | -3.0 ± 0.9 (2/3) | -1.0 ± 1.6 (0/3) | -1.9 ± 5.2 (1/3) |
| domain-shift 8× | endpoint | +0.40 ± 0.02 | **-6.6 ± 1.1 (3/3)** | -1.8 ± 0.9 (0/3) | -5.8 ± 3.7 (2/3) |
| LR ramp | breach | +0.35 ± 0.02 | **-5.1 ± 1.1 (3/3)** | -1.6 ± 1.7 (0/3) | **-8.5 ± 2.5 (3/3)** |
| LR ramp | endpoint | +0.56 ± 0.01 | **-9.1 ± 0.5 (3/3)** | -3.3 ± 1.3 (1/3) | **-15.6 ± 4.9 (3/3)** |
| high-LR | breach | +0.36 ± 0.04 | -4.0 ± 3.0 (2/3) | -1.5 ± 2.5 (1/3) | -5.9 ± 9.7 (1/3) |
| high-LR | endpoint | +0.77 ± 0.04 | **-13.4 ± 1.8 (3/3)** | **-5.6 ± 1.9 (3/3)** | **-26.9 ± 4.5 (3/3)** |

shrinks it to 10–30 steps across seeds (Appendix C). It is exactly as non-specific: both sequential detectors alarm on the no-damage duplication ramp and fire on the benign control at the same measurement as the one-sided level detector, adding two change-detector rows to the specificity matrix (Table 3).

## 7 Capability–NLL dissociation and the boundary of mechanism claims

Capability uses fixed 500-item subsamples of HellaSwag (Zellers et al., 2019), ARC-Challenge (Clark et al., 2018), and LAMBADA (Paperno et al., 2016) (identical items every checkpoint, paired deltas), alongside the fixed probe NLL (Databricks, 2023) and GSM8K NLL (Cobbe et al., 2021).

**Damage timing differs by regime** (clean campaigns, all three models). Under high-LR, capability damage is broad and *concurrent* with the NLL breach: HellaSwag is significantly down at the breach checkpoint in 3/3 OLMo seeds, and endpoints collapse (OLMo $-9.3 \pm 1.6$ pt, Qwen3-1.7B $-11.5 \pm 0.7$ pt HellaSwag, both 3/3). Extreme duplication largely reverses the ordering: at the breach checkpoint Qwen3-1.7B HellaSwag is significantly *positive* in 3/3 seeds ($+2.7 \pm 0.5$ pt) while probe NLL has crossed the damage gate, reaching $+0.81 \pm 0.33$ nats by endpoint—and only LAMBADA degrades reliably ($-5.2 \pm 0.4$ pt at the endpoint, 3/3). The pre-registered ordering "NLL breach precedes task-accuracy damage" holds for duplication and fails (in the informative direction) for high-LR: loss-based monitoring is conservative for the former and timely for the latter. LAMBADA is the most damage-sensitive task; the historical campaign shows the same pattern at larger duplication magnitudes (Appendix A).

**The development model carries capability ground truth** (Table 5). The clean battery covers every regime of the alarm-calibration model. High-LR endpoints lose $13.4 \pm 1.8$ pt HellaSwag and $26.9 \pm 4.5$ pt LAMBADA (3/3)—alarmed branches are genuinely damaged. Extreme duplication raises probe NLL by $+1.34 \pm 0.53$ nats while multiple-choice accuracy is unchanged within noise at the endpoint (HellaSwag $+0.3 \pm 0.4$ pt)—the study's starkest NLL/accuracy dissociation. The domain shift is a genuine trade (GSM8K NLL $-0.31$, HellaSwag $-6.6 \pm 1.1$ pt). On the gradual LR ramp, capability is significantly degraded at the damage-gate crossing (HellaSwag $-5.1 \pm 1.1$ pt, 3/3), so the LR-ramp spectral lead lands on trajectories with confirmed capability damage (sparse task checkpoints do not locate its onset)—while the duplication ramp's identical early alarms end with capability flat. The benign control's task metrics show no significant change in any of the nine cells (0/3 each) while its GSM8K target improves by $-0.38$ nats: the spectral, one-sided-NLL, and sequential alarms there are false alarms for task capability, not only the loss criteria. Earliness and harm remain separate dimensions.

**What the alignment data does and does not support (historical checkpoints)** (Appendix Table 15). Our pre-registered mechanism hypothesis—that rank-inversion damage adds tail variance in task-*irrelevant* directions, so head-aligned energy and gradient alignment should fall together—

is *not supported*: under extreme duplication head energy rises in OLMo ($0.124 \rightarrow 0.143$) and falls in Qwen3-1.7B ($0.141 \rightarrow 0.125$), while gradient alignment rises in both. Holding the *fork* head fixed resolves part of it: the apparent high-LR head-energy *rise* in Qwen ($0.141 \rightarrow 0.176$ against its moved head) is a *fall* against the frozen fork head ($\rightarrow$ 0.130), matching OLMo ($0.124 \rightarrow 0.095$)—for high-LR the scalar disagreement was readout drift, not geometry. (Gradient alignment: $\mathrm{tr}(\Sigma C_g)/(\mathrm{tr}\,\Sigma \cdot \mathrm{tr}\,C_g/d)$, overlap with the probe-loss gradient covariance $C_g$; 1 = isotropic; Appendix Table 15.) For duplication the split persists under the fixed head: a genuine representation-level difference between families. The one direction-stable channel is fork-relative eigenspace drift (healthy 2.6–3.4°, damaged 16–26°, both families, monotone in damage), with eigenvalue-gap diagnostics because near-degenerate spectra destabilize subspace angles (Appendix Table 16). We present drift as a candidate *correlate*, not a causal mechanism (that would require readout/representation interventions). This model-dependence of every scalar channel is the empirical face of Section 3: the information spectra discard (alignment) is exactly where the regimes differ.

## 8 Related work

**Collapse phenomena in representation learning.** The intuition that "collapse = rank falls" descends from a well-studied family: dimensional collapse in contrastive SSL (Jing et al., 2022), the doubly-exponential rank collapse of pure attention with depth (Dong et al., 2021), neural collapse of last-layer features in supervised training (Papyan et al., 2020), and the anisotropy of contextual embeddings (Ethayarajh, 2019). None claims that spectral scalars track *predictive degradation under post-training updates*—the alarm use case. Our results are consistent with all of these; what fails is the inference from motion direction to damage.

**Rank as representation quality.** RankMe (Garrido et al., 2023) connects effective rank to downstream quality of joint-embedding SSL representations; it claims no sequential safety guarantee for LLM updates, and our results do not refute it in its original setting. Tsitsulin et al. (2023) document task-dependence of unsupervised embedding-quality metrics. Our contribution is not "rank is imperfect" but the reproducible *wrong-signed* response under controlled damage, the fixed-fork design, and the equal-budget evaluation.

**Spectral monitors for degrading training.** SIGMA (Gu et al., 2026) monitors Gram-spectral envelopes in recursive training; matrix-entropy analyses (Zhao et al., 2026) track layer-wise $\alpha$-entropies; the Collapse Index (Kalinowski, 2026) builds a topological composite. We replay adapted scalars from the first two on stored spectra and observe the inversion; the third is not replayable from spectra.

**Recent post-training diagnostics.** Weight-direction monitoring (Zhong & Raghunathan, 2026), sequential plasticity analyses (Liu et al., 2026), and TRSP (Tao et al., 2026) address adjacent tasks. Appendix C distinguishes our adapted TRSP scalars from native implementations and states the missing comparisons; we claim no superiority in their original settings.

**Conformal methods and sequential change detection.** Our evaluation layer applies split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) at the trajectory level; exchangeability over same-recipe healthy trajectories gives the marginal guarantee. The sequential baselines are the classical change-point detectors—CUSUM (Page, 1954) and Page–Hinkley (Basseville & Nikiforov, 1993)—on the monitored loss stream; unlike that literature's standard average-run-length constraints, our budget controls a per-trajectory false-alarm probability, pricing all detectors on the same conformal scale.

## 9 Limitations

The historical probe exposure limits the original trajectories; the clean campaigns (Tables 21–22) replicate the sign, timing, specificity, and capability findings. Models are $\leq$1.7B; updates are full-parameter fine-tuning after one shared prefix, not multi-stage pipelines. The damaging duplication rung is extreme by construction (exploratory provenance; nine clean confirmations); intermediate rungs did not damage in our horizon, so continuous-damage evidence is the development-model LR ramp (like all timing evidence). The probe is one fixed instruction-tail distribution (GSM8K: NLL only); the benign control is one recipe, one model; loss-independent event definitions are open.

## Reproducibility statement

Every main-text table is script-generated from scored JSONs pinned by their own SHA-256 manifests together with the frozen plans, commands, and per-run split manifests—clean-campaign sources except the tables marked historical, which retain their original provenance and freeze manifest. A post-hoc data audit found the training/probe exposure that the structural artifact checker had missed—a passing structural audit is not a certificate of experimental validity—and Appendix C documents the audit, the repair, the ablation cells, and the scope of every supplementary analysis.

## AI use statement

In this work, we used generative AI tools (LLM research assistants, including ChatGPT/Codex and Claude Code) for the following tasks with required disclosure: helping develop the conceptual framework; formulating mathematical claims and assisting in the writing of their proofs (Proposition 1); proposing and refining hypotheses; designing and critiquing the experimental methodology; implementing methods; cleaning and reformatting datasets (training-pool and probe-split construction); and interpreting results. We did not use generative AI tools to generate synthetic data sets; translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: writing and editing code, creating figures, drafting and revising the manuscript, literature search and summarization, reference formatting, and assembling the supplementary artifact. We have reviewed all AI-assisted work: the human authors specified the research questions and the pre-registered protocols, gates, and frozen decisions before the corresponding data existed; every reported number is produced by the audited pipelines from raw logs (the pre-submission artifact audit passed with zero failures); and the post-hoc data audit and clean-replication campaigns that cross-checked the AI-implemented pipeline are documented, with their own manifests, in Appendix C. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## References


Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Databricks. Free Dolly: Introducing the world's first truly open instruction-tuned LLM, 2023. databricks-dolly-15k.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, 2023.

Yi Gu, Lingyou Pang, Xiangkun Ye, Tianyu Wang, Jianyu Lin, Carey E. Priebe, and Alexander Aue. SIGMA: Scalable spectral insights for LLM model collapse. *arXiv preprint arXiv:2601.03385*, 2026.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Alexander Kalinowski. Monitoring neural training with topology: A footprint-predictable collapse index. *arXiv preprint arXiv:2604.26984*, 2026.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

Yichen Liu, Mingyu Chen, Hao Wang, Xiaoran Xu, Chenxi Lin, Rui Zhang, Yutong Zhou, Yuxin Yang, Jiarui Wu, and Wei Sun. Representation collapse in sequential post-training of large language models. *arXiv preprint arXiv:2605.30524*, 2026. Work in progress.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of ACL*, 2016.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40), 2020.

Yiheng Tao, Kaiwen Cheng, Yao Lu, Chang Liu, and Jie Chen. TRSP: Official spectral diagnostic implementation, 2026. URL `https://github.com/Eziotao-tyd/TRSP`. Accessed 2026-09-20; pooled-spectrum diagnostic adaptation only.

Team OLMo, Pete Walsh, Luca Soldaini, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In *Topological, Algebraic and Geometric Learning Workshops (TAG-ML), PMLR 221*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, 2019.

Qinghua Zhao, Xueling Gong, Xinyu Chen, Zhongfeng Kang, and Xinlu Li. A layer-wise analysis of supervised fine-tuning. *arXiv preprint arXiv:2604.11838*, 2026.

Ziqian Zhong and Aditi Raghunathan. Watch the weights: Unsupervised monitoring and control of fine-tuned llms. In *International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ee64e62fd3ee763961ce3d6aed4971-Abstract-Conference.html`.

