DETECT LLM FINE-TUNING OVERFITTING EARLY

Anonymous authors
Paper under double-blind review

ABSTRACT

Four epochs out of fifty decide whether a fine-tuned model overfits, and four epochs are enough to catch it. Across twenty-two matched-pair comparisons spanning 1.5B to 32B parameters, an early check at epoch 4 calls the final test-set gap for every readable run. The remaining pair settles by epoch 10. Both thresholds were locked in advance. The results uncover a clear pattern: the reasoning-distilled checkpoints tested here overfit downstream tasks more than their standard instruction-tuned twins. Across four base models they show 1.20 to $4.60\times$ larger drift and 1.09 to $1.84\times$ wider final test gaps. In contrast, vendor reasoning post-training, domain adaptation, and reasoning SFT placed on top of an instruction model all fine-tune like the twin. What the separating checkpoints share is a raw base. Pairing models from the same base keeps the comparison clean by holding architecture, vocabulary and size constant. Using only raw training and validation loss curves, our metric splits the gap into two plain numbers. Roughness a is how far the training losses sit from a smooth descent, and drift c is how far the validation curve pulls away. Their ratio c/a pinpoints what kind of overfitting is happening. On an unseen 32B pair run across two seeds, the rule flags heavy overfitting at epoch 4, correctly anticipating final drifts over $4.4\times$. The rule is right on all twenty runs read at every horizon, at epoch 4 and at each later horizon the run reaches. No single-epoch reading of the same curves is right at both.

1 INTRODUCTION

Adapting a large language model to a new task now starts from a released checkpoint. Several are often built from one base model: an instruction release, a reasoning release, a domain release. They differ in how they were finished, not in architecture or size.

Suppose you hold a thousand labeled examples and two such candidates. From outside they are interchangeable, and the post-training that separates them is written in nothing you can inspect. Fine-tune the wrong one and the annotation budget is gone, hardest where labels are most expensive. *Which of two released checkpoints will overfit my data sooner, and can I tell before I pay for the run?*

Overfitting is normally legible only once a run is over. A small fine-tune drives the training loss to zero, numerically zero at the final epoch in every run here. The only question left is how much of what it kept transfers. The obvious quantity does not travel either: the train-validation gap carries the units of whichever checkpoint produced it, so 0.4 from one model and 0.6 from another cannot be compared. Weight-space measures avoid the unit problem but need the weights and extra passes, which a checkpoint chosen from a release page may not afford (Appendix G).

This paper takes a different route. Ratios of two comparable runs cancel the unit, so we build a *matched pair* (Section 3): the checkpoint is the only variable, and a ratio between its two members is unit-free. On that design we read an index from the loss log alone and ask how early the reading is fixed. The answer is four epochs of fifty.

Contributions. *(1) A four-epoch decision rule.* On all twenty-two matched-pair runs the early reading is the verdict the independent test gap delivers at the end, under two thresholds fixed in advance. It holds at epoch 4 on the twenty readable there and from epoch 10 on all of them (Section 5). On the twenty runs read at every horizon it is right on all twenty at epoch 4 and at every horizon from 10, each pair at its common horizon. No single-epoch reading of the same curves does both (Table A19). *(2) Theory for what the reading measures.* Theorem 1 gives the index in closed form. Proposition 1 pins the index between two increasing functions of c/a, with slack below 0.004 on

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.

Table 1: This paper against the closest work on predicting a fine-tune before it finishes. *Completed runs needed*: finished fine-tunes required before use; *new base model*: what must be redone for an unseen base.

	Zeng et al. (2025)	Luo et al. (2026)	This paper
What it predicts	which model scores higher	the final score	how much is memorized
Completed runs needed	a curve per candidate	1,300+	**none**
A new base model	refit the curve	a new meta-dataset	**the same two thresholds**
Controlled comparison	—	—	**ten matched pairs**
Readable at	after the fit	after a probe	**epoch** 4 **of** 50

every run here. Theorem 2 characterizes ratios of RMS gaps by a mediant bound; the fourth-epoch agreement below is an empirical finding under the evaluated protocol. Theorem 3, proved under squared loss and a linearized adapter, bounds the held-out excess a fine-tune of fixed capacity must pay. *(3) A 49-run matched-pair benchmark.* Twenty checkpoints, nine base models, four families (1.5B–32B) and four tasks across mathematics, finance and healthcare, with two seeds on all ten pairs. Per-epoch training, validation and test losses are released for every run. *(4) A post-training finding, tested out of sample.* Reasoning-distilled checkpoints overfit a later fine-tune more than the instruction-tuned twins their vendors released from the same base. The margin is 1.20 to $4.60\times$ the drift and 1.09 to $1.84\times$ the test gap on four base models from 8B to 32B. Vendor reasoning post-training and domain adaptation fine-tune like the twin. Six pair runs on three otherwise unused base models, fine-tuned after the thresholds were fixed, all fall where the rule predicts.

2 Related Work

The closest line predicts a fine-tune before it finishes. Zeng et al. (2025) fit an NTK-based scaling law to select models. It achieves up to 91.1% accuracy for 88.5% less compute, focusing on task performance rather than overfitting. Luo et al. (2026) predict a fine-tune's final score to an RMSE of 1.47 points on 370 held-out runs. Their setup requires over 1,300 completed runs trained within each target setting. Both lines predict how high a run will score (Table 1). This paper predicts how much train-held-out separation the post-training carries in. We use two fixed thresholds across nine base models and four families. Existing scalar measures quantify overfitting only on finished runs (Aburass, 2023; Rezaei et al., 2023; Fan et al., 2025). The pointwise index $\mathrm{ORI}(t)$ is taken from Fan et al. (2025). New here are its exact decomposition (Theorem 1), the prefix reading ORI-AUC over the retained segment, the matched-pair protocol, and the two-coordinate rule read at the fourth epoch. Carlsson et al. (2025) fix the scope of that reading. Driving a pre-trained model to near-zero training loss worsens validation loss, yet improves open-ended generation. A widening separation is thus adverse for exact-match fine-tuning, the setting here, rather than universally. Two further directions locate overfitting inside the network. Table A5 (Appendix G) scores them on this paper's axis: how many pair runs each calls correctly from the first four epochs.

3 Problem Setup

The challenge. Two released checkpoints, one dataset: which will overfit sooner during fine-tuning? A *matched pair* makes that challenge controlled. Checkpoints A and B share a base model, a tokenizer and a parameter count, differing only in post-training. Each pair sets a reasoning or domain release against the instruction-tuned twin from the same base (Table A16, Appendix Q). Both see identical data, learning rate, epoch budget and seed under the protocol below, so any difference between their loss trajectories is attributable to what the checkpoint brought with it. Every quantity below is an A/B ratio: loss units cancel, and values above one mean A overfits more.

Splits. Each run uses three disjoint splits of one task: 1,319 training examples, 1,319 validation examples, and an independent test split that never enters the index. The *independent test gap* is

$$G \;=\; L_{\mathrm{test}}(E) \;-\; L_{\mathrm{train}}(E), \tag{1}$$

its test loss minus its training loss at the final epoch, and the *test-gap ratio* of a pair is

$$\rho_G \;=\; G^{\mathrm{A}} \,/\, G^{\mathrm{B}}. \tag{2}$$

The index reads the training and validation curves only. Nothing being checked ever sees the test split, which is what makes it the referee.

Tasks. GSM8K (Cobbe et al., 2021) supplies grade-school math word problems, MedMCQA (Pal et al., 2022) medical multiple-choice questions, PubMedQA (Jin et al., 2019) research questions over abstracts, and FinQA (Chen et al., 2021) numerical questions over financial filings. Every target is a reasoning trace with an answer marker. Three tasks draw their splits from one labeled pool. PubMedQA instead draws train and validation from `pqa_artificial` and its test split from the 679 `pqa_labeled` instances, whose identifiers are removed from the pool first (Appendix Q). The 14B Qwen and Llama-3.1 pairs are fine-tuned on both GSM8K and FinQA, the HuatuoGPT pair on both MedMCQA and PubMedQA. That changes the task with the checkpoints held fixed, on three families.

Why this regime. Small-sample, multi-epoch fine-tuning on reasoning traces is how domain adaptation is done. LIMA (Zhou et al., 2023), s1 (Muennighoff et al., 2025) and LIMO (Ye et al., 2025) all fine-tune on 817 to 1,000 examples for up to 15 epochs. It is also where overfitting is most severe: the training questions are the 1,319 hardest of each task. The fifty-epoch horizon is the settled endpoint the $E = 4$ verdict is checked against.

Protocol. One protocol was fixed before the first run and kept for all forty-nine. It is LoRA (Hu et al., 2022) at rank 16 on all attention and feedforward projections, learning rate 10^{-4}, effective batch 16, 50 epochs. LoRA freezes every original weight and trains a small added matrix pair (the *adapter*) beside each, so the rank is the capacity the fine-tune is given, the same for both members. FinQA uses 1,024/256 examples at length 1,024. Before the first step the unmodified checkpoint is evaluated on the target format and its per-token loss recorded as the *pre-update mismatch* x_1. Remaining settings, hardware and the eligibility rule are in Appendix Q; a checkpoint is paired on a task only when its post-training data is disjoint from that task's source split.

4 The ORI Framework

The index, the curve, and the area. For the first t epochs let $\mathcal{T}$ and $\mathcal{V}$ be the empirical distributions of the training and validation losses. Write $Q_{\mu,\sigma}$ for the log-normal distribution and fix a floor $\sigma_{\min} > 0$ on its scale so the family is closed and the projection below is attained. The *ideal trajectory* is the $\mathcal{W}_2$ projection of $\mathcal{T}$ onto that family,

$$\mathcal{I} \;=\; \arg\min_{Q \in \{Q_{\mu,\sigma}\,:\,\sigma \geq \sigma_{\min}\}} \mathcal{W}_2(Q, \mathcal{T}), \tag{3}$$

sorted decreasing. It is the smooth descent this run would have made on population structure rather than on memorized examples, fitted to the run's own losses. Three distances follow,

$$a = \mathcal{W}_2(\mathcal{I}, \mathcal{T}), \qquad b = \mathcal{W}_2(\mathcal{I}, \mathcal{V}), \qquad c = \mathcal{W}_2(\mathcal{T}, \mathcal{V}), \tag{4}$$

the *roughness* a, the departure b, and the *drift* c. The pointwise index follows Fan et al. (2025). The index is

$$\mathrm{ORI}(t) = \max\Big(0,\; 1 - \frac{a}{b}\Big), \tag{5}$$

near 0 when training and validation depart from the ideal by the same amount. It is near 1 when the training losses hug the ideal while the validation losses run away from it. Figure 1 shows the three objects on one run. The following definition fixes the three objects the rest of the paper reads: one number per epoch, one curve per run, and one area per run.

Definition 1 (The three objects)**.** $\mathrm{ORI}(t)$*, Equation* (5)*, is computed from the first* t *epochs alone: one number per epoch. The* ORI curve *is the map* $t \mapsto \mathrm{ORI}(t)$ *over the retained segment: one curve per run.* ORI-AUC*, Equation* (7)*, is the normalized trapezoidal area under that curve: one number per run.*

A verdict at $E = 4$ reads ORI-AUC over the retained segment observed by epoch 4; the same area over the full run summarizes a finished run. Because a and b are recomputed at each t, $\mathrm{ORI}(4)$ uses four epochs of log and nothing else. The average runs over the *retained segment*, which opens at

$$\tau \;=\; \min\{\, t \leq E \;:\; L_{\mathrm{val}}(t) > L_{\mathrm{train}}(t) \,\}, \qquad s = \tau - 1, \qquad m = E - s, \tag{6}$$

the first epoch whose validation loss exceeds its training loss; before that the model is underfitting and there is nothing to measure. The area under the curve is the normalized trapezoidal mean

$$\text{ORI-AUC} \;=\; \frac{1}{m-1}\sum_{t=s+1}^{E-1}\frac{\text{ORI}(t)+\text{ORI}(t+1)}{2} \;\in\; (0,1). \tag{7}$$

Equation (5) is a ratio, so multiplying every loss by a constant leaves it unchanged. The cost is three sorts (Appendices H and M), and Table A6 collects the notation.

What the index reads. Whatever the index reports, it reports about a single quantity, the drift divided by the roughness. That quantity is

$$r_t \;=\; c_t \,/\, a_t. \tag{8}$$

The following proposition says the index is pinned between two simple functions of that ratio, so the index can move only when the ratio moves.

Proposition 1 (Ratio sandwich)**.** *Let $r_t = c_t/a_t$ with $c_t > a_t$. Then*

$$\frac{r_t-2}{r_t-1} \;\le\; \text{ORI}(t) \;\le\; \frac{r_t}{r_t+1}. \tag{9}$$

The window has width $2/(r_t^2-1)$.

The following corollary says that when one run's window lies entirely above the other's, that run also carries the larger ORI-AUC.

Corollary 1 (Ordering within a matched pair)**.** *ORI-AUC is a weighted mean of $\text{ORI}(t)$ over the retained segment (Appendix H). It therefore lies between the same weighted means of the two bounds in Equation* (9)*. For two runs whose intervals are disjoint, the run with the higher interval carries the larger ORI-AUC.*

So the index is not a heuristic. What it reads is the balance between drift and roughness, and nothing else.

Verified on this study. The final ratio lies between 23.8 and 557.9 on every run, so the window is below 0.004. `audit.py` checks the sandwich at every retained epoch of every released run (proof in Appendix I). One number indexes; three numbers diagnose. Read together as a quadratic form they also track the size of the final test gap, within this study at the fourth epoch and, from the tenth, across a held-out base model (Appendix C). A higher index can mean a smoother descent or a wider drift. The drift c gives the size, the roughness a says how much had to be rebuilt, and c/a names which produced it (Appendix L, Figure A2). The two coordinates report different things: the 32B pair is a drift reading, an 8B pair an index reading.

An exact decomposition. The three distances live in one Euclidean space. Sorting turns each $\mathcal{W}_2$ distance between one-dimensional measures into an ℓ_2 distance between the sorted vectors, so the triangle they form is genuine and the law of cosines applies exactly. The following theorem says the three distances close a triangle exactly, and it gives the index in closed form.

Theorem 1 (Decomposition, and ORI in closed form)**.** *Let γ be the correlation between the two residual vectors $\mathcal{I}-\mathcal{T}$ and $\mathcal{I}-\mathcal{V}$. Then*

$$c^2 \;=\; a^2 + b^2 - 2\gamma ab, \tag{10}$$

and on the unclipped branch the index is determined by r and γ alone,

$$\text{ORI} \;=\; 1 - \frac{-\gamma + \sqrt{\gamma^2 + r^2 - 1}}{r^2 - 1}. \tag{11}$$

It says the sandwich of Proposition 1 is the shadow of an identity. Once the residual correlation is known, the index is not bounded by r but fixed by it. *Checked against the runs.* Equation (10) is an identity; the closed form of Equation (11) recovers every measured ORI to 1.1×10^{-16}. And γ is nearly constant, lying in $[-0.293, +0.241]$ with mean -0.226. Substituting that single mean still predicts ORI to within 0.0001, thirty-five times tighter than the 0.0035 the sandwich allows.

A prefix–suffix decomposition of the gap ratio. Over the retained segment the drift is the root-mean-square of the per-epoch gaps $d_t = L_{\text{val}}(t) - L_{\text{train}}(t)$ when the two sorted orders coincide. The squared drift ratio is then a ratio of sums, $\rho_c^2 = \sum_t (d_t^{\text{A}})^2 / \sum_t (d_t^{\text{B}})^2$. By the rearrangement inequality

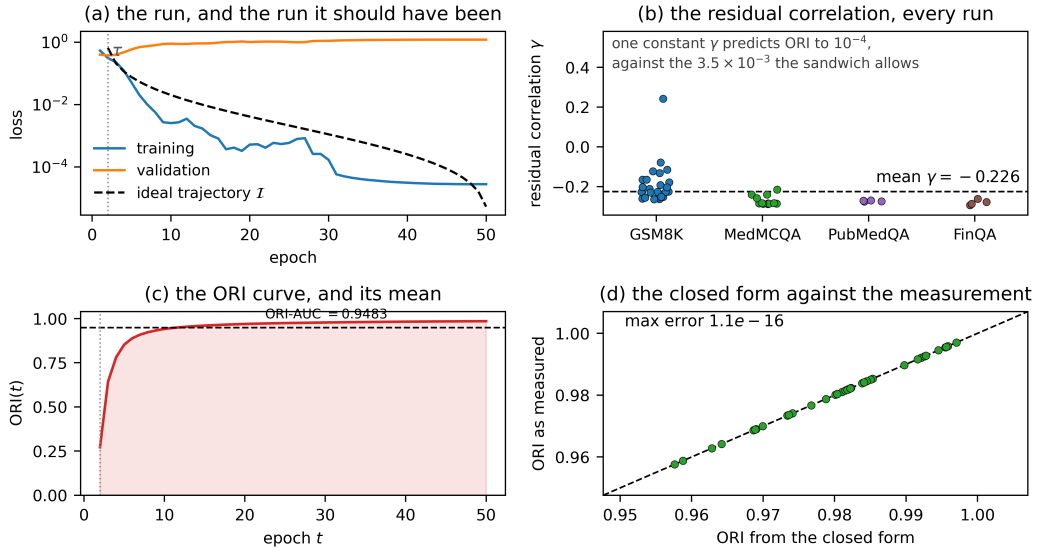

Figure 1: How the index is built. (a) One run with its ideal trajectory $\mathcal{I}$; dotted line τ. (b) Residual correlation γ, every run, by task. (c) The ORI curve, ORI-AUC dashed. (d) Closed form of Theorem 1 against measured ORI.

$c^2 \le \frac{1}{m}\sum_t d_t^2$, so the RMS gap bounds the drift from above; the theorem below concerns the ratio of RMS gaps (Appendix U). Split it at epoch k and write $\rho_c(k)$, $\rho_c(k{:}E)$ for the prefix and suffix ratios. The following theorem says the full-horizon ratio is an average of its early and late parts, so it cannot leave the range those two parts span.

Theorem 2 (Prefix–suffix mediant bound)**.** *For positive $d^{\mathrm{A}}, d^{\mathrm{B}}$, the full-horizon ratio lies between the two partial ratios,*

$$\min\{\rho_c(k),\, \rho_c(k{:}E)\} \;\le\; \rho_c(E) \;\le\; \max\{\rho_c(k),\, \rho_c(k{:}E)\}. \tag{12}$$

Consequently, if $\rho_c(k)$ and $\rho_c(k{:}E)$ lie on the same side of a threshold θ, then so does $\rho_c(E)$.

Read plainly. The reading is not an extrapolation; it is one term of an average that the remaining terms can only pull toward their own value. *The data agree.* On every full-horizon pair run the prefix and suffix RMS-gap ratios bracket the full-horizon ratio. On the two Qwen3 runs, where the suffix carries the separation (1.453, 1.440), the index flags the pair at $E = 4$ (proof in Appendix F). Once the drift is much larger than the roughness, the index stops resolving differences between runs, which is why the drift has to carry the size. The following proposition says how fast that resolution is lost.

Proposition 2 (Saturation, and why one coordinate cannot carry the size)**.** *On the unclipped branch $\partial\mathrm{ORI}/\partial r = O(r^{-2})$, so two runs whose ratios both exceed r_0 cannot differ in ORI by more than $1/(r_0 - 1)$.*

Borne out here. With r in the range above, no two runs can differ in ORI by more than 0.044. The 32B pair has a drift ratio of 4.596 against an index ratio of 0.996. The division of labor is forced: the drift is unbounded and carries the size; the index is bounded and carries the channel.

The decision rule. Write $\rho_c = c^{\mathrm{A}}/c^{\mathrm{B}}$ for the drift ratio and $\rho_I = \text{ORI-AUC}^{\mathrm{A}}/\text{ORI-AUC}^{\mathrm{B}}$ for the index ratio. Two points suffice to define Equation (7). We fixed $E = 4$ in advance so that every run's retained segment holds three points and the area rests on two trapezoids rather than one. The rule is

$$\text{flag at } E = 4 \iff \rho_c > 1.15 \;\text{ or }\; |\log \rho_I| > 0.034, \qquad \text{flag from } E = 10 \iff \rho_c > 1.15. \tag{13}$$

The index condition is two-sided because the index names the channel. A flagged pair *separates*; the member not flagged is the one to start from. Neither threshold is a free parameter, and both were fixed before the held-out pairs were launched and never adjusted. Both thresholds sit well outside the noise the runs show. Across ten pairs run on two seeds, the seeds' log drift ratios differ by a per-run standard deviation of 0.0189, seven times smaller than the 0.140 threshold. Their index readings differ by at most 0.022 against 0.034 (Table A8).

