PROBLEM OR PATH? CONTROLS FOR EVALUATING LLM REASONING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

High pooled correctness prediction and rising continuation success do not, by themselves, identify within-attempt information or a prefix-specific advantage. We examine two controls: within-problem evaluation and budget-indexed restart comparisons. On public R1-Distill-Qwen-7B trajectories, a separately versioned early-window probe reanalysis yields pooled AUROC 0.761 but pair-weighted within-problem AUROC 0.515 on the same 22-problem set at four tokens. Within-problem intervals include chance at all ten anchors, while post-hoc within-targeted probes reveal heterogeneous positives, including a failure-weighted mean of 0.556 at 32 tokens. A complementary truncation study covers 178 problem–model cells from 89 MATH problems and two small models. One cell receives a frozen prefix-limited label, although additional cells exhibit prefix advantages. An extra-batch check on all initial terminal candidates finds higher continuation success in all ten in-grid Ministral-3 comparisons against nominal-budget restart references, nine of which are interpolated. Selection, small branch counts, and unresolved numerical differences from archived probe reports limit interpretation. The results motivate problem-conditioned discrimination and explicit restart-budget controls, not a general absence of outcome information or prefix value.

1 INTRODUCTION

Two common measurements of language-model reasoning invite stronger interpretations than their designs identify. A probe may predict whether an attempt will succeed while primarily separating easier problems from harder ones. A truncated prefix may admit successful continuations because the remaining budget is sufficient, without outperforming a fresh attempt with an explicitly indexed budget. Neither possibility makes the measurement useless: pooled prediction can support workload triage, and conditional continuation success can provide a training target. The issue is which question each measurement answers.

We study this distinction through two empirical controls. The primary case uses repeated public R1-Distill-Qwen-7B attempts to compare pooled and within-problem correctness ranking using *the same fitted scores and evaluation rows*. The complementary case compares continuation from a model's own prefix with a measured restart-budget curve. Problem conditioning controls a source of between-instance variation; the restart comparison supplies a counterfactual reference for conditional success. These are different controls, not a joint causal decomposition of reasoning.

Our contributions are threefold. First, we make the evaluation target explicit with an elementary pair-count decomposition of pooled AUROC, and apply it separately to two overlapping panels. At four tokens, pooled and within-problem AUROC differ sharply even on the same 22-problem set: 0.761 versus 0.515. Second, we retain the qualification that different, within-targeted probes show heterogeneous positives; an average null for the pooled-trained probe is not an absence-of-information result. Third, we audit prefix advantages under a restart control, distinguishing frozen taxonomy labels from all observed advantages and confirming a positive extra-batch pattern conditional on an initial candidate screen.

This is a measurement case study, not a new decoding algorithm or a universal refutation of internal correctness signals. Related work already studies difficulty decoding, within-problem full-trace pre-

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.

diction, causal interventions, and restarting from bad prefixes. Our contribution is the concrete alignment of scores, comparison sets, budget references, and claim scope in these experiments. Small cohorts, selected problems, quantized states, and imperfect historical numerical reproducibility constrain what the results establish.

2 PREDICTION: DISTINGUISH POOLED FROM WITHIN-PROBLEM RANKING

2.1 THE EVALUATION TARGET

Let c denote a problem–model cell and $Y_{cj} \in \{0, 1\}$ the correctness of attempt j, with probe score s_{cj}. Define n_c^+ and n_c^- as its positive and negative sample counts. AUROC compares all positive–negative pairs, awarding 1 for correct score ordering, 0 for incorrect ordering, and $1/2$ for ties. Partitioning these pairs by whether their members share a cell gives

$$A_{\text{pooled}} = \lambda A_{\text{within}} + (1 - \lambda) A_{\text{between}}, \qquad \lambda = \frac{\sum_c n_c^+ n_c^-}{N^+ N^-}. \tag{1}$$

Here $N^+ = \sum_c n_c^+$ and $N^- = \sum_c n_c^-$. All terms must use the same samples and discordant-pair weights. Cells with only one outcome class contribute to pooled and between-cell ranking but not to within-cell pairs. The identity follows by partitioning the numerator and denominator of the pair average; it is not a novel theorem or a causal allocation of information.

Any cell-constant score $f(c)$ has within-cell AUROC $1/2$ when both outcomes exist. Under conditionally i.i.d. attempts, a score equal to the cell's success probability can nevertheless have high pooled AUROC across a heterogeneous problem population. With one attempt per cell, $\lambda = 0$: pooled evaluation alone cannot attribute discrimination to within-attempt differences. Between-cell ranking can itself reflect both problem-level and attempt-related information, so $1 - \lambda$ is not a fraction of the probe's information explained by difficulty.

Repeated attempts permit a direct comparison. We distinguish pair-weighted within AUROC, $\sum_c n_c^+ n_c^- A_c / \sum_c n_c^+ n_c^-$, from the historical failure-weighted mean, $\sum_c n_c^- A_c / \sum_c n_c^-$, and the problem-macro mean. These averages use only cells with defined within AUROC. Failure weighting emphasizes problems with many failures, whereas pair weighting emphasizes opportunities to compare opposite outcomes. Only pair weighting enters Eq. (1).

With equal numbers of attempts in K cells and both outcomes present globally, $\lambda \leq 1/K$. Thus, wherever the components are defined, $|A_{\text{pooled}} - A_{\text{between}}| \leq 1/K$: pooled ranking approaches between-cell ranking as the number of equally sampled cells grows, regardless of the within ranking. Equality $\lambda = 1/K$ holds when all empirical cell pass rates coincide. This elementary consequence does not identify between-cell information with difficulty, and the bound need not hold for unequal cell sizes (derivation in Appendix F.4).

2.2 PUBLIC TRAJECTORIES AND THE PROBE RECIPE

We use 128 MATH problems with 256 R1-Distill-Qwen-7B attempts each, from a public dump released alongside the data of Singhi et al. (2025); Singhi (2025). The MATH dump is not described in that paper. We reconstruct an early-window probe setting related to David (2025), whose reported early positive uses one trajectory per problem. This is not an exact replication on the original samples. No new generations are sampled for the public-dump analysis.

The frozen extraction set combines panel P, eight attempts per problem ($n = 1{,}024$), with panel W, all 256 attempts on 22 mid-band problems ($n = 5{,}632$). Their union contains 6,480 distinct trajectories; P and W overlap and are not independent experiments. W is selected by full-dump success counts from 14 to 242 before state extraction. It contains 94% of the dump's failures and 84% of its same-cell discordant pairs. This is outcome-based problem selection, despite being blind to probe scores, and restricts generalization to that pass-rate band.

We teacher-force observed tokens through a 4-bit model and store the mean final-layer hidden state over the last four generated tokens at each anchor, not a single-token state. Anchors are $4, 8, 16, 32, 64, 128, 192, 256, 384, 512$. Stored vectors have dimension 3,584 and are converted from float16 to float32 for fitting. Under teacher-forced replay, the 4-bit model's next-token argmax

Table 1: Current-environment $t = 4$ results. Each row is decomposed within its own sample set. PW: discordant-pair weighted; FW: failure weighted (not a term in the identity). The panels overlap in samples. Fold-local pooled AUROC is 0.8938 for P and 0.7326 for W, on different pair universes.

Panel	Rows	Pooled	Within PW	Between	Within FW	λ
P	1,024	0.8519	0.4744	0.8529	0.4333	0.00262
W	5,632	0.7610	0.5155	0.7675	0.4961	0.02563

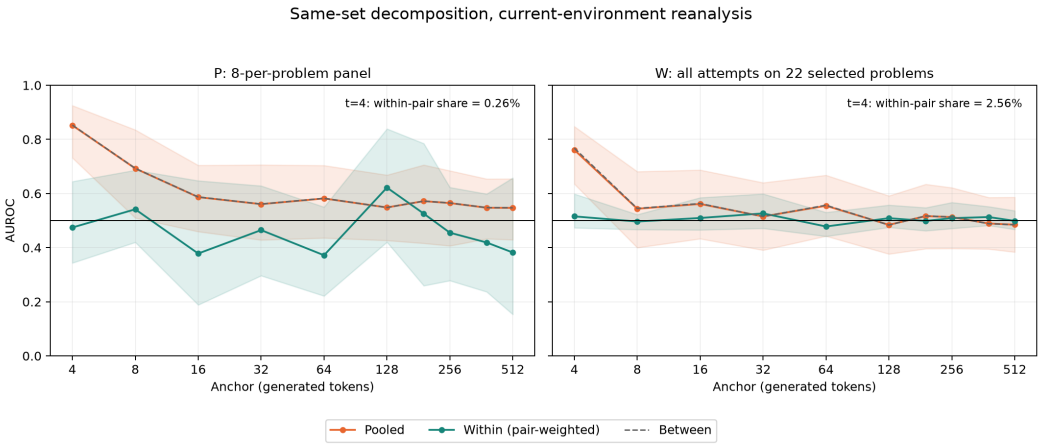

Figure 1: Same fitted probe, two overlapping panels; each panel uses the same rows for its pooled, within, and between metrics. Solid curves show pooled and pair-weighted within AUROC; dashed curves show between-problem AUROC. Shading denotes pointwise 95% fixed-OOF problem-cluster intervals. A high pooled score does not imply strong within-problem ranking here.

matches the observed token at a mean rate of 0.906 across trajectories. This measures token-level agreement, not hidden-state similarity to the original generating model; our probe findings therefore concern the replayed representations.

The union-trained probe uses five-fold StratifiedGroupKFold with problem groups, shuffle seed 20260831, training-fold standardization, PCA to at most 128 components, and class-balanced logistic regression (C=1, lbfgs, at most 2,000 iterations). Preprocessing is fitted on training rows only; each held-out problem receives out-of-fold (OOF) scores. No probe hyperparameter or seed search is conducted in the current reanalysis. Samples too short to provide an anchor are excluded at that anchor: 6,480 rows remain through 128 tokens, falling to 6,102 at 512. Later anchors therefore change the surviving population.

2.3 Same-set results and uncertainty

Table 1 and Figure 1 report a separately versioned current-environment reanalysis. Each panel is decomposed within its own rows using the union-trained OOF scores. At $t = 4$, W has pooled AUROC 0.7610 (95% CI [0.6351, 0.8477]), pair-weighted within AUROC 0.5155 [0.4733, 0.5985], and between-problem AUROC 0.7675 [0.6437, 0.8642]. Only 2.56% of its discordant pairs are within a problem. P has pooled AUROC 0.8519 [0.7319, 0.9250] and within AUROC 0.4744 [0.3434, 0.6443], with only 0.26% within pairs. Both identities hold before rounding. The wider P interval reflects its sparse within-problem comparison set.

We use 2,000 percentile bootstrap draws of problem clusters, holding OOF predictions fixed. Repeated clusters retain their original problem identity: multiplicity w_c contributes w_c^2 within pairs, but w_c to failure-weighted or macro summaries. Consequently these pair-bootstrap intervals differ from historical intervals computed by linearly resampling per-problem pair weights. They exclude training/refitting uncertainty and are pointwise, not simultaneous. We report observed pair fractions without a population confidence claim about λ (Appendix D.12).

Across all ten anchors, both pair-weighted and failure-weighted within intervals for the union-trained probe on W include chance. At $t = 4$, the failure-weighted mean is 0.4961; its different value is an estimand difference, not a contradiction. Restricting pooled comparisons to predictions from the same held-out fold gives 0.8938 on P and 0.7326 on W, while within AUROCs are unchanged. Thus cross-fold score calibration affects pooled ranking but does not remove the same-set contrast. Fold-local ranking uses a different pair universe and is reported as a sensitivity, not substituted for the primary estimate.

A post-hoc token-prefix check finds that 48.51% of W's within-problem opposite-outcome pairs share the same first four tokens. Restricting to different prefixes gives AUROC 0.5310 versus 0.5155 overall. By $t = 128$, same-prefix pairs account for only 0.0088%. Prefix repetition therefore contributes to early ties but does not alone explain the weak measured ranking; these conditional estimates have no new confidence intervals (Appendix F.5).

2.4 Positive residuals qualify the pooled-trained probe's null

Post-hoc probes target within-problem differences more directly. Probe A subtracts each problem's label-free mean state before problem-disjoint fitting; this uses other unlabeled attempts from the held-out problem and is transductive, not single-attempt online inference. Probe B fits a separate PCA-32 logistic probe within each problem using sample-disjoint folds. Probe C is an uncentered transfer fit on W alone. These differ from the union-trained probe in training population, preprocessing, or supervision and cannot be treated as the same fitted predictor.

In the current environment, A has failure-weighted AUROC 0.5263 [0.5020, 0.5498] at eight tokens and 0.5562 [0.5221, 0.5884] at 32 tokens. B has a failure-weighted mean of 0.558 at 64 tokens. At four tokens, three highlighted problems have B AUROCs approximately 0.92, 0.84, and 0.78; together they contain 84 of 2,178 failures. Two have much shorter failing than successful outputs, consistent with an early answer-without-reasoning mode but not identifying its cause. Full-output length is descriptive and is not an early prediction feature.

Figure 2 reports all 22 problems and all anchors, including undefined B scores after single-class attrition, rather than only these post-hoc examples. Full tables appear in Appendix D.12. Multiple anchors and problems were examined. Pointwise positives, and the archived 12-draw permutation sanity checks, do not establish a multiplicity-controlled onset time. These results qualify the null: within-problem information is detectable by some diagnostics in this dataset, even though the pooled-trained probe's weighted means do not detect it reliably. They do not quantify how much that probe's pooled score uses such information.

For B, few minority-class examples per problem also limit the stability of PCA-32 logistic fits and held-out AUROCs. Sample-disjoint evaluation does not remove fit variability or post-hoc selection effects; the highlighted values are finite-sample diagnostics, not established universal correctness directions.

Numerical reproducibility and backend sensitivity. Current scores differ from archived reports; the largest pooled anchor change is 0.0141 at $t = 8$. Archived row-level scores and environment records are unavailable. An audit localizes a warning to randomized PCA's matrix product. Changing NumPy 2.2.6 from Accelerate to OpenBLAS removes warnings in six float32/64 replays at $t = 4, 32, 512$, with other package versions fixed. At $t = 4$, OpenBLAS float32 gives W pooled/within AUROC 0.7573/0.5200 versus 0.7610/0.5155 originally, but the maximum union-row score change is 0.0805 and 3.48% of W opposite-outcome pair orderings change. This supports the aggregate contrast, not numerical equivalence or harmlessness of all warnings. Both versions and further thread/precision checks are retained, with the historical cause unresolved (Appendices D.10 and F).

Fixed probe and cohort sensitivities. An exploratory sweep crosses PCA dimensions 32/128/256 or no PCA, three fixed problem-disjoint splits, and $t = 4, 32, 512$. All 36 configurations converge; their pointwise W within intervals include chance. At $t = 4$, pooled/within ranges are 0.5684–0.7748/0.4955–0.5378: pooled discrimination is recipe/split-sensitive. Original-score leave-one-problem-out within AUROC is 0.4975–0.5206; restricting all anchors to 5,322 survivors through

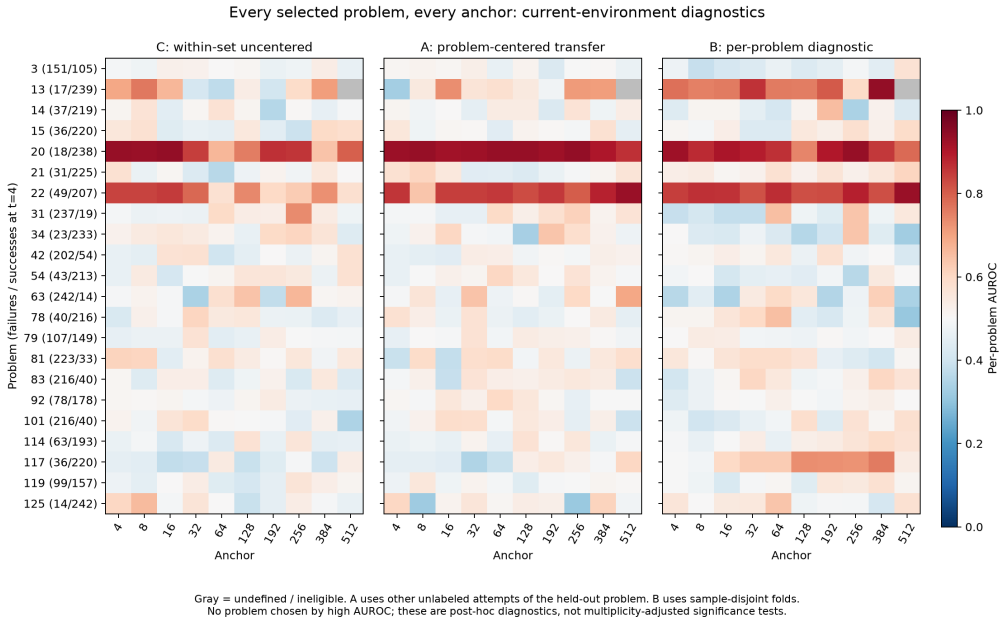

Figure 2: Within-problem heterogeneity, showing all 22 selected problems rather than only positive examples. A: problem-centered transfer; B: per-problem supervised fits; C: uncentered transfer on W. The shared color scale represents AUROC estimates, not adjusted significance; gray marks undefined or ineligible cells. A uses multiple unlabeled attempts of a held-out problem and B trains on labeled attempts from that problem. These diagnostics have different information access from the union-trained probe.