**Guide to the appendices.**

- **Appendix A (p. 12)—Full result tables.** The unaggregated per-branch versions of every main-text table: sign structure at the last and mid layers (Tables 6 and 7), replayed baseline monitors (Table 8), conformal per-branch detail (Table 9), the historical gradual-onset replay (Table 10), the clean sequential-alarm ordering (Table 11), the per-checkpoint capability battery (Tables 12, 13 and 14), and the alignment decomposition (Tables 15 and 16).
- **Appendix B (p. 20)—Protocol details and frozen decisions.** Models; training and probe configuration; frozen selection gates; the damage gate; gradual-onset, benign, sequential, and conformal designs; capability evaluation; identifiability certificates; artifact audit.
- **Appendix C (p. 22)—Post-review data audit and supplementary analyses.** The probe-exposure audit and repair: exposure-filtered endpoints (Table 17), independent capability events, the ablation grid (Table 18), diagnostic values (Table 19), the fresh text-disjoint confirmation (Table 20, Figure 3), the historical-vs-clean side-by-side (Table 21), clean capability endpoints, and the cross-model clean replication (Table 22).

## A FULL RESULT TABLES

The tables in this appendix are the unaggregated per-branch versions of the main tables, script-generated from the same frozen inputs, in four groups: sign structure at the last and mid layers (Tables 6 and 7); monitor replays and alarm timing (Tables 8, 9, 10 and 11); the per-checkpoint capability battery (Tables 12, 13 and 14); and the alignment decomposition (Tables 15 and 16). Table 10 is the historical gradual-onset replay behind Figure 2's clean counterpart. Throughout the appendices, bold marks one of the two frozen criteria restated in the captions: $\Delta$NLL at or beyond the $+0.30$-nat damage gate, or a task-metric decline of at least 5 points (the capability-event margin of Appendix C).

Table 6: Per-branch endpoint deltas, all models and seeds (layer $-1$). Bold $\Delta$NLL: at or beyond the frozen $+0.30$-nat damage gate.

| model | branch | dNLL | breach | rankme_raw | rankme_centred | k95 | eff_rank | xsim |
|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | domain_shift_seed0 | **+0.410** | 520 | +17.9 | +3.3 | 3 | +20.4 | -0.057 |
| Qwen3-0.6B | domain_shift_seed1 | **+0.391** | 590 | +23.6 | +3.8 | 6 | +4.7 | -0.081 |
| Qwen3-0.6B | domain_shift_seed2 | **+0.397** | 550 | +18.3 | +3.7 | 5 | +16.7 | -0.072 |
| Qwen3-0.6B | duplicate_data_seed0 | **+1.550** | 350 | +92.1 | +66.0 | 93 | +156.5 | -0.227 |
| Qwen3-0.6B | duplicate_data_seed1 | **+2.020** | 340 | +102.0 | +74.0 | 103 | +178.4 | -0.229 |
| Qwen3-0.6B | duplicate_data_seed2 | **+1.368** | 350 | +89.1 | +64.6 | 90 | +146.2 | -0.205 |
| Qwen3-0.6B | healthy_seed0 | -0.039 | – | +1.6 | -0.8 | -1 | -3.8 | -0.017 |
| Qwen3-0.6B | healthy_seed1 | -0.052 | – | +2.8 | -0.2 | 0 | -3.1 | -0.016 |
| Qwen3-0.6B | healthy_seed2 | -0.047 | – | +1.0 | -0.6 | -1 | -3.0 | -0.010 |
| Qwen3-0.6B | high_lr_seed0 | **+0.747** | 320 | +13.9 | -4.7 | -9 | +28.1 | -0.109 |
| Qwen3-0.6B | high_lr_seed1 | **+0.556** | 320 | +10.8 | -6.7 | -12 | +21.5 | -0.088 |
| Qwen3-0.6B | high_lr_seed2 | **+0.627** | 320 | -2.2 | -16.4 | -30 | +5.3 | -0.081 |
| Qwen3-1.7B | dup100_seed0 | -0.039 | – | +10.7 | +1.4 | 3 | -1.2 | -0.023 |
| Qwen3-1.7B | dup100_seed1 | -0.050 | – | +9.0 | +0.9 | 3 | -0.9 | -0.022 |
| Qwen3-1.7B | dup100_seed2 | -0.048 | – | +6.6 | -0.8 | -1 | -5.0 | -0.019 |
| Qwen3-1.7B | dup12_seed0 | +0.054 | – | +91.8 | +38.6 | 69 | +26.9 | -0.119 |
| Qwen3-1.7B | dup12_seed1 | +0.039 | – | +92.8 | +40.0 | 70 | +24.8 | -0.115 |
| Qwen3-1.7B | dup12_seed2 | +0.054 | – | +91.1 | +39.1 | 69 | +24.1 | -0.114 |
| Qwen3-1.7B | dup25_seed0 | -0.018 | – | +44.9 | +16.5 | 30 | +9.1 | -0.059 |
| Qwen3-1.7B | dup25_seed1 | -0.031 | – | +50.2 | +20.4 | 37 | +12.2 | -0.060 |
| Qwen3-1.7B | dup25_seed2 | -0.018 | – | +45.4 | +17.7 | 32 | +9.6 | -0.055 |
| Qwen3-1.7B | dup50_seed0 | -0.033 | – | +22.4 | +6.7 | 13 | +3.2 | -0.034 |
| Qwen3-1.7B | dup50_seed1 | -0.045 | – | +24.9 | +9.5 | 18 | +6.6 | -0.033 |
| Qwen3-1.7B | dup50_seed2 | -0.043 | – | +19.6 | +5.3 | 10 | -0.4 | -0.030 |
| Qwen3-1.7B | dupq8_seed0 | **+1.222** | 360 | +223.1 | +108.1 | 192 | +167.3 | -0.375 |
| Qwen3-1.7B | dupq8_seed1 | **+1.257** | 360 | +223.0 | +105.4 | 187 | +144.0 | -0.371 |
| Qwen3-1.7B | dupq8_seed2 | **+0.737** | 360 | +188.8 | +84.8 | 150 | +87.4 | -0.307 |
| Qwen3-1.7B | healthy_seed0 | -0.038 | – | +2.2 | -2.5 | -3 | -4.6 | -0.016 |
| Qwen3-1.7B | healthy_seed1 | -0.048 | – | +13.7 | +2.1 | 5 | +0.5 | -0.026 |
| Qwen3-1.7B | healthy_seed2 | -0.046 | – | +8.3 | +0.4 | 2 | -1.5 | -0.021 |
| Qwen3-1.7B | high_lr_seed0 | **+0.555** | 340 | +131.5 | +1.9 | 33 | +93.6 | -0.320 |
| Qwen3-1.7B | high_lr_seed1 | **+0.357** | 350 | +149.0 | +8.1 | 42 | +76.9 | -0.337 |
| Qwen3-1.7B | high_lr_seed2 | **+0.430** | 340 | +113.5 | -10.4 | 10 | +55.1 | -0.309 |
| OLMo-2-1B | dup100_seed0 | -0.037 | – | +7.7 | +6.0 | 10 | +8.8 | -0.006 |
| OLMo-2-1B | dup100_seed1 | -0.045 | – | +5.6 | +4.8 | 7 | +5.8 | -0.003 |
| OLMo-2-1B | dup100_seed2 | -0.043 | – | +8.0 | +6.2 | 10 | +8.2 | -0.005 |
| OLMo-2-1B | dup12_seed0 | -0.017 | – | +47.3 | +37.9 | 63 | +56.2 | -0.041 |
| OLMo-2-1B | dup12_seed1 | -0.023 | – | +48.0 | +38.5 | 63 | +54.9 | -0.041 |
| OLMo-2-1B | dup12_seed2 | -0.009 | – | +54.0 | +42.8 | 71 | +62.6 | -0.048 |
| OLMo-2-1B | dup25_seed0 | -0.032 | – | +26.9 | +21.2 | 36 | +30.4 | -0.024 |
| OLMo-2-1B | dup25_seed1 | -0.038 | – | +24.6 | +19.9 | 33 | +25.9 | -0.021 |
| OLMo-2-1B | dup25_seed2 | -0.029 | – | +28.5 | +22.4 | 37 | +30.8 | -0.025 |
| OLMo-2-1B | dup50_seed0 | -0.035 | – | +14.1 | +11.2 | 19 | +16.2 | -0.013 |
| OLMo-2-1B | dup50_seed1 | -0.042 | – | +13.2 | +10.8 | 17 | +14.4 | -0.010 |
| OLMo-2-1B | dup50_seed2 | -0.041 | – | +14.6 | +11.3 | 18 | +15.1 | -0.011 |
| OLMo-2-1B | dupq8_seed0 | **+0.884** | 380 | +234.8 | +200.3 | 296 | +517.0 | -0.150 |
| OLMo-2-1B | dupq8_seed1 | **+0.804** | 380 | +226.1 | +192.1 | 288 | +456.2 | -0.152 |
| OLMo-2-1B | dupq8_seed2 | **+0.455** | 390 | +172.8 | +142.9 | 222 | +291.5 | -0.139 |
| OLMo-2-1B | healthy_seed0 | -0.038 | – | +7.0 | +5.2 | 8 | +8.5 | -0.006 |
| OLMo-2-1B | healthy_seed1 | -0.044 | – | +5.1 | +4.2 | 6 | +5.4 | -0.003 |
| OLMo-2-1B | healthy_seed2 | -0.042 | – | +5.2 | +3.9 | 5 | +5.9 | -0.004 |
| OLMo-2-1B | high_lr_seed0 | **+0.423** | 390 | -64.0 | -19.6 | -47 | -41.6 | +0.174 |
| OLMo-2-1B | high_lr_seed1 | +0.194 | 390 | -33.6 | +4.2 | -5 | -29.4 | +0.141 |
| OLMo-2-1B | high_lr_seed2 | +0.262 | 420 | -44.2 | -6.5 | -24 | -31.7 | +0.144 |

Table 7: Per-branch endpoint deltas at the *mid* layer (same statistics as Table 6). The duplication inversion is layer-robust (rank rises at both layers in all nine damaged runs); the high-LR sign flip is *last-layer specific*—at the mid layer raw RankMe rises under high-LR in OLMo as well ($+24$ to $+44$), so the response sign depends on the monitored layer in addition to regime, statistic, and architecture. Bold as in Table 6.