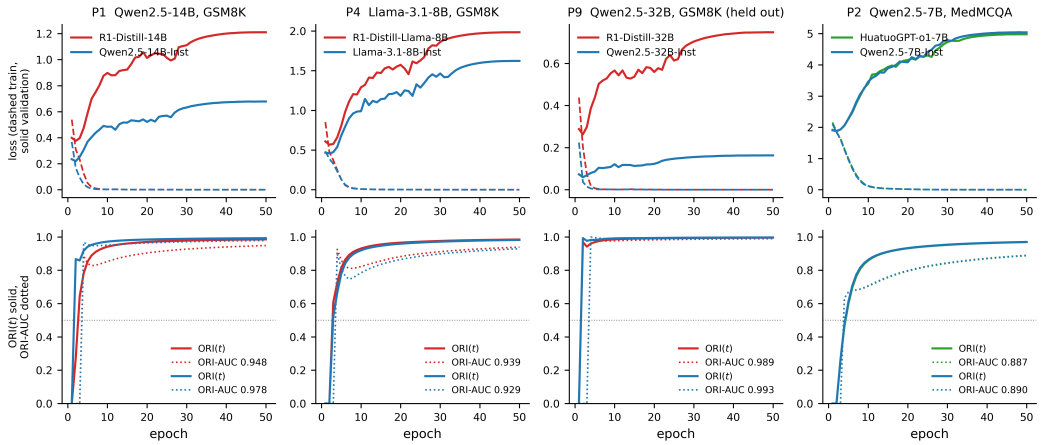

Figure 2: Four matched pairs; all thirteen pair-task combinations are in Figure A3. Top: training and validation loss of both members. Bottom: $\mathrm{ORI}(t)$, ORI-AUC in the legend. The 14B pair separates from epoch 2, the held-out 32B pair most widely of all, and the medical pair tracks its twin.

5 Experiments

Setup. Ten pairs over eleven pair-task combinations, each on two seeds, give twenty-two scored matched-pair runs across three tasks. Scoring takes the pairs carrying two seeds, a scope fixed in the pre-registration of September 19, 2026, so that every verdict rests on a replicate. The two single-seed FinQA comparisons are reported in full alongside them (Table 2, Appendix K). Read outside that scope they add one further correct call and one boundary case, twenty-three of twenty-four. That boundary case reflects what a single seed can resolve rather than a disagreement between the two readings. Its margin is an order of magnitude below the seed-to-seed movement of the same quantity (Appendix K). Figure 2 shows four pairs and Figure A3 (Appendix P) every pair.

The verdict is fixed by the fourth epoch. Grouped by the independent test gap at the full horizon, eight pair runs separate and fourteen match; the drift ratio at $E = 50$ places every run in the same group. The separating runs carry test-gap ratios of 1.087 to 1.841; the matching runs lie within 6.1% of unity. The boundary is a test-gap ratio of 1.05, pre-registered on September 19, 2026 before the 1.5B pair was run; any boundary in $(1.015, 1.087)$ gives the same grouping. The drift's verdict is the test gap's on all twenty-two runs, $r_{\log} = +0.930$ between the two log-ratios (95% bootstrap interval $[+0.895, +0.987]$). No single pair carries it: dropping each of the ten in turn leaves it between $+0.925$ and $+0.974$ (Appendix V). The reading does not need the full horizon. At $E = 4$ the two coordinates together give the full-horizon verdict on the twenty runs whose retained segment has opened. The PubMedQA pair opens one epoch later and reads the same verdict at $E = 5$. From $E = 10$ the drift ratio alone gives it on all twenty-two, with the two groups in bands that stay disjoint to $E = 50$ (Figure 3; Table A9).

Against the simple aggregates. Figure 4 makes the comparison. Over the twenty runs read at every horizon the frozen rule of Section 4 is right on all twenty at $E = 4$, and on every family taken alone. Each later horizon is read at the pair's common horizon: 50 for nineteen runs, 36 for P1 seed 1. Each gap statistic, handed its best achievable threshold after the fact, reaches fourteen, sixteen and eighteen. With every cut fixed in advance at the rule's own values, the rule reads twenty of twenty at $E = 4$, as do the validation-loss ratio and any gap-type coordinate paired with the index. Alone, the drift reads nineteen, the plain gap eighteen, the index seventeen. The rule is the reading right at $E = 4$ and at every later horizon. From $E = 10$ its drift reads twenty at $E = 10$, 15, 20, 30 and 50, where the single-epoch readings read nineteen at $E = 10$, 20 and 30. The RMS gap, the drift's upper bound in Theorem 2, gives the same numbers; the index carries P8 at $E = 4$ (Appendix U). The fixed cuts fire on none of the twelve matched runs and held on the four held-out runs and on the 1.5B pair (Table A19).

What each post-training regime brings. Table 2 reports the pairs, Table A1 groups them by regime and Appendix K gives every pre-update ratio. Reasoning-distilled checkpoints carry more drift and

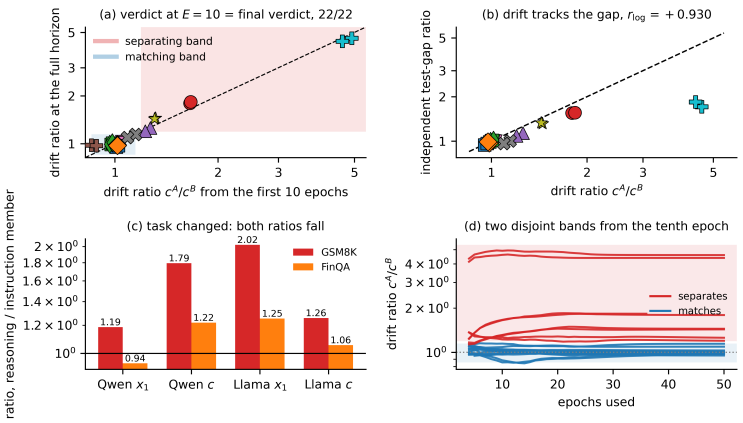

Figure 3: Twenty-two matched-pair runs. (a) Drift ratio at $E = 10$ against the full horizon, the two bands shaded; every run lies in its band on both axes. (b) Drift ratio against test-gap ratio: same verdict on all twenty-two, $r_{\log} = +0.930$. (c) Two pairs on two tasks. (d) Drift ratio against epochs used: disjoint bands from $E = 10$.

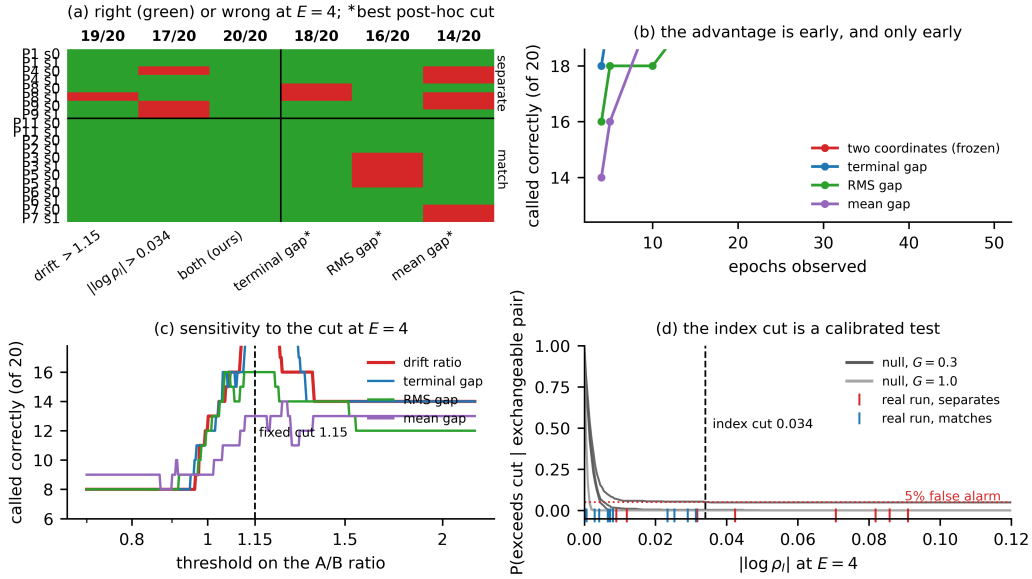

Figure 4: The rule against the simple aggregates, twenty pair runs. (a) Right (green) or wrong at $E = 4$ per run: the two coordinates alone and together under fixed cuts, each gap statistic under its best post-hoc cut. (b) Count against epochs observed. (c) Count against the cut at $E = 4$. (d) False-alarm rate of the index cut under the exchangeable-pair null, six cells; ticks mark the real runs.

a larger test gap than their twins on all eight scored runs and four base models, 1.20 to 4.60× and 1.09 to 1.84×, every run clearing the threshold. The two single-seed FinQA comparisons point the same way (Appendix K). Vendor reasoning post-training, domain adaptation and reasoning SFT on top of an instruction model all fine-tune like the twin. Seven of the ten A members are reasoning models and three of those seven fine-tune like their twin, one of them a reasoning SFT placed on an instruction model. What the separating runs share is a raw base. P7 is the control this contrast calls for: a reasoning SFT from another group, placed on the instruction model rather than the raw base, and it fine-tunes like its twin. A third arm settles what the instruction stage contributes. The raw base fine-tunes at least as well as the twin. The distilled member separates from that base more widely than from the twin, 2.09 and 2.22 against 1.795 in drift. The penalty is not a missing instruction stage. The same holds at 1.5B, on a pair run after every threshold was fixed and left unflagged at $E = 4$ (Appendix D).

Table 2: Matched pairs on the reasoning-versus-instruction axis (P1, P4, P5, P7, P8, P9), P1's protocol. Ratios are A/B at horizon E; *gap* is the independent test gap, never used by the index. Seeds agree to 4.7% on the drift ratio and to 0.013 on the index, the two coordinates the rule reads.

Pair	Task	Seed	E	x_1 ratio	c ratio	ORI-AUC (A / B)	gap ratio
P1	GSM8K	0	50	1.186	1.795	0.9483 / 0.9785	1.551
P1	GSM8K	1	36	1.187	1.833	0.9293 / 0.9704	1.558
P1	FinQA	0	24	**0.937**	1.221	0.9636 / 0.9398	1.319
P4	GSM8K	0	50	2.022	1.258	0.9388 / 0.9292	1.136
P4	GSM8K	1	50	2.024	1.199	0.9373 / 0.9287	1.087
P4	FinQA	0	50	1.253	1.056	0.9812 / 0.9772	1.053
P5	GSM8K	0	50	**3.595**	0.979	0.8044 / 0.8243	0.999
P5	GSM8K	1	50	**3.602**	0.971	0.7959 / 0.8134	1.010
P7	GSM8K	0	50	**0.953**	1.145	0.9879 / 0.9884	1.002
P7	GSM8K	1	50	**0.952**	1.093	0.9880 / 0.9884	0.964
P8	GSM8K	0	50	1.121	1.448	0.9313 / 0.9406	1.313
P8	GSM8K	1	50	1.122	1.434	0.9279 / 0.9391	1.332
P9	GSM8K	0	50	1.320	**4.596**	0.9894 / 0.9935	**1.706**
P9	GSM8K	1	50	1.318	**4.395**	0.9891 / 0.9952	**1.841**

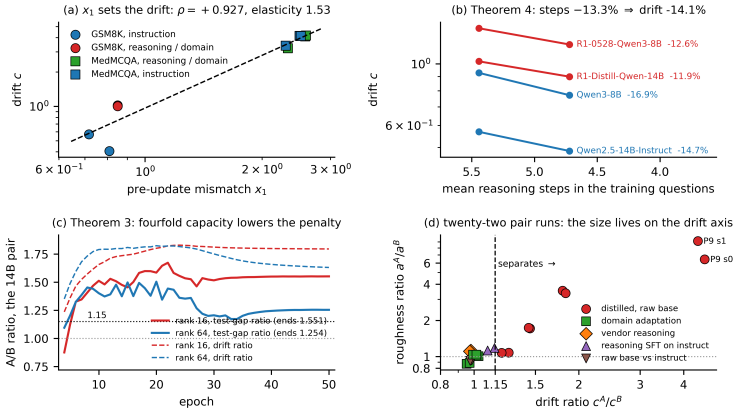

Figure 5: The mechanism and its tests. (a) The ten runs whose pre-update mismatch enters the pooled analysis: x_1 sets the drift, $\rho = +0.927$, elasticity 1.53. (b) Theorem 4: shorter chains, less drift, four members on the original and matched splits (red distilled, blue twin). (c) Theorem 3: the 14B pair at rank 16 and 64. (d) Roughness ratio against drift ratio, twenty-two pair runs; the separation lives on the drift axis.

The held-out pairs. Six pair runs on three base models no other pair uses, fine-tuned after every threshold and pooled statistic was fixed, all fall where the rule predicts. The widest is the 32B pair, flagged at the fourth epoch on both seeds, ending at 4.60 and $4.40\times$ the drift with test-gap ratios of 1.71 and 1.84.

Why the opening epochs are enough. A small fine-tune is not learning the task. Every base here met grade-school arithmetic, clinical text and financial filings in pre-training. What it changes is the format: a hard question admits many routes the model already carries, and the fine-tune fixes which one it emits, almost immediately. The training loss falls tenfold within four epochs and the two curves part company at epoch 2; later epochs scale both coordinates without turning a ratio around. Theorem 4 adds why this regime shows the effect: a k-step target pays the displacement k times. Rebuilding the GSM8K splits so that the mean step count falls by 13.3% lowers the drift by 14.1%, on all four members re-run that way. Raising the adapter rank fourfold lets the flagged member end closer to its twin. Appendix T tests one prediction of the account, Appendix W a protocol for the other.

Beyond fine-tuning. The index reads a loss triple, so any run that writes one is in scope. On a data-constrained pre-training simulation of fifteen runs the drift rises with the passes over the data and tracks the held-out gap, $\rho = +0.854$. A per-window alarm fixed in advance is silent on every healthy arm and fires on every seed of the most over-trained one (Appendix R).

Why a distilled checkpoint pays more. The account can be stated as a bound. Let D be the displacement the fine-tune must produce, $\mathcal{F}_r$ the maps a rank-r adapter can express, and let the training loss reach zero on n examples. The following theorem says the part of the displacement a rank-r adapter cannot represent becomes held-out error, shrinking with rank and growing with the displacement.

Theorem 3 (Displacement–capacity bound)**.** *Let Π_r be the projection onto $\mathcal{F}_r$. If the training loss reaches zero, the excess loss on held-out examples from the same target obeys*

$$\mathbb{E}\,[\,\textit{excess}\,] \;\geq\; \left\| D - \Pi_r D \right\|^2, \tag{14}$$

the part of the required displacement the adapter cannot represent, which is non-increasing in r and, when D grows along a fixed direction, non-decreasing in $\|D\|$.

Put concretely. What the adapter cannot represent is absorbed on the training examples, and a held-out split is where that shows. *Measured here.* Across runs x_1 sets the scale of both coordinates rather than the direction within a pair, so the bound is tested on its two other predictions. The drift should rise with the pre-update mismatch, and its elasticity in x_1 is $+1.53$ (Figure 5). Raising the adapter rank should lower the penalty. At rank 64 the same 14B pair ends at a test-gap ratio of 1.254 against 1.551 at rank 16 (Appendix T). Proof and assumptions: Appendix F. Controls for difficulty, overlap, sample composition and accelerator are in Appendix O.

6 Conclusion

Overfitting under fine-tuning is read early. Four epochs of fifty carry the verdict the independent test gap delivers at the end, on all twenty-two matched-pair runs. The thresholds were fixed in advance, and there is no false alarm on a matched pair. The theorems say what the index measures; the fourth-epoch agreement is the empirical finding, and it turns checkpoint choice into a rule read in the opening epochs. What separates the checkpoints that overfit more is a raw base narrowed by imitation; reasoning post-training laid over an instruction model fine-tunes like its twin. Because the reading needs nothing but a loss log, it applies wherever training and validation losses are written. Any fine-tuning pipeline that already logs both curves can add the check at no extra cost, and Appendix R shows the same index reading a pre-training regime. The reading is established where a pair carries two seeds and a margin. The thresholds clear the seed noise sevenfold, and the two groups stand 1.087 against 1.015 apart on the test gap. Pairs without a replicate, and pairs inside that interval, are where the next evidence belongs.

Where the reading applies, and what carries it there. Two boundaries mark the scope of the result, and the study answers each inside the same framework. The first is the opening epoch. The fourth-epoch reading needs a retained segment that has opened, which twenty of the twenty-two runs reach at $E = 4$ and the PubMedQA pair one epoch later. The rule reads that pair at $E = 5$ in the same two coordinates, and from $E = 10$ the drift alone returns the verdict on all twenty-two. The boundary therefore sets when the reading becomes available rather than whether it holds. The second is the reach of the theory. Theorem 2 governs ratios of RMS gaps rather than the Wasserstein drift itself, and Theorem 3 is proved under squared loss and a linearized adapter. Both are tested on the objects they do govern. The RMS gap gives the same verdict as the drift at every horizon (Appendix U). The capacity bound is confirmed on its two independent predictions: an elasticity of the drift in x_1 of $+1.53$ and a test-gap ratio that falls from 1.551 at rank 16 to 1.254 at rank 64 (Appendix T). Read outside the scored scope the two single-seed comparisons add one further correct call and one boundary case, twenty-three of twenty-four. The registered protocol of Appendix W extends the mismatch account to a second target format with no further fine-tuning.

What the early reading saves. The verdict arrives for a fraction of the price of the curves that confirm it. Reaching epoch 4 takes 7.7% of a run's wall clock across the fifty-epoch runs released with this paper, between 5.0 and 8.0% on any one of them. Those runs are 309 GPU-hours read to the end and 24 hours read at epoch 4. Choosing between two candidate checkpoints therefore costs under a tenth of the finished comparison, on the hardware the fine-tuning already uses.

Reproducibility Statement

Code and per-epoch logs are released as FINE-TUNE-4E at `https://anonymous.4open.science/r/Fine-tune-4E-8AA7`, where unzipping `data.zip` at the repository root gives `data/` and `scripts/audit.py` re-derives every number quoted in this paper from the logs on a CPU; Appendix Q lists every setting and Appendix N reports the seed agreement.

Ethics Statement

The study fine-tunes publicly released checkpoints on public benchmark data, involves no human subjects or personal data, and reads loss logs alone, so the weights and the data stay where they are; the eligibility rule, the hardware and the per-epoch wall-clock times are in Appendix Q.

The Use of Large Language Models

AI was used to polish the English of the manuscript: grammar, word choice and sentence length. All research work was conducted by the authors.

References

S. Aburass. Quantifying overfitting: Introducing the overfitting index. *arXiv preprint arXiv:2308.08682*, 2023.

Anonymous. SFT without overfitting: Analyzing the training dynamics of supervised fine-tuning. In *Under review at ICLR*, 2026. `https://openreview.net/forum?id=C3I7nIXKHc`.

Frank Bobe III, Gregory D. Vetaw, Chase Pavlick, Darshan Bryner, Matthew Cook, and Jose Salas-Vernis. Diagnosing generalization failures in fine-tuned LLMs: A cross-architectural study on phishing detection. *arXiv preprint arXiv:2601.10524*, 2026.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Fredrik Carlsson, Fangyu Liu, Daniel Ward, Murathan Kurfali, and Joakim Nivre. The hyperfitting phenomenon: Sharpening and stabilizing LLMs for open-ended text generation. In *International Conference on Learning Representations*, 2025.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1: Towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Y. Fan, H. Huang, and H. Han. Quantifying overfitting in deep learning. *Analysis and Applications*, 23:705–729, 2025.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2567–2577, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Yuxiang Luo, Haonan Long, Chen Wang, Qiqi Duan, Xiaotian Lin, Yanwei Xu, Yuyu Luo, Weikai Yang, and Nan Tang. TuneAhead: Predicting fine-tuning performance before full training begins. In *International Conference on Machine Learning*, 2026.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pp. 248–260, 2022.

Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade*, pp. 55–69. Springer, 1998.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. *International Conference on Learning Representations*, 2025.

H. Rezaei, H. Shakeri, and M. Daneshmand. Quantifying overfitting: Evaluating neural network performance through analysis of null space. *arXiv preprint arXiv:2305.19424*, 2023.

R. Roelofs. Measuring generalization and overfitting in machine learning. Technical Report UCB/EECS-2019-102, EECS Department, University of California, Berkeley, 2019. URL `https://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-102.html`.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, Cham, 2015.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

K. Tirumala, A. Markosyan, L. Zettlemoyer, and A. Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, pp. 38274–38290, 2022.

C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuohang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*, 2025.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Xinyue Zeng, Haohui Wang, Junhong Lin, Jun Wu, Tyler Cody, and Dawei Zhou. LENSLLM: Unveiling fine-tuning dynamics for LLM selection. In *International Conference on Machine Learning*, 2025.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Frequently Asked Questions

- *Q1: Why is this topic novel, and what is the contribution in one sentence?*
  The loss gap is in each model's own units and arrives at the end of the run. We make it unit-free and read it at epoch four. On twenty-two of twenty-two matched-pair runs, that early reading is the final verdict.
- *Q2: Why does detecting fine-tuning overfitting matter?*
  You pick a checkpoint before you can see whether it generalizes. Here the wrong pick costs up to 2.8 accuracy points, discovered only after fifty epochs. A verdict at epoch four saves the other forty-six.
- *Q3: Why is reading it in the early epochs hard?*
  At epoch four the raw gap points the wrong way on the 14B pair. There are only three retained points to read. And a gap in nats has no fixed threshold: its null cutoff moves from 0.056 to 1.06 with the noise.
- *Q4: How does the paper solve it?*
  Fine-tune matched pairs, so the checkpoint is the only variable. Read a unit-free index split into roughness a and drift c, with thresholds calibrated on a null. Then test on held-out pairs run after the thresholds were fixed (Appendices S and V).
- *Q5: Why can four epochs work at all? What is going on inside the model?*
  A small fine-tune does not teach the task; it fixes the output format. That collapse happens in the first epochs: training loss drops tenfold by epoch four and the two curves part at epoch 2. Later epochs only scale both coordinates.
- *Q6: What is the Overfitting Robustness Index?*
  Take the training curve, the validation curve, and the ideal smooth descent. Let a be the ideal-to-training distance and b the ideal-to-validation distance. Then $\mathrm{ORI} = \max(0, 1 - a/b)$, and ORI-AUC averages it over the run (Section 4).

- *Q7: What makes ORI-AUC a good instrument?*
  It is scale-free, bounded, and read from loss curves alone; four epochs suffice. Proposition 1 pins it to c/a with slack below 0.004. One threshold, 0.034, holds a 5% false-alarm rate at every noise level tested.

- *Q8: Does it order runs consistently?*
  Yes. Within each task it orders runs exactly as c/a does, rank agreement 1.000. Within every matched pair it orders the two members the same way, twenty-two of twenty-two (Appendix M).

- *Q9: What is the scope of the contribution?*
  A framework with five parts: the index, the a/c decomposition with Proposition 1, the matched-pair protocol, the pre-update mismatch x_1, and the calibrated threshold. Together they turn two loss logs into a verdict.

- *Q10: What exactly is "the drift"?*
  The $\mathcal{W}_2$ distance c between the training and validation curves over the retained epochs. A pair's drift ratio is the A member's c over the twin's.

- *Q11: What exactly is "the verdict"?*
  Separates: the flagged member overfits more. *Matches*: otherwise. From $E = 10$ the drift ratio alone decides; at $E = 4$ both coordinates do (Appendix J).

- *Q12: How early is early?*
  Epoch four, with both coordinates: drift ratio above 1.15 or index log-ratio above 0.034. From epoch ten the drift alone. Twenty of twenty at $E = 4$ and twenty-two of twenty-two from $E = 10$ (Figure 3a).

- *Q13: Which checkpoints overfit a fine-tune more?*
  Reasoning-distilled ones: the base trained by imitation alone. Four such pairs, eight scored runs, eight of eight in the same direction. The drift runs 1.20 to $4.60\times$ and the test gap 1.09 to $1.84\times$ (Table A1).

- *Q14: And which fine-tune like their twin?*
  Vendor reasoning post-training, domain adaptation, reasoning SFT on top of an instruction model, and the raw base itself. Their test gaps stay within 6.1% of unity on all fourteen runs and no drift ratio reaches the 1.15 cut (Table A1).

- *Q15: Why would distillation raise overfitting?*
  Imitation on one teacher's long traces leaves a narrow output distribution far from a terse target. A thousand examples close that distance by memorizing the specific traces.

- *Q16: How do you know that is the reason?*
  Two readings. The extra drift is validation loss that stayed up while both members reached a zero training floor. And the regimes that keep the base's breadth, vendor reasoning and domain adaptation, fine-tune like the twin, eight of eight (Table A1).

- *Q17: Why matched pairs?*
  They isolate post-training. Same base, tokenizer, size, data, adapter, learning rate, epochs and seed (Section 3). Whatever separates the curves came from the recipe.

- *Q18: How many matched pairs, and why those?*
  Ten: every released pair that passes the eligibility rule. They span four families, 1.5B to 32B, and four tasks. All ten carry two seeds. The seeds agree to 4.7% on the drift ratio and to 0.013 on the index, the two coordinates the rule reads (Appendix N).

- *Q19: Does any paired checkpoint overlap the task data?*
  One, HuatuoGPT-o1-7B, saw part of the MedMCQA source. We reran the pair on PubMedQA, outside that data. Same verdict: drift ratio 0.988 with overlap, 0.993 without (Appendix V).

- *Q20: Why fifty epochs on* $1{,}319$ *questions?*

That is how domain adaptation is done: LIMA, s1 and LIMO use 817 to 1,000 examples for up to 15 epochs. Fifty epochs is the settled endpoint the four-epoch verdict is checked against.

- *Q21: What makes drift-tracks-gap a prediction rather than a restatement?*
  The order of events. At $E = 4$ the 14B pair's test gap reads 0.875, the wrong way, while the drift already reads 1.236. The gap crosses unity at $E = 5$ and ends at 1.55.
- *Q22: Does the difficulty of the two splits affect the drift?*
  We rebuilt the splits to equal difficulty and reran. The drift ratio rose from 1.448 to 1.523 and the test-gap ratio from 1.313 to 1.404. Matching difficulty sharpens the separation (Appendix S).
- *Q23: Does it hold at $E = 5$ or $E = 10$?*
  At every horizon from 4 to 50 (Appendix J). The test gap itself takes until $E = 10$ on the 14B pair and $E = 26$ on the Llama pair to settle.
- *Q24: Loss is one thing. Does the flagged member answer fewer questions?*
  Yes. On GSM8K the 14B members start level, 81.65 against 82.11%, and the flagged one ends lower, 78.85 against 80.89%. The Llama-3.1 pair goes $60.1 \rightarrow 56.2\%$ against $69.3 \rightarrow 65.7\%$. The rule is checked against exact match, not against loss. The two can part ways on open-ended text. There a rising validation loss can sit next to writing that holds up, the hyperfitting effect. Every task here is exact-match. That is the regime the protocol is written for (Q37).
- *Q25: What does the index add over the raw validation gap?*
  One threshold that holds at every noise level. The raw gap's null cutoff moves from 0.056 to 1.06; the index's 0.034 holds 5% in every cell (Table A7).
- *Q26: How far apart are the two groups?*
  In drift ratio, matching runs sit at or below 1.148 and separating runs at or above 1.191 from $E = 10$ on. On the index at $E = 4$, 0.0006 to 0.031 against 0.012 to 0.091 (Appendix J).
- *Q27: What does the sign of the index mean?*
  It names the channel, not the size. In fine-tuning the separation comes through a rough training curve, so the index falls as the gap rises, $\rho = -0.916$. In pre-training it comes through drift, so the index rises, $\rho = +0.854$.
- *Q28: What does the index tell you that the loss gap does not?*
  The channel. The Qwen 14B member overfits through a rough descent, a ratio 3.5, and takes the lower index. The Llama and FinQA members overfit through drift on a smooth curve and take the higher one.
- *Q29: What does Proposition 1 guarantee?*
  That the index is the ratio c/a read through a monotone window of width $2/(r_t^2 - 1)$. On every run here that window is below 0.004. The proof, from the triangle inequality, is in Appendix F.
- *Q30: Why is the index scale-free, and why care?*
  Multiply every loss by a constant and a/b does not move (Equation (5)). A checkpoint that starts further from the target shows a bigger gap for reasons unrelated to generalization; the index removes that.
- *Q31: Why log-normal for the ideal trajectory?*
  The ideal trajectory is the $\mathcal{W}_2$ projection of the training losses onto the log-normal family (Appendix H). Optimization noise is multiplicative, and products of noise are log-normal in the limit.
- *Q32: Is x_1 a property of the model?*
  Of the model and the target together. Move the 14B pair from GSM8K to FinQA and its x_1 ratio goes from 1.186 to 0.937, across unity (Table A11).
- *Q33: Stratos is trained on R1 traces too. Why does it match?*
  Because of where the traces land. Stratos puts 17k traces on top of the instruction checkpoint, which keeps its breadth. The distilled members put hundreds of thousands onto a raw base with no instruction stage.

- *Q34: How well does the study replicate across seeds?*
  Two seeds on all ten pairs, agreeing to 4.7% on the drift ratio and to 0.013 on the index, partly on different accelerators (Appendix N). Every seed replicate gives the same verdict as its partner.
- *Q35: How are the statistics supported?*
  Every correlation carries a 95% bootstrap interval from $20{,}000$ resamples (Table A12). The primary statistic is a count, though: twenty-two of twenty-two verdicts, eight of eight distilled runs.
- *Q36: Why hold the protocol fixed?*
  So the checkpoint is the only variable. LoRA rank 16, learning rate 10^{-4}, batch 16, 50 epochs, fixed before the first run and kept for all forty-nine (Section 3).
- *Q37: What is the scope?*
  Supervised fine-tuning under one protocol: 1.5B to 32B, four families, four tasks, ten pairs. Every ordering is the same under any choice of reference run (Appendix M).
- *Q38: Does the index apply beyond supervised fine-tuning?*
  Wherever a run writes loss logs. On a fifteen-run pre-training simulation the drift tracks the test gap, $\rho = +0.854$, and a pre-set alarm fires only on the most over-trained budget (Appendix R).
- *Q39: What should a practitioner actually do?*
  Read both coordinates at epoch four, or the drift ratio alone from epoch ten. If the pair separates, the flagged member will overfit more: pick the other one. Expect a distilled checkpoint to be the flagged one.
- *Q40: What is released?*
  Everything behind every number: per-epoch training, validation and test losses for all forty-nine runs, the pre-update mismatches, the code, and the null simulation. `audit.py` recomputes every headline number on a CPU.
- *Q41: What would falsify the claim?*
  A matched pair passing the eligibility rule whose drift verdict at $E = 4$ disagrees with its test gap at $E = 50$. Twenty-two runs have agreed; the package lets anyone add a pair and check.
- *Q42: Is the drift just the train-validation gap averaged over epochs?*
  No. At $E = 5$ the drift separates the two groups and the mean or RMS gap does not. And it comes with a second coordinate, the roughness, which names the channel (Appendix U).
- *Q43: You only compare against the gap read at the last epoch. What about other baselines?*
  Three, in Appendix U. The gap read at the current epoch, what a patience rule uses, overlaps the groups at $E = 4$, 10 and 20. Mean and RMS gap work only from $E = 10$. Mean validation loss measures something else (Q44).
- *Q44: Why read a gap statistic rather than the validation loss alone?*
  A statistic that never references the training loss ranks the two members by held-out quality. The reading here is taken among checkpoints of comparable accuracy, where the gap is what remains: the 14B pair starts level at 81.65% against 82.11%.
- *Q45: What is the false-alarm rate of the rule you actually recommend?*
  The index threshold 0.034 holds a false-alarm rate at or below 5% in every cell of the simulated null. `pair_null.py` prints every rate (Table A7).
- *Q46: Does the roughness a measure the shape of the descent?*
  It is a $\mathcal{W}_2$ distance between loss distributions, not a shape. Epoch order enters only through where the retained segment opens. A run that is rebuilding visits losses a smooth descent never would, and a records that.
- *Q47: Your elasticities are pooled over two task clusters. Do they hold within a task?*
  Yes, and they are larger. Within GSM8K the elasticities of a and c in x_1 are 6.86 and 3.15; within MedMCQA, 6.71 and 2.10. Same sign and same conclusion as the pooled 2.50 and 1.53.

- *Q48: What does x_1 tell you, and does it pick the member that will overfit?*
  It sets the scale of both coordinates, not the verdict. One forward pass before training measures it. It reverses with the target: GSM8K to FinQA takes the 14B pair's x_1 ratio from 1.186 to 0.937. The verdict is read from the drift and the index.

- *Q49: Is the held-out pair really held out if one of its members appears in the pooled runs?*
  The thresholds are what is held out, fixed before either held-out pair ran. Qwen3-8B appeared earlier as a single pooled run; Table A21 repeats every pooled statistic without it, and every conclusion holds.

- *Q50: Does the index beat a plain train-validation gap at the fourth epoch?*
  Yes, head to head. Our rule, thresholds frozen, is right on all twenty runs at $E = 4$. Each gap statistic, given its best threshold after the fact, gets at most eighteen (Appendix U).

- *Q51: Are the separating and matching labels a cut on the statistic you then validate?*
  The groups are defined by the independent test gap at $E = 50$. The claim is that the drift at $E = 50$, and then at $E = 4$, lands every run in the same group. `audit.py` checks that comparison directly.

- *Q52: All four separating pairs are DeepSeek distillations. Is the claim about distillation or about one recipe?*
  The regime is defined by what the post-training did: imitation onto a raw base. Four base models, eight scored runs, eight of eight, and the two single-seed FinQA comparisons point the same way. Seven of the ten A members are reasoning models and three of them fine-tune like their twin, so "reasoning" is not the cause; the raw base is.

- *Q53: Where does the drift threshold 1.15 come from?*
  It is the null cutoff for the low-noise cells the real runs occupy, fixed before the held-out pairs were launched and never moved. Figure 4c plots the count against the cut: 1.15 sits on a plateau, not a spike.

- *Q54: Twenty-two runs, but how many independent units?*
  Ten pairs. The pair is the unit of evidence, and all ten agree. The seeds are there to show the reading reproduces. It does, to 4.7% on the drift ratio.

- *Q55: Does the threshold survive a change of adapter rank?*
  At rank 64, four times the capacity, the threshold fixed at rank 16 still flags the same pair at the fourth epoch (Appendix T).

- *Q55b: Does the reading depend on LoRA, or would full-parameter fine-tuning change it?*
  Every run here is LoRA. Rank 64 above is the capacity axis we varied. Two things do not care how the update is parameterized. The coordinates come from the loss curves. The pair cancels the units. What full-parameter training changes is the size of the displacement. Theorem 3 bounds that from below, and more capacity lowers the floor. Rank 16 to 64 moved the coordinates and left the verdict alone. Full-parameter is the next point on that axis. We did not run it.

- *Q55c: The protocol runs fifty epochs to a near-zero training loss. Does it read anything at one to three epochs, as a production SFT would?*
  The verdict is taken at epoch four. It uses the first 8% of the run and nothing after. So the reading already lives in the short regime. The fifty epochs supply the ground truth it is checked against. They do not produce it. A three-epoch run is read the same way at epoch four, against a smaller gap.

- *Q55d: Every dataset here is single-task. What happens on a mixed instruction corpus such as Alpaca or UltraFeedback?*
  Not run here. The prediction is checkable. Theorem 4 ties the penalty to chain length k, not to the task. Short answers carry a small k and should separate less. Long-form targets should behave like the traces here. A mixed corpus is a mixture of the two. Its drift is the length-weighted average, not a new effect. The reading does not change: loss curves in, and the pair cancels the unit. What a mixed corpus tests is effect size, not the instrument.

- *Q56: Why not just stop early instead?*
  The reading is taken at epoch four, before either member has been trained far. It answers a different question: which of the two checkpoints to start from.
- *Q57: What does the index add over the two distances?*
  The drift is unbounded and carries the size; the index is bounded and names the channel (Proposition 2). Reading them together is what gives the verdict at the first horizon.
- *Q58: Is the reading available when the two candidates come from different bases?*
  The protocol is written for a matched pair, where the ratio cancels the unit. A supporting comparison on unmatched bases is reported in Appendix J. One coordinate carries across a base change and one does not. The index is dimensionless, so ORI-AUC compares between any two runs. The drift is in nats, which is why the pair takes a ratio. With no twin, read the index against its own threshold of 0.034. Keep the drift as a within-model quantity. That is weaker than a matched pair. It is the form the protocol takes when no twin exists.
- *Q59: The rule joins two tests with OR. What is its false-alarm rate?*
  At most the sum of the two. Each test is read at its own fixed cut. On the twelve matched runs the joined rule fired zero times, and on the six runs launched after the thresholds were fixed, zero.
- *Q60: Does the drift beat the plain gap?*
  At the fourth epoch they tie. Paired with the index at the fixed cuts, the drift, the RMS gap and the plain gap each read twenty of twenty. The difference shows later. From epoch ten the drift and the RMS gap stay right at every horizon. The plain gap and the validation-loss ratio flip on two runs at epochs 10, 20 and 30. Averaging over the retained segment is what keeps the verdict still. Theorem 2 says why the drift and the RMS gap move together.
- *Q61: Could the labels just be tracking which vendor made the model?*
  No. Two controls separate the vendor from the axis. The same reasoning traces placed on an instruction model fine-tune like the twin (P7). A raw base with no distillation also fine-tunes like its instruction twin, at 14B and at 1.5B (Appendix D). Neither the trace source alone nor the raw base alone carries the excess.
- *Q62: Where is the line between separate and match, and when was it drawn?*
  A test-gap ratio of 1.05 on both seeds. It was written into the pre-registration file on September 19, 2026, before the 1.5B pair ran. Any line between 1.015 and 1.087 gives the same groups.
- *Q63: Do the three coordinates predict the final gap itself, not only which member is larger?*
  Yes, as a quadratic form. Fit G_E^2 on a^2, $\langle u, v\rangle$ and c^2 read at epoch four. Leave-one-run-out R^2 is 0.79 on 51 runs, 0.97 at epoch ten. Fitted on the pooled runs alone, it predicts the pair runs with R^2 0.73 and 0.94. At epoch ten it also holds out a whole base model, R^2 0.68. Appendix C gives the condition under which the form is exact.

B OVERFITTING BY POST-TRAINING REGIME

Grouped by what the post-training did, one regime separates from its twin and the rest fine-tune like it (Table A1).

Table A1: Overfitting under fine-tuning by post-training regime, each checkpoint against the twin released from its own base, instruction-tuned except in P8, under one protocol. Ranges span the pair runs in the row; ratios are A/B at the common horizon. †The two single-seed FinQA comparisons (P1: drift 1.221, gap 1.319; P4: 1.056, 1.053) are reported in Appendix K and not scored.

Regime of A	Pairs (runs)	x_1 ratio	c ratio	gap ratio	Verdict
Reasoning-distilled (R1), GSM8K	P1, P4, P8, P9 (8)	1.12–2.02	1.20–4.60	1.09–1.84	separates†
Vendor reasoning (Phi), GSM8K	P5 (2)	3.60	0.97–0.98	1.00–1.01	matches
Domain-adapted, MedMCQA	P2, P3, P6 (6)	1.00–1.05	0.95–1.03	0.94–1.01	matches
Instruct twin + reasoning SFT (Stratos), GSM8K	P7 (2)	0.95	1.09–1.15	0.96–1.00	matches
Raw base against its instruction twin, GSM8K	P11 (2)	0.79	0.98–0.98	0.97–0.99	matches

C A CONDITIONAL CONNECTION TO THE FINAL TEST GAP

Under one stated condition the three coordinates read at epoch k fix the squared final test gap, and the fitted form holds across the released runs (Table A2). Write $u_k = Q_{I_k} - Q_{T_k}$ and $v_k = Q_{I_k} - Q_{V_k}$ for the quantile departures of the training and validation distributions from the projected ideal at epoch k, so that $a_k = \|u_k\|$, $b_k = \|v_k\|$ and $c_k = \|u_k - v_k\|$ in $L^2(0,1)$, and write G_E for the independent test gap at the full horizon.