Table 2: Offline selection at $t = 4$ on W, equally weighting 22 problems. Expected correctness (%) averages exactly over uniform size-k candidate subsets; gain is in percentage points (pp). Intervals are paired, pointwise problem-cluster intervals on fixed OOF scores. Oracle uses correctness labels; it is not a deployable selector.

k	Random	Probe	Gain (pp)	95% CI (pp)	Oracle
2	61.33	61.74	+0.42	[-0.48, 1.63]	74.75
4	61.33	61.83	+0.50	[-0.99, 2.60]	83.25
8	61.33	61.57	+0.24	[-1.68, 2.59]	90.19

$t = 512$ gives pooled/within 0.7633/0.4959 at $t = 4$. Neither check supplies a second generating model or removes survival selection (Appendix F).

Does the pooled-trained score help select an attempt? For each W problem, we select the highest original OOF score from uniform size-k candidate subsets, averaging exactly over subsets and randomizing score ties. At $t = 4$, gains over random selection are $+0.42$, $+0.50$, and $+0.24$ percentage points for $k = 2, 4, 8$; paired problem-cluster intervals include zero despite label-informed oracle opportunities (Table 2). All ten anchors and common-survivor sensitivities retain this pointwise interval result. This offline evaluation is neither an equivalence test nor an online compute-saving policy (Appendix F).

3 CONTINUATION: DISTINGUISH CONDITIONAL SUCCESS FROM RESTART ADVANTAGE

3.1 DESIGN AND BUDGET ACCOUNTING

The complementary study uses 89 difficulty-targeted MATH problems (Hendrycks et al., 2021) and two 4-bit MLX models, Gemma-4 E4B and Ministral-3 3B, giving 178 problem–model cells. Development, supplement, and expansion cohorts contain 20, 20, and 49 problems. Screens were defined before the corresponding probe outcomes; research splits contain 54 training, 18 validation, and 17 test problems. Each cell has one instrumented base trajectory with a 16,384-token cap. Repeated models on the same problems do not make 178 independent problem draws.

Following truncation/resampling methods (Lanham et al., 2023; Bogdan et al., 2025), we truncate the model's own trajectory on an initial grid $64, 128, \ldots, 8{,}192$ and initially sample four continuations per anchor. Refinement adds intermediate anchors, including 32 and 16 for early crossings. The continuation reasoning cap is $B = 1{,}024$, with a separate 512-token answer reserve. The empirical solve rate $\hat{p}(t; B)$ estimates success conditional on that prefix. A budget-fit event $T_F(B)$ requires $\hat{p} \geq \tau = 0.75$ at an anchor and the next anchor; bisection refines crossing intervals, and absent crossings are censored.

These are operational brackets from noisy sampled rates, not confidence intervals for a unique latent transition. Appendix B gives the refinement limits and subsequent top-up order; neither refinement nor top-up guarantees monotonicity.

The 135 non-instant cells have four from-scratch attempts at each $C \in \{1{,}024, 2{,}048, 4{,}096, 8{,}192\}$; the 43 instant cells have only $C = 1{,}024$. The reference $\hat{R}(C)$ uses measured rates, log-budget interpolation, and endpoint proxies outside the measured grid. Define

$$\mathrm{adv}(t) = \hat{p}(t; B) - \hat{R}(t + B). \tag{2}$$

This matches *nominal reasoning budget*, not realized total tokens or compute. Both arms retain a separate answer reserve and a 16,384-token safety cap. The replayed prefix can include non-reasoning tokens; forced close tokens and early stopping affect actual consumption. FLOPs, latency, KV-cache reuse, and prefix-cost amortization across branches are not matched. Interpolated references are not new restart measurements at every $t + B$; budgets outside the grid require explicitly marked proxies.

The prefix-value time $T_V(\delta)$ requires a stable crossing of both $\hat{p} \geq 0.75$ and $\mathrm{adv} \geq \delta$, with primary $\delta = 0.5$. The frozen taxonomy gives precedence to instant events (interval upper bound ≤ 16) and budget-limited cells ($\hat{R}(4{,}096) \geq 0.75$) before the prefix-limited label. No-crossing, terminal, and unsolved labels distinguish other outcomes. Consequently a positive advantage and a prefix-limited classification are different events.

To harden noisy rates, all 148 ambiguous expansion anchors (1–3 successes out of four) received four additional attempts before classification. Initial terminal candidates—final-anchor success at least 3/4 but no stability anchor—received four further continuations; pooled success of at least 6/8 confirmed a terminal event without promoting it to T_V. Four-to-eight-attempt rates remain noisy. These rules and earlier gate failures are documented in staged protocol records; the revision's additional analyses are exploratory, not retrospectively preregistered (Appendix A).

3.2 LABELS, BUDGET RESPONSES, AND POSITIVE ADVANTAGES

Only one of 178 cells receives the prefix-limited label under the frozen precedence. However, there are three non-instant advantage crossings at $\delta = 0.5$, two of which are classified budget-limited, and four additional instant cells exhibit advantages. Sensitivity margins 0.25 and 0.75 give seven and two non-instant crossings. An upper-envelope sensitivity using bracketing restart rates agrees with all primary labels, but is not an assumption-free bound on the unknown between-grid success curve. The 1/178 count is not the prevalence of all prefix benefits.

A fixed-record sensitivity retaining available power-of-two anchors, including 16/32, removes 44 intermediate measurements: all 178 regime labels persist, but ten T_F anchors change and non-instant primary T_V events fall from three to two. Using only the original 64–8,192 grid also removes early

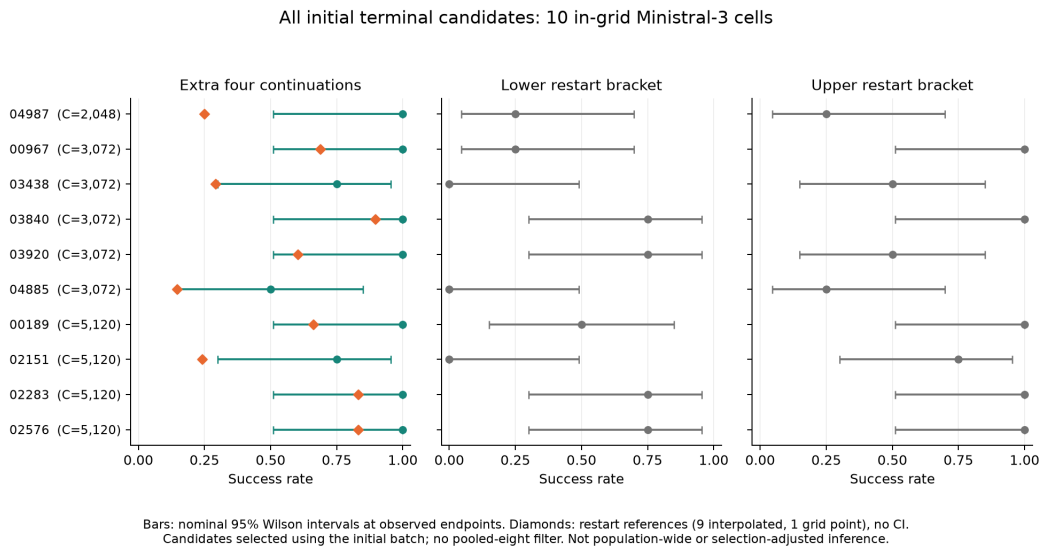

Figure 3: All ten in-grid initial Ministral-3 terminal candidates, evaluated on the extra four continuations. Bars are nominal 95% Wilson intervals for measured binomial endpoints; diamonds are restart reference point estimates. Right panels show measured restart brackets. Ten point-estimate wins are not ten significant differences: interpolation uncertainty and initial selection are not removed.

measurements and reclassifies all 43 instant cells by construction. These checks support broad-label stability under interior coarsening, not grid-invariant crossing times or advantages (Appendix F.6).

The models respond differently to restart budgets. Among non-instant expansion cells, Ministral-3's fraction reaching the success threshold rises from 0% to 19%, 45%, and 79% across the four budgets; Gemma-4 remains at 10–12%. Individual four-attempt rates can decrease, so monotonicity and a direction of bias for boundary proxies cannot be assumed. These are descriptive budget responses on selected small-model cohorts, not a causal test of overthinking or a general capability limit.

Among 13 pooled-confirmed Ministral-3 terminal candidates, nine have $t + B$ inside the measured restart grid. All nine continuation point estimates exceed the reference (eight interpolated, one measured-grid; median advantage 0.311). Four require an unmeasured 9,216-token restart and instead use an 8,192-token boundary proxy: two wins and two ties. The largest measured restart reaches the threshold in 11/13 cells and the prefix's pooled success rate in 9/13. This is consistent with budget savings in that selected subset, not evidence about asymptotic reachability or exact realized-token efficiency.

Extra-batch check without pooled-confirmation filtering. Selection on pooled continuation success favors positive differences. We therefore retain *all 16 initial Ministral-3 terminal candidates*, including those failing later pooled-eight confirmation, and evaluate only their extra four branches. Ten have in-grid nominal budgets; all ten extra-batch rates exceed the restart reference (nine interpolated, one measured-grid), with median advantage 0.346 and mean extra-batch success 0.900. The remaining six are boundary proxies. Gemma-4's single initial candidate has extra success 2/4 and ties its restart reference. This post-hoc check reduces dependence on pooled-eight selection but remains conditional on the initial four-branch screen, not an independent cohort.

The audit checks all 68 extra-branch outcomes and 1,376 expansion restart outcomes against their summaries. Appendix D.11 retains all 17 initial candidates and separates measured, interpolated, and proxy comparisons. The wide endpoint intervals in Figure 3 make clear why a count of wins is not a count of statistically established gains. Similarly, 55 of 148 enlarged ambiguous anchors remain intermediate after eight attempts; the unused second half also often has intermediate success. This supports probabilistic continuation values, not a violation of monotonicity or single-crossing assumptions.

4 SUPPORTING CONTROLS AND RELATED WORK

A limited confirmatory incremental-prediction test. A frozen test adds 15 summaries through 512 tokens to a question-only logistic baseline (human difficulty, topic, five text statistics, model identity; not an exhaustive difficulty control). Base-run correctness AUROC rises from 0.828 to 0.853, $\Delta = 0.026$ $[-0.054, +0.167]$, failing the predefined paired problem-cluster interval criterion. Only 33 cells from 17 problems and four failures enter this endpoint. Secondary restart-solvability on non-instant cells gives $\Delta = -0.090$ $[-0.213, +0.033]$; this is a different outcome. A timing target fails its power gate; a labeled post-hoc window sweep reuses the test split. These low-power nulls are not equivalence tests (Appendix E).

Difficulty references are not deployable baselines. A leave-one-out (LOO) rate reaches pooled AUROC 0.873 [0.870, 0.876] on 192,315 OpenR1-Math generations across 91,573 problems (Hugging Face Open-R1 team, 2025), 0.917 [0.875, 0.938] on 16,384 GPQA attempts (Rein et al., 2024; Singhi et al., 2025), and 0.981 [0.955, 0.991] on P using 255 other labels per problem. These trace-blind, label-informed references are not cost-matched predictors or upper bounds. Within a problem, LOO scores are lower for correct than incorrect attempts, giving AUROC exactly zero; cell-constant scores give 0.5. Cross-dataset comparisons are context, not matched counterevidence.

Correctness and difficulty probes. Internal-state studies target intermediate answers (Zhang et al., 2025), early steps (Yuan et al., 2026; David, 2025), and pre-generation states (Lugoloobi et al., 2026; Moreno Cencerrado et al., 2026). Yuan et al. (2026) include within-problem analyses for full-trace probes; Lugoloobi et al. (2026) explicitly interpret pre-generation activations as model-specific difficulty. Difficulty decoding and within-problem analysis are therefore not new in themselves. Our narrower contribution is the early-window, same-score, same-set comparison with explicit residual positives and evaluation weights. Later trajectory geometry and difficulty coupling (Sun et al., 2026; Gjølbye et al., 2026) concern related but distinct measurements.

Resampling, process values, and restart controls. Early answering and causal trace perturbations (Lanham et al., 2023; Bogdan et al., 2025), token-level forking and its sampling noise (Bigelow et al., 2025; 2026), and rollout-based process supervision (Wang et al., 2024; Luo et al., 2024; Setlur et al., 2025) predate this work. Wang et al. (2026) compare continuing truncated incorrect traces with re-solving. We add an explicit per-problem nominal-budget restart curve rather than claim the first counterfactual critique. Sequential-versus-parallel work (Snell et al., 2024; Ghosal et al., 2025; Sharma & Chopra, 2025) and the efficiency/reachability distinction of Wolf et al. (2026) motivate the comparison, but our selected finite-budget prefixes establish neither an asymptotic theorem nor a general ranking of inference strategies.

5 DISCUSSION AND LIMITATIONS

Attempt selection requires within-problem evaluation with explicit weighting and leakage controls; prefix advantage requires a specified restart policy, cost accounting, uncertainty, and selection path. Pooled prediction and rollout values answer different questions.

The activation case uses one quantized model, a linear-logistic probe family, selected mathematics, and anchor-dependent attrition. Problem conditioning is not a causal test; transductive and per-problem positives do not validate deployment. OOF-conditional intervals and the small recipe/split sweep omit broader training variability and multiplicity correction. Backend checks do not recover the historical environment or fully diagnose its mismatch (Appendices D.10 and F). Offline selection is not a deployment or equivalence test. Restart results use one base path per cell, small branch counts, interpolation, non-exact cost matching, and outcome-conditioned subsets. Stable taxonomy labels can mask grid-sensitive advantages (Appendix F.6). Neither nulls nor rare labels establish a universal absence of information or prefix value.

Future work should test nonlinear/multi-layer probes, repeated attempts beyond mathematics, and restart policies with measured costs and denser grids. Such extensions may reveal signals missed here. Our conclusion is narrower: pooled prediction can coexist with weak within-problem ranking, and continuation success is distinct from restart advantage. Both comparisons need residual positives and uncertainty.

AI USE STATEMENT

AI tools supported coding and writing and provided feedback on methodology, hypotheses, conceptual framing, mathematical claims, and result interpretation. The author reviewed this assistance, made the final research decisions, and takes responsibility for the paper.

ETHICS STATEMENT

This work studies benchmark solutions rather than human participants. Correctness discrimination on these selected benchmarks does not validate a safety mechanism or justify deployment in high-stakes settings. Reuse and redistribution of model outputs, weights, and benchmark content remain subject to their respective licenses. The distinction between prediction, conditional success, and causal utility is important to avoid overstating reliability.