| model | branch | dNLL | breach | rankme_raw | rankme_centred | k95 | eff_rank | xsim |
|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | domain_shift_seed0 | **+0.410** | 520 | +81.6 | +0.5 | 3 | +15.1 | -0.064 |
| Qwen3-0.6B | domain_shift_seed1 | **+0.391** | 590 | +107.1 | +3.9 | 9 | +20.4 | -0.069 |
| Qwen3-0.6B | domain_shift_seed2 | **+0.397** | 550 | +92.8 | -0.2 | 1 | +16.1 | -0.062 |
| Qwen3-0.6B | duplicate_data_seed0 | **+1.550** | 350 | +18.6 | +33.1 | 56 | +41.1 | -0.080 |
| Qwen3-0.6B | duplicate_data_seed1 | **+2.020** | 340 | +19.3 | +41.1 | 70 | +52.7 | -0.100 |
| Qwen3-0.6B | duplicate_data_seed2 | **+1.368** | 350 | +15.0 | +33.7 | 58 | +43.4 | -0.079 |
| Qwen3-0.6B | healthy_seed0 | -0.039 | – | +4.3 | -1.4 | -2 | -0.3 | -0.003 |
| Qwen3-0.6B | healthy_seed1 | -0.052 | – | +5.3 | -0.4 | 0 | +1.2 | -0.002 |
| Qwen3-0.6B | healthy_seed2 | -0.047 | – | +4.0 | -1.7 | -2 | -0.5 | +0.001 |
| Qwen3-0.6B | high_lr_seed0 | **+0.747** | 320 | +193.6 | +17.6 | 29 | +47.4 | -0.095 |
| Qwen3-0.6B | high_lr_seed1 | **+0.556** | 320 | +134.2 | +14.2 | 23 | +43.1 | -0.108 |
| Qwen3-0.6B | high_lr_seed2 | **+0.627** | 320 | +187.9 | +11.9 | 18 | +40.6 | -0.086 |
| Qwen3-1.7B | dup100_seed0 | -0.039 | – | +4.9 | -4.3 | -7 | -0.9 | +0.002 |
| Qwen3-1.7B | dup100_seed1 | -0.050 | – | +0.9 | -6.7 | -10 | -3.5 | +0.004 |
| Qwen3-1.7B | dup100_seed2 | -0.048 | – | +0.1 | -6.8 | -11 | -5.2 | +0.005 |
| Qwen3-1.7B | dup12_seed0 | +0.054 | – | +37.9 | +52.7 | 87 | +76.3 | -0.081 |
| Qwen3-1.7B | dup12_seed1 | +0.039 | – | +33.6 | +51.0 | 85 | +75.1 | -0.072 |
| Qwen3-1.7B | dup12_seed2 | +0.054 | – | +36.5 | +54.8 | 90 | +81.8 | -0.079 |
| Qwen3-1.7B | dup25_seed0 | -0.018 | – | +18.1 | +20.2 | 34 | +28.7 | -0.030 |
| Qwen3-1.7B | dup25_seed1 | -0.031 | – | +14.7 | +18.7 | 32 | +29.1 | -0.026 |
| Qwen3-1.7B | dup25_seed2 | -0.018 | – | +15.9 | +21.6 | 36 | +31.7 | -0.030 |
| Qwen3-1.7B | dup50_seed0 | -0.033 | – | +8.1 | +3.0 | 5 | +7.1 | -0.007 |
| Qwen3-1.7B | dup50_seed1 | -0.045 | – | +6.5 | +3.2 | 6 | +8.8 | -0.007 |
| Qwen3-1.7B | dup50_seed2 | -0.043 | – | +5.2 | +2.3 | 4 | +5.7 | -0.005 |
| Qwen3-1.7B | dupq8_seed0 | **+1.222** | 360 | +44.1 | +59.0 | 97 | +88.9 | -0.090 |
| Qwen3-1.7B | dupq8_seed1 | **+1.257** | 360 | +50.9 | +73.1 | 120 | +106.9 | -0.121 |
| Qwen3-1.7B | dupq8_seed2 | **+0.737** | 360 | +30.7 | +44.4 | 74 | +63.7 | -0.069 |
| Qwen3-1.7B | healthy_seed0 | -0.038 | – | +1.2 | -7.3 | -12 | -3.8 | +0.005 |
| Qwen3-1.7B | healthy_seed1 | -0.048 | – | -0.6 | -6.7 | -10 | -4.2 | +0.005 |
| Qwen3-1.7B | healthy_seed2 | -0.046 | – | -0.9 | -6.1 | -10 | -3.4 | +0.004 |
| Qwen3-1.7B | high_lr_seed0 | **+0.555** | 340 | +326.3 | +27.9 | 47 | +101.4 | -0.123 |
| Qwen3-1.7B | high_lr_seed1 | **+0.357** | 350 | +313.9 | +26.9 | 46 | +100.1 | -0.136 |
| Qwen3-1.7B | high_lr_seed2 | **+0.430** | 340 | +356.3 | +18.2 | 30 | +87.7 | -0.117 |
| OLMo-2-1B | dup100_seed0 | -0.037 | – | -4.3 | -4.1 | -7 | -8.9 | +0.002 |
| OLMo-2-1B | dup100_seed1 | -0.045 | – | -2.9 | -3.2 | -5 | -7.0 | +0.001 |
| OLMo-2-1B | dup100_seed2 | -0.043 | – | -3.4 | -3.3 | -5 | -7.6 | +0.001 |
| OLMo-2-1B | dup12_seed0 | -0.017 | – | +11.3 | +8.5 | 13 | +21.5 | -0.009 |
| OLMo-2-1B | dup12_seed1 | -0.023 | – | +10.0 | +7.2 | 12 | +18.3 | -0.008 |
| OLMo-2-1B | dup12_seed2 | -0.009 | – | +11.6 | +10.0 | 15 | +24.7 | -0.010 |
| OLMo-2-1B | dup25_seed0 | -0.032 | – | +2.2 | +1.3 | 2 | +3.5 | -0.003 |
| OLMo-2-1B | dup25_seed1 | -0.038 | – | +3.3 | +2.0 | 3 | +5.8 | -0.004 |
| OLMo-2-1B | dup25_seed2 | -0.029 | – | +4.5 | +3.8 | 6 | +9.5 | -0.004 |
| OLMo-2-1B | dup50_seed0 | -0.035 | – | -2.1 | -2.1 | -3 | -4.4 | -0.000 |
| OLMo-2-1B | dup50_seed1 | -0.042 | – | -1.0 | -1.4 | -2 | -2.7 | -0.001 |
| OLMo-2-1B | dup50_seed2 | -0.041 | – | -1.5 | -1.8 | -3 | -4.3 | +0.000 |
| OLMo-2-1B | dupq8_seed0 | **+0.884** | 380 | +32.3 | +28.6 | 44 | +75.9 | -0.025 |
| OLMo-2-1B | dupq8_seed1 | **+0.804** | 380 | +38.2 | +33.3 | 52 | +89.2 | -0.029 |
| OLMo-2-1B | dupq8_seed2 | **+0.455** | 390 | +31.6 | +27.3 | 42 | +72.2 | -0.023 |
| OLMo-2-1B | healthy_seed0 | -0.038 | – | -3.8 | -3.6 | -6 | -7.9 | +0.001 |
| OLMo-2-1B | healthy_seed1 | -0.044 | – | -4.7 | -4.4 | -7 | -10.0 | +0.002 |
| OLMo-2-1B | healthy_seed2 | -0.042 | – | -4.7 | -4.5 | -7 | -10.7 | +0.003 |
| OLMo-2-1B | high_lr_seed0 | **+0.423** | 390 | +24.2 | -3.1 | -3 | +50.3 | -0.030 |
| OLMo-2-1B | high_lr_seed1 | +0.194 | 390 | +44.2 | +8.5 | 16 | +70.4 | -0.036 |
| OLMo-2-1B | high_lr_seed2 | +0.262 | 420 | +40.3 | +6.8 | 13 | +71.9 | -0.038 |

Table 8: Per-branch replayed baseline monitors. Bold as in Table 6.

| model | branch | dNLL | sigma_gkf | entropy_a1 | entropy_a2 |
|---|---|---|---|---|---|
| Qwen3-0.6B | domain_shift_seed0 | **+0.410** | -248.3 | +0.519 | +0.359 |
| Qwen3-0.6B | domain_shift_seed1 | **+0.391** | -217.7 | +0.626 | +0.484 |
| Qwen3-0.6B | domain_shift_seed2 | **+0.397** | -228.2 | +0.569 | +0.430 |
| Qwen3-0.6B | duplicate_data_seed0 | **+1.550** | +1088.8 | +2.067 | +1.858 |
| Qwen3-0.6B | duplicate_data_seed1 | **+2.020** | +1130.1 | +2.222 | +1.994 |
| Qwen3-0.6B | duplicate_data_seed2 | **+1.368** | +1047.0 | +1.899 | +1.613 |
| Qwen3-0.6B | healthy_seed0 | -0.039 | +23.5 | +0.078 | +0.067 |
| Qwen3-0.6B | healthy_seed1 | -0.052 | +21.5 | +0.097 | +0.081 |
| Qwen3-0.6B | healthy_seed2 | -0.047 | +11.3 | +0.038 | +0.032 |
| Qwen3-0.6B | high_lr_seed0 | **+0.747** | -765.8 | +0.934 | +0.792 |
| Qwen3-0.6B | high_lr_seed1 | **+0.556** | -683.7 | +0.754 | +0.591 |
| Qwen3-0.6B | high_lr_seed2 | **+0.627** | -816.5 | +0.583 | +0.489 |
| Qwen3-1.7B | dup100_seed0 | -0.039 | +109.8 | +0.173 | +0.088 |
| Qwen3-1.7B | dup100_seed1 | -0.050 | +75.4 | +0.163 | +0.084 |
| Qwen3-1.7B | dup100_seed2 | -0.048 | +61.4 | +0.129 | +0.069 |
| Qwen3-1.7B | dup12_seed0 | +0.054 | +847.7 | +1.018 | +0.496 |
| Qwen3-1.7B | dup12_seed1 | +0.039 | +825.7 | +0.980 | +0.472 |
| Qwen3-1.7B | dup12_seed2 | +0.054 | +831.8 | +0.975 | +0.473 |
| Qwen3-1.7B | dup25_seed0 | -0.018 | +408.1 | +0.509 | +0.244 |
| Qwen3-1.7B | dup25_seed1 | -0.031 | +416.6 | +0.527 | +0.247 |
| Qwen3-1.7B | dup25_seed2 | -0.018 | +419.4 | +0.479 | +0.224 |
| Qwen3-1.7B | dup50_seed0 | -0.033 | +208.9 | +0.282 | +0.136 |
| Qwen3-1.7B | dup50_seed1 | -0.045 | +199.5 | +0.285 | +0.134 |
| Qwen3-1.7B | dup50_seed2 | -0.043 | +176.4 | +0.238 | +0.116 |
| Qwen3-1.7B | dupq8_seed0 | **+1.222** | +2447.0 | +3.257 | +2.154 |
| Qwen3-1.7B | dupq8_seed1 | **+1.257** | +2406.2 | +3.199 | +2.099 |
| Qwen3-1.7B | dupq8_seed2 | **+0.737** | +1954.3 | +2.550 | +1.530 |
| Qwen3-1.7B | healthy_seed0 | -0.038 | +34.7 | +0.104 | +0.059 |
| Qwen3-1.7B | healthy_seed1 | -0.048 | +109.5 | +0.211 | +0.105 |
| Qwen3-1.7B | healthy_seed2 | -0.046 | +72.0 | +0.147 | +0.076 |
| Qwen3-1.7B | high_lr_seed0 | **+0.555** | -1210.8 | +2.981 | +1.963 |
| Qwen3-1.7B | high_lr_seed1 | **+0.357** | -1074.0 | +3.067 | +2.032 |
| Qwen3-1.7B | high_lr_seed2 | **+0.430** | -1250.8 | +2.653 | +1.712 |
| OLMo-2-1B | dup100_seed0 | -0.037 | +48.4 | +0.075 | +0.052 |
| OLMo-2-1B | dup100_seed1 | -0.045 | +32.9 | +0.041 | +0.022 |
| OLMo-2-1B | dup100_seed2 | -0.043 | +49.3 | +0.074 | +0.050 |
| OLMo-2-1B | dup12_seed0 | -0.017 | +310.7 | +0.445 | +0.323 |
| OLMo-2-1B | dup12_seed1 | -0.023 | +312.6 | +0.445 | +0.326 |
| OLMo-2-1B | dup12_seed2 | -0.009 | +351.0 | +0.513 | +0.382 |
| OLMo-2-1B | dup25_seed0 | -0.032 | +177.4 | +0.257 | +0.186 |
| OLMo-2-1B | dup25_seed1 | -0.038 | +157.9 | +0.216 | +0.150 |
| OLMo-2-1B | dup25_seed2 | -0.029 | +184.0 | +0.266 | +0.190 |
| OLMo-2-1B | dup50_seed0 | -0.035 | +90.7 | +0.134 | +0.093 |
| OLMo-2-1B | dup50_seed1 | -0.042 | +83.8 | +0.115 | +0.076 |
| OLMo-2-1B | dup50_seed2 | -0.041 | +92.1 | +0.135 | +0.094 |
| OLMo-2-1B | dupq8_seed0 | **+0.884** | +1444.2 | +2.324 | +2.232 |
| OLMo-2-1B | dupq8_seed1 | **+0.804** | +1390.2 | +2.206 | +2.111 |
| OLMo-2-1B | dupq8_seed2 | **+0.455** | +1094.5 | +1.719 | +1.521 |
| OLMo-2-1B | healthy_seed0 | -0.038 | +43.8 | +0.077 | +0.057 |
| OLMo-2-1B | healthy_seed1 | -0.044 | +30.1 | +0.044 | +0.029 |
| OLMo-2-1B | healthy_seed2 | -0.042 | +33.5 | +0.055 | +0.039 |
| OLMo-2-1B | high_lr_seed0 | **+0.423** | -920.2 | -1.214 | -1.040 |
| OLMo-2-1B | high_lr_seed1 | +0.194 | -737.9 | -1.070 | -0.979 |
| OLMo-2-1B | high_lr_seed2 | +0.262 | -803.1 | -1.068 | -0.948 |

Table 9: Conformal per-branch detail: alarm, breach, both leads, censoring preserved, for the all-spectral union and both NLL variants. Bold as in Table 6.