The following proposition says that if one hidden state drives both curves and the final gap, then the squared final gap is a fixed weighted sum of the three coordinates read at epoch k.

Let z_t be the state of the run at epoch t, let F be the operator that advances that state by one epoch, let $\mathcal{U}$ and $\mathcal{V}$ read the two quantile departures off it, and let h read the final test gap off it. All four are fixed independently of the endpoints being evaluated.

Proposition 3 (Conditional endpoint identification)**.** *Suppose the shared state model $u_k = \mathcal{U}z_k$, $v_k = \mathcal{V}z_k$, $z_{t+1} = Fz_t$, $G_E = h^\top z_E$ holds, and put $Q_a = \mathcal{U}^*\mathcal{U}$, $Q_b = \mathcal{V}^*\mathcal{V}$, $Q_c = (\mathcal{U}-\mathcal{V})^*(\mathcal{U}-\mathcal{V})$ and $\ell = (F^{E-k})^\top h$. If for some scalars $\beta_1, \beta_2, \beta_3$ the quadratic forms $\ell\ell^\top$ and $\beta_1 Q_a + \beta_2 Q_b + \beta_3 Q_c$ agree on the states the runs visit, then*

$$G_E^2 = \beta_1 a_k^2 + \beta_2 b_k^2 + \beta_3 c_k^2, \tag{15}$$

and when $0 < \mathrm{ORI}(k) < 1$ one may substitute $b_k = a_k/(1 - \mathrm{ORI}(k))$.

Proof. The recurrence gives $G_E = \ell^\top z_k$, so $G_E^2 = z_k^\top \ell\ell^\top z_k = \beta_1 z_k^\top Q_a z_k + \beta_2 z_k^\top Q_b z_k + \beta_3 z_k^\top Q_c z_k = \beta_1 a_k^2 + \beta_2 b_k^2 + \beta_3 c_k^2$. The substitution is the definition of $\mathrm{ORI}(k)$. □

Measured here. Equation (15) says the squared final gap is a fixed quadratic form in the three coordinates read at epoch k. The basis $(a_k^2, \langle u_k, v_k\rangle, c_k^2)$ with $\langle u_k, v_k\rangle = (a_k^2 + b_k^2 - c_k^2)/2$, the shared departure of the two splits, spans the same forms. Table A2 fits it on every released run that carries a training, a validation and an independent test loss over at least thirty epochs. The run's retained segment must also be open at epoch k. That is 51 runs on nine base models, de-duplicated by file content. That sample is a superset of the twenty-two scored pair runs. It adds the pooled single-model runs and the difficulty-matched and rank-64 controls, each of which contributes one curve rather than a pair. All three coefficients are positive at both horizons. The leave-one-run-out R^2 is 0.79 at $k = 4$ and 0.97 at $k = 10$. Coefficients fitted on the ten pooled runs alone predict the 41 pair runs with R^2 0.73 and 0.94. At $k = 10$ the form also survives holding out a whole base model, R^2 0.68 over the nine folds. The drift alone, with an intercept, reaches 0.24 and 0.79. The cross-base column is an $E = 10$ result. What the fourth epoch supports is the pair decision of Equation (13). It does not support a numerical prediction of the gap on a base model the fit has not seen. The transfer of that decision across base models is measured separately, in Table A19 and on the held-out pairs. `endpoint_fit.py` regenerates the table.

Table A2: The squared final test gap as a quadratic form in the epoch-k coordinates, 51 runs on nine base models. *Leave-one-run-out*: each run predicted from the other 50. *Pooled → pairs*: coefficients fitted on the ten pooled runs, applied to the 41 pair runs. *Leave-one-base-out*: coefficients fitted on eight base models and applied to the ninth, summed over the nine folds; at $k = 4$ the fit is read within a base model. Every column after the coefficients is an R^2; the last is the drift alone with an intercept, leave-one-run-out. Coefficients in the basis $(a^2, \langle u, v\rangle, c^2)$.

k	$(\gamma_1, \gamma_2, \gamma_3)$	in-sample	one-run-out	pooled → pairs	one-base-out	c^2 alone
4	(45133, 51622, 123)	0.847	0.793	0.732	within base	0.236
10	(728, 524, 7.4)	0.973	0.967	0.935	0.684	0.793

D THE RAW BASE AS A THIRD ARM

A *third arm* is the base model both members of a pair were built from, fine-tuned under the same protocol and read against each of them. A distilled checkpoint differs from its instruction-tuned twin in two ways at once: it was narrowed by imitation, and it never received an instruction stage. Fine-tuning the raw base separates them. Qwen2.5-14B-Base is the model both members of P1 were built from, and it has no instruction stage either. Table A3 reports it on two seeds under P1's protocol, read at the full horizon $E = 50$.

Table A3: The 14B family on GSM8K under one protocol, read at $E = 50$. The raw base has no instruction stage, as the distilled member has none; ratios in the last two columns are taken against it.

Checkpoint	a	c	ORI-AUC	test gap	drift / base	gap / base
DeepSeek-R1-Distill-Qwen-14B	0.0149	1.0213	0.9483	1.2241	**2.09 / 2.22**	**1.73 / 1.72**
Qwen2.5-14B-Instruct	0.0042	0.5690	0.9785	0.7891	1.16 / 1.24	1.11 / 1.11
Qwen2.5-14B-Base, seed 0	0.0029	0.4898	0.9829	0.7092	—	—
Qwen2.5-14B-Base, seed 1	0.0030	0.4601	0.9814	0.7135	—	—

Three readings follow. The raw base fine-tunes at least as well as the instruction twin: drift ratios of 0.861 and 0.809 against it, and test-gap ratios of 0.898 and 0.904. The instruction stage is therefore not what protects a checkpoint here. The distilled member separates from the raw base more widely than from the twin, 2.09 and 2.22 against 1.795 in drift. The narrowing is therefore not a matter of what the distillation omitted. And the roughness ratio against the raw base, 5.192 and 4.951, is larger than the 3.532 against the twin. That is the coordinate Theorem 3 predicts should grow with the displacement the fine-tune must undo. The distilled checkpoint overfits a later fine-tune more than the base it was distilled from.

The same third-arm reading was repeated at 1.5B after every threshold was fixed, on two seeds under P1's protocol (Table A4). Qwen2.5-1.5B-Base and Qwen2.5-1.5B-Instruct fine-tune alike: drift ratios of 0.984 and 0.975, test-gap ratios of 0.970 and 0.989, index ratios within 0.005 in log at that horizon. The frozen rule leaves the pair unflagged at $E = 4$ on both seeds (drift ratios 1.032 and 0.991). At 1.5B as at 14B, the instruction stage neither adds nor removes overfitting under a later fine-tune.

Table A4: The 1.5B raw base against its instruction twin on GSM8K under P1's protocol, read at $E = 50$; ratios in the last two columns are taken against the raw base of the same seed.

Checkpoint	a	c	ORI-AUC	test gap	drift / base	gap / base
Qwen2.5-1.5B-Instruct, seed 0	0.0184	1.1340	0.9336	1.5094	1.016	1.031
Qwen2.5-1.5B-Instruct, seed 1	0.0181	1.1023	0.9327	1.4811	1.026	1.012
Qwen2.5-1.5B-Base, seed 0	0.0171	1.1160	0.9374	1.4645	—	—
Qwen2.5-1.5B-Base, seed 1	0.0168	1.0743	0.9361	1.4642	—	—

E WHY MULTI-STEP TARGETS AMPLIFY THE COST

The targets in this study are reasoning traces, and a trace of k steps factorizes: its log likelihood is a sum of k per-step terms. A checkpoint distilled to imitate one teacher commits to a particular continuation at each step, so wherever the target's step differs the displacement is paid again.

The following theorem says that a target of k steps makes the checkpoint pay its displacement k times, so a longer chain costs more at the same adapter rank.

Theorem 4 (Chain length amplifies the displacement)**.** *Let the target be a k-step trace and let the per-step displacement between the checkpoint's preferred continuation and the target's have expected size δ, independently across steps. Then the trace-level displacement satisfies $\mathbb{E}\,\|D_k\|^2 = k\,\delta^2$, and by Theorem 3 the held-out excess grows linearly in k at fixed adapter rank.*

In words. The penalty is not a property of the task but of how many decisions the target asks the model to get right in a row. It is why this regime, small-sample fine-tuning on multi-step traces, is where the effect is visible at all. It is also why the effect matters in settings such as clinical question answering, where the target is an expert's worked explanation rather than a label.

Tested directly. Appendix S rebuilds the GSM8K splits so that training and validation carry the same difficulty. That lowers the mean reasoning-step count of the training questions from 5.443 to 4.72, a drop of 13.3%, with the task, the protocol and the checkpoints held fixed. The drift falls on all four members of the two pairs re-run this way: -12.6% and -16.9% on the Qwen3 pair, -11.9% and -14.7% on the 14B pair. The mean is -14.1% against the -13.3% change in step count. The elasticity is close to one, which is what Theorem 4 predicts.

F PROOFS

Throughout, the retained segment has m epochs, and $\mathcal{T}, \mathcal{V}, \mathcal{I}$ are empirical measures with m atoms of mass $1/m$ each. We write $T^*, V^*, I^* \in \mathbb{R}^m$ for the three loss vectors sorted in decreasing order, and $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^m$.

A LEMMA USED THROUGHOUT: THE THREE DISTANCES LIVE IN ONE EUCLIDEAN SPACE

The following lemma says that once each set of losses is sorted, the transport distance between them is the ordinary Euclidean distance between the two sorted lists.

Lemma 1 (One-dimensional $\mathcal{W}_2$ is an ℓ_2 distance between sorted vectors)**.** *For two empirical measures $\mu = \frac{1}{m}\sum_i \delta_{x_i}$ and $\nu = \frac{1}{m}\sum_i \delta_{y_i}$ on $\mathbb{R}$, with $x_{(1)} \geq \cdots \geq x_{(m)}$ and $y_{(1)} \geq \cdots \geq y_{(m)}$ the sorted atoms,*

$$\mathcal{W}_2^2(\mu,\nu) \;=\; \frac{1}{m}\sum_{i=1}^{m} \big(x_{(i)} - y_{(i)}\big)^2. \tag{16}$$

Why it matters here. This identity is what lets the three distances be read off a loss log at all. `audit.py` recomputes them this way and checks the resulting bound at every retained epoch of all 57 released runs.

Proof. Every coupling of two uniform empirical measures with m atoms is a doubly stochastic matrix, and by the Birkhoff–von Neumann theorem the transport cost is minimized at a permutation. Among permutations the monotone one minimizes $\sum_i (x_i - y_{\pi(i)})^2$: if π pairs $x_{(i)}$ with $y_{(j)}$ and $x_{(i')}$ with $y_{(j')}$ where $i < i'$ but $j > j'$, then swapping the two targets changes the cost by $-2\,(x_{(i)} - x_{(i')})(y_{(j')} - y_{(j)}) \leq 0$, so uncrossing never increases the cost, and repeating until no crossing remains reaches the monotone coupling. Its cost is the right-hand side of Equation (16). □

Consequently

$$a = \tfrac{1}{\sqrt{m}}\,\|I^* - T^*\|, \qquad b = \tfrac{1}{\sqrt{m}}\,\|I^* - V^*\|, \qquad c = \tfrac{1}{\sqrt{m}}\,\|T^* - V^*\|, \tag{17}$$

and since I^* is already sorted by construction, all three are distances in the same space $\mathbb{R}^m$. The common factor $1/\sqrt{m}$ cancels from every ratio below.

PROPOSITION 1 (RATIO SANDWICH)

Proof. By Equation (17) and the triangle inequality in $\mathbb{R}^m$,

$$|c - a| \;\leq\; b \;\leq\; c + a. \tag{18}$$

Under the hypothesis $c > a$ the left-hand side is $c - a > 0$, so $b > 0$ and

$$\frac{a}{c+a} \;\leq\; \frac{a}{b} \;\leq\; \frac{a}{c-a}. \tag{19}$$

Write $r = c/a > 1$ and divide numerator and denominator by a: $\frac{1}{r+1} \le \frac{a}{b} \le \frac{1}{r-1}$. Since $\mathrm{ORI} = 1 - a/b$ on the branch where $a < b$,

$$1 - \frac{1}{r-1} \;=\; \frac{r-2}{r-1} \;\le\; \mathrm{ORI} \;\le\; 1 - \frac{1}{r+1} \;=\; \frac{r}{r+1}, \tag{20}$$

which is Equation (9). The width of the window is

$$\frac{r}{r+1} - \frac{r-2}{r-1} \;=\; \frac{r(r-1) - (r-2)(r+1)}{(r+1)(r-1)} \;=\; \frac{2}{r^2-1}. \tag{21}$$

Two remarks. First, the lower bound is non-negative exactly when $r \ge 2$, so for $r \ge 2$ the clipping in $\max(0, \cdot)$ is inactive and the sandwich describes the index itself. On every run of this study the final-epoch ratio $r_E = c_E/a_E$ satisfies $r_E \ge 23.8$. Second, both bounds are attained: the upper when $I^* - T^*$ and $I^* - V^*$ point the same way, the lower when they point opposite ways (Theorem 1 below makes this exact). □

COROLLARY 1 (ORDERING WITHIN A MATCHED PAIR)

Proof. Over a retained segment with epochs $s+1, \dots, E$ the trapezoidal average of Equation (7) is $\sum_j w_{m,j}\,\mathrm{ORI}(s+j)$ with $w_{m,1} = w_{m,m} = \frac{1}{2(m-1)}$, $w_{m,j} = \frac{1}{m-1}$ otherwise, so $w_{m,j} \ge 0$ and $\sum_j w_{m,j} = 1$. Applying Equation (9) at each j and taking the weighted sum,

$$\sum_j w_{m,j}\,\frac{r_{s+j}-2}{r_{s+j}-1} \;\le\; \text{ORI-AUC} \;\le\; \sum_j w_{m,j}\,\frac{r_{s+j}}{r_{s+j}+1}. \tag{22}$$

If run A's interval lies entirely above run B's, the left end of A's exceeds the right end of B's, so $\text{ORI-AUC}^{\mathrm{A}} > \text{ORI-AUC}^{\mathrm{B}}$. □

THEOREM 1 (DECOMPOSITION, AND ORI IN CLOSED FORM)

Proof. Set $u = I^* - T^*$ and $v = I^* - V^*$, so that $T^* - V^* = v - u$. By Equation (17), $a = \|u\|/\sqrt{m}$, $b = \|v\|/\sqrt{m}$ and

$$c^2 \;=\; \frac{1}{m}\|v-u\|^2 \;=\; \frac{1}{m}\big(\|u\|^2 + \|v\|^2 - 2\langle u, v\rangle\big) \;=\; a^2 + b^2 - 2\gamma ab, \qquad \gamma := \frac{\langle u, v\rangle}{\|u\|\,\|v\|} \in [-1, 1], \tag{23}$$

which is Equation (10); γ is the cosine of the angle between the two residual vectors, the uncentered correlation named in the statement. For the closed form, regard Equation (10) as a quadratic in b:

$$b^2 - 2\gamma a\, b + (a^2 - c^2) = 0 \quad\Longrightarrow\quad b = a\Big(\gamma \pm \sqrt{\gamma^2 + r^2 - 1}\Big), \qquad r = c/a. \tag{24}$$

For $r > 1$ the square root exceeds $|\gamma|$, so the root with the minus sign is negative and is rejected; the admissible root is $b = a\,(\gamma + S)$ with $S := \sqrt{\gamma^2 + r^2 - 1}$. Then on the unclipped branch

$$\mathrm{ORI} \;=\; 1 - \frac{a}{b} \;=\; 1 - \frac{1}{\gamma + S} \;=\; 1 - \frac{S - \gamma}{S^2 - \gamma^2} \;=\; 1 - \frac{-\gamma + \sqrt{\gamma^2 + r^2 - 1}}{r^2 - 1}, \tag{25}$$

which is Equation (11). Finally, $\gamma = +1$ gives $S = r$ and $\mathrm{ORI} = r/(r+1)$; $\gamma = -1$ gives $\mathrm{ORI} = (r-2)/(r-1)$. The two ends of the sandwich of Proposition 1 are therefore the two extreme alignments of the residuals, and the index moves monotonically between them as γ runs from -1 to $+1$, since $\partial\mathrm{ORI}/\partial\gamma = (1 + \gamma/S)/(\gamma + S)^2 > 0$. □

THEOREM 2 (PREFIX DETERMINACY)

Over the retained segment write $d_t = L_{\mathrm{val}}(t) - L_{\mathrm{train}}(t) > 0$. When the sorted orders of the training and validation losses coincide on that segment, the monotone coupling of Lemma 1 is the identity, so

$$c^2 = \frac{1}{m}\sum_t d_t^2 \qquad \text{and, for a pair,} \qquad \rho_c(E)^2 = \frac{\sum_t (d_t^{\mathrm{A}})^2}{\sum_t (d_t^{\mathrm{B}})^2}. \tag{26}$$

The following lemma says that a ratio of two sums lies between the smallest and the largest of the term-by-term ratios. It cannot escape the range its parts occupy.

Lemma 2 (Mediant). *For positive reals α_i, β_i, $\min_i \alpha_i/\beta_i \le \big(\sum_i \alpha_i\big)/\big(\sum_i \beta_i\big) \le \max_i \alpha_i/\beta_i$.*

Seen in the runs. The bracketing it gives is visible on the data. On the twenty full-horizon pair runs the prefix and the suffix RMS-gap ratios fall on the same side of $\theta = 1.15$ on eighteen, and the full-horizon ratio lies between them on all twenty.

Proof. Let $\underline{q} = \min_i \alpha_i/\beta_i$. Then $\alpha_i \ge \underline{q}\,\beta_i$ for every i; summing gives $\sum_i \alpha_i \ge \underline{q} \sum_i \beta_i$. The upper bound is the same argument with $\overline{q} = \max_i \alpha_i/\beta_i$. □

Proof of Theorem 2. Split the segment at epoch k and apply Lemma 2 with two blocks, $\alpha_1 = \sum_{t\le k}(d_t^{\mathrm{A}})^2$, $\beta_1 = \sum_{t\le k}(d_t^{\mathrm{B}})^2$, and α_2, β_2 the corresponding suffix sums, so that $\alpha_1/\beta_1 = \rho_c(k)^2$ and $\alpha_2/\beta_2 = \rho_c(k{:}E)^2$. By Equation (26) the ratio of the totals is $\rho_c(E)^2$, hence

$$\min\{\rho_c(k)^2, \rho_c(k{:}E)^2\} \le \rho_c(E)^2 \le \max\{\rho_c(k)^2, \rho_c(k{:}E)^2\}, \tag{27}$$

and taking the increasing map $x \mapsto \sqrt{x}$ gives Equation (12). If both partial ratios exceed θ the total does; if both fall below, so does the total. In either case the verdict at k is the verdict at E. □

Remark. The theorem is a statement about ratios of sums of squared per-epoch gaps and needs no assumption on the loss distributions. By the rearrangement inequality the drift obeys $c^2 \le \frac{1}{m}\sum_t d_t^2$, so the RMS gap is its identity-coupling upper bound. On the twenty full-horizon pair runs the prefix and suffix RMS-gap ratios fall on the same side of $\theta = 1.15$ on eighteen. The two Qwen3 runs are the exception. There the prefix RMS gap has not yet crossed the threshold at $E = 4$ while the suffix has $(1.453, 1.440)$. The index reads the channel on those two, and the $E = 4$ verdict is right on both.