REPRODUCIBILITY STATEMENT

Section 2 specifies the probe and evaluation targets; Appendix D.12 documents the full reanalysis and uncertainty convention. Appendices B, D.9, and D.11 describe generation, exclusions, and the candidate audit. Appendix A summarizes staged protocol records and corrections. Appendix F reports additional numerical, probe, cohort, and offline-selection sensitivities. The accompanying anonymous code supplement provides a CPU-only replay from stored OOF scores and numeric branch outcomes, including metric tests and generated tables/figures. It does not bundle model weights, activation arrays, or raw solutions, and does not independently regrade answers or reproduce original generation. Archived and current results remain separate; the historical numerical mismatch is unresolved (Appendix D.10).

REFERENCES

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.07961.

Eric Bigelow, Amir Zur, Satchel Grant, et al. Forking fast: Efficiently estimating uncertainty dynamics in text generation. *arXiv preprint arXiv:2608.19611*, 2026.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

Joey David. Temporal predictors of outcome in reasoning language models. *arXiv preprint arXiv:2511.14773*, 2025.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.15602.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, et al. Inverse scaling in test-time compute. *Transactions on Machine Learning Research*, 2025. arXiv:2507.14417.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, et al. Does thinking more always help? mirage of test-time scaling in reasoning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.04210.

Anders Gjølbye, Lars Kai Hansen, and Sanmi Koyejo. Reasoning models don't just think longer, they move differently. *arXiv preprint arXiv:2605.15454*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

Hugging Face Open-R1 team. Openr1-math-220k. `https://huggingface.co/datasets/open-r1/OpenR1-Math-220k`, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

William Lugoloobi, Thomas Foster, William Bankes, and Chris Russell. LLMs encode their failures: Predicting success from pre-generation activations. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2602.09924.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting LLM answer accuracy from question-only linear probes. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI*, 2026. arXiv:2509.10625.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2311.12022.

Amrith Setlur, Chirag Nagpal, Adam Fisch, et al. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.08146.

Aman Sharma and Paras Chopra. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute. *arXiv preprint arXiv:2511.02309*, 2025.

Nishad Singhi. math128 r1-distill-qwen-7b solution dump: 128 MATH problems $\times$ 256 samples at a 32k-token budget. Hugging Face dataset, `https://huggingface.co/datasets/nishadsinghi/math128_solutions_r1_distill_qwen_7b_32K_tokens`, 2025. released alongside the data of Singhi et al. (COLM 2025); not described in that paper.

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for llm reasoning. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.01005; source of the GPQA-diamond sample dump analyzed in Section 4.3.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. LLM reasoning as trajectories: Step-specific representation geometry and correctness signals. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2604.05655.

Peiyi Wang, Lei Li, Zhihong Shao, et al. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of ACL*, 2024. arXiv:2312.08935.

Pinzheng Wang, Shuli Xu, Juntao Li, Yu Luo, Dong Li, Jianye Hao, and Min Zhang. Re^2: Unlocking LLM reasoning via reinforcement learning with re-solving. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.07197.

Yotam Wolf, Noam Wies, and Amnon Shashua. When does in-context search help? a sampling-complexity theory of reflection-driven reasoning. *arXiv preprint arXiv:2607.06720*, 2026.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, Yi Nian, and Yue Zhao. Hidden error awareness in chain-of-thought reasoning: The signal is diagnostic, not causal. In *ICML 2026 Workshop on Mechanistic Interpretability*, 2026. arXiv:2605.09502.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.05419.

A PROTOCOL RECORDS AND INTERPRETATION CORRECTIONS

The staged protocols and their subsequent outcome reports were recorded in dated, committed amendments. This appendix summarizes the records in chronological order and records how every gate resolved, including the failures. The identifying repository link is omitted for double-blind review; this summary alone does not independently authenticate the commit chronology. Exploratory follow-ups and corrections of interpretation are not preregistered findings.

Breakthrough forecasting protocol. Froze the base probe design on the development cohort: anchor grid, $m = 4$ continuations, reasoning budget $B = 1{,}024$ plus a separate 512-token answer reserve, threshold $\tau = 0.75$ with next-anchor stability, interval censoring, bisection refinement, the research splits (assigned before any outcome), and the within-trajectory forecasting target $P(T_F \leq t + k \mid \text{features through } t)$ whose power gate Section 4 later reports.

Probe sensitivity and supplement (A1–A3). Froze three responses to the development cohort's censoring pattern: A1 terminal replication (threshold-censored cells gain four branches; a terminal event requires pooled $\geq 6/8$); A2 budget sensitivity (re-probing at a 4,096-token continuation budget with paired seeds); A3 a screened 20-problem supplement cohort. *Resolution:* A2 falsified the naive labels for Ministral-3 (16-token prefixes solve at the larger budget) while Gemma-4's early crossings changed less — the observation that forced the T_F/T_V distinction.

Restart-controlled breakthrough (A5). Froze the restart curves $R(C)$, the advantage definition at matched nominal reasoning budget (accounting clarified in Section 3), $T_V(\delta)$ with primary $\delta = 0.5$, the regime taxonomy and its precedence order, and the conservative-interpolation sensitivity. *Pilot resolution:* the pilot gate required $8/8$ agreement and scored $7/8$; recorded as a failure by the frozen wording, with the single permitted enlargement resolving the borderline cell *against* the prior expectation.

Cohort expansion (waves and gates). The cohorts were collected in staged *waves*; the expansion cohort is wave 3, and released artifact files keep the wave-numbered names. Froze the staged expansion: screens blind to subsequent probe outcomes, wave gates with numeric thresholds, and split assignment before generation. *Gate resolutions:* G1 (construct validity) failed for Ministral-3 per A2; G2 (wave-1 interior-event yield ≥ 6) failed at 4, with the diagnosis recorded — the screen selected on 16K terminal solvability while probes test 1,024-token continuations — and the wave-3 screen was repaired accordingly before any wave-3 outcome existed.

Ambiguity enlargement (A5.1). Froze the forward-looking rule that every 1–$3/4$ probe cell is enlarged to eight attempts before labels are derived, with thresholds applied to pooled rates. *Outcome:* labels were trimmed, not inflated (six events removed, one added, project-wide).

Confirmatory early-signal test. Froze endpoints, feature sets, model class, folds, power gates, and the success criterion; committed before the single test-split evaluation. The same document's addendum labels the forecast-point sweep ($t \in \{128, 256, 1{,}024, 2{,}048\}$) as openly post-hoc. *Outcome:* both endpoints failed the success criterion; the horizon endpoint stayed below its power gate and was not fit.

External re-analysis protocol (reconstruction). Froze a faithful reconstruction of the closest published early-window positive (model, data construction, probe recipe) plus the added controls. Superseded before execution by the public-dump design below, with more repeated samples, no new generation, and a within-problem control; the apparatus was retained in the archived research materials.

Public-dump probe dissection. Froze, after sample verification but before any state extraction: the trajectory selection (all 256 samples of the 22 mid-band problems plus 8 per problem elsewhere), the teacher-forcing extraction, the probe recipe, both evaluation metrics with their reference rows, and the reading key for either outcome. *Outcome:* recorded in the same document; reported in Section 2. A within-problem robustness check run after reporting is labeled post-hoc in the same document and in Appendix D.7.

Interpretation corrections. The initial public-dump plan expected within-AUROC 0.5 for the empirical LOO reference. Its outcome record and the current analysis correct this to 0: the LOO construction ranks incorrect attempts above correct ones within a problem. A cell-constant score, not the empirical LOO score, provides the 0.5 reference. The current pooled LOO result uses the full verification table rather than only extracted samples. Later within-targeted positives also limit the initial broad no-signal interpretation; the $t = 8$ result rules out treating $t \approx 32$ as an established onset. These corrections preserve the historical records and do not turn post-hoc analyses into confirmatory evidence.

B Extended restart methods

B.1 Setup

We study Gemma-4 E4B and Ministral-3 3B, two small open-weight models with distinct architectures and training pipelines (the latter an explicitly reasoning-tuned release)[1] — run locally in thinking mode, in 4-bit MLX quantization, with full token-level instrumentation (per-token logits and final-layer hidden states). Problems come from the MATH benchmark (Hendrycks et al., 2021). Because breakthrough measurement is only informative on problems a model sometimes solves, cohorts were selected by intermediate-difficulty screens defined before the corresponding probe outcomes were observed. These screens use separate success measurements, not the eventual probe scores; they are difficulty-targeted, not outcome-free sampling. We use a development cohort (20 problems), a supplement cohort (20), and an expansion cohort (49) whose screen was repaired after a recorded gate failure (App. A). Together they yield 89 problems and $89 \times 2 = 178$ problem–model cells. Each cell has one instrumented base trajectory generated with a 16,384-token budget. Research splits (54 train / 18 validation / 17 test problems) were assigned before any outcome was observed. Section 4 distinguishes the single confirmatory test evaluation from the subsequent post-hoc window sweep on that same held-out split.

B.2 Truncation probe and the budget-fit time T_F

The probe follows the truncation-and-resampling tradition (Lanham et al., 2023; Bogdan et al., 2025): truncate the model's *own* trajectory at an anchor t on the initial log-spaced grid $64, 128, 256, 512, 1{,}024, 2{,}048, 4{,}096, 8{,}192$, and sample $m = 4$ continuations from the truncated prefix under a fixed budget of $B = 1{,}024$ reasoning tokens plus a 512-token answer reserve, with deterministic per-branch seeds. The per-anchor solve rate $\hat{p}(t; B)$ estimates conditional continuation success, not by itself an advantage over restarting. A threshold event is the first anchor with $\hat{p} \geq \tau$ ($\tau = 0.75$) that remains above threshold at the next anchor; crossings are refined by bisection and recorded as intervals, and trajectories whose curve never crosses are right-censored at their final anchor. We call the resulting time $T_F(B)$, the *budget-fit time*: the first observed stable crossing under that continuation budget. As Section 3.2 shows, T_F is what an uncontrolled probe reports as a breakthrough — and it conflates two very different things.

All three cohort manifests use this initial grid and allow at most two refinement rounds. The 32- and 16-token measurements arise from refining early crossings, not a uniform initial grid beginning at 16. After sorting the measured anchors, the algorithm finds the earliest adjacent pair whose rates both meet the threshold; the bracket runs from the preceding anchor (or zero) to the first anchor of that pair. It probes the integer midpoint $\lfloor (t_{\text{lower}} + t_{\text{upper}})/2 \rfloor$ with four continuations, then rederives the earliest stable pair from all available anchors. Refinement stops if there is no stable crossing, the bracket is at most 16 tokens wide, or the midpoint is nonpositive or already measured. The resulting bracket describes this operational search, not a confidence interval with coverage for a true transition time. Small-sample fluctuations can change both the bracket and which midpoint is queried; stable adjacent empirical rates do not establish a monotone underlying curve or a unique crossing.

[1] Checkpoints `google/gemma-4-E4B-it` and `mistralai/Ministral-3-3B-Reasoning-2512`, run as their `mlx-community` 4-bit conversions.

B.3 RESTART CONTROL AND THE PREFIX-VALUE TIME T_V

To separate "the prefix carries value" from "the budget became sufficient," we measure a problem–model cell's *restart curve* $R(C)$: the solve rate of from-scratch attempts (empty prefix, same prompt) at reasoning budgets $C \in \{1{,}024, 2{,}048, 4{,}096, 8{,}192\}$, four attempts each for the 135 non-instant cells. The 43 instant cells retain only the 1,024-token restart measurement; their advantages use that endpoint as a proxy, not measurements at $t + 1{,}024$. For multi-point panels, $\hat{R}$ interpolates linearly in probability against log budget. For $C_L < C < C_U$ between adjacent measured budgets,

$$\widehat{R}(C) = (1 - w)\widehat{R}(C_L) + w\widehat{R}(C_U), \qquad w = \frac{\log C - \log C_L}{\log C_U - \log C_L}. \tag{3}$$

No logarithm of a success rate is taken, so zero rates are allowed. Measured grid rates are used exactly at grid points. Outside the grid, the implementation clamps to the nearest endpoint; these comparisons are explicitly labeled boundary proxies, not measured matched-budget restarts. Interpolation introduces uncertainty about the between-grid curve in addition to binomial uncertainty at its endpoints. The comparison matches *nominal reasoning budget*: the t reasoning tokens in the retained prefix plus B newly allocated reasoning tokens against a restart cap $C = t + B$. Both arms have a separate 512-token answer reserve and a 16,384-token overall safety cap. The replayed prefix also retains any generated non-reasoning tokens before the anchor; forced reasoning-close tokens can be inserted, and either sampled stage may end early. Consequently $t + B$ is not the realized total generated-token count. We do not match FLOPs, latency, KV-cache reuse, or amortized prefix cost across branches. The budget-indexed *advantage* is $\mathrm{adv}(t) = \hat{p}(t; B) - \hat{R}(t + B)$, and the prefix-value time $T_V(\delta)$ is the earliest stable anchor with $\hat{p} \geq \tau$ and $\mathrm{adv} \geq \delta$ (primary margin $\delta = 0.5$; 0.25 and 0.75 reported as sensitivity, along with a conservative variant that takes the upper envelope of the bracketing grid values of $\hat{R}$). This envelope is a sensitivity calculation, not an assumption-free upper bound on the unmeasured restart curve. In-grid comparisons use measured grid rates when available and interpolated estimates otherwise, not fresh measurements at every $t+B$. Cells are then classified by frozen precedence: *instant* (event interval upper bound ≤ 16 tokens), *budget-limited* ($\hat{R}(4{,}096) \geq \tau$: a restart solves it), *prefix-limited* (a $\delta = 0.5$ crossing exists), *no-crossing* (T_F exists but no advantage crossing), *terminal* (a replicated event at the final anchor with no stability anchor; annotated, never promoted to T_V), and *unsolved*. Because instant and budget-limited labels take precedence, an advantage crossing and a prefix-limited label are not equivalent. Failure to receive that label does not imply that every prefix has zero value.

B.4 NOISE HARDENING

With $m = 4$, a $3/4$ crossing occurs with probability 0.31 under a true success rate of 0.5, so single-shot labels are optimistic. Two frozen rules harden them. First, threshold-censored trajectories (final-anchor rate $\geq \tau$, no stability anchor available) get four additional branches, and a terminal event is recorded only if the pooled rate reaches $\geq 6/8$ (under $p = 0.5$ this has probability 0.14). Second, every ambiguous anchor-level cell (1–3 successes of 4) across the expansion cohort was enlarged to eight attempts before labels were derived (148 anchor-level cells). Both rules cut in both directions in practice: replication confirmed 13 of 16 candidates for one model and rejected the other model's single candidate, and pooling removed six events while adding one. The quoted binomial probabilities concern a fixed endpoint before selection; they are not selection-adjusted error rates for a scan over anchors or problems.

Ambiguity enlargement is a separate pass after the initial grid and refinement measurements. It examines every recorded expansion-cohort anchor, including any refinement midpoint, rather than only the original grid. The algorithm adds four branches at each ambiguous anchor and rederives labels from the pooled rates. It does not apply enlargement inside the midpoint search or run a new refinement search after pooling.

B.5 PRE-REGISTRATION DISCIPLINE

The staged design rules were recorded in committed amendments before their corresponding evaluations; gate failures (two cohort gates and one pilot gate) are recorded as failures and resolved by amendment, never by relabeling. The confirmatory prediction protocol of Section 4 — endpoints, feature sets, model class, power gates, and success criterion — was committed before the single

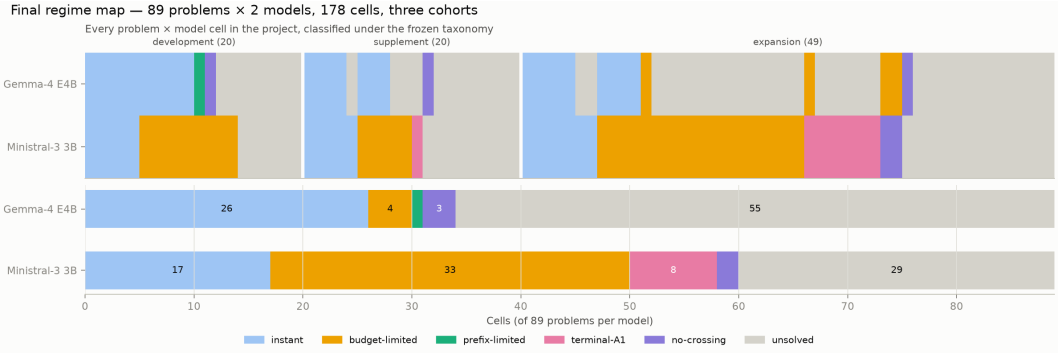

Figure 4: The final regime map: 89 problems × 2 models classified under the frozen rules (top: per-cell strip grouped by cohort; bottom: per-model totals). One cell of 178 receives the prefix-limited label; other labels can also contain positive prefix advantages.

test-split evaluation, and every post-hoc analysis is labeled as such in the amendment that reports it. Appendix A gives the full timeline.