| pool | family | alpha | branch | alarm | breach | lead | nll_alarm | lead_nll | dNLL |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | 310 | 0 | **+0.410** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | 310 | 0 | **+0.391** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | 310 | 0 | **+0.397** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 320 | 350 | 30 | 320 | 0 | **+1.550** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 320 | 340 | 20 | 320 | 0 | **+2.020** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 320 | 350 | 30 | 320 | 0 | **+1.368** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | 310 | 0 | **+0.747** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | 310 | 0 | **+0.556** |
| Qwen3-0.6B | rank_family | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | 310 | 0 | **+0.627** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | 310 | 0 | **+0.410** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | 310 | 0 | **+0.391** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | 310 | 0 | **+0.397** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 320 | 350 | 30 | 320 | 0 | **+1.550** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 320 | 340 | 20 | 320 | 0 | **+2.020** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 320 | 350 | 30 | 320 | 0 | **+1.368** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | 310 | 0 | **+0.747** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | 310 | 0 | **+0.556** |
| Qwen3-0.6B | entropy_sigma | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | 310 | 0 | **+0.627** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | 310 | 0 | **+0.410** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | 310 | 0 | **+0.391** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | 310 | 0 | **+0.397** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 320 | 350 | 30 | 320 | 0 | **+1.550** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 320 | 340 | 20 | 320 | 0 | **+2.020** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 320 | 350 | 30 | 320 | 0 | **+1.368** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | 310 | 0 | **+0.747** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | 310 | 0 | **+0.556** |
| Qwen3-0.6B | xsim | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | 310 | 0 | **+0.627** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | 310 | 0 | **+0.410** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | 310 | 0 | **+0.391** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | 310 | 0 | **+0.397** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 320 | 350 | 30 | 320 | 0 | **+1.550** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 320 | 340 | 20 | 320 | 0 | **+2.020** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 320 | 350 | 30 | 320 | 0 | **+1.368** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | 310 | 0 | **+0.747** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | 310 | 0 | **+0.556** |
| Qwen3-0.6B | all | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | 310 | 0 | **+0.627** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | – | – | **+0.410** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | – | – | **+0.391** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | – | – | **+0.397** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 320 | 350 | 30 | – | – | **+1.550** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 320 | 340 | 20 | – | – | **+2.020** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 320 | 350 | 30 | – | – | **+1.368** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | – | – | **+0.747** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | – | – | **+0.556** |
| Qwen3-0.6B | nll_up | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | – | – | **+0.627** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/domain_shift_seed0 | 310 | 520 | 210 | – | – | **+0.410** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/domain_shift_seed1 | 310 | 590 | 280 | – | – | **+0.391** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/domain_shift_seed2 | 310 | 550 | 240 | – | – | **+0.397** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/duplicate_data_seed0 | 330 | 350 | 20 | – | – | **+1.550** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/duplicate_data_seed1 | 330 | 340 | 10 | – | – | **+2.020** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/duplicate_data_seed2 | 330 | 350 | 20 | – | – | **+1.368** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/high_lr_seed0 | 310 | 320 | 10 | – | – | **+0.747** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/high_lr_seed1 | 310 | 320 | 10 | – | – | **+0.556** |
| Qwen3-0.6B | nll_abs | +0.050 | Qwen3-0.6B/high_lr_seed2 | 310 | 320 | 10 | – | – | **+0.627** |

Table 10: **Gradual-onset damage, historical campaign** (pre-registered ramps, 3 seeds each, scored once against the frozen historical thresholds; step ranges over seeds; censoring preserved). Clean-campaign counterparts: Figure 2 and Table 11. Bold as in Table 6.

| gradual regime | ΔNLL end | damage (step) | spectral alarm | NLL alarm | lead vs. $\text{NLL}_{\uparrow}$ | vs. $\text{NLL}_{\lvert\cdot\rvert}$ |
|---|---|---|---|---|---|---|
| LR ramp $1\times \rightarrow 15\times$ | **+0.53 ± 0.09** | 3/3 (730–760) | 400–410 | 470–520 | 60–110 | 170–190 |
| duplication ramp $p(t)=t/600$ | +0.06 ± 0.00 | 0/3 (–) | 410–450 | 460–490 | 40–50 | 100–160 |

Table 11: Sequential-detector alarm ordering, clean campaign (per-regime step ranges over 3 seeds; "–" = no alarm within the horizon). Numeric companion to Figure 2: the earliest spectral family retains a 30–90 step lead over both sequential loss-stream detectors on damaging ramps, but every detector above the last row alarms on the no-damage duplication ramp and the benign control alike. The historical ordering is qualitatively identical (Table 21).

| detector | LR ramp (damaging) | dup. ramp (no damage) | benign control |
|---|---|---|---|
| earliest spectral family | 390–410 | 420–440 | 320–330 |
| CUSUM (EMA baseline, one-sided) | 440–460 | 430–450 | 320 |
| Page–Hinkley (one-sided) | 460–480 | 460–490 | 320 |
| NLL one-sided (frozen) | 490–500 | 480–530 | 320 |
| NLL two-sided (frozen) | 560–570 | 630–700 | – |
| damage gate (+0.30 nats) | 750–780 | – | – |

Table 12: Capability metrics per checkpoint and seed, part 1 of 3 (including Dolly and GSM8K NLL columns; GSM8K is NLL, not solve accuracy). Bold task deltas: decline of at least 5 points, the capability-event margin of Appendix C.

| checkpoint | dolly_nll | gsm8k_nll | hswag | arc_c | lambada | d_hswag | d_arc_c | d_lambada |
|---|---|---|---|---|---|---|---|---|
| OLMo-2-0425-1B s0:fork | +2.211 | +1.252 | +0.626 | +0.416 | +0.678 | +0.000 | +0.000 | +0.000 |
| OLMo-2-0425-1B s0:healthy_seed0/endpoint | +2.173 | +1.248 | +0.630 | +0.406 | +0.686 | +0.004 | -0.010 | +0.008 |
| OLMo-2-0425-1B s0:dup100_seed0/endpoint | +2.174 | +1.247 | +0.632 | +0.408 | +0.686 | +0.006 | -0.008 | +0.008 |
| OLMo-2-0425-1B s0:dup50_seed0/endpoint | +2.176 | +1.248 | +0.630 | +0.408 | +0.690 | +0.004 | -0.008 | +0.012 |
| OLMo-2-0425-1B s0:dup25_seed0/endpoint | +2.179 | +1.255 | +0.636 | +0.410 | +0.688 | +0.010 | -0.006 | +0.010 |
| OLMo-2-0425-1B s0:dup12_seed0/endpoint | +2.194 | +1.270 | +0.642 | +0.418 | +0.682 | +0.016 | +0.002 | +0.004 |
| OLMo-2-0425-1B s0:dupq8_seed0/breach | +2.630 | +1.627 | +0.654 | +0.398 | +0.658 | +0.028 | -0.018 | -0.020 |
| OLMo-2-0425-1B s0:dupq8_seed0/endpoint | +3.095 | +1.908 | +0.624 | +0.370 | +0.652 | -0.002 | -0.046 | -0.026 |
| OLMo-2-0425-1B s0:high_lr_seed0/breach | +2.531 | +1.508 | +0.580 | +0.350 | +0.656 | -0.046 | **-0.066** | -0.022 |
| OLMo-2-0425-1B s0:high_lr_seed0/endpoint | +2.634 | +1.676 | +0.534 | +0.350 | +0.640 | **-0.092** | **-0.066** | -0.038 |
| OLMo-2-0425-1B s1:fork | +2.209 | +1.254 | +0.626 | +0.414 | +0.678 | +0.000 | +0.000 | +0.000 |
| OLMo-2-0425-1B s1:healthy_seed1/endpoint | +2.165 | +1.249 | +0.628 | +0.404 | +0.684 | +0.002 | -0.010 | +0.006 |
| OLMo-2-0425-1B s1:dup100_seed1/endpoint | +2.164 | +1.246 | +0.626 | +0.412 | +0.692 | +0.000 | -0.002 | +0.014 |
| OLMo-2-0425-1B s1:dup50_seed1/endpoint | +2.167 | +1.251 | +0.626 | +0.410 | +0.686 | +0.000 | -0.004 | +0.008 |
| OLMo-2-0425-1B s1:dup25_seed1/endpoint | +2.171 | +1.257 | +0.630 | +0.416 | +0.690 | +0.004 | +0.002 | +0.012 |
| OLMo-2-0425-1B s1:dup12_seed1/endpoint | +2.186 | +1.278 | +0.648 | +0.410 | +0.690 | +0.022 | -0.004 | +0.012 |
| OLMo-2-0425-1B s1:dupq8_seed1/breach | +2.586 | +1.610 | +0.640 | +0.398 | +0.662 | +0.014 | -0.016 | -0.016 |
| OLMo-2-0425-1B s1:dupq8_seed1/endpoint | +3.013 | +1.918 | +0.618 | +0.378 | +0.654 | -0.008 | -0.036 | -0.024 |
| OLMo-2-0425-1B s1:high_lr_seed1/breach | +2.515 | +1.685 | +0.568 | +0.332 | +0.588 | **-0.058** | **-0.082** | **-0.090** |
| OLMo-2-0425-1B s1:high_lr_seed1/endpoint | +2.403 | +1.764 | +0.540 | +0.348 | +0.546 | **-0.086** | **-0.066** | **-0.132** |
| OLMo-2-0425-1B s2:fork | +2.207 | +1.254 | +0.630 | +0.408 | +0.680 | +0.000 | +0.000 | +0.000 |
| OLMo-2-0425-1B s2:healthy_seed2/endpoint | +2.166 | +1.250 | +0.622 | +0.408 | +0.686 | -0.008 | +0.000 | +0.006 |
| OLMo-2-0425-1B s2:dup100_seed2/endpoint | +2.164 | +1.248 | +0.630 | +0.410 | +0.682 | +0.000 | +0.002 | +0.002 |
| OLMo-2-0425-1B s2:dup50_seed2/endpoint | +2.167 | +1.251 | +0.632 | +0.414 | +0.684 | +0.002 | +0.006 | +0.004 |
| OLMo-2-0425-1B s2:dup25_seed2/endpoint | +2.178 | +1.260 | +0.642 | +0.418 | +0.690 | +0.012 | +0.010 | +0.010 |
| OLMo-2-0425-1B s2:dup12_seed2/endpoint | +2.199 | +1.282 | +0.652 | +0.422 | +0.686 | +0.022 | +0.014 | +0.006 |
| OLMo-2-0425-1B s2:dupq8_seed2/breach | +2.524 | +1.525 | +0.650 | +0.400 | +0.670 | +0.020 | -0.008 | -0.010 |
| OLMo-2-0425-1B s2:dupq8_seed2/endpoint | +2.663 | +1.606 | +0.638 | +0.378 | +0.668 | +0.008 | -0.030 | -0.012 |
| OLMo-2-0425-1B s2:high_lr_seed2/breach | +2.508 | +1.637 | +0.574 | +0.372 | +0.622 | **-0.056** | -0.036 | **-0.058** |
| OLMo-2-0425-1B s2:high_lr_seed2/endpoint | +2.469 | +1.787 | +0.542 | +0.348 | +0.480 | **-0.088** | **-0.060** | **-0.200** |

Table 13: Capability metrics per checkpoint and seed, part 2 of 3. Bold as in Table 12.