PROPOSITION 2 (SATURATION)

Proof. Two arguments, one from the window and one from the closed form. From Proposition 1, for $r \ge r_0 > 2$ the index lies in $[\frac{r_0-2}{r_0-1},\, 1)$, an interval of length $1/(r_0-1)$, which is what makes the drift the coordinate that carries the size. From Equation (25), with $S = \sqrt{\gamma^2 + r^2 - 1}$ and $\partial S/\partial r = r/S$,

$$\frac{\partial\,\mathrm{ORI}}{\partial r} \;=\; \frac{\partial}{\partial r}\Big(1 - \frac{1}{\gamma + S}\Big) \;=\; \frac{r}{S\,(\gamma+S)^2} \;\le\; \frac{r}{(r-1)^3} \;=\; O(r^{-2}), \tag{28}$$

using $S \ge \sqrt{r^2-1} \ge r-1$ and $\gamma + S \ge S - 1 \ge r - 2$ for $r \ge 3$ (the constant is immaterial). On this study $r_E \in [23.8, 557.9]$, so $1/(r_0-1) \le 0.044$: the index cannot resolve differences of drift beyond that, and the drift has to carry the size. □

THEOREM 3 (DISPLACEMENT AND CAPACITY)

The bound needs the fine-tune to be modeled, and we state the model rather than leave it implicit. Let P be the input distribution of the target and $\mathcal{H} = L^2(P)$ the Hilbert space of square-integrable output displacements, with $\|g\|^2 = \mathbb{E}_{x\sim P}\, g(x)^2$.

Assumption 1. *(i) The fine-tuned predictor is $f = f_0 + \Delta$, where f_0 is the released checkpoint and Δ the adapter's contribution; the contributions a rank-r adapter can express form a closed linear subspace $\mathcal{F}_r \subseteq \mathcal{H}$ under a first-order expansion in the adapter parameters, with $\mathcal{F}_r \subseteq \mathcal{F}_{r'}$ for $r \le r'$. (ii) The target is $f^\star$ and the required displacement is $D = f^\star - f_0 \in \mathcal{H}$; its norm is the quantity the pre-update mismatch x_1 measures. (iii) The loss is squared error, $\ell(f, x) = (f(x) - f^\star(x))^2$, the training loss reaches zero on the n training inputs, and held-out inputs are drawn from the same P.*

The following lemma says that nothing in the subspace gets closer to the target than the projection does. Whatever the adapter cannot represent is a floor on the error.

Lemma 3 (Projection). *For a closed subspace $\mathcal{F} \subseteq \mathcal{H}$ with orthogonal projection Π, every $g \in \mathcal{F}$ satisfies $\|D - g\|^2 = \|D - \Pi D\|^2 + \|\Pi D - g\|^2 \ge \|D - \Pi D\|^2$.*

Seen in the runs. Raising the adapter rank raises what the adapter can represent and lowers the floor. At rank 64 the 14B pair ends at a test-gap ratio of 1.254, against 1.551 at rank 16 (Appendix T).

Proof. $D - g = (D - \Pi D) + (\Pi D - g)$, and the first summand is orthogonal to $\mathcal{F}$ while the second lies in it; expand the square. □

Proof of Theorem 3. By Assumption 1(iii) the excess loss on a held-out input x is $(f(x)-f^\star(x))^2 = (\Delta(x) - D(x))^2$, so

$$\mathbb{E}\,[\text{excess}] \;=\; \mathbb{E}_{x\sim P}\big(\Delta(x) - D(x)\big)^2 \;=\; \|\Delta - D\|^2. \tag{29}$$

Since $\Delta \in \mathcal{F}_r$, Lemma 3 with $g = \Delta$ gives $\|\Delta - D\|^2 \geq \|D - \Pi_r D\|^2$, which is Equation (14). Zero training loss fixes Δ on the n training inputs but not elsewhere; the bound holds for every $\Delta \in \mathcal{F}_r$, interpolating or not, and it is attained exactly when $\Delta = \Pi_r D$. Monotonicity in r: $\mathcal{F}_r \subseteq \mathcal{F}_{r'}$ implies $\|D - \Pi_{r'} D\| = \inf_{g\in\mathcal{F}_{r'}} \|D - g\| \leq \inf_{g\in\mathcal{F}_r} \|D - g\| = \|D - \Pi_r D\|$. Monotonicity in D: along a ray, $\|\lambda D - \Pi_r(\lambda D)\| = \lambda\|D - \Pi_r D\|$ for $\lambda \geq 0$, and in general $D \mapsto \|D - \Pi_r D\|$ is 1-Lipschitz because $I - \Pi_r$ is a contraction. □

Assumption 1(i) is the first-order reading of a low-rank adapter and (iii) idealizes the observed behavior, since the training loss in every run of this study reaches zero. For a loss that is only locally quadratic, such as cross-entropy near its minimum, the same argument gives the bound up to the strong-convexity constant. The bound buys two falsifiable predictions, and both are measured here. First, the drift rises with the pre-update mismatch, at an elasticity of $+1.53$ (Figure 5). Second, raising the adapter rank is consistent with a lower penalty. The same 14B pair ends at a test-gap ratio of 1.254 at rank 64 against 1.551 at rank 16 (Appendix T). The bound is proved under squared loss and a linearized adapter. It does not by itself fix the size of that change.

THEOREM 4 (CHAIN LENGTH)

Proof. Model a k-step target as a displacement that decomposes over steps, $D_k = \sum_{i=1}^{k} \delta_i$ with $\delta_i \in \mathcal{H}$ the disagreement between the checkpoint's preferred continuation and the target's at step i. The hypothesis of Theorem 4 is that the δ_i are independent with $\mathbb{E}\,\delta_i = 0$ and $\mathbb{E}\|\delta_i\|^2 = \delta^2$; independence and zero mean give $\mathbb{E}\langle\delta_i, \delta_j\rangle = 0$ for $i \neq j$, so

$$\mathbb{E}\,\|D_k\|^2 \;=\; \sum_{i=1}^{k} \mathbb{E}\|\delta_i\|^2 + 2\sum_{i<j} \mathbb{E}\langle\delta_i, \delta_j\rangle \;=\; k\,\delta^2. \tag{30}$$

Because $I - \Pi_r$ is linear, the same computation applies to the unrepresentable part: $\mathbb{E}\,\|D_k - \Pi_r D_k\|^2 = \sum_i \mathbb{E}\|(I - \Pi_r)\delta_i\|^2 = k\,\delta_\perp^2$ with $\delta_\perp^2 := \mathbb{E}\|(I - \Pi_r)\delta_1\|^2$. Taking expectations in Equation (14) gives $\mathbb{E}[\text{excess}] \geq k\,\delta_\perp^2$, linear in k at fixed rank. □

G QUANTITIES USED TO JUDGE OVERFITTING

Two directions outside the early-prediction line of Section 1 bear on this paper. One locates overfitting inside the network (Anonymous, 2026; Bobe III et al., 2026; Chu et al., 2025). The other supplies the quantities practitioners read today: early stopping on a held-out loss (Prechelt, 1998), sharpness (Keskar et al., 2017; Foret et al., 2021), memorization (Zhang et al., 2017; Carlini et al., 2021; Shokri et al., 2017; Tirumala et al., 2022), distillation (Hinton et al., 2015), the train-validation gap under distribution shift (Roelofs, 2019), weight-space measures (Jiang et al., 2020), calibration (Guo et al., 2017) and learning dynamics (Ren & Sutherland, 2025). Table A5 places them beside the index of this paper.

Table A5 places ORI-AUC beside the six quantities in current use, along the four properties Section 4 names.

H THE INDEX IN FULL

The index needs three ingredients: the ideal trajectory, the retained segment, and the area under the curve. This appendix gives each in full, with the cost of computing it.

The ideal trajectory. Let $\ell_1, \ldots, \ell_m$ be the training losses on the retained segment, $\ell_{(1)} \leq \cdots \leq \ell_{(m)}$ their order statistics, and $q_j = j/(m+1)$ the plotting positions. The ideal trajectory is the

Table A5: Quantities used to judge overfitting. *Log-only*: from loss curves alone. *Scale-free*: invariant to rescaling the loss. *Online*: from a prefix of the run. *Pre-update*: before the first step. *Verdict at* $E = 4$: of twenty pair runs, how many the quantity calls as the test gap does, each row under its own best threshold, each gap statistic under its best threshold chosen after the fact, our rule under thresholds fixed in advance. A dash marks a quantity a loss log does not supply. References in row order: Roelofs (2019); patience; Carlini et al. (2021); Shokri et al. (2017); Tirumala et al. (2022); Jiang et al. (2020); Guo et al. (2017).

Quantity	Log-only	Scale-free	Online	Pre-update	Verdict at $E = 4$
Train-validation loss gap	✓		✓		$18/20$
Patience-based early stopping	✓		✓		$18/20$
Memorization and membership tests		✓			—
Sharpness, path and spectral norms					—
Calibration error		✓	✓		—
ORI-AUC (this work)	✓	✓	✓		$18/20$
with the decomposition a, c	✓	✓	✓		$\mathbf{20/20}$
with the pre-update mismatch x_1	✓	✓	✓	✓	$\mathbf{20/20}$

The rows of this paper are the only $20/20$ reached under thresholds fixed in advance (Appendix U).

log-normal curve closest to the training losses in $\mathcal{W}_2$: with $Q_{\mu,\sigma}$ the log-normal quantile function,

$$(\mu^\star, \sigma^\star) = \arg\min_{\mu,\ \sigma>0} \sum_{j=1}^{m} \big(\ell_{(j)} - Q_{\mu,\sigma}(q_j)\big)^2, \tag{31}$$

and the ideal trajectory is $\mathcal{I} = \big(Q_{\mu^\star,\sigma^\star}(q_m), \ldots, Q_{\mu^\star,\sigma^\star}(q_1)\big)$, the fitted quantiles in decreasing order, so that $\mathcal{I}$ is a monotone descent. The fit starts at the median and the normalized median absolute deviation of $\log \ell$ and is refined by Nelder–Mead. The log-normal reference is motivated by the multiplicative structure of optimization noise, whose products are log-normal in the limit. A run whose ordered training losses are exactly log-normal coincides with its own ideal trajectory and carries $a = 0$.

Table A6: Notation.

$\ell_t,\ v_t$	training and validation loss at epoch t; E epochs in the run
$\mathcal{T},\ \mathcal{V}$	empirical distributions of the training and validation losses on the retained segment
$\mathcal{I}$	ideal trajectory: log-normal $\mathcal{W}_2$ projection of the training losses, in decreasing order
$a,\ b,\ c$	$\mathcal{W}_2(\mathcal{I}, \mathcal{T})$ roughness; $\mathcal{W}_2(\mathcal{I}, \mathcal{V})$; $\mathcal{W}_2(\mathcal{T}, \mathcal{V})$ drift
$\tau,\ s,\ m$	first epoch with $v_t > \ell_t$; $s = \tau - 1$; retained length $m = E - s$
$\mathrm{ORI}(t)$	$\max(0,\ 1 - a_t/b_t)$ on the retained segment
ORI-AUC, ORI-AUC	trapezoidal average of $\mathrm{ORI}(t)$ over the segment, and over its first k epochs
x_1	per-token loss of the unmodified checkpoint on the target format, before any update
$r_t,\ r_E$	the ratio c_t/a_t at retained epoch t, and its value at the final epoch
$w_{m,j},\ o_j,\ \underline{o}_j,\ \overline{o}_j$	trapezoid weights; ORI at the j-th retained epoch and its two bounds (Prop. 1)
$\ell_{(j)},\ q_j,\ Q_{\mu,\sigma}$	ordered retained training losses; plotting positions $j/(m+1)$; log-normal quantile function
$\rho,\ r_{\log}$	Spearman rank correlation; Pearson correlation of logarithms
A, B	the reasoning-distilled or domain-adapted member of a pair; the twin released from the same base

The retained segment. Overfitting is measured only where a test gap exists. Let τ be the first epoch at which the validation loss exceeds the training loss, put $s = \tau - 1$, and retain epochs $s + 1, \ldots, E$, a segment of length $m = E - s$. All three distances of Equation (4) are computed on the retained segment. In every run of this paper $\tau = 2$, except the PubMedQA pair of Appendix V, which has $\tau = 3$; so $m = E - \tau + 1$. A run whose validation loss never exceeds its training loss has no retained segment. The index then reports that no overfitting has been observed, which is the correct report for such a run.

The trajectory average. ORI-AUC is the trapezoidal mean of the per-epoch index over the retained segment,

$$\text{ORI-AUC} \;=\; \sum_{j=1}^{m} w_{m,j}\, o_j, \qquad w_{m,j} = \begin{cases} \dfrac{1}{2(m-1)}, & j \in \{1, m\}, \\ \dfrac{1}{m-1}, & \text{otherwise}, \end{cases} \tag{32}$$

where o_j is the index at the j-th retained epoch. The weights sum to one, so ORI-AUC lies in $[0, 1]$ whenever the o_j do.

Cost. In one dimension the 2-Wasserstein distance between two empirical distributions of equal size is the L^2 distance between their sorted vectors. Each of a, b, c therefore costs one sort. The whole index over a 50-epoch run is microseconds, and it needs the loss curves alone.

The earliest computable horizon. Two retained points define the trapezoidal average. We fixed $E = 4$ in advance so that every run's segment holds three points when $\tau = 2$ and the area rests on two trapezoids rather than one. Appendix J reads $E = 4$ for that reason.

I PROPOSITION 1 IN FULL

We restate the quantile isometry, which is what makes the one-dimensional case exact.

The following proposition says that reading two distributions through their quantile functions turns the transport distance into a plain distance between two curves.

Proposition 4 (Quantile isometry)**.** *For distributions μ, ν on $\mathbb{R}$ with finite second moment and quantile functions $Q_\mu, Q_\nu : (0, 1) \to \mathbb{R}$,*

$$\mathcal{W}_2(\mu, \nu) \;=\; \|Q_\mu - Q_\nu\|_{L^2(0,1)}.$$

Seen in the runs. The closed form this isometry makes available recovers every measured ORI to 1.1×10^{-16} on the released runs. This is standard (Villani, 2009; Santambrogio, 2015). The isometry means the space of one-dimensional distributions under $\mathcal{W}_2$ is isometric to a convex subset of the Hilbert space $L^2(0, 1)$, so $\mathcal{W}_2$ is a metric and the triangle inequality is available. The exponent 2 is what buys this: no other $\mathcal{W}_p$ gives a Hilbert structure.

Proof of Proposition 1. At every retained epoch t the three distances are $a_t = \mathcal{W}_2(\mathcal{I}, \mathcal{T})$, $b_t = \mathcal{W}_2(\mathcal{I}, \mathcal{V})$ and $c_t = \mathcal{W}_2(\mathcal{T}, \mathcal{V})$ on the window of the first t retained epochs. By the triangle inequality $b_t \le a_t + c_t$ and $b_t \ge c_t - a_t$, the latter positive when $c_t > a_t$. Hence

$$\frac{a_t}{a_t + c_t} \;\le\; \frac{a_t}{b_t} \;\le\; \frac{a_t}{c_t - a_t}, \qquad 1 - \frac{a_t}{c_t - a_t} \;\le\; 1 - \frac{a_t}{b_t} \;\le\; 1 - \frac{a_t}{a_t + c_t},$$

and dividing numerator and denominator by a_t gives $(r_t - 2)/(r_t - 1) \le \text{ORI}(t) \le r_t/(r_t + 1)$ with $r_t = c_t/a_t$; for $r_t \le 2$ the lower bound is at most zero and the truncation $\max(0, \cdot)$ in Equation (5) takes over. The upper bound $r/(r + 1)$ has derivative $1/(r + 1)^2 > 0$ and the lower bound $(r - 2)/(r - 1) = 1 - 1/(r - 1)$ has derivative $1/(r - 1)^2 > 0$, so both increase in r. Their difference is

$$\frac{r}{r+1} - \frac{r-2}{r-1} \;=\; \frac{r(r-1) - (r-2)(r+1)}{(r+1)(r-1)} \;=\; \frac{2}{r^2 - 1}.$$

□

Proof of Corollary 1. With the weights $w_{m,j}$ of Appendix H, which are positive and sum to one, $\sum_j w_{m,j}\, \underline{o}_j \le \text{ORI-AUC} \le \sum_j w_{m,j}\, \overline{o}_j$, where $\underline{o}_j$ and $\overline{o}_j$ are the two bounds of Equation (9) at the j-th retained epoch. If the interval of run A lies above the interval of run B, every value ORI-AUC_A can take exceeds every value ORI-AUC_B can take. □

The index is the ratio c/a read through a monotone window of width $2/(r^2 - 1)$. On every run of this study the final ratio lies between 23.8 and 557.9, so the width at the final epoch is below 0.004. Table A13 lists the ratio for the twelve pooled runs. The 14B pair's intervals are disjoint on both seeds and on FinQA: $[0.932, 0.955]$ against $[0.977, 0.979]$ on seed 0 of GSM8K. That is the hypothesis of

Corollary 1. On every one of the twenty-two pair runs the ordering of ORI-AUC is the ordering of the final ratio.

J THE PAIR VERDICT IS IN THE LOGS BY THE FOURTH EPOCH

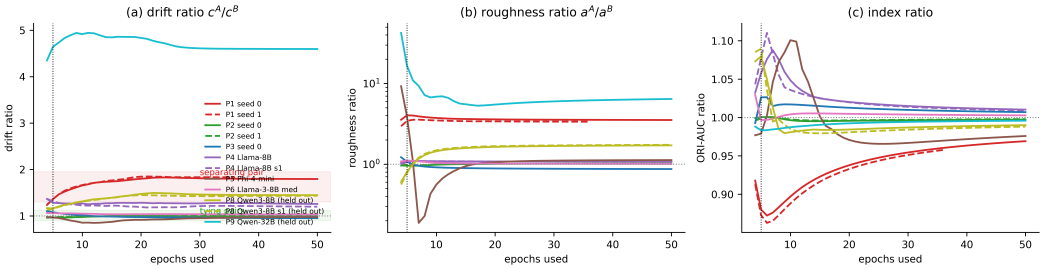

Figure A1: The pair verdict against epochs observed, $E = 4$ to 50. Red: the separating 14B pair (solid seed 0, dashed seed 1); green: the two HuatuoGPT pairs; blue: the BioMistral pair; dotted line $E = 5$. (a) Drift ratio, bands shaded. (b) Roughness ratio, log axis. (c) ORI-AUC ratio. Each panel's verdict at $E = 4$ is its verdict at $E = 50$.

Table A9 recomputes each pair's verdict from a prefix of the run, for three statistics: the drift ratio c^A/c^B, the ORI-AUC ratio, and the ratio of raw validation gaps $(v_E - \ell_E)^A/(v_E - \ell_E)^B$. The last is what a patience rule reads. Figure A1 plots the first continuously. The drift ratio places the separating and the matching pairs in disjoint bands at every horizon from $E = 10$, one fifth of the budget. The two coordinates together do so from $E = 4$. The separating pairs are the Qwen-14B, the Llama-8B and the two held-out pairs. The matching pairs are the four Qwen and Mistral medical pair runs, the two Phi seeds, the Llama-3 medical pair, the Stratos pair and the 1.5B raw-base pair. The ORI-AUC ratio reads the channel rather than the size (Section 5) and the raw validation-gap ratio reads the size in nats; both blocks are tabulated for the record.

The horizon $E = 4$ is where the comparison begins because it is where the index begins. A retained segment of three points is the first on which the trapezoidal average is defined. Nothing is chosen in starting there. What the index adds at that horizon is a decision rule with a stated false-alarm rate. The index is bounded in $[0, 1]$ and reads "no difference" as unity by construction. A null distribution for it can therefore be simulated directly at every retained length rather than assumed. The index is bounded and calibrated, which makes it the instrument for the first horizon, where a decision to stop, switch or continue is cheapest.

The null is simulated directly. Table A7 draws exchangeable pairs: two members sharing one systematic gap and differing only in independent noise. Each draw is conditioned on both members carrying a retained segment at $E = 4$, as every real run does. The table records the 95th and 99th percentiles of $|\log(\text{ratio})|$ for each statistic, across gap levels and noise levels. The index's 95% threshold lies between 0.0014 and 0.034 across the six cells. The index is the bounded coordinate, and one threshold serves it at every noise level; the quantities carried in nats are read against the cell that matches the run. This is what the dimensionless coordinate buys. One index threshold holds the false-alarm rate at or below 5% in every cell, at 0.034. At $E = 4$ the two coordinates are read together, the drift for the size and the index for the channel. From $E = 10$ the drift ratio carries the reading alone (Table A9). The index therefore carries the calibrated rule and the drift ratio the size, which is the division of labor of Section 5. The index threshold is the calibrated coordinate, and the two thresholds are calibrated against the noise the runs actually show. Every pair was run on two seeds, which gives a direct, model-free reading of that noise at $E = 4$ (Table A8). Across the ten pairs the two seeds' log drift ratios differ by 0.3% to 5.0%, a per-run standard deviation of 0.0189; the drift threshold $\log 1.15 = 0.140$ lies seven such deviations away. The two seeds' index readings differ by at most 0.022, inside the 0.034 threshold on every pair. On every pair both seeds return the same verdict at $E = 4$, and the same verdict at the full horizon. Appendix U compares the drift against three simpler aggregates of the same gap. The threshold is calibrated at $E = 4$, the horizon at which the decision is taken. Later horizons are read from Table A9 directly.