C Extended restart results

C.1 Threshold crossings and prefix-limited classifications

The restart control changes what the probe data mean. In the development cohort, Ministral-3 exhibited apparently clean mid-trajectory crossings — until a budget-sensitivity re-probe showed the same problems solved from 16-token prefixes once the continuation budget was raised to 4,096 tokens. This shows that crossing times depend on the continuation budget; it does not show that the intervening prefix has zero value. Applied to all 178 cells, the frozen taxonomy yields the map in Figure 4: for Ministral-3, 33 cells are budget-limited, 17 instant, 8 terminal, 2 no-crossing, and 29 unsolved; for Gemma-4, 26 instant, 4 budget-limited, 3 no-crossing, 55 unsolved — and exactly *one* prefix-limited cell in the entire dataset (a development problem with $T_V = 896$ and advantage 0.75). Decomposed: the advantage crossing itself occurs in three cells at the primary margin; the frozen precedence classifies the other two as budget-limited because a 4,096-token restart already solves them, and four *instant* cells additionally show an advantage at their ≤ 16-token anchors. The 98 expansion cells, collected after the rules were frozen, contributed zero new prefix-limited cases. Loosening the margin to $\delta = 0.25$ adds four crossings, tightening to 0.75 keeps two, and the conservative $\hat{R}$ envelope agrees with the primary labels on all 178 cells. The 1-of-178 count therefore describes the frozen classification, not the frequency of any positive prefix advantage. Detection also depends on the 0.75 success threshold, the 0.5 advantage margin, next-anchor stability, and four to eight attempts. This difficulty-targeted cohort contains one base trajectory per cell; the 89 problems repeat across two models, so cells are not independent and no population-prevalence interval is implied.

C.2 Different responses to increased restart budgets

The restart curves themselves carry the cleanest model-level finding (Figure 5). On expansion cells that are not instant, the share of problems Ministral-3 solves from scratch climbs with budget — $0\% \rightarrow 19\% \rightarrow 45\% \rightarrow 79\%$ across $1{,}024 \rightarrow 8{,}192$ tokens — while Gemma-4 stays flat at 10–12%. Individual four-attempt curves are noisy: 8 of Ministral-3's 42 show at least one decrease between consecutive budgets (all of a single attempt, 0.25; two cells revert across the τ threshold), against 18 of Gemma-4's 40, five with drops of at least 0.5 and one threshold reversal. These are noisy empirical decreases, not by themselves causal diagnoses of overthinking or capability limits; they motivate comparison with test-time inverse scaling (Gema et al., 2025). Ministral-3 is more responsive to the measured restart budgets, while Gemma-4's aggregate threshold rate changes little. The two models agree on a collapsed regime class for only 42 of 89 problems: the budget response differs between these models and should not be hidden by averaging them together.

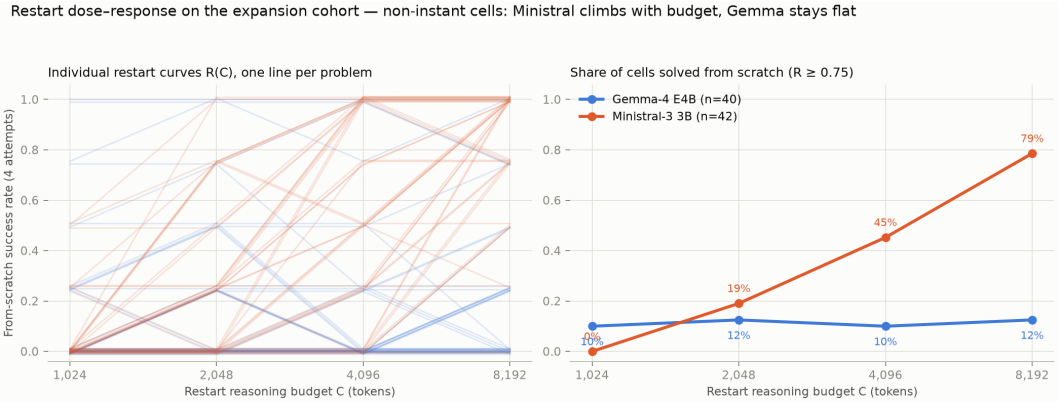

Figure 5: Restart dose–response on non-instant expansion cells. Individual curves (left) and share solved from scratch (right): Ministral-3 climbs $0\% \rightarrow 79\%$; Gemma-4 stays flat.

C.3 Prefix advantages in a selected threshold-censored subset

The 13 threshold-censored Ministral-3 expansion cells — trajectories whose crossing was replicated only at the final anchor; under the frozen precedence seven carry the terminal label, five budget-limited, and one no-crossing — form a selected subset for the comparison in Figure 6. They were retained after pooled continuation success reached threshold, so the comparison is conditional on that selection. In the nine cells whose nominal matched budget $t + B$ lies inside the restart grid, continuation has a higher point estimate than the restart reference in all nine (eight interpolated and one measured-grid comparison; median advantage $+0.31$). These are not nine independently significant advantages or direct restart measurements at every matched budget. The four cells anchored at $t = 8{,}192$ require $\hat{R}(9{,}216)$, which was not measured; substituting $\hat{R}(8{,}192)$ makes them proxy comparisons whose bias direction is unknown — restart curves are not uniformly monotone — and on these proxies the prefix wins two and ties two. The largest measured restart reaches τ in 11 of 13 cells and matches the prefix's own pooled success rate in 9 of 13. This is consistent with *budget savings* in the selected subset: the prefix reaches a success threshold under a smaller nominal reasoning budget than some measured restarts. It is not an exact realized-token cost comparison or a population claim about reasoning. Selection on high continuation success and small branch counts limit the inference. In two cells no measured restart reaches τ, leaving larger budgets open; binary correctness also compares success rates, not solution content. The sampling-complexity framework of Wolf et al. (2026) motivates the distinction between efficiency and reachability, but these finite-budget prefix experiments do not test its asymptotic conclusions. Our contribution to sequential-versus-parallel comparisons (Snell et al., 2024; Ghosal et al., 2025; Sharma & Chopra, 2025) is a per-prefix, budget-indexed control with its selection conditions stated.

Extra-batch check on all initial candidates. A post-hoc audit retains all 16 Ministral-3 terminal candidates selected using their initial four continuations, without filtering on the pooled eight-attempt outcome. Ten have nominal matched budgets inside the restart grid. On these ten, the additional four continuations alone exceed the restart reference in 10/10 cases (nine interpolated, one measured-grid; median advantage $+0.346$). This reduces the particular dependence induced by selecting the 13 pooled-confirmed cells, but remains conditional on the initial terminal-candidate screen. It is not ten statistically significant gains or a population-wide estimate. The other six require boundary proxies. Gemma-4's single initial candidate ties the in-grid restart estimate on its extra batch. All candidates and nominal endpoint uncertainty appear in Appendix D.11.

C.4 A persistent band of intermediate solvability

Enlarging all 148 ambiguous anchor-level cells to eight attempts placed 55 in the low-success band ($\leq 2/8$) and 38 in the high-success band ($\geq 6/8$) — and left *55 intermediate after eight attempts* (3–5 of 8; Figure 7). Because these cells were selected on their first four attempts, pooled rates partly reflect selection; the independent second half of attempts, however, reproduces the band —

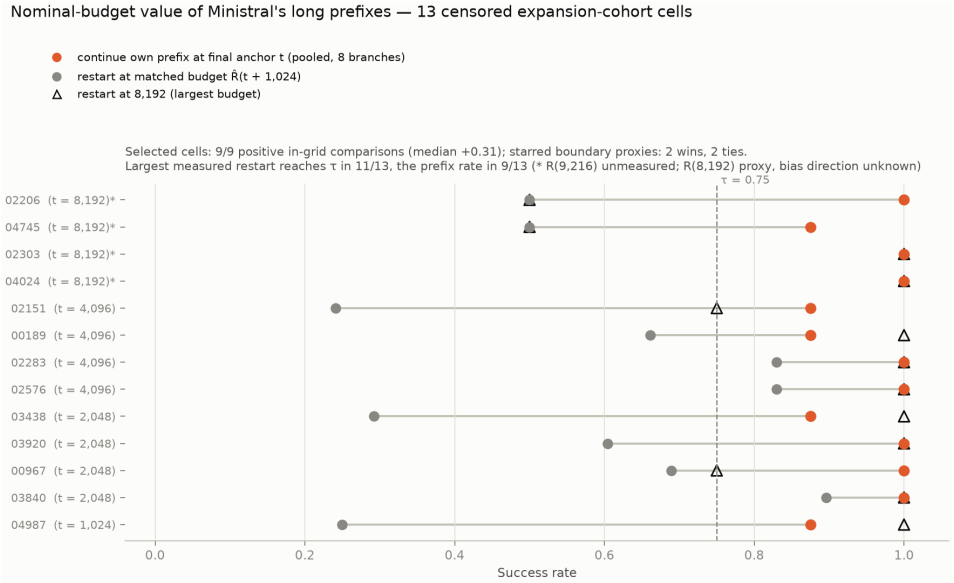

Figure 6: Nominal-budget comparison on 13 selected threshold-censored cells: continuation has the higher point estimate in all nine in-grid restart comparisons (eight interpolated, one measured-grid; starred boundary proxies: two wins, two ties); the largest measured restart reaches τ in 11/13 and the prefix's own rate in 9/13. Counts are conditional on selection and do not establish nine statistically significant gains or exact realized-token savings.

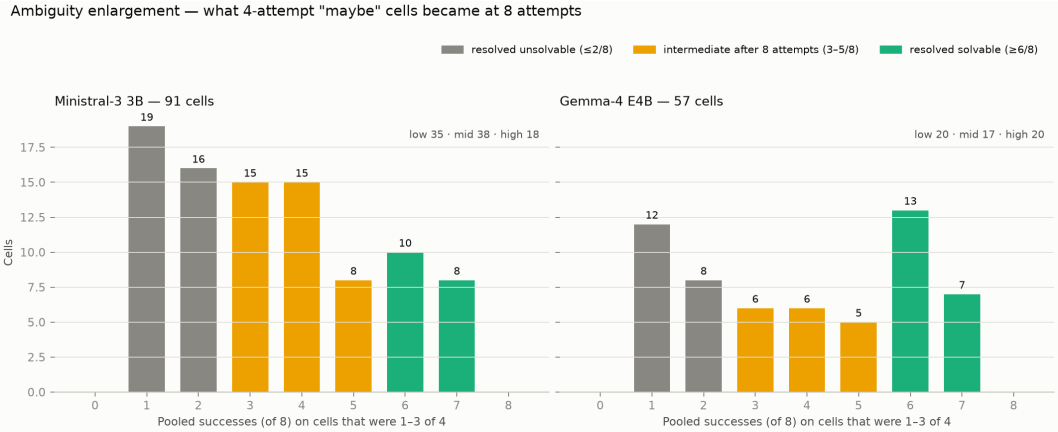

Figure 7: What ambiguous 1–3/4 cells became at eight attempts: about a third remain intermediate, a pattern the independent second half of attempts reproduces.

58% of enlarged cells land at 1–3 of 4 on branches never used for selection. The independent-half results support treating continuation outcomes probabilistically, although eight-attempt intervals remain wide and the sample is still conditional on first-half ambiguity. Intermediate probabilities alone do not establish non-monotonicity: a smooth monotone curve can also pass through them. Testing the assumptions of single-crossing process-labeling schemes (Luo et al., 2024) or monotone-degradation accounts (Gan et al., 2025) requires evidence about the shape of individual curves, not just this distribution.

D ADDITIONAL RESULTS AND NUMERICAL AUDITS

D.1 EARLY-WINDOW SIGNAL DEFINITIONS

For each generated token t let p_t be the model's next-token distribution over vocabulary V at the position that emits token x_t, and $h_t \in \mathbb{R}^d$ the final-layer hidden state at that position. The instrumented token-level series are:

$$\tilde{H}_t = -\frac{1}{\log|V|} \sum_{v \in V} p_t(v)\, \log p_t(v) \qquad \text{normalized entropy}$$

$$u_t = -\log p_t(x_t) \qquad \text{surprisal}$$

$$\mu_t = p_t^{(1)} - p_t^{(2)} \qquad \text{top-1/top-2 probability margin}$$

$$\mathrm{KL}_t = \mathrm{KL}\big(p_t \,\big\|\, p_{t-1}\big) \qquad \text{successive KL divergence}$$

$$\mathrm{JS}_t = \tfrac{1}{2}\,\mathrm{KL}\big(p_t \,\big\|\, m_t\big) + \tfrac{1}{2}\,\mathrm{KL}\big(p_{t-1} \,\big\|\, m_t\big), \qquad \text{successive JS divergence}$$

$$\nu_t = \|h_t\| \qquad \text{hidden-state norm}$$

$$\rho_t = \|h_t - h_{t-1}\| / \|h_{t-1}\| \qquad \text{relative hidden step}$$

$$\gamma_t = 1 - \cos\big(h_t,\, h_{t-1}\big) \qquad \text{cosine drift}$$

Here $m_t = (p_t + p_{t-1})/2$. Over the forecast window ($t \le 512$), each series is reduced by simple statistics: the mean and standard deviation; the *robust slope*, a Theil–Sen estimator (median of pairwise slopes, evaluated on at most 512 evenly spaced points for bounded cost); and for surprisal the maximum one-step rise $\max_t(u_{t+1} - u_t)$. Spectral summaries are computed on the JS_t series: robust-standardize by median and median absolute deviation, linearly interpolate to 256 equally spaced points on a normalized $[0, 1]$ index using `numpy.interp`, apply a Hann window, take the real FFT power spectrum with the DC term removed, and normalize to π_f; then

$$\mathrm{SpecEnt} = -\tfrac{1}{\log F} \textstyle\sum_f \pi_f \log \pi_f, \qquad \mathrm{Centroid} = \textstyle\sum_f f\, \pi_f, \qquad \mathrm{LowRatio} = \textstyle\sum_{f \le 0.10} \pi_f.$$

The resampling first removes nonfinite observations and uses their remaining ordered indices, not Fourier-domain resampling. Fewer than 63 finite observations yield missing spectral features; an MAD below 10^{-12} yields zero summaries. These rules are implemented in the supplement's `reasonbench/features/spectral.py`. The fifteen frozen summaries of Section 4 are: $\mathrm{mean}(\tilde{H})$, $\mathrm{sd}(\tilde{H})$, $\mathrm{slope}(\tilde{H})$; $\mathrm{mean}(u)$, $\max\text{-rise}(u)$; $\mathrm{mean}(\mu)$; $\mathrm{mean}(\mathrm{JS})$, $\mathrm{sd}(\mathrm{JS})$, $\mathrm{slope}(\mathrm{KL})$; $\mathrm{mean}(\rho)$, $\mathrm{mean}(\gamma)$, $\mathrm{slope}(\nu)$; and SpecEnt, Centroid, LowRatio of JS_t. The public-dump probes of Section 2 reuse the unnormalized entropy, u_t, and top-1 probability series from teacher-forced forwards (the normalization constant is fixed within a model and immaterial after standardization).

D.2 THE BUDGET SENSITIVITY THAT MOTIVATED THE RESTART CONTROL

Figure 8 shows the A2 experiment behind Section 3.2: development-cohort cells re-probed at a 4,096-token continuation budget with paired seeds. Ministral-3's apparent mid-trajectory crossings dissolve — 16-token prefixes solve at the larger budget — while Gemma-4's early crossings remain pinned, showing different sensitivity to this budget change, not by itself validating a causal breakthrough mechanism.