| checkpoint | dolly_nll | gsm8k_nll | hswag | arc_c | lambada | d_hswag | d_arc_c | d_lambada |
|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B s0:fork | +2.584 | +1.403 | +0.474 | +0.332 | +0.466 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s0:lr_ramp_seed0/breach | +2.884 | +1.633 | +0.412 | +0.340 | +0.398 | **-0.062** | +0.008 | **-0.068** |
| Qwen3-0.6B s0:lr_ramp_seed0/endpoint | +3.202 | +1.872 | +0.378 | +0.284 | +0.282 | **-0.096** | -0.048 | **-0.184** |
| Qwen3-0.6B s0:dup_ramp_seed0/endpoint | +2.638 | +1.458 | +0.492 | +0.352 | +0.450 | +0.018 | +0.020 | -0.016 |
| Qwen3-0.6B s1:fork | +2.583 | +1.401 | +0.474 | +0.328 | +0.440 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s1:lr_ramp_seed1/breach | +2.888 | +1.618 | +0.398 | +0.322 | +0.424 | **-0.076** | -0.006 | -0.016 |
| Qwen3-0.6B s1:lr_ramp_seed1/endpoint | +3.027 | +1.861 | +0.420 | +0.260 | +0.276 | **-0.054** | **-0.068** | **-0.164** |
| Qwen3-0.6B s1:dup_ramp_seed1/endpoint | +2.636 | +1.466 | +0.480 | +0.334 | +0.450 | +0.006 | +0.006 | +0.010 |
| Qwen3-0.6B s2:fork | +2.579 | +1.403 | +0.474 | +0.330 | +0.456 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s2:lr_ramp_seed2/breach | +2.906 | +1.649 | +0.410 | +0.318 | +0.318 | **-0.064** | -0.012 | **-0.138** |
| Qwen3-0.6B s2:lr_ramp_seed2/endpoint | +3.105 | +1.963 | +0.366 | +0.282 | +0.230 | **-0.108** | -0.048 | **-0.226** |
| Qwen3-0.6B s2:dup_ramp_seed2/endpoint | +2.640 | +1.457 | +0.486 | +0.340 | +0.452 | +0.012 | +0.010 | -0.004 |
| Qwen3-0.6B s0:fork | +2.584 | +1.403 | +0.474 | +0.332 | +0.466 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s0:healthy_seed0/endpoint | +2.545 | +1.383 | +0.468 | +0.320 | +0.468 | -0.006 | -0.012 | +0.002 |
| Qwen3-0.6B s0:duplicate_data_seed0/breach | +3.195 | +1.779 | +0.510 | +0.352 | +0.424 | +0.036 | +0.020 | -0.042 |
| Qwen3-0.6B s0:duplicate_data_seed0/endpoint | +4.134 | +2.477 | +0.496 | +0.336 | +0.416 | +0.022 | +0.004 | **-0.050** |
| Qwen3-0.6B s0:high_lr_seed0/breach | +2.966 | +1.605 | +0.422 | +0.296 | +0.310 | **-0.052** | -0.036 | **-0.156** |
| Qwen3-0.6B s0:high_lr_seed0/endpoint | +3.319 | +2.086 | +0.328 | +0.284 | +0.258 | **-0.146** | -0.048 | **-0.208** |
| Qwen3-0.6B s0:domain_shift_seed0/breach | +2.918 | +1.134 | +0.426 | +0.296 | +0.474 | -0.048 | -0.036 | +0.008 |
| Qwen3-0.6B s0:domain_shift_seed0/endpoint | +2.975 | +1.095 | +0.420 | +0.318 | +0.396 | **-0.054** | -0.014 | **-0.070** |
| Qwen3-0.6B s1:fork | +2.583 | +1.401 | +0.474 | +0.328 | +0.440 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s1:healthy_seed1/endpoint | +2.530 | +1.384 | +0.476 | +0.334 | +0.466 | +0.002 | +0.006 | +0.026 |
| Qwen3-0.6B s1:duplicate_data_seed1/breach | +2.920 | +1.618 | +0.504 | +0.376 | +0.414 | +0.030 | +0.048 | -0.026 |
| Qwen3-0.6B s1:duplicate_data_seed1/endpoint | +4.602 | +2.997 | +0.460 | +0.346 | +0.374 | -0.014 | +0.018 | **-0.066** |
| Qwen3-0.6B s1:high_lr_seed1/breach | +2.946 | +1.564 | +0.434 | +0.322 | +0.350 | -0.040 | -0.006 | **-0.090** |
| Qwen3-0.6B s1:high_lr_seed1/endpoint | +3.132 | +2.194 | +0.354 | +0.240 | +0.236 | **-0.120** | **-0.088** | **-0.204** |
| Qwen3-0.6B s1:domain_shift_seed1/breach | +2.897 | +1.144 | +0.442 | +0.310 | +0.354 | -0.032 | -0.018 | **-0.086** |
| Qwen3-0.6B s1:domain_shift_seed1/endpoint | +2.953 | +1.101 | +0.432 | +0.300 | +0.408 | -0.042 | -0.028 | -0.032 |
| Qwen3-0.6B s2:fork | +2.579 | +1.403 | +0.474 | +0.330 | +0.456 | +0.000 | +0.000 | +0.000 |
| Qwen3-0.6B s2:healthy_seed2/endpoint | +2.533 | +1.386 | +0.478 | +0.328 | +0.458 | +0.004 | -0.002 | +0.002 |
| Qwen3-0.6B s2:duplicate_data_seed2/breach | +3.162 | +1.818 | +0.490 | +0.356 | +0.430 | +0.016 | +0.026 | -0.026 |
| Qwen3-0.6B s2:duplicate_data_seed2/endpoint | +3.947 | +2.418 | +0.466 | +0.316 | +0.424 | -0.008 | -0.014 | -0.032 |
| Qwen3-0.6B s2:high_lr_seed2/breach | +3.051 | +1.569 | +0.444 | +0.298 | +0.406 | -0.030 | -0.032 | **-0.050** |
| Qwen3-0.6B s2:high_lr_seed2/endpoint | +3.225 | +2.249 | +0.316 | +0.268 | +0.194 | **-0.158** | **-0.062** | **-0.262** |
| Qwen3-0.6B s2:domain_shift_seed2/breach | +2.882 | +1.143 | +0.434 | +0.336 | +0.444 | -0.040 | +0.006 | -0.012 |
| Qwen3-0.6B s2:domain_shift_seed2/endpoint | +2.964 | +1.100 | +0.418 | +0.300 | +0.332 | **-0.056** | -0.030 | **-0.124** |

Table 14: Capability metrics per checkpoint and seed, part 3 of 3. Bold as in Table 12.

| checkpoint | dolly_nll | gsm8k_nll | hswag | arc_c | lambada | d_hswag | d_arc_c | d_lambada |
|---|---|---|---|---|---|---|---|---|
| Qwen3-1.7B s0:fork | +2.264 | +1.219 | +0.588 | +0.432 | +0.594 | +0.000 | +0.000 | +0.000 |
| Qwen3-1.7B s0:healthy_seed0/endpoint | +2.226 | +1.197 | +0.604 | +0.448 | +0.602 | +0.016 | +0.016 | +0.008 |
| Qwen3-1.7B s0:dup100_seed0/endpoint | +2.226 | +1.203 | +0.604 | +0.438 | +0.594 | +0.016 | +0.006 | +0.000 |
| Qwen3-1.7B s0:dup50_seed0/endpoint | +2.231 | +1.212 | +0.604 | +0.444 | +0.596 | +0.016 | +0.012 | +0.002 |
| Qwen3-1.7B s0:dup25_seed0/endpoint | +2.246 | +1.248 | +0.616 | +0.444 | +0.590 | +0.028 | +0.012 | -0.004 |
| Qwen3-1.7B s0:dup12_seed0/endpoint | +2.318 | +1.413 | +0.622 | +0.452 | +0.564 | +0.034 | +0.020 | -0.030 |
| Qwen3-1.7B s0:dupq8_seed0/breach | +2.602 | +1.762 | +0.616 | +0.438 | +0.534 | +0.028 | +0.006 | **-0.060** |
| Qwen3-1.7B s0:dupq8_seed0/endpoint | +3.487 | +2.710 | +0.590 | +0.374 | +0.514 | +0.002 | **-0.058** | **-0.080** |
| Qwen3-1.7B s0:high_lr_seed0/breach | +2.565 | +1.360 | +0.564 | +0.442 | +0.482 | -0.024 | +0.010 | **-0.112** |
| Qwen3-1.7B s0:high_lr_seed0/endpoint | +2.819 | +1.609 | +0.460 | +0.356 | +0.424 | **-0.128** | **-0.076** | **-0.170** |
| Qwen3-1.7B s1:fork | +2.260 | +1.221 | +0.590 | +0.436 | +0.580 | +0.000 | +0.000 | +0.000 |
| Qwen3-1.7B s1:healthy_seed1/endpoint | +2.212 | +1.207 | +0.608 | +0.448 | +0.588 | +0.018 | +0.012 | +0.008 |
| Qwen3-1.7B s1:dup100_seed1/endpoint | +2.210 | +1.203 | +0.612 | +0.452 | +0.586 | +0.022 | +0.016 | +0.006 |
| Qwen3-1.7B s1:dup50_seed1/endpoint | +2.216 | +1.214 | +0.608 | +0.446 | +0.600 | +0.018 | +0.010 | +0.020 |
| Qwen3-1.7B s1:dup25_seed1/endpoint | +2.229 | +1.245 | +0.616 | +0.462 | +0.588 | +0.026 | +0.026 | +0.008 |
| Qwen3-1.7B s1:dup12_seed1/endpoint | +2.299 | +1.435 | +0.612 | +0.476 | +0.584 | +0.022 | +0.040 | +0.004 |
| Qwen3-1.7B s1:dupq8_seed1/breach | +2.738 | +1.925 | +0.628 | +0.434 | +0.520 | +0.038 | -0.002 | **-0.060** |
| Qwen3-1.7B s1:dupq8_seed1/endpoint | +3.517 | +2.781 | +0.588 | +0.390 | +0.510 | -0.002 | -0.046 | **-0.070** |
| Qwen3-1.7B s1:high_lr_seed1/breach | +2.604 | +1.361 | +0.546 | +0.378 | +0.540 | -0.044 | **-0.058** | -0.040 |
| Qwen3-1.7B s1:high_lr_seed1/endpoint | +2.617 | +1.713 | +0.464 | +0.334 | +0.366 | **-0.126** | **-0.102** | **-0.214** |
| Qwen3-1.7B s2:fork | +2.258 | +1.220 | +0.586 | +0.438 | +0.580 | +0.000 | +0.000 | +0.000 |
| Qwen3-1.7B s2:healthy_seed2/endpoint | +2.212 | +1.201 | +0.598 | +0.444 | +0.592 | +0.012 | +0.006 | +0.012 |
| Qwen3-1.7B s2:dup100_seed2/endpoint | +2.210 | +1.194 | +0.604 | +0.442 | +0.588 | +0.018 | +0.004 | +0.008 |
| Qwen3-1.7B s2:dup50_seed2/endpoint | +2.215 | +1.208 | +0.604 | +0.446 | +0.586 | +0.018 | +0.008 | +0.006 |
| Qwen3-1.7B s2:dup25_seed2/endpoint | +2.240 | +1.246 | +0.608 | +0.450 | +0.586 | +0.022 | +0.012 | +0.006 |
| Qwen3-1.7B s2:dup12_seed2/endpoint | +2.312 | +1.425 | +0.614 | +0.452 | +0.578 | +0.028 | +0.014 | -0.002 |
| Qwen3-1.7B s2:dupq8_seed2/breach | +2.572 | +1.644 | +0.618 | +0.426 | +0.570 | +0.032 | -0.012 | -0.010 |
| Qwen3-1.7B s2:dupq8_seed2/endpoint | +2.995 | +2.055 | +0.614 | +0.400 | +0.552 | +0.028 | -0.038 | -0.028 |
| Qwen3-1.7B s2:high_lr_seed2/breach | +2.568 | +1.411 | +0.548 | +0.396 | +0.410 | -0.038 | -0.042 | **-0.170** |
| Qwen3-1.7B s2:high_lr_seed2/endpoint | +2.689 | +1.696 | +0.454 | +0.360 | +0.342 | **-0.132** | **-0.078** | **-0.238** |

Table 15: **Alignment decomposition at endpoint** (layer $-1$; seed mean $\pm$ SD). Head energy = fraction of representation second moment in the top-128 right-singular subspace of the unembedding, computed against the checkpoint's own head ($W$) and against the frozen fork head. Drift = principal-angle drift of the top-32 eigenspace from fork.