A pair whose members fine-tune equivalently is reported that way at every horizon. The Mistral-7B pair reads *matches* at every horizon under all three statistics, on both seeds. This is the behavior a

Table A7: Exchangeable-pair null at $E = 4$: 95th and 99th percentiles of $|\log(\text{ratio})|$ for the ORI-AUC, drift and raw-gap ratios, by systematic gap G and noise σ, 600 accepted pairs per cell, both members conditioned on $\tau = 2$; *accept* is the fraction of draws meeting it. The index's 95th percentile is at or below 0.0341 in every cell; the threshold 0.034 is that value rounded down. The drift ratio's 95% cutoff runs from 0.050 to 0.723 and the raw gap's from 0.056 to 1.06. `pair_null.py` regenerates the table.

			ORI-AUC ratio		drift ratio		raw-gap ratio	
G	σ	accept	95%	99%	95%	99%	95%	99%
0.30	0.02	1.00	0.0044	0.0062	0.145	0.190	0.177	0.242
0.30	0.05	0.74	0.0055	0.0112	0.374	0.488	0.481	0.618
0.30	0.10	0.33	0.0341	0.695	0.723	0.939	1.065	1.788
1.00	0.02	1.00	0.0015	0.0019	0.050	0.063	0.056	0.075
1.00	0.05	1.00	0.0014	0.0018	0.125	0.157	0.133	0.168
1.00	0.10	0.95	0.0015	0.0020	0.246	0.337	0.282	0.363

Table A8: Seed replication of the two coordinates at $E = 4$, ten pairs, two independent fine-tuning seeds each. Drift is $\log \rho_c$ and index is $|\log \rho_I|$; the last column is the same verdict from both seeds. The drift threshold is $\log 1.15 = 0.140$, the index threshold 0.034.

pair	drift, $\log \rho_c$		$\lvert\Delta\rvert$	index, $\lvert\log \rho_I\rvert$		$\lvert\Delta\rvert$	verdict
P1	$+0.212$	$+0.218$	0.006	0.086	0.091	0.005	separates
P4	$+0.312$	$+0.314$	0.003	0.032	0.042	0.011	separates
P8	$+0.156$	$+0.125$	0.031	0.071	0.082	0.011	separates
P9	$+1.470$	$+1.421$	0.049	0.012	0.009	0.003	separates
P2	-0.038	-0.026	0.012	0.001	0.004	0.004	matches
P3	$+0.100$	$+0.107$	0.007	0.007	0.025	0.018	matches
P5	-0.024	$+0.002$	0.026	0.023	0.031	0.008	matches
P6	$+0.062$	$+0.051$	0.011	0.029	0.007	0.022	matches
P7	$+0.127$	$+0.094$	0.033	0.003	0.001	0.002	matches
P11	$+0.032$	-0.009	0.040	0.007	0.008	0.002	matches

band verdict should have: what Corollary 1 orders is a margin, and the reported verdict acts on that margin.

The separating behavior belongs to the post-training regime and is present at both parameter counts. The last row of Table A9 compares a 7B reasoning checkpoint with a 7B instruction checkpoint on the same task. It separates by 1.166 at $E = 4$ and 3.214 at $E = 50$, more strongly than the 14B pair, and both members again carry $\tau = 2$. Its two members are built from different bases, so it enters as a supporting comparison.

Table A9: The pair verdict from the first E epochs, twenty pair runs on nine base models. *Drift ratio* c^A/c^B carries the verdict; *ORI-AUC ratio* names the channel; *raw validation-gap ratio* is what a patience rule reads. From $E = 10$ the two groups occupy disjoint bands. The last row of each block is a 7B reasoning against a 7B instruction checkpoint on unmatched bases, for support. Blank cells lie beyond a run's length.

Pair	E=4	E=5	E=10	E=15	E=20	E=30	E=40	E=50	group
Drift ratio c^A/c^B									
P1 seed 0	1.236	1.359	1.657	1.742	1.801	1.811	1.799	1.795	separates
P1 seed 1	1.243	1.360	1.665	1.764	1.846	1.834			separates
P4 seed 0	1.366	1.284	1.270	1.275	1.283	1.270	1.272	1.258	separates
P4 seed 1	1.369	1.313	1.226	1.222	1.191	1.212	1.210	1.199	separates
P8 seed 0	1.169	1.162	1.314	1.392	1.474	1.462	1.450	1.448	separates
P8 seed 1	1.134	1.145	1.304	1.388	1.450	1.426	1.430	1.434	separates
P9 seed 0	4.350	4.639	4.917	4.846	4.835	4.610	4.598	4.596	separates
P9 seed 1	4.143	4.426	4.611	4.554	4.429	4.395	4.396	4.395	separates
P2 seed 0	0.963	0.965	0.975	0.997	1.000	0.992	0.989	0.988	matches
P2 seed 1	0.974	0.973	0.985	0.993	0.996	1.002	1.007	1.008	matches
P3 seed 0	1.105	1.080	1.010	0.987	0.974	0.962	0.954	0.952	matches
P3 seed 1	1.113	1.107	1.024	0.998	0.990	0.981	0.974	0.971	matches
P5 seed 0	0.976	0.966	0.854	0.871	0.915	0.952	0.970	0.979	matches
P5 seed 1	1.002	0.987	0.884	0.858	0.906	0.935	0.959	0.971	matches
P6 seed 0	1.064	1.062	1.048	1.036	1.037	1.036	1.033	1.031	matches
P6 seed 1	1.052	1.065	1.051	1.039	1.029	1.024	1.019	1.017	matches
P7 seed 0	1.136	1.138	1.148	1.110	1.102	1.124	1.143	1.145	matches
P7 seed 1	1.098	1.090	1.087	1.103	1.111	1.092	1.093	1.093	matches
P11 seed 0	1.032	1.029	1.017	1.010	0.996	0.994	0.987	0.984	matches
P11 seed 1	0.991	1.049	1.017	0.994	0.963	0.969	0.974	0.975	matches
margin between the two groups									
	−0.002	0.008	0.078	0.112	0.079	0.088	0.067	0.054	
ORI-AUC ratio									
P1 seed 0	0.918	0.880	0.897	0.923	0.938	0.954	0.963	0.969	separates
P1 seed 1	0.913	0.873	0.888	0.916	0.933	0.951			separates
P4 seed 0	1.032	1.058	1.048	1.025	1.020	1.015	1.012	1.010	separates
P4 seed 1	1.043	1.078	1.036	1.025	1.020	1.014	1.011	1.009	separates
P8 seed 0	1.073	1.080	0.981	0.984	0.984	0.987	0.989	0.990	separates
P8 seed 1	1.085	1.090	0.985	0.979	0.981	0.984	0.986	0.988	separates
P9 seed 0	0.988	0.983	0.987	0.990	0.992	0.994	0.995	0.996	separates
P9 seed 1	0.991	0.981	0.983	0.987	0.989	0.992	0.993	0.994	separates
P2 seed 0	0.999	1.000	0.996	0.995	0.996	0.996	0.997	0.997	matches
P2 seed 1	0.996	0.999	0.998	0.996	0.996	0.997	0.998	0.998	matches
P3 seed 0	0.993	1.026	1.017	1.015	1.013	1.010	1.008	1.007	matches
P3 seed 1	0.975	0.979	1.016	1.016	1.014	1.011	1.009	1.008	matches
P5 seed 0	0.977	0.979	1.101	1.002	0.971	0.967	0.972	0.976	matches
P5 seed 1	0.969	0.964	1.138	1.025	0.982	0.972	0.975	0.978	matches
P6 seed 0	1.029	0.997	1.004	1.005	1.005	1.004	1.003	1.003	matches
P6 seed 1	0.993	0.989	1.000	1.004	1.004	1.003	1.003	1.002	matches
P7 seed 0	0.997	0.997	0.998	0.999	0.999	0.999	0.999	0.999	matches
P7 seed 1	1.000	0.999	0.999	0.999	0.999	1.000	1.000	1.000	matches
P11 seed 0	0.993	0.999	1.015	1.009	1.008	1.006	1.005	1.004	matches
P11 seed 1	0.992	1.000	1.008	1.008	1.007	1.005	1.004	1.004	matches
Raw validation-gap ratio									
P1 seed 0	1.315	1.537	1.860	1.793	1.945	1.757	1.785	1.785	separates
P1 seed 1	1.334	1.523	1.802	1.833	1.965	1.836			separates
P4 seed 0	1.336	1.221	1.309	1.291	1.331	1.271	1.233	1.223	separates
P4 seed 1	1.306	1.278	1.155	1.288	1.080	1.286	1.179	1.173	separates
P8 seed 0	1.083	1.157	1.425	1.493	1.656	1.422	1.441	1.443	separates
P8 seed 1	1.058	1.157	1.385	1.478	1.524	1.423	1.445	1.447	separates
P9 seed 0	4.993	5.166	4.701	4.787	4.551	4.539	4.591	4.590	separates
P9 seed 1	4.765	4.896	4.879	4.782	4.365	4.403	4.390	4.397	separates
P2 seed 0	0.961	0.968	0.977	1.010	1.017	0.987	0.987	0.988	matches
P2 seed 1	0.967	0.971	0.996	1.015	0.977	1.023	1.012	1.012	matches
P3 seed 0	1.090	1.057	0.992	0.990	0.950	0.928	0.946	0.947	matches
P3 seed 1	1.113	1.116	1.005	0.976	0.968	0.959	0.963	0.963	matches
P5 seed 0	0.995	0.952	0.785	0.909	0.932	0.984	0.998	0.997	matches
P5 seed 1	1.023	0.958	0.811	0.915	0.956	0.991	0.995	0.994	matches
P6 seed 0	1.045	1.057	1.031	1.031	1.025	1.021	1.025	1.025	matches
P6 seed 1	1.032	1.084	1.070	1.011	1.030	1.012	1.012	1.012	matches
P7 seed 0	1.116	1.150	1.167	1.106	1.097	1.190	1.153	1.148	matches
P7 seed 1	1.070	1.086	1.102	1.081	1.066	1.092	1.094	1.095	matches
P11 seed 0	1.044	1.028	1.005	0.987	0.986	0.975	0.977	0.978	matches
P11 seed 1	0.950	1.113	1.009	0.922	0.922	0.997	0.977	0.976	matches
Supporting comparison, unmatched bases, drift ratio									
R1-Dist-7B / Qwen-7B-Inst	1.166	1.355	2.300	2.723	2.845	3.098	3.194	3.214	separates

Table A10: Matched pairs P2, P3 and P6 on MedMCQA, three families, and *P2 repeated on PubMedQA, where its A member has no documented overlap. A is the domain-adapted member, B the instruction-tuned twin; ratios are A/B. Drift and gap lie within 6.1% of unity and the index within 1% on both seeds; the pre-update ratio is within 4.9% in every row. P2's overlap and the PubMedQA re-run: Appendices Q and V.

Pair	A / B (seed)	x_1 ratio	a ratio	c ratio	ORI-AUC (A/B)	gap
P2	Huatuo-o1-7B / Qwen2.5-7B-Inst (0)	1.034	1.018	0.988	0.887 / 0.890	0.987
P2	Huatuo-o1-7B / Qwen2.5-7B-Inst (1)	1.032	1.017	1.008	0.885 / 0.887	1.010
P2*	Huatuo-o1-7B / Qwen2.5-7B-Inst (0)	1.048	1.037	0.993	0.928 / 0.931	0.998
P3	BioMistral-7B / Mistral-7B-Inst (0)	1.013	0.872	0.952	0.941 / 0.934	0.940
P3	BioMistral-7B / Mistral-7B-Inst (1)	1.000	0.886	0.971	0.929 / 0.922	0.964
P6	OpenBioLLM-8B / Llama-3-8B-Inst (0)	1.048	1.009	1.031	0.903 / 0.900	1.001
P6	OpenBioLLM-8B / Llama-3-8B-Inst (1)	1.049	1.009	1.017	0.901 / 0.899	1.006

K PRE-UPDATE RATIOS OF EVERY PAIR RUN

The one boundary case, and why a replicate settles it. The Llama-3.1 pair on FinQA is the single comparison whose two readings fall on opposite sides of their cuts. Its test-gap ratio of 1.053 clears the 1.05 grouping line by 0.3%, while its drift ratio of 1.056 stays well inside the matched band below 1.15. That margin is smaller than what one run can resolve. Across the ten two-seed pairs the test-gap ratio moves by 0.4% to 7.3% between seeds, and on GSM8K this very pair moves 4.3%. A 0.3% margin measured once therefore does not place a pair on one side of the line. This is the resolution limit a replicate removes, and it is why the pre-registered scope scores only pairs that carry one.

Table A11 lists, for each of the twenty-two matched-pair runs, the pre-update ratio measured on the unmodified checkpoints next to the roughness, drift and test-gap ratios measured after fine-tuning. Figure 3 plots the drift and test-gap columns.

L COMPONENT DECOMPOSITION OF EVERY RUN

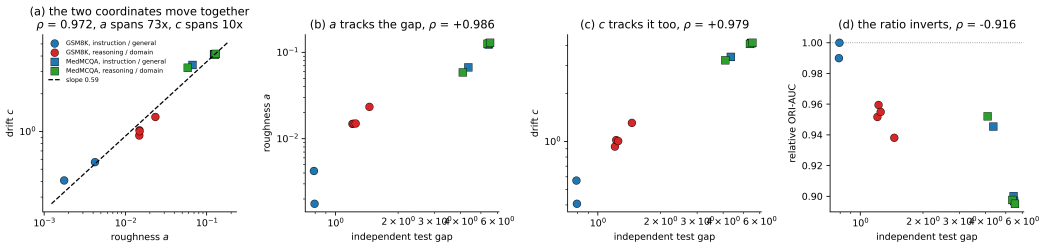

Figure A2: The decomposition against the independent test gap, twelve runs, color by post-training regime, marker by task. (a) Drift against roughness, log axes: collinear, $r_{\log} = +0.986$. (b) Roughness against the gap, $\rho = +0.986$. (c) Drift against the gap, $\rho = +0.979$. (d) Relative ORI-AUC against the gap: the index falls as the gap rises, $\rho = -0.916$. The components carry the magnitude and the index the scale-free verdict (Table A12).

Table A13 lists the roughness, the drift and their ratio for each of the twelve single-task runs, sorted by the ratio. It also gives the rank of each run under the ratio and under the index. Within each task family the two rankings coincide at every position; the four positions at which they cross are where the two task families interleave.

M THE RELATIVE SCALE IS THE SAME UNDER ANY REFERENCE

Relative ORI-AUC divides by the value of a fixed reference run, so a different reference multiplies every entry by a constant and preserves every ordering. We confirmed this numerically: taking each of the twelve runs in turn as the reference gives twelve rank orderings, and all twelve are identical.

Table A11: What is set before the first gradient step, twenty-two scored pair runs on three tasks and nine base models, ordered by pre-update ratio; the two single-seed FinQA comparisons are listed below the rule, reported and not scored. The pre-update ratio is measured on the unmodified checkpoints, the other columns at each pair's common horizon. The drift ratio gives the test-gap verdict on all twenty-two ($r_{\log} = +0.930$, 95% interval $[+0.895, +0.987]$). The last two rows repeat x_1 on the 256 hardest questions.

Pair	Task	x_1 ratio (pre-update)	a ratio	c ratio	gap ratio
P5 seed 1	GSM8K	**3.602**	1.106	0.971	1.010
P5 seed 0	GSM8K	**3.595**	1.126	0.979	0.999
P4 seed 1	GSM8K	2.024	1.075	1.199	1.087
P4 seed 0	GSM8K	2.022	1.078	1.258	1.136
P1 seed 1	GSM8K	1.187	3.363	1.833	1.558
P1 seed 0	GSM8K	1.186	3.532	1.795	1.551
P8 seed 0	GSM8K	1.121	1.713	1.448	1.313
P8 seed 1	GSM8K	1.122	1.740	1.434	1.332
P9 seed 0	GSM8K	1.320	6.435	4.596	1.706
P9 seed 1	GSM8K	1.318	9.129	4.395	1.841
P6 seed 1	MedMCQA	1.049	1.009	1.017	1.006
P6 seed 0	MedMCQA	1.048	1.009	1.031	1.001
P2 seed 0	MedMCQA	1.034	1.018	0.988	0.987
P2 seed 1	MedMCQA	1.032	1.017	1.008	1.010
P3 seed 0	MedMCQA	1.013	0.872	0.952	0.940
P3 seed 1	MedMCQA	1.000	0.886	0.971	0.964
P2 PubMedQA s0	PubMedQA	1.048	1.037	0.993	0.998
P2 PubMedQA s1	PubMedQA	1.050	1.039	1.015	1.015
P7 seed 0	GSM8K	0.953	1.182	1.145	1.002
P7 seed 1	GSM8K	0.952	1.126	1.093	0.964
P11 seed 0	GSM8K	0.786	0.933	0.984	0.970
P11 seed 1	GSM8K	0.786	0.924	0.975	0.989
P4, single seed	FinQA	1.253	**0.908**	1.056	1.053
P1, single seed	FinQA	**0.937**	**0.788**	1.221	1.319
P1, hardest 256, GSM8K		1.236	within 4.2% of the full-set 1.186		
P1, hardest 256, FinQA		0.942	within 0.5% of the full-set 0.937		

Table A12: Spearman correlation of each quantity with the independent test gap over the twelve single-task runs (x_1 rows: the ten that recorded it), with 95% bootstrap intervals from 20,000 resamples. On these runs the separation grows through the roughness channel, so the index falls as the gap rises.

Quantity	Spearman ρ vs. test gap	95% interval	Reads
Roughness a	$+0.986$	$[+0.887, +1.000]$	size
Drift c	$+0.979$	$[+0.870, +1.000]$	size
Ratio c/a	-0.965	$[-1.000, -0.806]$	composition
Relative ORI-AUC	-0.916	$[-1.000, -0.650]$	composition, on one scale
Pre-update mismatch x_1	$+0.952$ (vs. a)	$[+0.761, +1.000]$	sets a before the run
	$+0.927$ (vs. c)	$[+0.612, +1.000]$	

The reference used in the main text is one fixed run, which places the twelve pooled entries in $(0, 1]$. That is a presentational choice, and nothing in the paper depends on it.

Table A13: Component decomposition, sorted by the ratio c/a. *Rank* c/a and *rank ORI* are the positions of the run under the two orderings. Within each task family the two agree at every position, $\rho = 1.000$ on six runs apiece. Pooled, the families occupy overlapping bands of c/a and interleave at four positions, marked †, which is the whole of the difference between the pooled $\rho = +0.965$ and the within-family 1.000.

Task	Checkpoint (seed)	a	c	c/a	relative	rank c/a	rank ORI
GSM8K	Qwen2.5-7B-Inst (0)	0.0018	0.4066	230.8	1.000	1	1
GSM8K	Qwen2.5-14B-Inst (0)	0.0042	0.5690	134.7	0.990	2	2
GSM8K	R1-Distill-Qwen-14B (0)	0.0149	1.0213	68.5	0.959	3	3
GSM8K	R1-Distill-Qwen-14B (1)	0.0150	1.0050	67.1	0.955	4	4
GSM8K	Qwen3-8B (0)	0.0148	0.9282	62.7	0.952	5	6†
GSM8K	R1-Distill-Qwen-7B (0)	0.0234	1.3067	55.9	0.938	6	8†
MedMCQA	BioMistral-7B (0)	0.0585	3.2276	55.2	0.952	7	5†
MedMCQA	Mistral-7B-Inst (0)	0.0671	3.3909	50.5	0.945	8	7†
MedMCQA	Qwen2.5-7B-Inst (0)	0.1234	4.1434	33.6	0.900	9	9
MedMCQA	HuatuoGPT-o1-7B (0)	0.1256	4.0947	32.6	0.898	10	10
MedMCQA	Qwen2.5-7B-Inst (1)	0.1272	4.1287	32.5	0.897	11	11
MedMCQA	HuatuoGPT-o1-7B (1)	0.1293	4.1634	32.2	0.895	12	12

N SEED AGREEMENT

Table A14 gives the fourteen checkpoints fine-tuned under two seeds.