D.3 PER-TRAJECTORY INTERVALS AND LABEL HARDENING

Figure 9 shows what the noise-hardening rules changed; Figure 10 shows the breakthrough interval or censoring bound of every expansion-cohort trajectory. Terminal replication confirmed 13 of Ministral-3's 16 threshold-censored candidates and rejected Gemma-4's single candidate; pooling ambiguous cells to eight attempts moved labels on $7 + 6$ of $49 + 49$ trajectories, dominated by losses (six events removed, one added; Gemma-4's changes were mostly interval shifts).

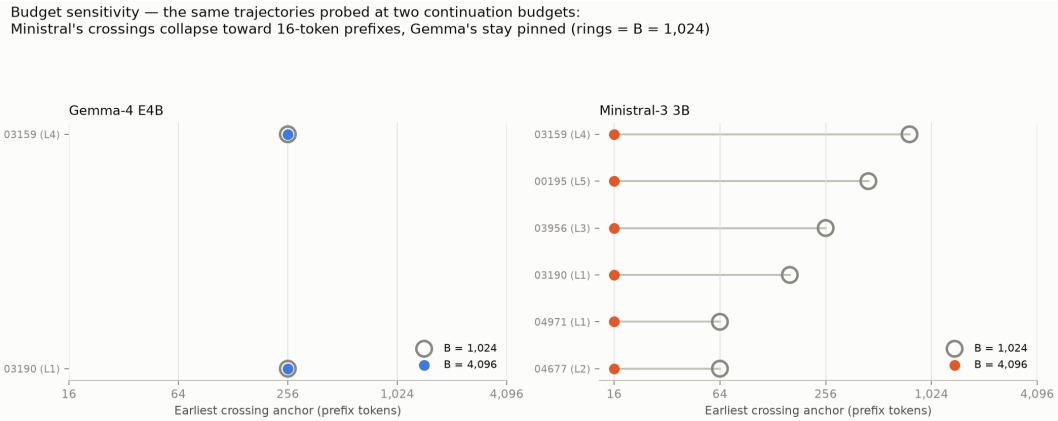

Figure 8: Budget sensitivity (amendment A2, App. A): the same cells probed at $B = 1{,}024$ versus $B = 4{,}096$. Ministral-3's earliest crossings move toward 16-token prefixes; Gemma-4's change less. These are reasoning budgets with separate answer reserves, not realized-token cost matches.

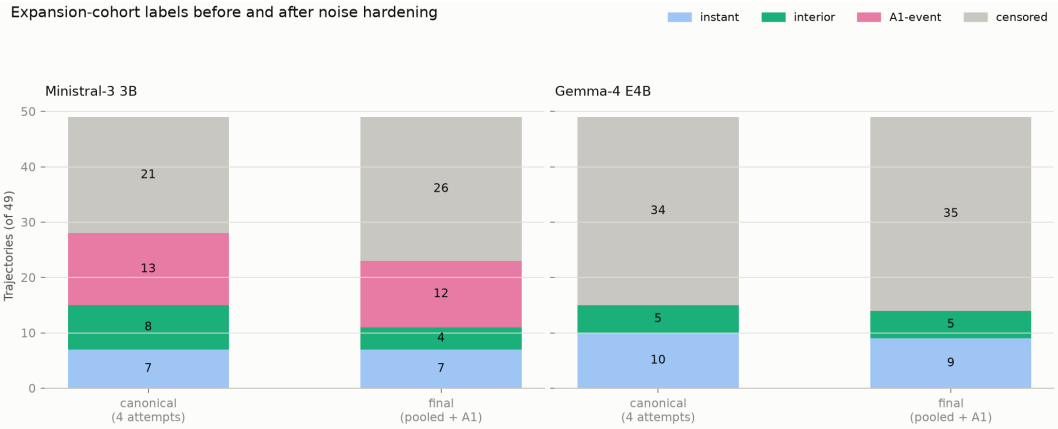

Figure 9: Expansion-cohort labels before and after noise hardening (pooling + terminal replication): trimmed, not inflated.

D.4 Sensitivity of T_V labels

Independent-half check for the intermediate band: on branches 4–7 alone — never used for selection — the 148 enlarged cells land at $0/1/2/3/4$ successes in $37/37/28/21/25$ cells; 58% fall in the intermediate 1–3 range. This describes outcomes on branches not used for selection, conditional on first-half ambiguity; it is not a test of monotonicity or a fully specified selection-null model.

Across all 178 cells: $\delta = 0.5$ yields 3 prefix-value events, $\delta = 0.25$ yields 7 (the additions all sit inside budget-limited cells of the budget-elastic model), $\delta = 0.75$ keeps 2. The conservative variant of $\hat{R}$ (upper envelope of the bracketing grid values) agrees with the primary labels on 178 of 178 cells. Cross-model agreement on collapsed regime classes is 42 of 89 problems.

D.5 Confirmatory test: level-free variants and per-model detail

Per-model test-split descriptives (no per-model claims are made): on the primary endpoint the baseline scores 1.00 for Gemma-4 (n=16) and 0.69 for Ministral-3 (n=17, early signals 0.76); on the secondary endpoint 1.00 versus 0.73 (n=11, 14). The horizon endpoint's power arithmetic: 21 interior events project-wide, 4 in the test split; the frozen gate (five problem groups per class per fold) was not met and no model was fit.

Table 3: Held-out test AUROCs for both endpoints, with the difficulty level dropped from both feature sets (level-free). No variant meets the success criterion.

Endpoint	Variant	Baseline	+Early signals	ΔAUROC [95% CI]
Primary	main	0.828	0.853	$+0.026\ [-0.054, +0.167]$
Primary	level-free	0.914	0.879	$-0.035\ [-0.156, +0.024]$
Secondary	main	0.875	0.785	$-0.090\ [-0.213, +0.033]$
Secondary	level-free	0.840	0.771	$-0.070\ [-0.174, +0.056]$

D.6 Public-dump dissection: full anchor table

Table 4: Hidden-state probe on public R1-Distill-7B trajectories at every anchor. The rows column counts trajectories entering the problem-disjoint out-of-fold prediction; the *pooled* metric is evaluated on the 8-per-problem pooled subset (n=1,024 at t=4, shrinking with window attrition) and the *within* metrics on the 22 mid-band problems (n=5,632; 2,178 failures; at t=512 attrition leaves 21 problems with both outcome classes and 2,114 failures), under the pre-registered failure-count weighting and discordant-pair weighting. The latter matches the within metric in Eq. (1) only for this same sample set; the displayed pooled column uses a different set. References at t=4: full-dump LOO pass-rate pooled AUROC 0.981 [0.955,0.991] (255 other attempts per problem, evaluated on the same pooled set); stream-summary probe pooled 0.647, within 0.490. The within-AUROC of the LOO score itself is mechanically 0 (within a problem it takes two values, and the incorrect attempt's is always higher), so the neutral reference for "no within-attempt information" is 0.5.

t	P pooled [95% CI]	W within FW [95% CI]	W within PW [95% CI]	rows
4	0.849 [0.735,0.919]	0.496 [0.466,0.527]	0.515 [0.481,0.562]	6,480
8	0.678 [0.503,0.835]	0.481 [0.461,0.502]	0.498 [0.480,0.515]	6,480
16	0.586 [0.461,0.706]	0.495 [0.467,0.526]	0.510 [0.478,0.552]	6,480
32	0.561 [0.422,0.702]	0.525 [0.494,0.555]	0.526 [0.490,0.568]	6,480
64	0.582 [0.442,0.706]	0.481 [0.451,0.510]	0.478 [0.452,0.508]	6,480
128	0.548 [0.425,0.664]	0.515 [0.491,0.542]	0.509 [0.485,0.538]	6,480
192	0.572 [0.421,0.706]	0.499 [0.466,0.525]	0.499 [0.475,0.527]	6,478
256	0.565 [0.406,0.684]	0.530 [0.476,0.591]	0.509 [0.483,0.543]	6,476
384	0.547 [0.432,0.650]	0.509 [0.482,0.537]	0.513 [0.494,0.536]	6,294
512	0.547 [0.428,0.656]	0.497 [0.468,0.526]	0.499 [0.477,0.521]	6,102

The historical numerical check reports reproduction in float64 with zero NaN out-of-fold predictions. Observed-token/top-1 agreement has mean 0.906 (10th percentile 0.859). This measures agreement with sampled dump tokens, not hidden-state fidelity to a full-precision model.

The revision-time check in Appendix D.10 does not establish exact numerical reproduction of that archived run.

D.7 Post-hoc within-problem robustness checks

Run after all frozen endpoints were reported and labeled post-hoc in the amendment, these checks ask whether within-attempt information could exist that the pooled-trained probe of Section 2 cannot see. Both use the stored states of the 22-problem within set (n=5,632). **A, problem-centered transfer:** each problem's mean state is subtracted (label-free) and the probe (standardization, PCA ≤ 128, balanced logistic) is trained on problem-disjoint folds and scored within held-out problems. Centering includes the held-out problem's other unlabeled trajectories, making this a transductive diagnostic. **B, per-problem supervised diagnostic:** a separate probe (PCA 32) per problem with sample-disjoint folds; this finite fit is not a mathematical upper bound on linearly decodable information. An uncentered re-run on the within set alone gives a qualitatively similar null ($0.515\ [0.488, 0.547]$ at t=4 against $0.496\ [0.466, 0.527]$, which trains on both sets). Table 5 reports failure-weighted means with problem-clustered bootstrap CIs.

The signal is heterogeneous. Three problems (ids 20, 22, and 13, with 18, 49, and 17 failures — 84 of 2,178) have high diagnostic AUROCs at t=4 (0.92, 0.84, 0.78; centered transfer 0.93, 0.86, 0.33);

across all ten anchors the diagnostic stays above 0.74 for problems 20 and 22, while problem 13 dips to 0.60 at t=256 and loses its failures to attrition by t=512; per-problem diagnostic AUROCs are correlated across anchors (Spearman 0.71 between t=32 and 64). In problems 20 and 22 the failing samples are four to six times shorter than the successes (median 2,364 versus 15,321 and 1,299 versus 5,168 characters), consistent with an early answer-without-reasoning mode. Problem 13 has comparable lengths and observed-token agreement, suggesting a different failure mode without identifying its cause. These three examples were highlighted post-hoc, not selected independently of the diagnostic results. The failure-weighted average emphasizes different problems than a macro or pair-weighted average. The centered probe has pointwise intervals above chance at $t = 8$ as well as $t = 32$ and some later anchors (Table 5); there is no validated first-signal time. Heterogeneous positives qualify the frozen probe's null without determining the source of its pooled score.

Table 5: Post-hoc within-problem checks on the 22-problem within set: failure-weighted mean within-problem AUROC of a problem-centered transfer probe (A) and of per-problem supervised diagnostics (B), the largest single-problem diagnostic AUROC, and the number of problems above 0.6. Within-problem label-permutation nulls (12 draws at $t \in \{4, 32, 64, 128\}$) span 0.47–0.54 for the A means, 0.44–0.54 for the B means, and 0.53–0.76 for the B maxima. These are limited sanity checks, not multiple-comparison-adjusted significance tests. Intervals are pointwise.

t	A: centered [95% CI]	B: per-problem [95% CI]	B max	B > 0.6
4	0.490 [0.464,0.523]	0.476 [0.435,0.529]	0.924	3
8	0.526 [0.502,0.550]	0.479 [0.449,0.519]	0.873	3
16	0.486 [0.451,0.534]	0.481 [0.437,0.536]	0.908	4
32	0.556 [0.522,0.588]	0.519 [0.478,0.560]	0.894	5
64	0.532 [0.502,0.565]	0.558 [0.523,0.593]	0.868	7
128	0.524 [0.504,0.548]	0.519 [0.489,0.556]	0.824	4
192	0.526 [0.498,0.556]	0.491 [0.452,0.543]	0.900	5
256	0.536 [0.503,0.570]	0.531 [0.493,0.576]	0.935	5
384	0.529 [0.516,0.546]	0.532 [0.493,0.576]	0.939	6
512	0.521 [0.466,0.580]	0.507 [0.457,0.549]	0.932	3

D.8 Post-hoc forecast-point sweep: full table

Table 6 lists both endpoints at every forecast point.

Table 6: ΔAUROC (early signals $-$ baseline) at every forecast point; * marks the pre-registered evaluation, all others post-hoc. Larger prefixes also shrink and select the sample (only runs surviving to t enter).

t	Primary Δ [95% CI]	n	Secondary Δ [95% CI]	n
128	-0.083 $[-0.281, +0.016]$	34	-0.118 $[-0.274, +0.048]$	25
256	$+0.092$ $[-0.032, +0.384]$	34	-0.035 $[-0.133, +0.053]$	25
512*	$+0.026$ $[-0.054, +0.167]$	33	-0.090 $[-0.213, +0.033]$	25
1,024	-0.125 $[-0.409, 0.000]$	24	-0.137 $[-0.287, +0.018]$	22
2,048	-0.133 $[-0.364, 0.000]$	13	-0.259 $[-0.800, 0.000]$	12

D.9 Reproducibility details

Generation. Both models sample at temperature 0.6, top-p 0.95, top-k 20 (no repetition or presence penalties), thinking mode enabled, prompt version v1 (a fixed task instruction requesting a final boxed answer), 16,384-token cap for base trajectories. Probe branches use deterministic seeds derived as $\mathrm{sha256}(\text{run}, \text{anchor}, \text{branch})$; restart attempts hash the problem, model, budget, and branch. Model checkpoints have resolved revisions recorded in the study's readiness manifest (`mlx-community` 4-bit conversions of the checkpoints named in Section 3).

Prompt construction. In both instrumented models, the user message is the problem text followed by two newline characters and the literal task instruction:

```
Solve the problem carefully.  Put only the final
answer inside \boxed{}.
```

Gemma uses the system message `You are a careful mathematical assistant.` and its checkpoint chat template with `enable_thinking=True`.

Ministral uses the checkpoint's official `SYSTEM_PROMPT.txt`, not that generic system message. Its `[THINK]` and `[/THINK]` segment is represented as a closed structured thinking block between the surrounding text blocks; no extra thinking-template switch is added.

The readiness record identifies the system-prompt source as `mistralai/Ministral-3-3B-Reasoning-2512`, file `SYSTEM_PROMPT.txt`, and the following revisions:

System-prompt source:
`4a36357c811bf511a7b625d132e12f22408aac91`
Gemma MLX weights:
`475b9088d29754a3379866cf5aeb6b41acd313c2`
Ministral MLX weights:
`2cd2087aad40c28747f8ede17851de6035f12b16`

These identify recorded artifacts, not a newly verified reconstruction of every historical download. The supplement contains the message-building code in `reasonbench/instrumentation/adapters/` and prompt-source loading in `reasonbench/generation/mlx_modeling.py`; the external Ministral system-prompt file must be obtained at the stated revision. This prompt specification concerns the instrumented continuation/restart study, not the original generation prompts of the public R1 dump. Both adapters render messages through the checkpoint chat template with `add_generation_prompt=True` before tokenization. The external system-prompt bytes were not newly retrieved or validated for this documentation update; the recorded revision is supplied for retrieval.

Verification. A final answer is extracted from the generated text; extraction failure scores as incorrect. Extracted answers are compared to the reference by numeric equivalence first, then symbolic equivalence. GPQA answers compare the extracted option letter ("final answer is X") against the key; unparseable samples (3.0%) score incorrect.

Cohort screens. Development and supplement problems were drawn from a 100-problem level-balanced pool by rules fixed before the corresponding probe outcomes; the expansion screen selected problems with 1–5 verified successes out of 6 fresh 3,072-token attempts (2 models $\times$ 3 seeds).

Classifiers. The instrumented-cohort analyses (confirmatory test and sweep) use ℓ_2 logistic regression with $C = 0.1$, liblinear, class-weight balanced, at most 5,000 iterations, on standardized features. Numeric preprocessing uses median imputation and a single per-feature `StandardScaler` fitted on the combined training rows of both models, not separate scalers per model. Categorical features, including model identity, are imputed and one-hot encoded. Within cross-validation these transformations are fitted only on the training fold; the final held-out test fit uses train plus validation. This avoids test leakage but does not establish that pooled scaling is optimal for architecture-specific feature distributions. The public-dump probe battery uses scikit-learn defaults ($C = 1$, lbfgs) with class-weight balanced, at most 2,000 iterations, standardization, and PCA to at most 128 components — following the probe recipe being reconstructed. Transfer-probe folds use StratifiedGroupKFold with problems as groups; the per-problem diagnostic instead uses sample-disjoint StratifiedKFold within each problem. Confidence intervals are percentile bootstraps over 2,000 problem-clustered draws (1,000 for the post-hoc checks of Appendix D.7), conditional on the fitted out-of-fold predictions rather than refitting models in each draw.