| model | regime | headE fork | headE end (own $W$) | headE end (fork $W$) | grad-align | drift (°) |
|---|---|---|---|---|---|---|
| OLMo-2-1B | healthy | $0.124 \pm 0.000$ | $0.123 \pm 0.000$ | $0.124 \pm 0.000$ | $2.16 \pm 0.01$ | $2.6 \pm 0.2$ |
| OLMo-2-1B | dup. 12.5% | $0.124 \pm 0.000$ | $0.127 \pm 0.001$ | $0.127 \pm 0.001$ | $2.24 \pm 0.01$ | $5.4 \pm 0.7$ |
| OLMo-2-1B | dup. extreme ($n_u$=8) | $0.124 \pm 0.000$ | $0.143 \pm 0.003$ | $0.143 \pm 0.003$ | $2.67 \pm 0.12$ | $20.2 \pm 4.2$ |
| OLMo-2-1B | high-LR | $0.124 \pm 0.000$ | $0.113 \pm 0.001$ | $0.095 \pm 0.000$ | $1.87 \pm 0.05$ | $21.3 \pm 0.8$ |
| Qwen3-1.7B | healthy | $0.141 \pm 0.000$ | $0.138 \pm 0.001$ | $0.139 \pm 0.001$ | $1.18 \pm 0.01$ | $3.4 \pm 0.1$ |
| Qwen3-1.7B | dup. 12.5% | $0.141 \pm 0.000$ | $0.133 \pm 0.001$ | $0.133 \pm 0.001$ | $1.34 \pm 0.02$ | $6.8 \pm 0.0$ |
| Qwen3-1.7B | dup. extreme ($n_u$=8) | $0.141 \pm 0.000$ | $0.125 \pm 0.000$ | $0.125 \pm 0.000$ | $1.99 \pm 0.15$ | $16.5 \pm 2.1$ |
| Qwen3-1.7B | high-LR | $0.141 \pm 0.000$ | $0.176 \pm 0.007$ | $0.130 \pm 0.003$ | $3.06 \pm 0.10$ | $25.6 \pm 0.6$ |

Table 16: Task-alignment decomposition per checkpoint (both models, seed 0 shown per checkpoint kind; full JSONs in supplementary).

| checkpoint | NLL | headE_k128 | grad_align | drift_deg |
|---|---|---|---|---|
| Qwen3-1.7B s0:fork | +2.268 | +0.141 | +1.155 | – |
| Qwen3-1.7B s0:healthy_seed0/endpoint | +2.229 | +0.139 | +1.172 | +3.5 |
| Qwen3-1.7B s0:dup12_seed0/endpoint | +2.318 | +0.132 | +1.359 | +6.8 |
| Qwen3-1.7B s0:dupq8_seed0/breach | +2.606 | +0.127 | +1.728 | +10.0 |
| Qwen3-1.7B s0:dupq8_seed0/endpoint | +3.507 | +0.125 | +2.143 | +17.2 |
| Qwen3-1.7B s0:high_lr_seed0/breach | +2.564 | +0.111 | +1.773 | +18.5 |
| Qwen3-1.7B s0:high_lr_seed0/endpoint | +2.818 | +0.173 | +3.045 | +26.3 |
| Qwen3-1.7B s1:fork | +2.264 | +0.142 | +1.141 | – |
| Qwen3-1.7B s1:healthy_seed1/endpoint | +2.214 | +0.137 | +1.196 | +3.3 |
| Qwen3-1.7B s1:dup12_seed1/endpoint | +2.301 | +0.132 | +1.327 | +6.7 |
| Qwen3-1.7B s1:dupq8_seed1/breach | +2.736 | +0.126 | +1.720 | +11.6 |
| Qwen3-1.7B s1:dupq8_seed1/endpoint | +3.515 | +0.125 | +1.991 | +18.2 |
| Qwen3-1.7B s1:high_lr_seed1/breach | +2.593 | +0.115 | +1.948 | +19.8 |
| Qwen3-1.7B s1:high_lr_seed1/endpoint | +2.603 | +0.171 | +2.968 | +25.3 |
| Qwen3-1.7B s2:fork | +2.261 | +0.141 | +1.158 | – |
| Qwen3-1.7B s2:healthy_seed2/endpoint | +2.215 | +0.138 | +1.186 | +3.3 |
| Qwen3-1.7B s2:dup12_seed2/endpoint | +2.313 | +0.133 | +1.334 | +6.8 |
| Qwen3-1.7B s2:dupq8_seed2/breach | +2.580 | +0.128 | +1.620 | +10.8 |
| Qwen3-1.7B s2:dupq8_seed2/endpoint | +3.015 | +0.125 | +1.840 | +14.2 |
| Qwen3-1.7B s2:high_lr_seed2/breach | +2.562 | +0.117 | +1.811 | +18.4 |
| Qwen3-1.7B s2:high_lr_seed2/endpoint | +2.689 | +0.184 | +3.163 | +25.2 |
| OLMo-2-0425-1B s0:fork | +2.221 | +0.124 | +2.152 | – |
| OLMo-2-0425-1B s0:healthy_seed0/endpoint | +2.181 | +0.124 | +2.163 | +2.4 |
| OLMo-2-0425-1B s0:dup12_seed0/endpoint | +2.201 | +0.126 | +2.248 | +4.7 |
| OLMo-2-0425-1B s0:dupq8_seed0/breach | +2.637 | +0.137 | +2.640 | +17.3 |
| OLMo-2-0425-1B s0:dupq8_seed0/endpoint | +3.112 | +0.145 | +2.792 | +24.2 |
| OLMo-2-0425-1B s0:high_lr_seed0/breach | +2.530 | +0.096 | +1.846 | +17.1 |
| OLMo-2-0425-1B s0:high_lr_seed0/endpoint | +2.635 | +0.113 | +1.817 | +21.8 |
| OLMo-2-0425-1B s1:fork | +2.219 | +0.124 | +2.152 | – |
| OLMo-2-0425-1B s1:healthy_seed1/endpoint | +2.172 | +0.123 | +2.152 | +2.8 |
| OLMo-2-0425-1B s1:dup12_seed1/endpoint | +2.193 | +0.127 | +2.234 | +5.2 |
| OLMo-2-0425-1B s1:dupq8_seed1/breach | +2.581 | +0.138 | +2.526 | +14.0 |
| OLMo-2-0425-1B s1:dupq8_seed1/endpoint | +3.012 | +0.144 | +2.668 | +20.7 |
| OLMo-2-0425-1B s1:high_lr_seed1/breach | +2.511 | +0.099 | +2.000 | +17.1 |
| OLMo-2-0425-1B s1:high_lr_seed1/endpoint | +2.388 | +0.114 | +1.886 | +21.7 |
| OLMo-2-0425-1B s2:fork | +2.217 | +0.123 | +2.146 | – |
| OLMo-2-0425-1B s2:healthy_seed2/endpoint | +2.173 | +0.123 | +2.153 | +2.6 |
| OLMo-2-0425-1B s2:dup12_seed2/endpoint | +2.205 | +0.127 | +2.247 | +6.2 |
| OLMo-2-0425-1B s2:dupq8_seed2/breach | +2.535 | +0.137 | +2.492 | +14.2 |
| OLMo-2-0425-1B s2:dupq8_seed2/endpoint | +2.676 | +0.139 | +2.545 | +15.9 |
| OLMo-2-0425-1B s2:high_lr_seed2/breach | +2.509 | +0.096 | +1.812 | +18.0 |
| OLMo-2-0425-1B s2:high_lr_seed2/endpoint | +2.474 | +0.112 | +1.904 | +20.3 |

## B PROTOCOL DETAILS AND FROZEN DECISIONS

**Models.** Qwen3-0.6B (28 layers, $d$=1024), Qwen3-1.7B, and OLMo-2-0425-1B (base, 16 layers, $d$=2048); every run's manifest records the model revision, config SHA-256, and tokenizer-vocabulary SHA-256, and the resolved layer indices.

**Training and probe (identical across all campaigns).** Full fine-tuning, bf16, AdamW with weight decay 0, gradient clip 1.0, batch $4 \times 512$ tokens. Prefix: $\tau = 300$ healthy steps (warmup 20, measured every 25), with a weights+optimizer checkpoint and a SHA-256 `fork.json` at step 300; every branch verifies its parent hash against this file. Branches: 600 steps from the fork, warmup 0, measured every 10 (61 measurements); the data-order RNG is seed+1000, shared across all branches of a seed so that regimes differ only in the intervention. Probe: a fixed 256-sequence instruction-tail slice (deterministic seed-0 shuffle; historical nonzero-seed training overlap is disclosed in Section 2), max length 256, tokenized by each model's own tokenizer, padding masked to $-100$ in the loss. Hidden states are taken at layers `-1` and `mid` (resolving to $[-1, 8]$ on OLMo-2-1B and $[-1, 14]$ on Qwen3); the per-token RMS normalization applied before covariance estimation is an explicit, model-independent diagnostic choice, never auto-selected from the architecture. Both centred and raw (uncentred) spectra are logged at every measurement.

**Frozen selection gates (loss-only; spectra never feed any gate).** Base learning rate per model is chosen by a frozen ladder $\{10^{-5}, 3{\times}10^{-6}, 3{\times}10^{-5}\}$: the first rung whose healthy prefix satisfies (a) endpoint probe NLL $\leq$ step-0 value and (b) no measurement exceeding step-0 $+0.10$ nats is frozen for every branch of that model (both OLMo-2-1B and Qwen3-1.7B passed at $10^{-5}$). Failed prefixes are kept, renamed, never overwritten. The duplication ladder uses nested pools ($n_u \in \{2400, 1200, 600, 300\}$, verified $12.5\% \subset 25\% \subset 50\% \subset 100\%$). High-LR escalates $15\times \rightarrow 25\times \rightarrow 40\times$ until the first breaching rung ($15\times$ breached on all models). The extreme-duplication rung ($n_u$=8) was added *after* the OLMo seed-0 pilot as a loss-only severity escalation matched to the Qwen3-0.6B extreme-duplication condition; the protocol change-log entry predates every confirmation seed reported here (six fresh model–seeds).

**Damage gate and checkpoints.** Damage = probe NLL $\geq$ branch step-0 value $+0.30$ nats; each branch saves a weights-only checkpoint at the first crossing (`breach`) and at step 900 (`endpoint`). The gate is a severity definition, pre-registered before any monitored comparison, and is never used as a detector.

**Gradual-onset and replicate design.** Frozen before implementation: the LR ramp multiplies the base learning rate by $1 + (15 - 1)\, t/600$ at branch step $t$; the duplication ramp draws each training batch from the frozen $n_u$=8 pool with probability $p(t)$=$t/600$ (mixture RNG seed+2000), else from the full corpus; both on the development model, seeds 0–2, from the same forks, with breach/endpoint checkpoints, in a separate root that never enters calibration. A ramp that fails to breach by step 600 is reported censored; no post-hoc escalation. Alongside them, checkpointed configuration-level replicates of the development model's original damage branches were run (the originals saved no checkpoints), with the disclosure frozen in advance that capability outcomes link to alarms at the configuration level. Observed replicate fidelity is high: endpoint $\Delta$NLL matches the originals to $\pm 0.001$ nats on extreme duplication, and the replicates reproduce the original zero-lead alarm pattern (all monitors at steps 310–320) and 3/3-silent healthy runs under the frozen thresholds.

**Benign control design.** Frozen before any benign run: mode domain-shift, lr-mult 2 (mid rung of the frozen $\{1, 2, 4, 8\}$ ladder; rung 8 is the damaging regime of Section 5), seeds 0–2 forked from the same prefixes, 600 steps measured every 10, written to a separate root so it can never enter calibration or damage-branch enumeration. Benign criteria (both required): endpoint GSM8K probe NLL $\leq$ fork $-0.05$; source probe NLL $<$ fork $+0.30$ at every measurement. Scoring: the frozen $\alpha$=0.05 thresholds and scale splits, applied once. Both possible outcomes (alarms mean change-not-damage detection; silence extends FA specificity) were assigned their reading in the protocol log before launch. In the historical campaign benignity was certified by the two probe NLLs only (no checkpoints were saved); the clean campaign saved endpoint checkpoints, and the capability battery confirms the benign label at the task level (Table 5).

**Sequential-detector design.** Frozen before implementation and before computing any detector statistic on any trajectory: both detectors read the probe-NLL delta-from-fork stream $x_t$ ($t = 0..60$, every 10 steps), the same single channel as the NLL families, in raw nats. CUSUM: causal EMA

baseline $b_0 = x_0$, $b_t = 0.7\, b_{t-1} + 0.3\, x_{t-1}$; $C_t = \max(0,\, C_{t-1} + x_t - b_t - \kappa)$. Page–Hinkley: $m_t$ the running mean of $x_{0..t}$, $U_t = \sum_{i \le t}(x_i - m_i - \kappa)$, $\mathrm{PH}_t = U_t - \min_{i \le t} U_i$ (the same slack $\kappa$ serves as the Page–Hinkley tolerance). Slack $\kappa = 1.5\hat{\sigma}$, with $\hat{\sigma}$ the median over the three scale seeds of the standard deviation of first differences of the healthy stream ($4.4 \times 10^{-4}$ nats historical, $5.1 \times 10^{-4}$ clean). Thresholds: trajectory score $= \max_t$ of the statistic, split conformal on the unchanged calibration seeds ($n_{\mathrm{cal}}$=19, $k$=19, $\alpha = 0.05$)—the identical recipe that froze the NLL families; healthy calibration trajectories generate zero positive detector mass at this slack, so both thresholds are exactly $0.00$. Exactly two configurations were run (no parameter sweeps); the erasure criterion and both interpretations were assigned in the protocol log before scoring. The matched-threshold control (pre-registered 2026 protocol entry) scores the one-sided statistic $\max(x_t, 0)/s$ against the frozen two-sided threshold; its false-alarm rate is bounded by the two-sided detector's conformal-valid rate because the one-sided score is pointwise smaller.