Table A14: Seed agreement for fourteen of the checkpoints run twice; differences are absolute and *agree* is the relative difference in drift. *Both seeds read at $E = 36$. The index agrees within 0.013 and the drift within 6.8%, the two coordinates the rule reads, on runs whose ORI-AUC spans 0.796 to 0.974.

Checkpoint	ORI-AUC		$\lvert\Delta\rvert$	c		agree
	s0	s1		s0	s1	
R1-Distill-Qwen-14B (GSM8K)	0.9483	0.9438	0.0045	1.0213	1.0050	1.6%
Qwen2.5-14B-Instruct (GSM8K)*	0.9738	0.9704	0.0034	0.5211	0.5024	3.6%
HuatuoGPT-o1-7B (MedMCQA)	0.8872	0.8847	0.0024	4.0947	4.1634	1.7%
Qwen2.5-7B-Instruct (MedMCQA)	0.8896	0.8866	0.0030	4.1434	4.1287	0.4%
R1-0528-Qwen3-8B (GSM8K)	0.9313	0.9279	0.0034	1.3439	1.3128	2.3%
Qwen3-8B (GSM8K)	0.9406	0.9391	0.0015	0.9282	0.9153	1.4%
R1-Distill-Llama-8B (GSM8K)	0.9388	0.9373	0.0015	1.6327	1.5725	3.8%
Llama-3.1-8B-Instruct (GSM8K)	0.9292	0.9287	0.0005	1.2983	1.3114	1.0%
Phi-4-mini-reasoning (GSM8K)	0.8044	0.7959	0.0085	1.5514	1.5068	2.9%
Phi-4-mini-instruct (GSM8K)	0.8243	0.8134	0.0109	1.5850	1.5521	2.1%
BioMistral-7B (MedMCQA)	0.9410	0.9294	0.0116	3.2276	3.4473	6.8%
Mistral-7B-Instruct (MedMCQA)	0.9344	0.9220	0.0124	3.3909	3.5515	4.7%
OpenBioLLM-8B (MedMCQA)	0.9027	0.9006	0.0021	4.1644	4.1335	0.7%
Llama-3-8B-Instruct (MedMCQA)	0.9002	0.8986	0.0016	4.0399	4.0652	0.6%

O CONTROLS

None of the four confounders checked moves a verdict. Difficulty-matched splits raise the separation rather than create it (Appendix S). One paired checkpoint's post-training data touches its task's source split, so the pair was run again on a fourth task outside that data. Every quantity lands within 2.8% and the reading is the same. Every pooled statistic is repeated without it and without the one pooled checkpoint that later joined a held-out pair (Appendix V). Seed agreement across accelerators is in Appendix N.

P EVERY RUN

Table A15 lists the twelve single-task fifty-epoch runs that enter the pooled analyses of Section 5.

Table A15: The twelve single-task fifty-epoch runs of the pooled analyses; †P8's run was measured after those analyses were fixed, so the x_1 statistics use the other ten. x_1 is the pre-update mismatch; a, c the terminal roughness and drift; ORI-AUC the trajectory average, *relative* to the reference 0.9884 (Qwen2.5-7B-Instruct, GSM8K). *Gap* is test minus training loss at the end, never used by the index. Every run opens its retained segment at epoch 2; the PubMedQA pair at epoch 3, so it is read from $E = 5$.

Task	Checkpoint (seed)	x_1	a	c	ORI-AUC	relative	gap
GSM8K	Qwen2.5-7B-Inst (0)	0.8096	0.0018	0.4066	0.9884	**1.000**	0.794
	Qwen2.5-14B-Inst (0)	0.7168	0.0042	0.5690	0.9785	0.990	0.789
	R1-Distill-Qwen-14B (0)	0.8500	0.0149	1.0213	0.9483	0.959	1.224
	R1-Distill-Qwen-14B (1)	0.8496	0.0150	1.0050	0.9438	0.955	1.252
	Qwen3-8B (0)	0.9439†	0.0148	0.9282	0.9406	0.952	1.208
	R1-Distill-Qwen-7B (0)		0.0234	1.3067	0.9272	0.938	1.457
MedMCQA	BioMistral-7B (0)	2.3313	0.0585	3.2276	0.9410	0.952	4.093
	Mistral-7B-Inst (0)	2.3011	0.0671	3.3909	0.9344	0.945	4.355
	Qwen2.5-7B-Inst (0)	2.4985	0.1234	4.1434	0.8896	0.900	5.449
	Qwen2.5-7B-Inst (1)	2.5159	0.1272	4.1287	0.8866	0.897	5.478
	HuatuoGPT-o1-7B (0)	2.5827	0.1256	4.0947	0.8872	0.898	5.378
	HuatuoGPT-o1-7B (1)	2.5958	0.1293	4.1634	0.8847	0.895	5.534

Figure A4 shows the three loss curves and the per-epoch index for all twelve pooled single-task runs. The independent test curve is plotted for reference and never enters the index.

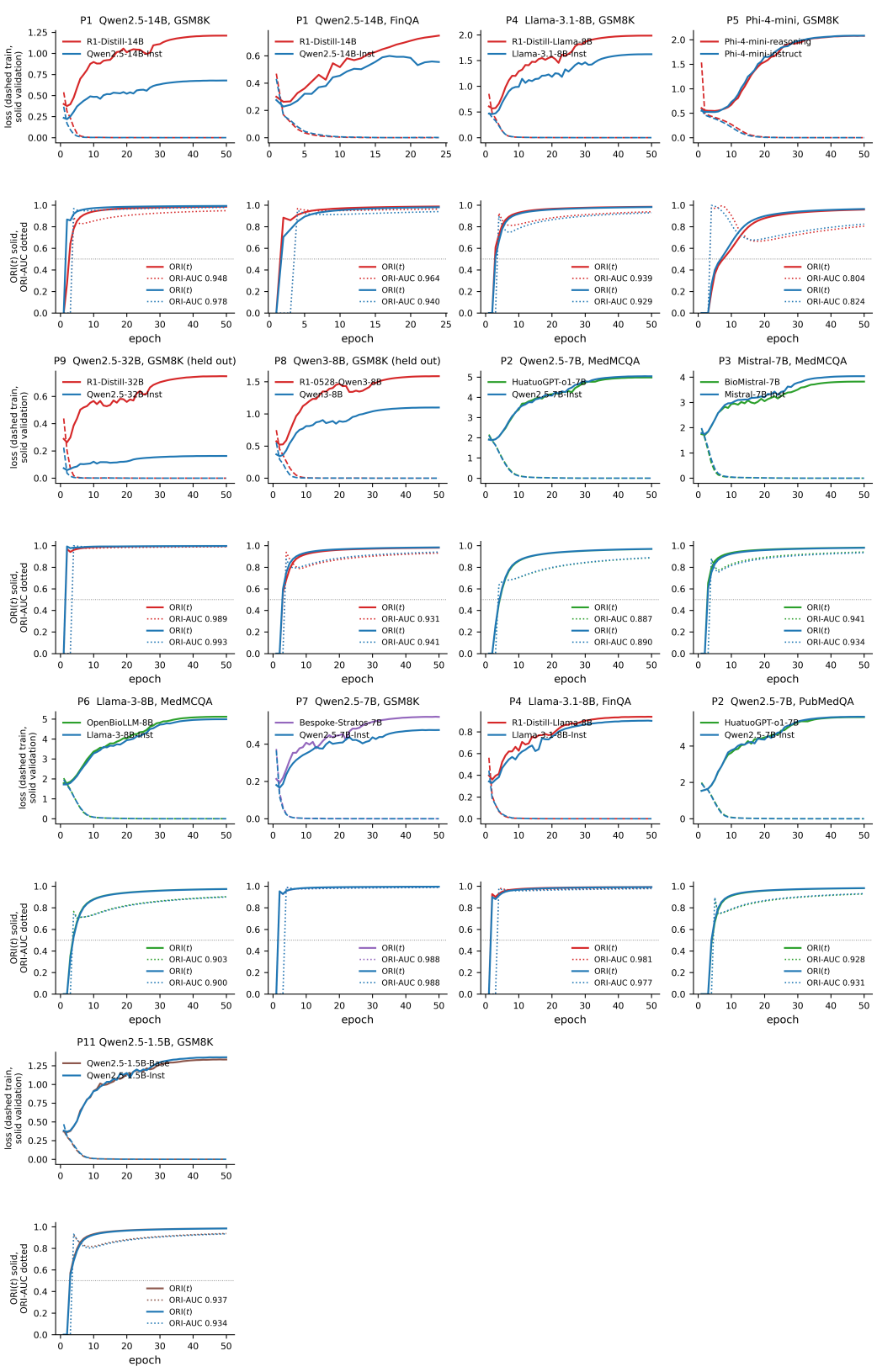

Figure A3: All thirteen pair-task combinations. Top of each block: training (dashed) and validation (solid) loss of both members. Bottom: the per-epoch index of Equation (5), with ORI-AUC in the legend.

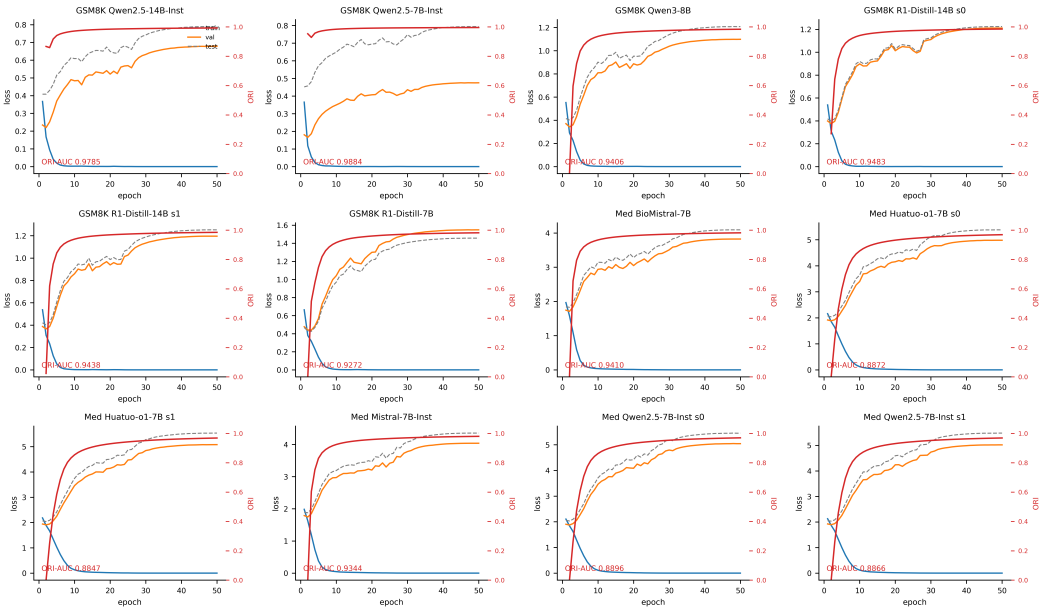

Figure A4: All twelve runs. Left axis: training (blue), validation (orange) and independent test (gray, dashed) loss. Right axis, red: the per-epoch index of Equation (5) over the retained segment, with its trajectory average printed in each panel. Top row and the first two panels of the middle row are GSM8K; the remainder are MedMCQA.

Q REPRODUCIBILITY

Table A16: The twenty checkpoints, four model families, and the twenty-seven checkpoint-task combinations behind the forty-nine runs. Rows between rules share a base model, tokenizer and parameter count and differ only in post-training; the last row is a single pooled run. *Regime*: the post-training that produced the checkpoint. *Seeds*: independent fine-tuning runs.

Pair	Checkpoint	Base	Par.	Post-training	Task	Sd.
P1	DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	14B	reasoning-distilled	GSM8K	2
P1	Qwen2.5-14B-Instruct	Qwen2.5-14B	14B	instruction-tuned	GSM8K	2
P1	DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	14B	reasoning-distilled	FinQA	1
P1	Qwen2.5-14B-Instruct	Qwen2.5-14B	14B	instruction-tuned	FinQA	1
P4	DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	8B	reasoning-distilled	GSM8K	2
P4	Llama-3.1-8B-Instruct	Llama-3.1-8B	8B	instruction-tuned	GSM8K	2
P4	DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	8B	reasoning-distilled	FinQA	1
P4	Llama-3.1-8B-Instruct	Llama-3.1-8B	8B	instruction-tuned	FinQA	1
P11	Qwen2.5-1.5B-Base	Qwen2.5-1.5B	1.5B	raw base	GSM8K	2
P11	Qwen2.5-1.5B-Instruct	Qwen2.5-1.5B	1.5B	instruction-tuned	GSM8K	2
P5	Phi-4-mini-reasoning	Phi-4-mini	3.8B	reasoning post-training	GSM8K	2
P5	Phi-4-mini-instruct	Phi-4-mini	3.8B	instruction-tuned	GSM8K	2
P6	Llama3-OpenBioLLM-8B	Llama-3-8B	8B	medical continued training	MedMCQA	2
P6	Llama-3-8B-Instruct	Llama-3-8B	8B	instruction-tuned	MedMCQA	2
P2	HuatuoGPT-o1-7B	Qwen2.5-7B	7B	medical reasoning	MedMCQA	2
P2	Qwen2.5-7B-Instruct	Qwen2.5-7B	7B	instruction-tuned	MedMCQA	2
P2	HuatuoGPT-o1-7B	Qwen2.5-7B	7B	medical reasoning	PubMedQA	2
P2	Qwen2.5-7B-Instruct	Qwen2.5-7B	7B	instruction-tuned	PubMedQA	2
P3	BioMistral-7B	Mistral-7B	7B	medical cont. pre-training	MedMCQA	2
P3	Mistral-7B-Instruct-v0.1	Mistral-7B	7B	instruction-tuned	MedMCQA	2
P7	Bespoke-Stratos-7B	Qwen2.5-7B	7B	reasoning SFT of instruct	GSM8K	2
P7	Qwen2.5-7B-Instruct	Qwen2.5-7B	7B	instruction-tuned	GSM8K	2
P9	DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B	32B	reasoning-distilled	GSM8K	2
P9	Qwen2.5-32B-Instruct	Qwen2.5-32B	32B	instruction-tuned	GSM8K	2
P8	DeepSeek-R1-0528-Qwen3-8B	Qwen3-8B-Base	8B	reasoning-distilled	GSM8K	2
P8	Qwen3-8B	Qwen3-8B-Base	8B	vendor reasoning	GSM8K	2
–	DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	7B	reas.-distilled	GSM8K	1

Fine-tuning. LoRA rank 16, $\alpha = 32$, dropout 0.05, applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. AdamW, learning rate 10^{-4}, cosine schedule with a linear warmup over the first 3% of steps, weight decay 0, gradient clipping at 1.0, bfloat16, maximum sequence length 512. The effective batch is 16 in every run. The micro-batch and accumulation that realize it follow the memory of the checkpoint: 8×2 for the 14B models, 4×4 for the 7–8B GSM8K models and 16×1 for the MedMCQA models. Fifty epochs over $1{,}319$ training examples with $1{,}319$ validation examples entering the index. The independent test split never enters it and holds the $1{,}319$ GSM8K test questions, the $1{,}740$ MedMCQA questions, the $1{,}097$ FinQA test questions or the 679 expert-annotated PubMedQA instances. Hardware: one GPU per run, an A100 40GB, an RTX 6000 Ada 48GB, or an RTX PRO 6000 Blackwell 96GB. The seed replicates of Appendix N were in part run on different accelerators from their first seeds, so the agreement they report covers hardware as well as seed. The two Meta checkpoints, Llama-3.1-8B-Instruct and Llama-3-8B-Instruct, were loaded from the `unsloth` and `NousResearch` mirrors, which carry the released weights unchanged. Every other checkpoint was loaded from its original repository. The exact repository of every run is recorded in its `_config.json`.

Splits. PubMedQA training and validation questions are drawn from the $211{,}269$ instances of the `pqa_artificial` configuration. Filtering to long answers between 200 and $1{,}200$ characters leaves $136{,}090$. The independent test split is the 679 instances of the expert-annotated `pqa_labeled` configuration that pass the same filter. Every `pqa_labeled` identifier is removed from the training pool before the draw. GSM8K questions are drawn from the training split by descending number of reasoning steps, so that the same $1{,}319$ hardest questions are used for every checkpoint. Appendix S reads the held-out pair again under training and validation splits of matched difficulty. Validation is a disjoint draw from the remainder under a fixed seed and test is the official GSM8K test split.

MedMCQA is filtered to questions carrying an explanation between 200 and 1,200 characters and drawn the same way. Both tasks use one terse target format, identical across every checkpoint.

Before the first optimizer step, the unmodified checkpoint is evaluated on the target-format solutions and the per-token loss is recorded as x_1. The subsample check of Table A11 repeats the measurement on the 256 hardest of those questions, a fivefold reduction in sample size restricted to the hardest end of the difficulty range. Those 256 questions are the first 256 of the ranked 1,319, so the check establishes that the ratio is stable under sample size and difficulty composition. The released `x1_probe.py` carries a `--skip` option that ranks 1,575 questions and drops the 1,319 used for fine-tuning, which measures the same ratio on 256 items no run has seen.

Verification. The released package carries `audit.py`, which recomputes every headline number, every numeric cell of Tables A15 and A9, and the intervals of Table A12 from the per-epoch logs. It prints a pass or a mismatch per number. All pass. Intervals are 95% percentile bootstrap intervals from 20,000 resamples of runs, or of pair runs for the pair-level correlations, with a fixed seed. Two figures in this appendix are regenerated by `make_supp_figs.py` from the same logs.

Accuracy check. The 14B and Llama-3.1 pairs were re-run under the protocol above, with greedy generation on the full GSM8K test set after epochs 4 and 50. The set holds 1,319 questions, generation is capped at 400 new tokens, and scoring is exact match on the final answer. The results: R1-Distill-Qwen-14B 81.65% then 78.85%, Qwen2.5-14B-Instruct 82.11% then 80.89%; R1-Distill-Llama-8B 60.1% then 56.2%, Llama-3.1-8B-Instruct 69.3% then 65.7%. In both pairs the drift ratio read at $E = 4$ names the member whose accuracy falls further. The logs are released with the run logs.

Checkpoint eligibility and documented overlap. A checkpoint is paired on a task when its documented post-training data is disjoint from the task's source split, and a documented overlap is reported either way. The rule admits Bespoke-Stratos-7B on GSM8K, whose mathematics comes from the AIME, MATH and Olympiad subsets of NuminaMath. It leaves OpenThinker-7B out on GSM8K. The OpenThoughts-114k mixture draws its mathematics from NuminaMath-CoT, a collection that contains 7,345 GSM8K problems. That checkpoint has therefore already trained on the questions a GSM8K fine-tune draws its training and validation sets from, and GSM8K is the task on which this paper reports separation. One paired checkpoint's post-training data touches its task's source split, and the direction of that effect is known. HuatuoGPT-o1-7B reports that its verifiable medical problems were built from the training splits of MedQA-USMLE and MedMCQA (Chen et al., 2024). Its A member therefore enters the MedMCQA fine-tune having seen part of that split. Prior exposure lowers the separation a member can show, so it can only move a pair toward *matches*. That is the verdict P2 receives. The pair is marked in Table A10 and run again on a task outside that data in Appendix V. Appendix V repeats every pooled statistic and every pair claim with P2 removed. DeepSeek reports the size and domain composition of the R1 distillation set, Microsoft the Phi-4-mini-reasoning mixture, and BioMistral and OpenBioLLM their medical corpora. None of these documents places a GSM8K, MedMCQA or FinQA split in the data. The matched pairs also carry their own check. The two 14B members score 81.65 and 82.11% on the official GSM8K test set at $E = 4$, level to within half a point. The separation that follows therefore grows during the fine-tune.