Exclusions. Fifteen of 178 cells whose base run ended before the 512-token forecast window (10 train, 4 validation, 1 test) are excluded from the confirmatory endpoints, per the frozen inclusion rule.

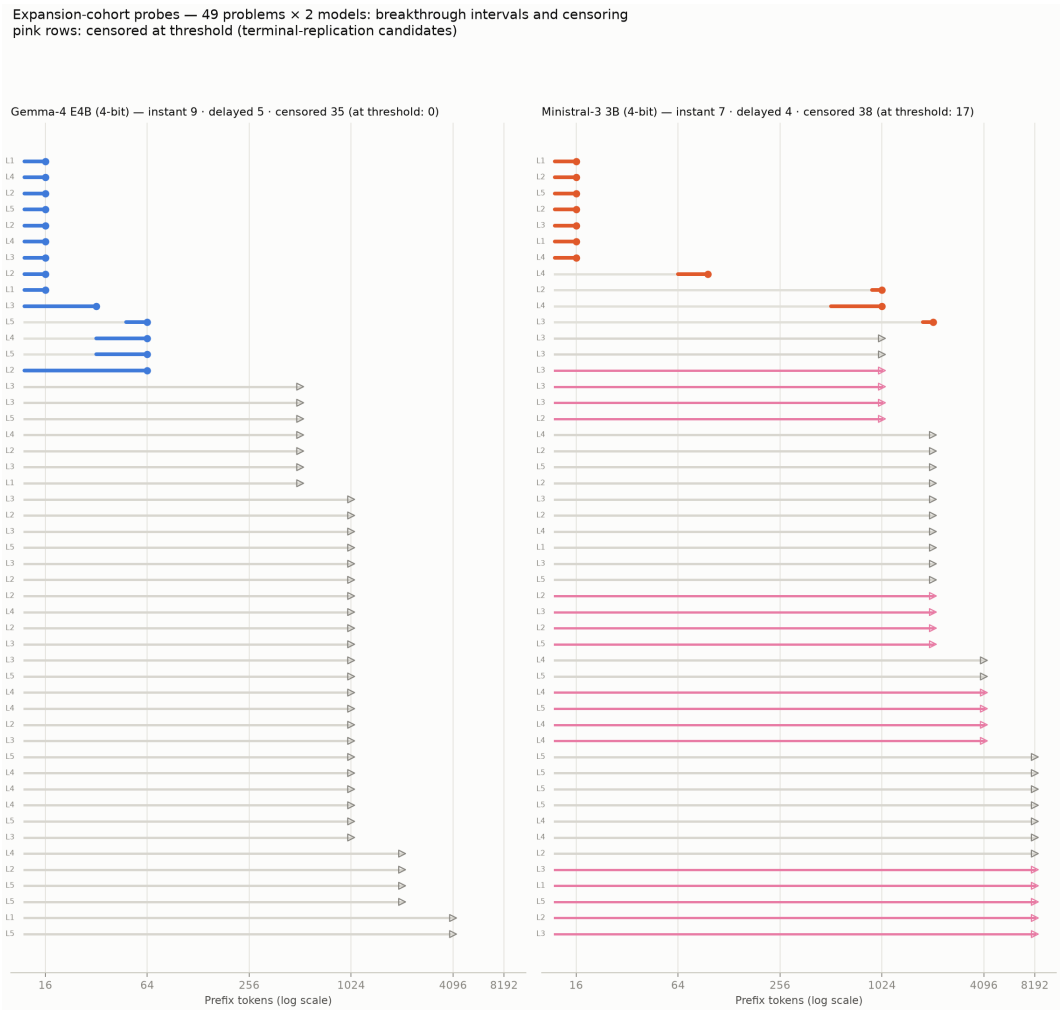

Figure 10: Expansion cohort, one row per trajectory: breakthrough intervals (solid), censoring bounds (arrows), and threshold-censored replication candidates (highlighted).

D.10 REVISION-TIME NUMERICAL REPRODUCTION AUDIT

A CPU-only audit of the stored activations uses the original float32 pipeline function, index order, grouped five-fold splitter, and seed 20260831. All 6,480 eligible $t = 4$ state vectors are finite and have dimension 3,584; stored float16 values are converted to float32. Labels match the full verification table. Training and test problems are disjoint, and each row receives one finite out-of-fold prediction. The environment is Python 3.12.11, NumPy 2.2.6, pandas 3.0.5, and scikit-learn 1.9.0, with Apple's Accelerate BLAS. The pilot takes 4.77 seconds with peak RSS 1.11 GiB.

The rerun yields pooled AUROC 0.8519 with cluster-bootstrap interval [0.7297, 0.9221], failure-weighted within AUROC 0.4961 [0.4661, 0.5271], and pair-weighted within AUROC 0.5155 [0.4818, 0.5636]. These differ from the archived four-decimal results 0.8494, 0.4958, and 0.5146. Calling the original OOF function directly produces bit-identical predictions to the audit wrapper; reducing the thread setting from four to one also leaves predictions unchanged. A separate float64 diagnostic produces pooled AUROC 0.8556 and failure-weighted within AUROC 0.4976, with finite predictions. It is not a replacement for the frozen float32 analysis.

Matrix-multiplication warnings occur in these replays and are retained in the audit records; finite outputs alone do not establish their cause or numerical equivalence. The archived per-row OOF scores, fold assignments, and execution-environment record are unavailable, so the historical discrepancy is unresolved. The rerun does not satisfy the four-decimal reproduction criterion set before this audit. We initially paused the full sweep, then documented an execution-specification addendum to proceed as a separately versioned current-environment reanalysis (Appendix D.12). We have not replaced the archived tables or claimed exact historical reproduction. This unresolved discrepancy limits numerical reproducibility even though the $t = 4$ weighted-mean interpretation is similar.

D.11 ALL-CANDIDATE EXTRA-BATCH RESTART AUDIT

The restart audit reconstructs all 178 cell labels and their primary and sensitivity crossings from 1,112 anchor records. Of these, 148 anchor records have ambiguity top-ups. The label counts and the selected 13-cell comparison reproduce the archived summaries: nine in-grid wins, median advantage 0.3113, four boundary proxies (two wins, two ties), 11 cells reaching the success threshold at the largest restart budget, and nine matching the prefix's pooled rate. These statements concern different selection and measurement units, not independent samples.

For a less outcome-dependent supplementary check, retain every initial terminal candidate: 16 for Ministral-3 and one for Gemma-4, including candidates that subsequently fail pooled-eight confirmation. Selection requires no initial stable crossing and a final initial-batch rate of at least 3/4. Evaluate only branch indices 4–7, whose outcomes were not used in this initial selection. We verified all 68 extra-branch outcomes against their summaries, together with 1,376 wave-3 restart outcomes. This is a post-hoc extra-batch sensitivity conditional on the initial candidate pool, not an independent new cohort or a population estimate.

Ten Ministral-3 candidates have $t + B$ within the restart grid; all ten extra-batch rates exceed the restart reference (nine interpolated, one measured-grid), with mean extra-batch success 0.900 and median advantage 0.3464. The six remaining candidates require a 9,216-token restart outside the measured grid and are reported as proxies, not added to that denominator. Gemma-4's one candidate has extra success 2/4 and ties its restart reference. Table 7 retains all 17 candidates; Figure 3 displays the ten in-grid Ministral-3 comparisons and measured restart brackets.

Table 7: All 17 initial terminal candidates (G4: Gemma-4; M3: Ministral-3). Status indicates pooled-eight confirmation, not inclusion in the extra-batch check. Proxy rows use the largest measured restart budget (8,192) instead of the unmeasured 9,216. Rates are descriptive, with four extra continuations per candidate.

Model	Problem	$t + B$	First	Extra	Pooled	$\hat{R}$	Status
G4	04606	2,048	3/4	2/4	5/8	0.500	no
M3	04987	2,048	3/4	4/4	7/8	0.250	yes
M3	00967	3,072	4/4	4/4	8/8	0.689	yes
M3	03438	3,072	4/4	3/4	7/8	0.292	yes
M3	03840	3,072	4/4	4/4	8/8	0.896	yes
M3	03920	3,072	4/4	4/4	8/8	0.604	yes
M3	04885	3,072	3/4	2/4	5/8	0.146	no
M3	00189	5,120	3/4	4/4	7/8	0.661	yes
M3	02151	5,120	4/4	3/4	7/8	0.241	yes
M3	02283	5,120	4/4	4/4	8/8	0.830	yes
M3	02576	5,120	4/4	4/4	8/8	0.830	yes
M3	02206	9,216	4/4	4/4	8/8	0.500	yes / proxy
M3	02273	9,216	3/4	1/4	4/8	1.000	no / proxy
M3	02303	9,216	4/4	4/4	8/8	1.000	yes / proxy
M3	02985	9,216	3/4	2/4	5/8	0.000	no / proxy
M3	04024	9,216	4/4	4/4	8/8	1.000	yes / proxy
M3	04745	9,216	3/4	4/4	7/8	0.500	yes / proxy

D.12 Current-environment full sweep and heterogeneity

The reanalysis runs the ten previously specified anchors without hyperparameter, seed, or permutation search. The $t = 4$ union-trained OOF scores are reused from the audited pilot; nine remaining union fits and all A/B/C diagnostics use the original recipes in the current environment. Runtime is 109.50 seconds, peak RSS 1.29 GiB, with thread settings capped at four. All eligible state vectors have the expected dimension and finite values; fitted parameters and predictions are finite, and no convergence failure is recorded. Matmul warnings remain in the machine-readable fit metadata; these checks do not establish that numerical warnings are harmless or explain the historical mismatch.

The extraction implementation averages the last four generated-token hidden states before storing each anchor vector; it is not a single-token state. All same-set quantities and intervals are computed from persistent OOF scores. Original problem identities are retained in cluster bootstrap draws: a problem drawn w_c times contributes w_c^2 within pairs, while failure-weighted and macro summaries use w_c. Thus the pair-bootstrap within interval need not equal the legacy bootstrap of per-problem AUCs with linear pair weights. All intervals are pointwise, conditional on fitted predictions, and omit training/refitting uncertainty. Bootstrap pair-fraction distributions depend on this cluster-multiplicity convention; we report observed pair fractions, not a population confidence claim about the fraction of reasoning information.

Tables 8 and 9 retain the old/new comparison and both full panels. Failure-weighted within intervals on W contain chance at all ten anchors. The A/B/C summaries are shown in Table 10; all B summary fields reproduce the archived reporting precision at every anchor, while A/C and union-trained discrepancies remain in the version-comparison records. No new permutation runs were performed; historical permutation ranges are not a newly calibrated null for this version.

Table 11 and Figure 2 include every selected problem. B scores for problem 13 are undefined at $t = 512$ because its surviving attempts are single-class; the output records the exclusion rather than filling it with chance. Current B32/B64 per-problem AUROCs have Spearman correlation 0.7143 (22 problems), and A32/B32 correlation is 0.5121. Length diagnostics independently recover median failure/success character counts of 2,363.5/15,321 for problem 20 and 1,299/5,168 for problem 22; problem 13 has 982/1,016 characters. These are full-output character counts, not tokens or early prediction features, and do not identify the cause of the diagnostic separation.

Table 8: Archived versus separately versioned current-environment results at every anchor. New same-set quantities were not available in the archived aggregate report; no historical OOF equivalence is claimed.

t	P pooled (old / new)	P within PW	W pooled	W within FW (old / new)
4	0.8494 / 0.8519	0.4744	0.7610	0.4958 / 0.4961
8	0.6784 / 0.6925	0.5417	0.5436	0.4815 / 0.4800
16	0.5860 / 0.5866	0.3782	0.5614	0.4954 / 0.4951
32	0.5610 / 0.5607	0.4647	0.5151	0.5252 / 0.5253
64	0.5815 / 0.5816	0.3718	0.5546	0.4812 / 0.4813
128	0.5484 / 0.5484	0.6218	0.4841	0.5155 / 0.5155
192	0.5718 / 0.5717	0.5256	0.5165	0.4993 / 0.4993
256	0.5645 / 0.5646	0.4551	0.5129	0.5297 / 0.5297
384	0.5475 / 0.5474	0.4189	0.4891	0.5088 / 0.5089
512	0.5468 / 0.5468	0.3819	0.4847	0.4967 / 0.4967

Table 9: Full current-environment same-set results. Intervals use 2,000 problem-cluster draws at fixed OOF scores; repeated clusters retain their original problem identity. No refitting uncertainty or multiple-testing correction is included.

t	Panel	Pooled [95% CI]	Within PW [95% CI]	Between [95% CI]
4	P	0.852 [0.732, 0.925]	0.474 [0.343, 0.644]	0.853 [0.734, 0.926]
4	W	0.761 [0.635, 0.848]	0.515 [0.473, 0.598]	0.767 [0.644, 0.864]
8	P	0.693 [0.510, 0.835]	0.542 [0.420, 0.686]	0.693 [0.510, 0.837]
8	W	0.544 [0.400, 0.681]	0.496 [0.467, 0.522]	0.545 [0.396, 0.692]
16	P	0.587 [0.459, 0.704]	0.378 [0.189, 0.647]	0.587 [0.460, 0.707]
16	W	0.561 [0.433, 0.687]	0.510 [0.465, 0.585]	0.563 [0.428, 0.695]
32	P	0.561 [0.428, 0.706]	0.465 [0.296, 0.628]	0.561 [0.429, 0.707]
32	W	0.515 [0.391, 0.640]	0.526 [0.471, 0.599]	0.515 [0.383, 0.645]
64	P	0.582 [0.436, 0.704]	0.372 [0.221, 0.550]	0.582 [0.436, 0.705]
64	W	0.555 [0.443, 0.668]	0.478 [0.442, 0.530]	0.557 [0.440, 0.676]
128	P	0.548 [0.427, 0.668]	0.622 [0.421, 0.839]	0.548 [0.426, 0.669]
128	W	0.484 [0.377, 0.591]	0.509 [0.475, 0.557]	0.483 [0.369, 0.596]
192	P	0.572 [0.416, 0.706]	0.526 [0.259, 0.785]	0.572 [0.416, 0.706]
192	W	0.517 [0.396, 0.634]	0.499 [0.462, 0.548]	0.517 [0.391, 0.643]
256	P	0.565 [0.407, 0.684]	0.455 [0.278, 0.623]	0.565 [0.407, 0.685]
256	W	0.513 [0.397, 0.621]	0.509 [0.471, 0.567]	0.513 [0.390, 0.628]
384	P	0.547 [0.433, 0.654]	0.419 [0.237, 0.598]	0.548 [0.433, 0.656]
384	W	0.489 [0.395, 0.585]	0.513 [0.482, 0.552]	0.489 [0.389, 0.588]
512	P	0.547 [0.429, 0.654]	0.382 [0.153, 0.658]	0.547 [0.431, 0.655]
512	W	0.485 [0.384, 0.586]	0.499 [0.467, 0.536]	0.484 [0.377, 0.590]

Table 10: Current-environment A/B/C failure-weighted means and pointwise 95% intervals (1,000 problem-cluster draws). A is transductive; B is sample-disjoint within problem. All B summary fields match the archived report at its reporting precision; A/C differences are retained in the version comparison records.

t	C: within-set uncentered	A: centered transfer	B: per-problem
4	0.517 [0.491, 0.549]	0.490 [0.464, 0.523]	0.476 [0.434, 0.528]
8	0.509 [0.479, 0.551]	0.526 [0.502, 0.550]	0.479 [0.449, 0.519]
16	0.514 [0.486, 0.548]	0.486 [0.450, 0.534]	0.481 [0.437, 0.536]
32	0.514 [0.463, 0.557]	0.556 [0.522, 0.588]	0.519 [0.478, 0.560]
64	0.507 [0.464, 0.543]	0.532 [0.502, 0.565]	0.558 [0.523, 0.593]
128	0.523 [0.492, 0.560]	0.524 [0.504, 0.548]	0.519 [0.489, 0.555]
192	0.491 [0.462, 0.518]	0.526 [0.498, 0.556]	0.491 [0.452, 0.543]
256	0.542 [0.488, 0.607]	0.536 [0.503, 0.570]	0.531 [0.493, 0.576]
384	0.517 [0.500, 0.533]	0.529 [0.515, 0.546]	0.532 [0.493, 0.576]
512	0.489 [0.449, 0.526]	0.520 [0.466, 0.580]	0.507 [0.457, 0.548]

Table 11: All 22 problems: current-environment diagnostics, ordered by problem ID, not performance. Class counts refer to $t = 4$; the full ten-anchor coverage and exclusions accompany the row-level outputs.