**Conformal design.** Clean campaign: 32 fresh fixed-probe-v2 trajectories, scale $= \{32, 33, 34\}$, calibration $= \{35, \ldots, 53\}$, test $= \{54, \ldots, 63\}$, thresholds at new output paths. Historical campaign: seeds 0–31 with scale $= \{0, 1, 2\}$, calibration $= \{3, \ldots, 21\}$, test $= \{22, \ldots, 31\}$; overlap and missing seeds are hard errors in both. The NLL-detector direction (one-sided upward primary, two-sided secondary) was frozen on 2026-09-15, before any test-split trajectory existed. The test split was initially scored with the final frozen thresholds; the post-hoc sensitivity analysis in Appendix C reuses it.

**Capability evaluation.** Fixed 500-item subsamples per task (deterministic seed-0 shuffle), identical items at every checkpoint. Multiple choice: length-normalized log-likelihood scoring with the standard harness conventions; LAMBADA: teacher-forced greedy exact match over the target word's tokens plus target NLL. Paired bootstrap: $B = 10^4$ resamples of item indices shared between fork and checkpoint, fixed RNG seed.

**Identifiability certificates.** Four numerical checks accompany Proposition 1: (1) raw/centred spectrum identity between $Z$ and $ZQ^\top$ (and singular values of the data matrices); (2) mean loss gap exactly $a$ with accuracy $1 \to 0$; (3) joint-rotation invariance of every logit; (4) boundedness $|w^\top h| \le \|w\|\|h\|$ with per-sample loss gap $\le 2a$ at fixed scale.

**Artifact audit.** The pre-submission audit re-derives, from raw logs: branch completeness per model×seed; first/last-NLL and measurement-count consistency between logs and result records; parent-hash lineage of every branch; recomputed breach steps; healthy-pool completeness and measurement-grid integrity; conformal split disjointness, threshold-order recomputation, and held-out coverage; capability coverage of every on-disk checkpoint; and table freshness. It exits nonzero on any failure and passed with zero failures and zero warnings at freeze, with a SHA-256 manifest over all analysis inputs and generated tables.

## C POST-REVIEW DATA AUDIT AND SUPPLEMENTARY ANALYSES

These analyses were specified after inspection of the original results. They are exploratory and reuse historical trajectories; they do not constitute an untouched-test replication. Their complete machine-readable records, including failed training smoke attempts, are retained separately from the original freeze.

**Training/probe separation.** The probe is the last 256 rows of the seed-0 Dolly shuffle. The original training loader excluded the last 256 rows of each training seed's shuffle instead. Across seeds 1–31, 248–254 probe rows remained eligible for training; reconstruction of 300 prefix and 600 healthy-branch updates (batch 4) found 46–72 distinct probe rows actually drawn. Seed 0 was row-disjoint. The corrected implementation excludes every exact formatted probe-text hash from all training pools, including duplicate copies; checks across 32 seeds pass. Branches must carry a matching versioned probe manifest from a newly trained prefix. Exact-text exclusion does not assert absence of paraphrases or pretraining exposure. Different duplication pools also change content and sequence-length distributions; equal update counts do not isolate a causal effect of diversity. Token-matched and length-stratified pool controls remain necessary for that claim.

**Exposure-filtered endpoint evaluation.** We reconstructed the union of training draws for the three development seeds and conservatively excluded the first 300 duplication-pool texts, along with exact matches in GSM8K training data. Of the original 256 probe sequences, 159 survived; the first 64 in fixed order were evaluated at the three forks and 18 endpoints, with length 256 and microbatch 1. Table 17 reports token-weighted NLL changes. Paired sequence bootstrap intervals (20,000 resamples) have lower bounds above $+0.30$ for duplication, high-LR, domain shift, and LR ramp in all three seeds; duplication-ramp upper bounds remain below $+0.30$. This supports endpoint loss degradation on a post-training-unexposed subset. Retrospective subset selection and the absence of clean intermediate histories prevent transferring historical alarm calibration. Different batch aggregation also prevents interpreting old/new NLL changes as the causal effect of probe exposure.

| Regime | Seed 0 | Seed 1 | Seed 2 |
|---|---|---|---|
| domain_shift_v2 | **+0.3851** | **+0.3568** | **+0.3668** |
| duplicate_data | **+1.6583** | **+2.0852** | **+1.4152** |
| healthy | -0.0392 | -0.0380 | -0.0376 |
| high_lr | **+0.7512** | **+0.7869** | **+0.7608** |
| dup_ramp | +0.0540 | +0.0658 | +0.0639 |
| lr_ramp | **+0.6616** | **+0.6033** | **+0.5912** |

Table 17: Exposure-filtered endpoint minus fork NLL (nats/token), 64 sequences. These are new inference measurements on historical checkpoints. Bold: at or beyond the frozen $+0.30$-nat damage gate.

**Independent capability endpoints.** We applied a post-hoc rule to the existing same-run task checkpoints: a decrease of at least 5 percentage points on HellaSwag normalized accuracy, ARC-C normalized accuracy, or LAMBADA exact match, together with a one-sided paired exact binomial test and Holm correction across the three tasks within each run ($0.05$). Ten of 18 endpoints meet this rule: healthy 0/3, extreme duplication 2/3, high-LR 3/3, domain shift 2/3, LR ramp 3/3, duplication ramp 0/3. The spectral union, both NLL scores, and the specified sequential loss baselines detect all ten, but also alarm on five of the eight endpoints without a confirmed event. The latter are not proven false positives: three tasks cannot certify absence of damage. Threshold sensitivities of 2 and 10 points yield 11/18 and 7/18 events. Item pairing assumes the evaluator's fixed item order; legacy bitstrings contain no item IDs. Sparse checkpoints do not identify the first capability-loss time or establish capability early warning.

**Layer, channel, and time-resolution ablations.** We evaluated 117 cells: three layer choices (middle/final/both), thirteen channel choices (ten individual original channels, stable rank, spectral flatness, and the original spectral union), and observation spacings 10/20/50 steps. Each cell refits its

scales and conformal threshold using the same 3/19/10 trajectory split and horizon 600 at $\alpha = .05$. This absorbs multiplicity within a cell, not selection across 117 cells; no best cell is selected. Table 18 gives specified 10-step slices; the artifact contains every cell and alarm time.

| Layers | Channel | Healthy /10 | Abrupt /9 | Dup-ramp /3 | Benign /3 |
|---|---|---|---|---|---|
| Final | all | 0 | 9 | 3 | 3 |
| Middle | all | 0 | 9 | 3 | 3 |
| Both | rankme_raw_id | 0 | 9 | 3 | 3 |
| Both | rankme_raw_xf | 0 | 9 | 3 | 3 |
| Both | rankme_centred | 0 | 9 | 3 | 3 |
| Both | eff_rank | 1 | 9 | 3 | 3 |
| Both | stable_rank | 0 | 9 | 3 | 3 |
| Both | spectral_flatness | 0 | 9 | 3 | 3 |
| Both | all | 0 | 9 | 3 | 3 |

Table 18: Historical per-cell alarm counts. Abrupt denotes the nine original development-model stress branches. Reused test trajectories and probe exposure limit interpretation.

The union's LR-ramp alarms at the final layer are 410/460/410, at the middle layer 400/410/420, and at both layers 400/410/410 (absolute steps; fork 300). Layer union is therefore not uniformly earlier after separate calibration. Across all cells, healthy alarm counts range from 0 to 1 out of 10. Stable rank and flatness at both layers alarm on all nine abrupt branches and all three benign adaptations: adding these scalars does not fix damage specificity in the replay. The additional CUSUM/PH sensitivity grid spans four false-alarm levels, three EMA coefficients, and four slack multipliers. CUSUM coefficient $0.5$ with slack multiplier $1.5$ alarms on LR ramp at 420/440/420, leaving 20/30/10 steps relative to the spectral union, rather than a universal 20–40-step residual. These are descriptive operating points.

**Additional diagnostic definitions and coverage.** For stored raw second-moment eigenvalues $e_i \geq 0$, the new metrics are stable rank $\sum_i e_i / \max_i e_i$ and spectral flatness $\mathrm{GM}(\sqrt{e_i}) / \mathrm{AM}(\sqrt{e_i})$, with $10^{-12}$ numerical regularization as in the inspected TRSP code. Our pooled-probe aggregation differs from its per-prompt evaluation; this is a metric adaptation. Its propagation metric needs paired noisy forward passes and cannot be recovered from spectra. Likewise, Collapse Index requires topological activation information and Watch the Weights requires weight directions plus activation projections. Neither is reproduced by the stored-spectrum replay. Independent GPU inference on 16 probe sequences (length 128) additionally measured full-vocabulary KL and exact relative parameter L2 drift at saved checkpoints. Without benign checkpoint coverage and calibrated sequential trajectories, these endpoint diagnostics establish neither a competitive monitoring baseline nor a damage decision rule.

**SIGMA adaptation and numerical diagnostic values.** The replay fixes $\delta = 10^{-4} d^{-1} \sum_i \lambda_i(G_0)$ separately per branch and layer, with $G_0 = Z_0^\top Z_0$, and uses $\hat{\beta} = 0$. For fixed probe token count $n$ and dimension $d$,

$$G_{\mathrm{KF}} = \sum_i \log(\lambda_i + \delta) - d \log \delta, \qquad U_{\mathrm{LLN}} = \sum_i \log(\lambda_i + \delta) - d \log n.$$

Their difference is the time-independent constant $d \log(n/\delta)$; their fork-relative drifts therefore coincide. The artifact verifies this identity numerically for all endpoints and intermediate observations used in Table 19 and records every ridge value. This redundancy is specific to our adaptation, not evidence that the two native SIGMA tracks are always interchangeable.

**Fresh text-disjoint paired confirmation.** We rebuilt the Qwen3-0.6B seed-1 prefix from pretrained weights using the fixed text-disjoint split: 300 prefix updates followed by two 600-update branches, healthy and extreme duplication. Both branches load the same weights and optimizer state, verified by checkpoint hash; their complete fork measurements are numerically identical even though the branches ran on separate RTX 3090 GPUs. Effective batch size is four, accumulated as four token-weighted microbatches, with training length 512. The fixed 256-sequence probe uses length 256 and batch size two. There are 61 measurements per branch at ten-step intervals.

| Regime | Layer | $\Delta(\Delta H)$ | $\Delta$stable rank | $\Delta$flatness |
|---|---|---|---|---|
| healthy | Final | $+2.586 \pm 0.822$ | $+0.051 \pm 0.021$ | $+0.001 \pm 0.000$ |
| healthy | Middle | $+2.414 \pm 1.594$ | $+0.000 \pm 0.000$ | $+0.004 \pm 0.001$ |
| domain_shift_v2 | Final | $-13.106 \pm 2.971$ | $+0.372 \pm 0.065$ | $+0.008 \pm 0.003$ |
| domain_shift_v2 | Middle | $-7.825 \pm 5.369$ | $+0.008 \pm 0.001$ | $+0.076 \pm 0.009$ |
| duplicate_data | Final | $-179.941 \pm 13.175$ | $+2.209 \pm 0.319$ | $+0.052 \pm 0.004$ |
| duplicate_data | Middle | $-85.676 \pm 10.414$ | $+0.000 \pm 0.000$ | $+0.035 \pm 0.004$ |
| high_lr | Final | $+10.499 \pm 17.322$ | $+0.569 \pm 0.161$ | $-0.007 \pm 0.005$ |
| high_lr | Middle | $-40.978 \pm 7.043$ | $+0.017 \pm 0.005$ | $+0.127 \pm 0.017$ |
| dup_ramp | Final | $-14.846 \pm 0.445$ | $+0.085 \pm 0.018$ | $+0.007 \pm 0.000$ |
| dup_ramp | Middle | $-42.919 \pm 12.091$ | $+0.000 \pm 0.000$ | $+0.018 \pm 0.004$ |
| lr_ramp | Final | $+13.282 \pm 34.771$ | $+0.479 \pm 0.293$ | $-0.002 \pm 0.011$ |
| lr_ramp | Middle | $-31.478 \pm 3.989$ | $+0.009 \pm 0.001$ | $+0.090 \pm 0.005$ |
| benign_ds2m2 | Final | $-20.080 \pm 2.499$ | $+0.175 \pm 0.010$ | $+0.009 \pm 0.001$ |
| benign_ds2m2 | Middle | $+5.205 \pm 3.119$ | $+0.001 \pm 0.000$ | $+0.015 \pm 0.002$ |

Table 19: Historical endpoint-minus-fork diagnostic changes, mean and sample standard deviation across three seeds. Delta-H uses transformed centred covariance; stable rank and flatness use raw uncentred spectra.

| Regime | $\Delta$NLL | Breach | $\Delta$RankMe final | middle |
|---|---|---|---|---|
| healthy | -0.0452 | – | +2.28 | +4.51 |
| duplicate_data | **+2.0488** | 340 | +102.26 | +17.97 |

Table 20: New fixed-probe-v2 confirmation, seed 1. Breach is the first absolute step satisfying the stated $+0.30$ NLL rule; "–" is no breach within the horizon. No historical conformal threshold is transferred. Bold: at or beyond the $+0.30$-nat gate.