R THE INDEX ON PRE-TRAINING LOGS

The index reads a loss triple and nothing else, so a pre-training run that writes one is in scope; "epoch" in the definitions is any checkpoint index. This appendix places the index in the data-constrained pre-training regime of Muennighoff et al. (2023) at small scale, because it answers the question of scope directly. A byte-level GPT (6 layers, width 256, context 256, 4.8M parameters) is trained on WikiText-2 with a fixed budget of 4×10^7 tokens. The unique data is $4 \times 10^7 / R$ bytes for repetition factors $R \in \{4, 8, 16, 32, 64\}$. Every run therefore spends the same compute and differs only in how many passes it makes over its data. Forty checkpoints per run record the training loss, the held-out loss that enters the index and the loss on the official test split, which never enters it. There are three seeds per R, fifteen runs.

The drift c rises monotonically with the number of passes, from 0.020 at four passes to 0.411 at sixty-four. It tracks the independent test gap, which rises from 0.046 to 0.704 nats per byte

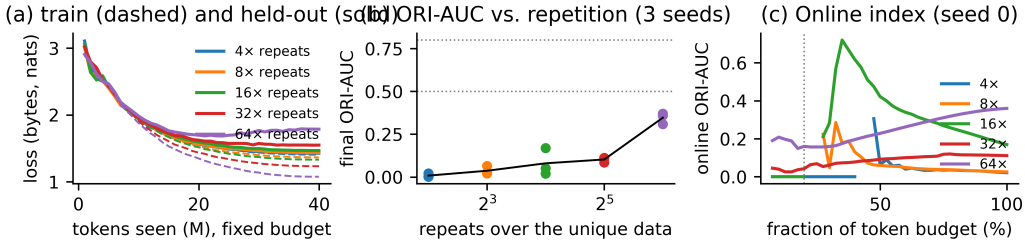

Figure A5: Data-constrained pre-training simulation, fifteen runs. (a) Training (dashed) and held-out (solid) loss against tokens seen for repetition factors $R = 4$ to 64 at a fixed token budget. (b) Final ORI-AUC against R, one point per seed. (c) Online ORI-AUC as a function of the fraction of the budget; the dotted line marks 20%.

over the same range (Figure A5). Across the fifteen runs $\rho(\text{ORI-AUC}, \text{test gap}) = +0.854$ and $\rho(\text{ORI-AUC}, \log R) = +0.873$. Here the overfitting grows through the drift channel, a smooth descent with the validation curve rising, so the index rises with the gap. The two signs, -0.916 in fine-tuning and $+0.854$ here, are the two channels of Section 5 read from the same ratio c/a. A pre-training run is a single trajectory with no twin to compare against, so the reading it needs is absolute. From this run alone, is the departure large enough to act on? Table A17 answers with the per-window index against a threshold of 0.5 fixed before the experiment. The scored horizons are those with retained length $m \geq 3$, as in Appendix H. On them $\text{ORI}(t)$ crosses the threshold in all three seeds at sixty-four passes, at checkpoints 27 to 28 of 40, where the final test gap is 0.70 nats. It stays below the threshold throughout at four, eight, sixteen and thirty-two passes. Over $3{,}600$ pure-noise trajectories the alarm $\max_t \text{ORI}(t) \geq 0.5$ fires zero times. The fine-tuning verdicts of the main text and this pre-training reading use the same index, the same threshold logic and the same inputs.

Table A17: Absolute alarm from a single pre-training run: per-window $\text{ORI}(t)$ against the pre-set threshold 0.5 on the scored horizons ($m \geq 3$), three seeds per repetition factor R. *At* is the checkpoint of the crossing, out of forty.

R	final test gap (nats/byte)	seeds with $\text{ORI}(t) \geq 0.5$	at
4	0.046	0/3	
8	0.072	0/3	
16	0.150	0/3	
32	0.327	0/3	
64	0.704	**3/3**	27–28

S A DIFFICULTY-MATCHED VALIDATION SPLIT

The training questions are the $1{,}319$ with the most reasoning steps and the validation questions are drawn from the remainder. The two splits therefore differ in difficulty: 5.443 steps on average against 2.687. This appendix rebuilds the splits so that they do not, and reads the pair again. GSM8K holds $1{,}238$ questions of five reasoning steps or more, so two matched sets of $1{,}319$ sit at a mean of 4.72 steps. `finetune_reasoning.py --match_difficulty` sorts the hardest $2{,}638$ questions by step count. It then walks them in blocks of two, alternating which set receives which member, so the two histograms agree. Table A18 gives the result. The measured difference in mean reasoning steps falls from 2.756 to 0.001. The protocol is otherwise P1's, unchanged, and the pair is the held-out Qwen3 pair.

Removing the difficulty difference raises the drift ratio, from 1.448 to 1.523 on the held-out Qwen3-8B pair and from 1.795 to 1.853 on the 14B pair. The independent test-gap ratio rises with it, from 1.313 to 1.404 on the Qwen3-8B pair and from 1.551 to 1.587 on the 14B pair. The reading on the original splits is the conservative one. The roughness ratio is unmoved, 1.712 against 1.713, so the channel of Section 5 is the same one, and the verdict is *separates* at every horizon from $E = 5$. The pre-update mismatch shows why the two splits can be compared. Under the matched split the untouched checkpoint scores 1.104 on the training questions and 1.097 on the validation questions,

Table A18: The held-out Qwen3 pair under the two splits. *Original*: training questions are the hardest 1,319, validation is drawn from the remainder. *Matched*: training and validation are drawn from one pool so that their reasoning-step histograms agree. Ratios are A/B at $E = 50$.

Quantity	Original	Matched
mean reasoning steps, train / validation	5.443 / 2.687	4.721 / 4.722
pre-update validation-training difference, A	0.284	−0.007
a ratio	1.713	1.712
c ratio	1.448	**1.523**
ORI-AUC ratio	0.990	0.989
independent test-gap ratio	1.313	**1.404**
drift ratio at $E = 10$	1.314	1.354

seven thousandths of a nat apart. Under the original split the gap is 0.284. What the matched split removes is present before the fine-tune starts, and what the drift reports grows during it.

T WHY THE OPENING EPOCHS ARE ENOUGH

The opening epochs carry the verdict because they are where the output distribution collapses onto one format. This appendix completes the account of Section 5.

A fine-tune of a released checkpoint is not the acquisition of a task. Every base here met grade-school arithmetic, clinical text and financial filings in pre-training, and both members of a pair met the same ones. What $1{,}319$ examples change is the *format*: which of the modes the model already carries gets emitted, and how the trace is laid out. The loss curve does not record learning. It records a distribution collapsing onto one format. That reframing is what makes an early reading possible, and it fixes three things.

The collapse is early and monotone. The training loss falls by an order of magnitude within four epochs, and every run but one separates its two curves at epoch 2 (Figure 2). Later epochs scale both coordinates; they do not turn a ratio around. A ratio is therefore settled once the collapse is under way, which is why the reading at $E = 4$ survives to $E = 50$ with a growing margin.

What differs between the members is fixed before the first step. Post-training decides how far the output distribution sits from the target format, which is x_1 and costs one forward pass, and how concentrated that distribution is. A reasoning-distilled checkpoint is the base trained by imitation alone on long traces from one teacher (DeepSeek-AI, 2025). Its mass sits on a narrow set of continuations. Its twin went through preference and reinforcement stages on broad data (Xu et al., 2025), so its mass is spread. Both must reach the same terse target, and the narrow one has further to move. The regimes that fine-tune like the twin are the ones that kept the base's breadth. A 17k-trace reasoning SFT of an instruction checkpoint (P7) leaves it fine-tuning like itself.

And the target admits many paths. Every target here is a multi-step trace of several hundred tokens, not a label. One worked solution is one derivation among many that reach the same answer, so the supervision is many-to-one. The loss is taken on a single path while the task admits a family of them. A model fitting a many-to-one target on $1{,}319$ examples need not learn the task; it can learn a path. Committing to a path is cheap, it lowers the training loss as fast as learning the task would, and it is available from the first epoch. That is why the two curves separate at epoch 2 rather than late, and why a few epochs are enough to see which member is committing. A narrower starting distribution has fewer paths to choose among, and it commits sooner.

The two coordinates read the two consequences of that move. A model that must rebuild its output distribution spends its epochs at losses a smooth descent would not visit, so its retained losses sit further from the log-normal reference. One that re-weights modes it already has does not. So a measures *how much of the model the fine-tune had to rebuild*, and c measures whether the move transfers, opening at the epoch the model begins fitting example-specific structure. One coordinate is set by the checkpoint, the other by the data, and the index reads their ratio (Proposition 1).

In words: overfitting under a small fine-tune is largely decided before the fine-tune starts, by how much of the checkpoint must be rebuilt to reach the target format. The opening epochs are enough because they are where the rebuilding happens.

The pre-update mismatch moves both coordinates, and it moves the roughness faster. Elasticities of 2.50 and 1.53 leave c/a an elasticity of -0.97. That is why the index is read between two members meeting the same target, rather than across checkpoints with different targets. When the target is itself a long trace the same distilled model starts closer and descends more smoothly (FinQA, Section 5), so the roughness follows the target and the checkpoint together.

Why a rebuild generalizes worse. Two reasons, and they are the reasons the two coordinates exist. Capacity: the adapter is fixed at rank 16 for both members. A checkpoint needing a small correction learns a map that transfers. One asking the same capacity for a large transformation on $1{,}319$ examples produces it one example at a time. Shape: a run that fits population structure spends its epochs on a smooth descent, and its losses follow the reference closely. A run that is rebuilding visits losses off that descent, and the distance a records it. Those excursions are progress on structure the next batch does not share. That is what a measures, and it is why the ideal trajectory is a monotone descent rather than a smoothed curve.

The capacity half is testable, and we tested it. The account above says the adapter rank is part of the cost, so raising it should let the distilled member finish the move it could not finish at rank 16.

We reran the 14B pair with the adapter rank at 64 and the alpha at 128, every other setting of P1's protocol held, including the seed. The prediction holds. With four times the capacity the flagged member ends closer to its twin, a test-gap ratio of 1.254 against 1.551. The threshold fixed at rank 16 still flags the pair at the fourth epoch.

The chain is then short enough to check at every link. Post-training fixes the distance and the concentration of the output distribution, measured by x_1 and by the regime contrast of Table A1. Those fix how much has to be rebuilt, measured by the roughness a from the opening epochs. The rebuild produces gains that do not transfer, measured by the drift c, which opens at epoch 2 in all but one run. The ratio of two members' drifts is therefore set by quantities fixed before the first gradient step, so the opening epochs carry the whole reading. The account commits to a direction and the data follow it. Distance and concentration set the size of the effect, not the parameter count. The largest pair in the study carries the widest margins, and the medical pairs, whose members start at equal distance, carry none.

U THE DRIFT AGAINST SIMPLER AGGREGATES OF THE SAME GAP

The drift c is a distance between the distributions of training and validation losses over the retained segment. Three simpler statistics summarize the same gap. The first is the gap read once at the current epoch, which is what a patience rule uses; the others are its mean and RMS over the epochs seen. They are averages of per-epoch differences. The drift is a distance between the two loss distributions, and the optimal coupling that defines it is the identity only when the two sorted orders agree. After τ the validation loss rises while the training loss falls, so the two orders are reversed. The drift is then strictly smaller than the RMS gap. The two are different statistics, which is why the four readings stay distinct at a short prefix. Each is scored over the twenty pair runs with a reading at every horizon.

Read head to head, the two horizons separate. At $E = 4$ the rule of Section 5, with both thresholds fixed in advance, gives the full-horizon verdict on all twenty runs. It does so on every family taken alone: four of four on Llama, two of two on Mistral, two of two on Phi and twelve of twelve on Qwen. Each gap statistic was then handed its *best achievable* threshold, searched over these same twenty runs after the fact. At $E = 4$ the mean of the per-epoch gap reaches fourteen, the RMS sixteen and the terminal gap eighteen. Among the statistics that measure a train-validation gap, the drift is the one that separates the two groups at $E = 5$, the earliest epoch count any recipe of Section 3 uses. `scripts/headtohead.py` prints the comparison at $E = 4$, 5, 10 and 50.

The drift also carries the roughness with it. The mean gap is one number; (a, c) is two, and the second coordinate names the channel of Section 5 and carries the held-out Qwen3 pair's second seed at $E = 4$.

Model selection under equal conditions. The practical question is which of two candidates to fine-tune. Every reading here picks the less-overfitting member on all eight separating runs. Validation loss, the train–validation gap, the drift and the rule agree on direction, so direction is not where the readings differ. The comparison that matters is whether a reading can also call a matched pair a tie, and at what false-alarm rate. The threshold must not be tuned on the runs being scored. Table A19 gives every reading the same tie mechanism, a band $|\log S| \le \tau$ on its ratio S. The band is calibrated out of family. For each of the four model families, τ is the largest $|\log S|$ observed on the matched runs of the other three, so the in-fold false-alarm rate is zero. The held-out family is then scored, and the four folds are summed. Truth is the independent test gap. For transparency the table also reports the best single band chosen after the fact on all twenty runs. A post-hoc band lets validation loss reach twenty of twenty. An out-of-family calibration does not reproduce that number. The frozen rule is scored with the thresholds of Equation (13), fixed before the held-out pairs were run. On the matched pairs it fires on none of the ten. The first two columns give every reading the rule's own fixed cuts, 1.15 on its ratio and 0.034 on the index. They are read at $E = 4$ and then at the five later horizons $E = 10$, 15, 20, 30 and 50. The rule uses two coordinates at $E = 4$ and the drift alone from $E = 10$. The two-coordinate variants below it swap the size coordinate and keep that schedule. Averaging over the retained segment, as the drift and the RMS gap do, is what keeps the late verdict still. The single-epoch readings flip on P4 seed 1 and P7 seed 0 at $E = 10$, 20 and 30. `model_select.py` regenerates the table.

Table A19: Model selection with a tie output for every reading, twenty pair runs (eight separate, twelve match), truth from the independent test gap. *Fixed cuts*: every reading at the rule's own thresholds, at $E = 4$ and at the five later horizons $E = 10, 15, 20, 30, 50$, each pair read at its common horizon (two-coordinate readings switch to their size coordinate alone from $E = 10$, as the rule does). *Out-of-family*: the tie band is calibrated on the matched runs of the other three model families at zero in-fold false alarms and applied to the held-out family, summed over four folds. *Post-hoc*: the best single band on all twenty, reported for transparency only. The rule uses the frozen thresholds of Equation (13).

	fixed cuts 1.15 / 0.034		out-of-family		post-hoc
Reading	E=4 correct/20	$E\geq10$ horizons/5	correct /20	false alarms /12	correct /20
two-coordinate rule (drift, index)	20	5	—	—	—
with the RMS gap instead	20	5	—	—	—
with the plain gap instead	20	2	—	—	—
with the validation-loss ratio	20	2	—	—	—
drift alone	19	5	19	1	19
validation-loss ratio	20	2	17	3	20
RMS gap ratio	18	5	18	1	18
train–validation gap ratio	18	2	17	1	18
index alone	17	0	17	1	18

V WITH AND WITHOUT THE DOCUMENTED OVERLAP

Leave one pair out. The correlation between the drift ratio and the test-gap ratio is not carried by any one pair. Dropping each of the ten in turn and recomputing $r_{\log}$ over the remaining runs gives

pair dropped	P1	P2	P3	P4	P5	P6	P7	P8	P9	P11
$r_{\log}$	$+0.960$	$+0.925$	$+0.928$	$+0.930$	$+0.928$	$+0.929$	$+0.936$	$+0.942$	$+0.974$	$+0.928$

against $+0.930$ on all twenty-two. The widest-separating pair, the held-out 32B one, is not the pair that sets it: without P9 the correlation is $+0.974$, and without both P9 and P1 it is $+0.931$. One paired checkpoint carries a documented overlap with its task's source data: HuatuoGPT-o1-7B built its verifiable problems from the MedQA-USMLE and MedMCQA training splits (Chen et al., 2024). That gives the study something a matched-pair design rarely gets, a controlled test of what prior exposure does to the reading. We fine-tuned the same pair a second time on PubMedQA, a task outside that data, under the same protocol, and read the same quantities (Table A20).

Table A20: The same matched pair on two tasks, with and without a documented overlap with the task's source split. Ratios are A/B at $E = 50$, two seeds per member. Every quantity agrees within 2.8% across the four runs; the verdict is the same in all, and both drift ratios straddle unity.

P2, HuatuoGPT-o1-7B	MedMCQA (overlap)		PubMedQA (none)	
against Qwen2.5-7B-Instruct	seed 0	seed 1	seed 0	seed 1
pre-update x_1 ratio	1.034	1.032	1.048	1.050
roughness ratio a	1.018	1.017	1.037	1.039
drift ratio c	0.988	1.008	0.993	1.015
ORI-AUC ratio	0.997	0.998	0.997	0.997
independent test-gap ratio	0.987	1.010	0.998	1.015
verdict	matches	matches	matches	matches

The reading is the same on both tasks. The domain-adapted member fine-tunes like its twin whether or not it has seen the task's source split. The four ratios agree to within 1.9% across a change of task, of target format and of exposure. Prior exposure is therefore not what produced the match on MedMCQA, and the index reports the same thing on a task where no exposure exists. The PubMedQA pair also extends the study to a fourth task and a second medical one. The medical axis now carries a change of target with the checkpoints held fixed, as the reasoning axis does. Its first positive gap appears at epoch 3 rather than epoch 2. Its retained segment opens one epoch later, and its earliest reading is $E = 5$, where the drift ratio is 0.948. The retained segment is set by the run, not by the analyst.

Table A21 repeats every pooled statistic of Section 5 on the full sample and again with the two MedMCQA runs of P2 removed. Every conclusion holds under both. With P2 the three medical pairs carry drift and test-gap ratios within 6.1% of unity on six runs and three families. Without it the two remaining pairs carry them within 6.1% on four runs and two families. The pair count changes from twenty-two to twenty, the separating runs are untouched at eight, and the opening-epoch verdict holds on every remaining run under both readings.

Table A21: Every pooled statistic with and without P2, whose A member overlaps the MedMCQA training split. *Full*: the twelve pooled runs; *without P2*: minus the two HuatuoGPT-o1-7B runs; *without Qwen3-8B*: minus the run whose checkpoint later became P8's B member. The x_1 rows read the same ten runs in that column. Every conclusion holds under all three. `pooled_sensitivity.py` regenerates the table.

Statistic	Full (12; x_1 10)	Without P2 (10; x_1 8)	Without Qwen3-8B (11; x_1 10)
$\rho(x_1, a)$	$+0.952$	$+0.929$	$+0.952$
$\rho(x_1, c)$	$+0.927$	$+0.929$	$+0.927$
elasticity of a in x_1	2.50	2.45	2.50
elasticity of c in x_1	1.53	1.54	1.53
elasticity of c/a in x_1	-0.97	-0.91	-0.97
$r_{\log}(a, c)$	$+0.986$	$+0.982$	$+0.988$
$\rho(\text{ORI-AUC}, c/a)$	$+0.965$	$+0.939$	$+0.973$
$\rho(\text{ORI-AUC}, \text{gap})$	-0.916	-0.867	-0.955
$\rho(a, \text{gap})$	$+0.986$	$+0.988$	$+0.982$
$\rho(c, \text{gap})$	$+0.979$	$+0.964$	$+0.973$
relative scale, smallest	0.895	0.897	0.895

W THE FORMAT-FLIP PROTOCOL

The mismatch account of Section 5 extends to a second target format under a protocol fixed in advance. The package carries the generation script `make_cot_targets.py`, the probe `x1_probe.py` and the trainer option that consumes the targets. Where the ordering within a pair is set by distance from the *target format*, replacing the terse target with a long chain-of-thought target moves the ordering with it:

$$\text{terse targets: } x_1^{\text{reason}} > x_1^{\text{instruct}}, \qquad \text{long targets: } x_1^{\text{reason}} < x_1^{\text{instruct}}. \tag{33}$$

The first inequality is measured, at 1.236 on the hardest 256 questions. For the second, long targets are generated for the same 256 questions, filtered to those whose final answer verifies against gold. The probe of Appendix Q is then rerun; no fine-tuning is required. Targets are generated *twice*, once by each member of the pair, because targets written by one member are in that member's style and would favor it. The two generators bias in opposite directions, so the reading that appears under both is a property of the format. The released package carries the generation script, the probe, and the exact commands.