Problem	Fail	Success	C4	A4	B4	B32	B64
3	151	105	0.489	0.511	0.475	0.433	0.466
13	17	239	0.695	0.329	0.777	0.863	0.759
14	37	219	0.512	0.512	0.436	0.452	0.418
15	36	220	0.559	0.559	0.509	0.432	0.430
20	18	238	0.934	0.927	0.924	0.894	0.865
21	31	225	0.576	0.576	0.533	0.522	0.575
22	49	207	0.839	0.859	0.844	0.818	0.868
31	237	19	0.489	0.489	0.386	0.380	0.650
34	23	233	0.522	0.478	0.503	0.476	0.427
42	202	54	0.463	0.463	0.507	0.513	0.553
54	43	213	0.461	0.461	0.488	0.450	0.461
63	242	14	0.493	0.493	0.361	0.547	0.594
78	40	216	0.424	0.576	0.512	0.599	0.650
79	107	149	0.459	0.459	0.501	0.486	0.482
81	223	33	0.611	0.389	0.555	0.588	0.589
83	216	40	0.509	0.491	0.409	0.605	0.554
92	78	178	0.509	0.509	0.437	0.497	0.527
101	216	40	0.517	0.483	0.478	0.463	0.488
114	63	193	0.465	0.465	0.496	0.537	0.461
117	36	220	0.455	0.455	0.499	0.625	0.626
119	99	157	0.475	0.475	0.498	0.493	0.509
125	14	242	0.602	0.602	0.562	0.551	0.639

E SUPPORTING PREDICTION EXPERIMENTS

E.1 A PRE-REGISTERED INCREMENTAL-PREDICTION TEST

Can early features predict outcomes beyond a specified question-only baseline? The frozen primary endpoint is eventual success of the 16K-token instrumented run; the secondary endpoint is restart-solvability, $\hat{R}(4{,}096) \geq \tau$, on non-instant cells, not that run's correctness. We compare a question-only baseline with the same baseline plus fifteen early-window summaries of entropy, surprisal, divergences, hidden-state geometry, and spectral dynamics through 512 tokens (Appendix D.1). Logistic models use problem-grouped cross-validation on train+validation. The frozen success criterion is that the paired ΔAUROC's problem-clustered 95% CI exclude zero. The baseline includes human difficulty level, topic, five text statistics (character and whitespace-token counts and numeric, operator and equation token counts), and model identity. It is available before generation but does not exhaust model-specific difficulty; pre-generation activation probes provide a stronger alternative (Lugoloobi et al., 2026).

Neither endpoint met the criterion (Figure 11). Primary: baseline AUROC 0.83 versus 0.85 with early signals, $\Delta = +0.026$ $[-0.054, +0.167]$; secondary: 0.88 versus 0.78, $\Delta = -0.090$ $[-0.213, +0.033]$. A level-free sensitivity variant also fails the criterion. The primary test contains 33 cells from 17 problems, with only four failures; the interval leaves meaningful positive effects open. The timing endpoint has four test-split interior events, fails its power gate, and is not fit. No equivalence margin was preregistered: these are absences of detected gain, not demonstrations of equivalence or universal unpredictability.

Before the 512-observed-token eligibility filter, the held-out cohort contains 34 cells: 30 successful base runs and four failures. The filter removes one successful Gemma run with 491 observed tokens, leaving 29 successes and the same four failures. Thus the scarcity of failures in this endpoint is already present in the selected held-out base runs, not caused by filtering out failing runs at this length cutoff.

After the single confirmatory evaluation, a labeled post-hoc sweep reused the test split at $t \in \{128, 256, 1{,}024, 2{,}048\}$. Including the 512-token result, 8 of 10 point estimates are negative and all intervals include zero (Figure 12); sample attrition and low power limit interpretation. Per-model descriptives also differ: baseline AUROC is 1.0 for Gemma-4 ($n = 16/11$ for the two endpoints) and 0.69–0.73 for Ministral-3. These small subgroups do not establish that difficulty fully determines one model's outcomes or that remaining variation in the other is intrinsically unpredictable.

E.2 HIGH LABEL-INFORMED DIFFICULTY REFERENCES ON PUBLIC CORPORA

On 192,315 DeepSeek-R1 generations across 91,573 problems (Hugging Face Open-R1 team, 2025), the leave-one-out pass rate reaches AUROC 0.873 $[0.870, 0.876]$ (problem-clustered bootstrap; estimator frozen before evaluation). For 92% of problems it uses a single other attempt's label. It reads neither the target trace nor the question, but is *label-informed*, not a deployable question-only predictor or a mathematical performance ceiling. Dataset curation restricts generalization; no guaranteed direction of bias is established here.

This reference lies within the 0.79–0.95 range reported by internal-state probe studies (Zhang et al., 2025; Yuan et al., 2026; Sun et al., 2026; David, 2025). These numbers concern different datasets and protocols, so the comparison is contextual, not matched evidence against any particular published result (Figure 13). On 16,384 public GPQA-diamond (Rein et al., 2024) samples (64 questions $\times$ 256 attempts, Llama-3.3-70B (Singhi et al., 2025)), the corresponding reference is 0.917 $[0.875, 0.938]$. This extends the reference calculation beyond mathematics, not the activation-probe dissection.

E.3 AN EARLY-PROBE RECONSTRUCTION WITH TWO EVALUATION SETS

We reconstruct an early-window probe setup on public trajectories: 128 MATH problems with 256 R1-Distill-Qwen-7B samples each (Singhi, 2025), released alongside the data of Singhi et al. (2025) but not described in that paper. No new generations are sampled. We teacher-force the dump's tokens through a 4-bit model, with observed-token/top-1 agreement 0.906 (protocol frozen before

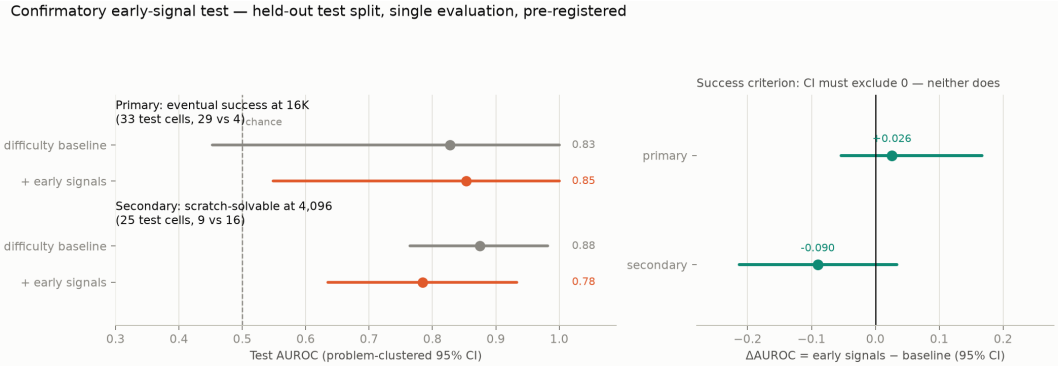

Figure 11: The confirmatory test: single evaluation on the held-out split. Neither endpoint's ΔAUROC CI excludes zero; the primary interval still permits meaningful positive effects.

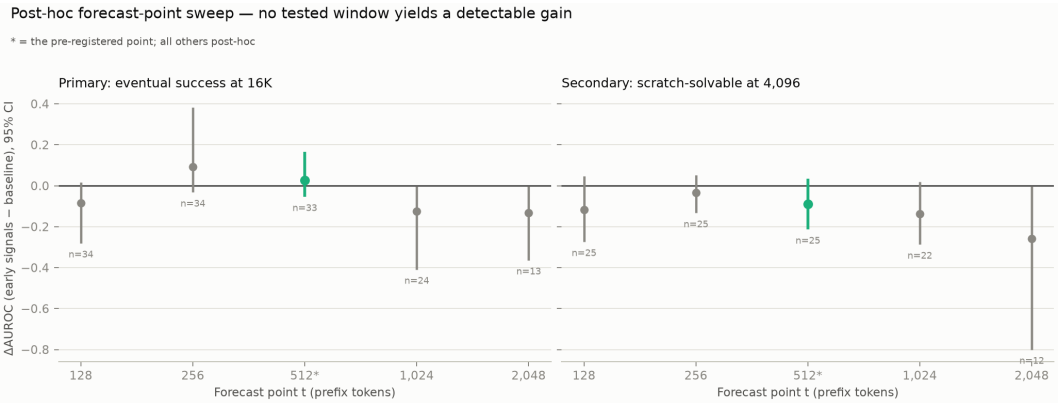

Figure 12: Post-hoc forecast-point sweep of the frozen feature family. No tested window yields a detectable gain; longer windows also select the surviving trajectories.

extraction). This compares model argmax predictions with observed tokens, not full-precision with quantized activations; it does not validate hidden-state fidelity.

A probe of the mean hidden state over the last four tokens at anchor $t = 4$ yields archived pooled AUROC 0.849 $[0.735, 0.919]$ on the frozen 8-per-problem set ($n = 1{,}024$), comparable to the 0.84 of David (2025). This is a reconstruction in a related regime, not an exact replication on the original samples. Training uses problem-disjoint out-of-fold prediction on the union of the selected samples.

The values reported here and in Figure 14 are the archived analysis. A revision-time rerun on the stored states gives pooled AUROC 0.8519 and failure-weighted within AUROC 0.4961, versus archived 0.8494 and 0.4958. It does not meet our four-decimal reproduction criterion. We retain both records rather than silently replace the archived results; the numerical audit and unresolved scope of this discrepancy are reported in Appendix D.10.

The within evaluation uses all 256 samples of 22 mid-band problems ($n = 5{,}632$), selected by full-dump pass rate before state extraction. These problems contain 94% of failures and 84% of same-cell discordant pairs in the dump. For the same fitted probe, within AUROC is 0.496 $[0.466, 0.527]$ with preregistered failure-count weighting and 0.515 $[0.481, 0.562]$ with discordant-pair weighting. Both intervals include chance at all ten anchors through $t = 512$; attrition leaves 21 problems with both classes at the last anchor (Appendix D). These intervals describe weighted means for this probe on this selected set, not individual problems or all decodable information.

The pooled and within evaluation sets differ. Only pair weighting corresponds to Eq. (1), and it cannot be substituted into a decomposition of the other evaluation set. The contrast shows that

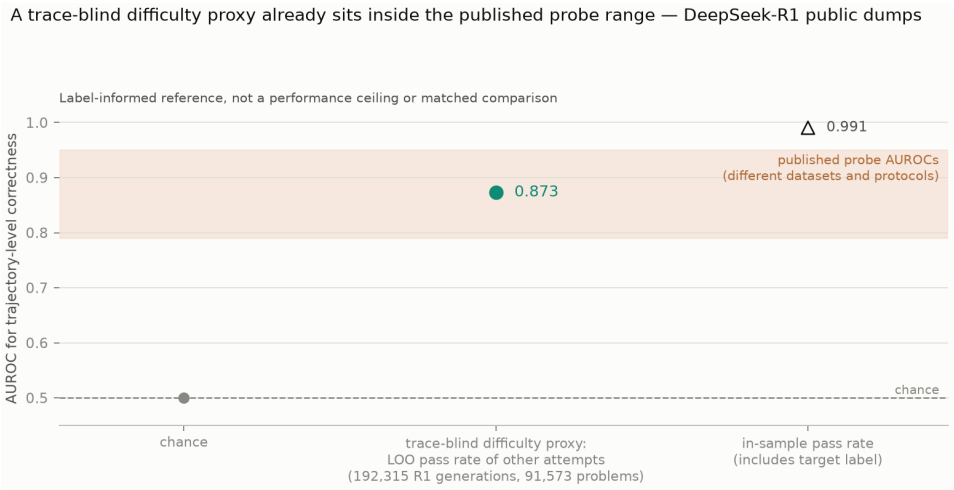

Figure 13: A trace-blind, label-informed LOO reference reaches AUROC 0.873 on public R1 generations. The published range concerns other datasets and protocols; neither the LOO nor the non-LOO in-sample reference is a deployable question-only baseline or an upper bound on probe performance.

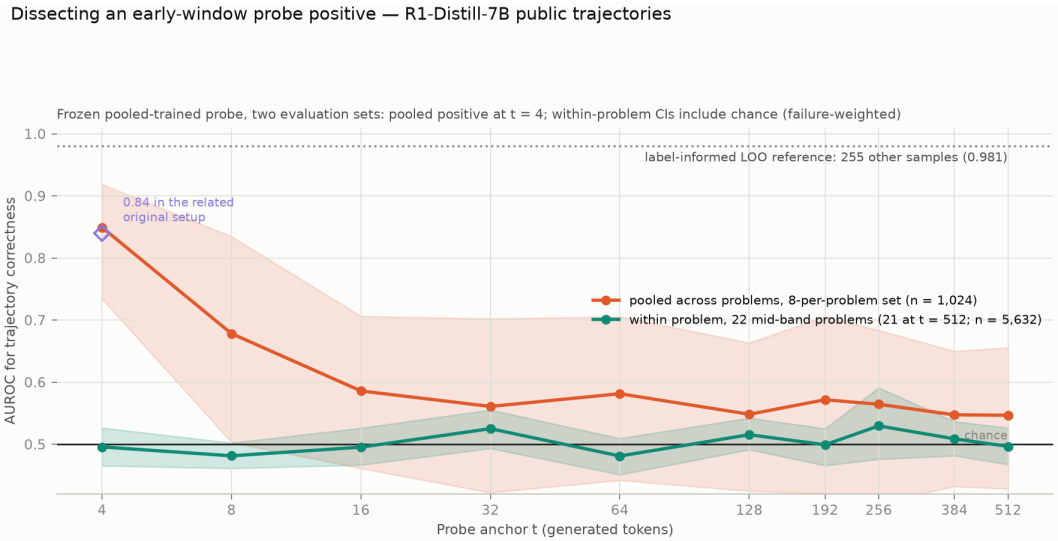

Figure 14: Frozen pooled-trained probe, two evaluation sets, 6,480 trajectories entering out-of-fold prediction. All within-problem intervals include chance; the post-hoc within-targeted probes are different fits and exhibit heterogeneous positives.

high pooled performance need not transfer to strong within-problem discrimination here; it does not quantify the fraction of pooled AUROC caused by difficulty.

As a separate reference, LOO scoring from the full 32,768-sample verification table (255 other attempts per problem) yields pooled AUROC 0.981 $[0.955, 0.991]$ on the same pooled subset. With sample composition fixed through $t = 128$, the probe's pooled AUROC falls to 0.586 at $t = 16$ (CI includes 0.5) and roughly 0.55 at later anchors. Decaying prompt-related information is a possible explanation, not a mechanism established by this performance curve.

Within-targeted positives and their limitations are reported in Section 2 and Appendices D.7 and D.12. The archived union-trained probe null is not an absence-of-information result.