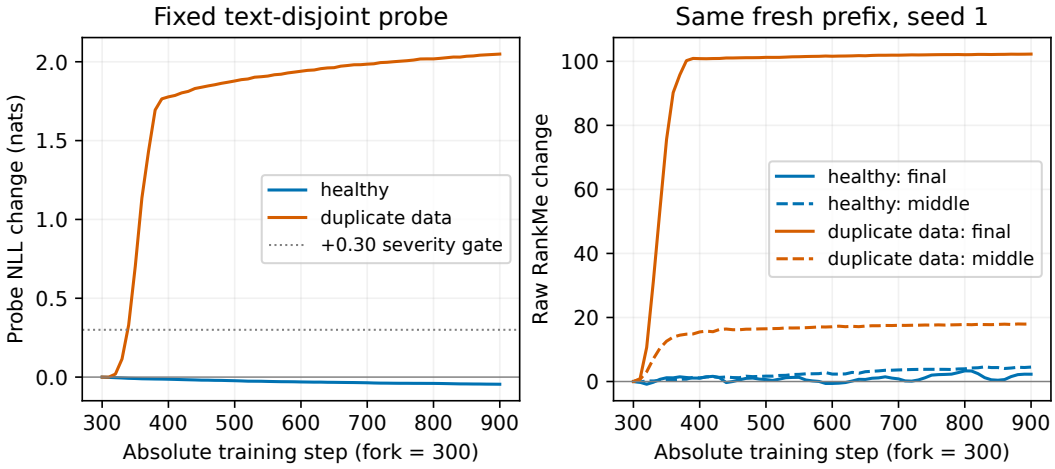



Figure 3: Complete clean paired trajectories from the same fresh prefix. The horizontal dotted line is a severity reference, not a calibrated alarm. These curves are one seed, with no between-seed uncertainty claim.

The healthy branch changes probe NLL by $-0.0452$, whereas extreme duplication changes it by $+2.0488$ (Table 20; full paired trajectories in Figure 3). Raw RankMe increases at both monitored layers under duplication despite this loss increase. This reproduces the direction mismatch in a fresh, text-disjoint paired run. The healthy and duplication branches process 374,603 and 537,300 non-padding target tokens respectively; equal update counts do not imply equal token budgets. The result does not replace clean multi-seed calibration/test trajectories or cross-model replication. It also does not estimate the causal effect of removing probe exposure: data order, floating-point accumulation, and probe batch aggregation differ from the historical campaign. Independent task endpoints were not evaluated for these newly trained branches.

**Clean-protocol recalibration and replay (fixed-probe-v2, 32 fresh seeds).** After the analyses above, the full monitoring design of Section 6 was re-executed on a fresh campaign trained with the corrected split: 32 new prefixes and healthy branches (seeds 32–63; every training pool excludes the exact formatted probe texts, verified per run with `overlap_count`=0), plus the duplication, high-LR, domain-shift, LR-ramp, duplication-ramp, and benign branches on seeds 32–34. The split (scale 32–34, calibration 35–53, test 54–63), the six families and their directions, the sequential-detector parameters, the $+0.30$-nat gate, and both primary readouts were frozen in the protocol log before the queue finished; nothing was tuned on the new data, and no historical threshold was reused. Table 21 reports the two protocols side by side. The structure of the historical result replicates: zero held-out false alarms for the spectral union, both NLL variants, and both sequential detectors; a degenerate $0.00$ one-sided NLL threshold; the same alarm ordering on damaging LR ramps (earliest spectral 390–410 $\rightarrow$ CUSUM 440–460 $\rightarrow$ one-sided NLL 490–500 $\rightarrow$ gate 750–780), with the spectral-before-NLL readout positive on 3/3 and the sequential-erasure readout again negative on 3/3; the same early spectral alarms on a duplication ramp that never damages the model; and a benign control (3/3 criteria met) on which every monitor except the two-sided NLL variant alarms. Differences are disclosed: two individual families (entropy/Sigma and cross-batch similarity) each record one false alarm on the ten clean test trajectories (the union records zero), the two-sided NLL variant misses one of three duplication-ramp branches, and all thresholds differ numerically from the historical ones, which remain non-transferable. This section's replication covers the development model and this recipe; the cross-model clean sign replication and the clean capability battery were subsequently completed and are reported in the main text (Tables 1 and 5).

Table 21: Historical (probe-exposed) protocol vs. clean fixed-probe-v2 replication, side by side (development model; ranges over 3 seeds; "–" = censored). Script-generated from the frozen scored JSONs of both campaigns.

| quantity | historical (exposed probe) | clean (fixed-probe-v2) |
|---|---|---|
| held-out healthy FA, all-spectral | 0/10 | 0/10 |
| held-out healthy FA, NLL one-sided | 0/10 | 0/10 |
| held-out healthy FA, worst single family | 0/10 | 1/10 |
| held-out healthy FA, CUSUM | 0/10 | 0/10 |
| held-out healthy FA, Page–Hinkley | 0/10 | 0/10 |
| NLL one-sided threshold | 0.00 | 0.00 |
| LR ramp: damage gate crossed | 3/3 (730–760) | 3/3 (750–780) |
| LR ramp: earliest spectral alarm | 400–410 | 390–410 |
| LR ramp: CUSUM alarm | 430–450 | 440–460 |
| LR ramp: NLL one-sided alarm | 470–520 | 490–500 |
| spectral alarms $\geq$10 steps before NLL | 3/3 | 3/3 |
| sequential alarm within 10 steps of spectral | 0/3 | 0/3 |
| dup. ramp: endpoint $\Delta$NLL | +0.05..+0.06 | +0.03..+0.11 |
| dup. ramp: earliest spectral alarm | 400–450 | 420–440 |
| benign control: criteria met | 3/3 | 3/3 |
| benign: all-spectral alarms | 3/3 | 3/3 |
| benign: NLL two-sided alarms | 0/3 | 0/3 |

**Clean capability endpoints.** The pre-registered task battery (fixed 500-item subsamples; paired bootstrap over shared items, $B{=}10^4$) was run once at the fork, saved breach, and endpoint checkpoints of all 21 clean branches (Table 5 (main text)). Every historical dissociation reappears under the repaired split: healthy endpoints are flat; high-LR collapses broadly by the endpoint (HellaSwag $-13.4 \pm 1.8$, LAMBADA $-26.9 \pm 4.5$ pt, 3/3 significant) while its breach checkpoint, one to two

measurements after the fork, is only mildly degraded; the LR ramp is already significantly degraded at its damage-gate crossing (HellaSwag $-5.1 \pm 1.1$, LAMBADA $-8.5 \pm 2.5$ pt, 3/3), so the clean spectral lead of Table 21 precedes genuine incoming capability damage; extreme duplication raises probe NLL by $+1.34 \pm 0.53$ nats with multiple-choice accuracy flat even at the endpoint (only LAMBADA declines, $-5.0 \pm 1.6$ pt)—the sharpest NLL/accuracy dissociation; and the domain shift is a genuine trade (GSM8K NLL $-0.31$ at the source tasks' expense). The two cells frozen with both-way readings resolve on the specificity side: the benign control's task metrics are flat to slightly positive (no significant decline in any of the nine task–seed cells) while its GSM8K target improves by $-0.38$ nats—so the spectral, one-sided-NLL, CUSUM, and Page–Hinkley alarms on these runs are false alarms with respect to task capability, not only with respect to the loss criteria—and the duplication ramp's endpoints are flat, matching its no-damage reading. This battery is descriptive, feeds no gate, and covers the development model only.

**Cross-model clean replication.** Finally, the cross-model sign-structure claims were re-executed under the repaired split (pre-registered targets frozen before the campaign was generated): fresh 300-step prefixes and healthy / extreme-duplication / high-LR branches for Qwen3-1.7B and OLMo-2-1B, seeds 32–34, base LR carried over from the historical frozen ladder gates. All three targets replicate exactly (Table 22). (a) The duplication inversion holds 3/3 on both models (probe $\Delta$NLL $+0.31$ to $+1.08$ nats with raw last-layer RankMe rising $+139$ to $+225$), which together with the development model's clean branches makes the inversion 9/9 model–seeds under the repaired protocol, matching the historical 9/9. (b) The high-LR sign flip reproduces seed-exactly: last-layer raw RankMe falls in all three OLMo seeds ($-36$ to $-52$) and rises in all three Qwen3-1.7B seeds ($+113$ to $+141$); OLMo's *middle* layer rises under high-LR ($+26$ to $+31$), so the flip remains a last-layer phenomenon, as in the historical data. (c) All six healthy branches stay below the damage gate with endpoint $\Delta\text{NLL} \approx -0.05$. The frozen capability battery on the new checkpoints reproduces the cross-model capability patterns as well: under extreme duplication, Qwen3-1.7B HellaSwag is significantly *above* fork at the breach checkpoint (3/3, $+2.7 \pm 0.5$ pt) and multiple-choice accuracy stays flat-to-positive at both models' endpoints despite $+0.31$ to $+1.11$ nats of probe-NLL damage, while high-LR degrades broadly by the endpoint (Qwen3-1.7B HellaSwag $-11.5 \pm 0.7$, LAMBADA $-27.1 \pm 5.5$; OLMo $-9.3 \pm 1.6$, $-18.5 \pm 2.7$ pt). No conformal claim is made for these models (no healthy pool at scale); the development-model thresholds were never applied to them.

Table 22: Cross-model sign structure, historical (probe-exposed) vs. clean fixed-probe-v2 campaigns (endpoint probe $\Delta$NLL range over 3 seeds; signs = per-seed sign of the endpoint raw last-layer RankMe delta). Script-generated; historical values read from the frozen analysis JSONs. Bold $\Delta$NLL ranges: every seed at or beyond the $+0.30$-nat gate.

| model | regime | historical (exposed) | | clean (fixed-probe-v2) | |
|---|---|---|---|---|---|
| | | $\Delta$NLL | $\Delta$RankMe signs | $\Delta$NLL | $\Delta$RankMe signs |
| Qwen3-1.7B | healthy | -0.05..-0.04 | +++ | -0.05 | +++ |
| Qwen3-1.7B | duplication (extreme) | **+0.74..+1.26** | +++ | **+0.44..+1.08** | +++ |
| Qwen3-1.7B | high-LR | **+0.36..+0.56** | +++ | **+0.49..+0.51** | +++ |
| OLMo-2-1B | healthy | -0.04 | +++ | -0.05 | +++ |
| OLMo-2-1B | duplication (extreme) | **+0.46..+0.88** | +++ | **+0.31..+0.57** | +++ |
| OLMo-2-1B | high-LR | +0.19..+0.42 | −−− | **+0.33..+0.38** | −−− |