Table 12: All numerical sensitivities: NumPy Accelerate (Acc) versus OpenBLAS (OB), float32/64. Other package versions and original folds are fixed. $\max|\Delta s|$ is over union rows, and flips are W correct–incorrect pair-order changes (%), both relative to original float32 scores. Warning counts cover fit and prediction; all fits converge with finite outputs.

t	BLAS	Bits	W pool	W within	P pool	$\max\|\Delta s\|$	Flips	Warn.
4	Acc	32	0.7610	0.5155	0.8519	0.0000	0.00	13110
4	Acc	64	0.7530	0.5192	0.8556	0.1560	1.93	21018
32	Acc	32	0.5151	0.5265	0.5607	0.0000	0.00	4302
32	Acc	64	0.5152	0.5265	0.5612	0.0027	0.03	4104
512	Acc	32	0.4847	0.4988	0.5468	0.0000	0.00	1878
512	Acc	64	0.4847	0.4987	0.5469	0.0010	0.02	1506
4	OB	32	0.7573	0.5200	0.8508	0.0805	3.48	0
4	OB	64	0.7554	0.5192	0.8638	0.1519	0.43	0
32	OB	32	0.5152	0.5264	0.5609	0.0029	0.02	0
32	OB	64	0.5152	0.5263	0.5612	0.0027	0.04	0
512	OB	32	0.4847	0.4987	0.5468	0.0010	0.02	0
512	OB	64	0.4847	0.4987	0.5469	0.0010	0.02	0

F STORED-DATA NUMERICAL, ROBUSTNESS, AND SELECTION AUDIT

The following revision-time analyses were specified before their execution, after the earlier results were known. They are exploratory sensitivities, not new preregistered confirmations. No new model generations, activation extraction, or raw-answer verification were performed. Original and current-environment outputs were preserved. The anonymous supplement includes the specification, all 48 new OOF score files (240 folds), metadata, and executable fixed-score replay.

F.1 NUMERICAL LOCALIZATION AND A SECOND NUMPY BLAS BUILD

Raising divide/overflow/invalid events as exceptions localizes the first $t = 4$ warning to the product AQ in the randomized range finder used by PCA. The captured finite float32 operands have shapes $5{,}216 \times 3{,}584$ and $3{,}584 \times 138$. In isolation, the Accelerate product again emits all three warning types despite finite output. Its relative Frobenius difference from a float64, non-BLAS `einsum` reference is 1.06×10^{-6}; the OpenBLAS float32 product is warning-free with difference 4.35×10^{-7}. The float64 Accelerate replay of these same operands still warns but matches that reference in this check; the OpenBLAS float64 relative difference is 1.75×10^{-15}. These checks narrow the warning to a reproducible numerical operation/build interaction, not an identified source-code defect or a validation of every downstream operation. The activation-derived operand arrays are retained locally but are not redistributed in the supplement.

For the pipeline sensitivity, Python 3.12.11, NumPy 2.2.6, SciPy 1.18.0, scikit-learn 1.9.0, labels, folds, and seeds are fixed. Only the NumPy wheel is swapped from its Accelerate build to a macOS 11 arm64 wheel linked to OpenBLAS 0.3.29; SciPy is unchanged, so this is not an all-library BLAS swap. Each build fits float32 and float64 versions at $t = 4, 32, 512$. All 12 configurations converge with finite parameters and predictions. All six OpenBLAS runs are warning-free; the six Accelerate runs retain matrix warnings. No run replaces the original float32 primary estimate.

Table 12 reports all outcomes, including score and pair-order changes. At $t = 4$, the float32 backend swap changes W pooled AUROC by -0.0037 and within AUROC by $+0.0045$, but is not prediction-identical: the largest union-row score difference is 0.0805 and 3.48% of within-problem opposite-outcome pairs change ordering or tie status. The archived environment and OOF rows remain unavailable; these new checks do not establish the cause of the historical discrepancies documented in Appendix D.10.

F.2 FIXED RECIPE, SPLIT, AND EVALUATION-COHORT SENSITIVITIES

The sensitivity grid crosses anchors $4, 32, 512$, PCA dimensions $32, 128, 256$ or no PCA, and split seeds $20260831, 20260913, 20260914$: 36 configurations, 180 folds. All use the original eligible union population, train-fold-only standardization, five problem-disjoint folds, balanced logistic re-

Table 13: All 12 fixed probe/split sensitivities at $t = 4$, original float32 environment. S0/S1/S2 are seeds 20260831/20260913/20260914. Intervals are pointwise, fixed-OOF, pair-weighted W intervals, not adjusted across settings. No configuration is selected as a replacement primary.

PCA	Split	P pool	W pool	W within	95% CI	W FW
32	S0	0.8323	0.6442	0.5196	[0.482, 0.598]	0.5179
32	S1	0.7496	0.5684	0.4955	[0.468, 0.536]	0.5046
32	S2	0.7752	0.6049	0.5097	[0.471, 0.563]	0.4929
128	S0	0.8519	0.7610	0.5155	[0.473, 0.598]	0.4961
128	S1	0.8277	0.6611	0.5223	[0.475, 0.616]	0.4961
128	S2	0.8499	0.6716	0.5064	[0.460, 0.600]	0.4872
256	S0	0.8823	0.7748	0.5258	[0.482, 0.619]	0.5142
256	S1	0.8574	0.6564	0.5319	[0.486, 0.624]	0.5193
256	S2	0.8515	0.6863	0.5204	[0.462, 0.611]	0.5122
None	S0	0.8617	0.7677	0.5255	[0.478, 0.620]	0.4989
None	S1	0.8544	0.6710	0.5378	[0.491, 0.629]	0.5201
None	S2	0.8508	0.6841	0.5209	[0.466, 0.611]	0.5124

Table 14: All 12 fixed probe/split sensitivities at $t = 32$, original float32 environment. S0/S1/S2 are seeds 20260831/20260913/20260914. Intervals are pointwise, fixed-OOF, pair-weighted W intervals, not adjusted across settings. No configuration is selected as a replacement primary.

PCA	Split	P pool	W pool	W within	95% CI	W FW
32	S0	0.5758	0.5023	0.4877	[0.441, 0.532]	0.4959
32	S1	0.5342	0.4679	0.5060	[0.462, 0.565]	0.5045
32	S2	0.5139	0.4761	0.4989	[0.449, 0.570]	0.5046
128	S0	0.5607	0.5151	0.5265	[0.471, 0.599]	0.5253
128	S1	0.5299	0.5224	0.5272	[0.472, 0.598]	0.5159
128	S2	0.5177	0.4436	0.5224	[0.464, 0.609]	0.5029
256	S0	0.5235	0.5720	0.5239	[0.476, 0.594]	0.5108
256	S1	0.5638	0.6222	0.5145	[0.464, 0.573]	0.4981
256	S2	0.5214	0.4860	0.4915	[0.440, 0.550]	0.4689
None	S0	0.4890	0.5127	0.4989	[0.462, 0.537]	0.4873
None	S1	0.4843	0.5026	0.4965	[0.454, 0.544]	0.4750
None	S2	0.4430	0.5027	0.4887	[0.448, 0.541]	0.4652

gression with $C = 1$, `lbfgs`, and 2,000 maximum iterations. PCA randomness is split seed plus fold. The original float32 environment is retained, including its warnings. All fits converge; no configuration is selected for a better result. This varies linear-probe recipes, not model families or datasets.

Tables 13, 14, and 15 retain every configuration. All 36 W within pointwise intervals include chance. At $t = 4$, pooled W AUROC spans 0.5684–0.7748 and within AUROC spans 0.4955–0.5378. At $t = 32$, the respective ranges are 0.4436–0.6222 and 0.4877–0.5272; at $t = 512$, 0.4564–0.5308 and 0.4920–0.5314. Thus pooled magnitudes vary substantially with the recipe/split, whereas these checks do not establish a within-ranking positive. This is not equivalence or a simultaneous null across configurations.

For cohort sensitivity, we use recorded original scores without refitting. Leaving each of the 22 W problems out of evaluation at $t = 4$ yields within AUROC 0.4975–0.5206. A second analysis restricts every anchor to the same 5,322 W trajectories available at $t = 512$, giving $t = 4$ pooled/within AUROC 0.7633/0.4959. Table 16 retains all anchors. The common cohort remains conditioned on survival; it is not a fresh representative sample, and evaluation-only omission does not measure training sensitivity to removing that problem from other folds.

Table 15: All 12 fixed probe/split sensitivities at $t = 512$, original float32 environment. S0/S1/S2 are seeds 20260831/20260913/20260914. Intervals are pointwise, fixed-OOF, pair-weighted W intervals, not adjusted across settings. No configuration is selected as a replacement primary.

PCA	Split	P pool	W pool	W within	95% CI	W FW
32	S0	0.5748	0.4746	0.4920	[0.457, 0.529]	0.4899
32	S1	0.5758	0.4679	0.4956	[0.461, 0.528]	0.4984
32	S2	0.5383	0.4564	0.4987	[0.464, 0.535]	0.4944
128	S0	0.5468	0.4847	0.4988	[0.467, 0.536]	0.4967
128	S1	0.5577	0.4886	0.5064	[0.476, 0.548]	0.4931
128	S2	0.5394	0.4874	0.4944	[0.462, 0.527]	0.4897
256	S0	0.5724	0.4849	0.5106	[0.475, 0.548]	0.5019
256	S1	0.5790	0.4784	0.5117	[0.477, 0.554]	0.4998
256	S2	0.5336	0.4779	0.5065	[0.476, 0.541]	0.5007
None	S0	0.5948	0.5195	0.5300	[0.496, 0.571]	0.5168
None	S1	0.6132	0.5308	0.5314	[0.492, 0.578]	0.5156
None	S2	0.5821	0.5172	0.5244	[0.485, 0.570]	0.5088

Table 16: Evaluation-only sensitivities on recorded scores. LOPO omits each W problem in turn without refitting; ranges are not confidence intervals. Common cohort uses the same 5,322 W trajectories surviving to 512 tokens at every anchor. It remains survival-conditioned.

t	LOPO within	LOPO pool–within	Common pool	Common within	95% CI
4	0.498–0.521	0.220–0.273	0.7633	0.4959	[0.466, 0.529]
8	0.492–0.500	0.009–0.075	0.5391	0.4999	[0.473, 0.523]
16	0.493–0.517	0.021–0.081	0.5487	0.4937	[0.457, 0.529]
32	0.514–0.534	-0.047–0.020	0.5172	0.5181	[0.469, 0.570]
64	0.469–0.484	0.052–0.097	0.5470	0.4810	[0.446, 0.529]
128	0.501–0.515	-0.048–0.005	0.4788	0.5080	[0.473, 0.552]
192	0.490–0.505	0.002–0.056	0.5103	0.5003	[0.462, 0.548]
256	0.500–0.514	-0.018–0.040	0.4997	0.5109	[0.475, 0.570]
384	0.509–0.517	-0.045–0.005	0.4891	0.5081	[0.479, 0.543]
512	0.495–0.504	-0.032–0.007	0.4847	0.4988	[0.467, 0.536]

F.3 Exact offline selection among recorded attempts

For each W problem with n eligible attempts, draw a uniform size-k subset without replacement, $k \in \{2, 4, 8\}$, and select its highest original union-probe OOF score. Ties are resolved uniformly. We compute the expectation over all subsets exactly: for a tied-score group of size m with ℓ lower-scored attempts, the probability the selected attempt belongs to that group is

$$\frac{\binom{\ell+m}{k} - \binom{\ell}{k}}{\binom{n}{k}}.$$

Its contribution to correctness is this probability times the group's empirical correct fraction. Binomial coefficients with upper argument below k are zero. The random-selection comparator has expected correctness n^+/n; the label-informed oracle is $1-\binom{n^-}{k}/\binom{n}{k}$. Exhaustive small-subset tests independently check this implementation, including ties, constant scores, and pure-label problems. Candidate subsets are not restricted to opposite-outcome pairs.

We average gains equally across all 22 problems, including single-class survivors whose gain is zero. Every problem has at least eight candidates; none is excluded for candidate count. Paired percentile intervals resample problems 2,000 times (seed 20260913), holding scores and empirical within-problem distributions fixed. Failure-weighted gains are also supplied in the numeric supplement. These intervals omit refitting and within-problem sampling uncertainty and are not adjusted over windows or candidate sizes.

At $t = 4$, random correctness is 61.33%; probe selection gives 61.74%, 61.83%, and 61.57% for $k = 2, 4, 8$, respectively. Gains are $+0.42$ $[-0.48, +1.63]$, $+0.50$ $[-0.99, +2.60]$, and $+0.24$ $[-1.68, +2.59]$ percentage points. The corresponding label-informed oracle expectations

Table 17: All offline selection gains on the anchor-eligible W cohort. Random correctness is in percent; each candidate-size cell reports gain and paired pointwise 95% interval in percentage points. All 22 problems enter, including single-class survivors. All 30 intervals include zero; this is not an equivalence test.

t	Random (%)	$k = 2$: gain [CI]	$k = 4$: gain [CI]	$k = 8$: gain [CI]
4	61.33	+0.42 [-0.48, 1.63]	+0.50 [-0.99, 2.60]	+0.24 [-1.68, 2.59]
8	61.33	-0.10 [-0.56, 0.35]	-0.01 [-0.91, 0.88]	+0.20 [-1.23, 1.67]
16	61.33	+0.26 [-0.65, 1.39]	+0.56 [-1.24, 2.66]	+0.70 [-1.78, 3.27]
32	61.33	+0.71 [-0.30, 1.80]	+1.09 [-0.83, 3.13]	+1.30 [-1.22, 3.99]
64	61.33	-0.59 [-1.33, 0.17]	-0.83 [-2.24, 0.50]	-0.67 [-2.80, 1.28]
128	61.33	+0.24 [-0.44, 0.90]	+0.25 [-0.93, 1.34]	+0.16 [-1.67, 1.85]
192	61.34	-0.04 [-0.66, 0.72]	-0.03 [-1.04, 1.14]	-0.05 [-1.52, 1.47]
256	61.36	+0.24 [-0.51, 1.04]	+0.42 [-0.85, 1.74]	+0.59 [-1.42, 2.52]
384	61.68	+0.43 [-0.14, 1.04]	+0.59 [-0.42, 1.67]	+0.55 [-0.99, 2.23]
512	62.39	-0.01 [-0.58, 0.54]	+0.09 [-1.04, 1.14]	+0.32 [-1.50, 2.07]

Table 18: All offline selection gains on the common 512-token-survivor W cohort. Random correctness is in percent; each candidate-size cell reports gain and paired pointwise 95% interval in percentage points. All 22 problems enter, including single-class survivors. All 30 intervals include zero; this is not an equivalence test.

t	Random (%)	$k = 2$: gain [CI]	$k = 4$: gain [CI]	$k = 8$: gain [CI]
4	62.39	+0.00 [-0.57, 0.60]	-0.27 [-1.11, 0.59]	-0.57 [-1.68, 0.48]
8	62.39	-0.01 [-0.40, 0.37]	+0.12 [-0.75, 0.96]	+0.34 [-1.05, 1.80]
16	62.39	-0.05 [-0.74, 0.60]	-0.14 [-1.55, 1.03]	-0.14 [-2.18, 1.64]
32	62.39	+0.54 [-0.31, 1.32]	+0.72 [-0.89, 2.15]	+0.79 [-1.34, 2.78]
64	62.39	-0.46 [-1.16, 0.23]	-0.61 [-1.95, 0.60]	-0.46 [-2.55, 1.40]
128	62.39	+0.21 [-0.42, 0.81]	+0.23 [-0.89, 1.25]	+0.26 [-1.44, 1.88]
192	62.39	-0.03 [-0.61, 0.65]	+0.07 [-0.87, 1.11]	+0.25 [-1.00, 1.60]
256	62.39	+0.32 [-0.36, 1.01]	+0.68 [-0.47, 1.82]	+0.99 [-0.73, 2.62]
384	62.39	+0.23 [-0.24, 0.74]	+0.19 [-0.65, 1.08]	+0.08 [-1.33, 1.67]
512	62.39	-0.01 [-0.58, 0.54]	+0.09 [-1.04, 1.14]	+0.32 [-1.50, 2.07]

are 74.75%, 83.25%, and 90.19%. These oracle gaps describe opportunities in the empirical candidate population, not achievable performance by a deployable score. Tables 17 and 18 retain all ten anchors for the eligible and common-survivor cohorts; all 60 gain intervals include zero. The analysis does not establish that every alternative probe lacks selection value, nor that the measured gains are practically equivalent to zero. It evaluates selection offline, not a tested early-stopping or compute-saving controller.

F.4 AN EQUAL-CELL-SIZE CONSEQUENCE OF THE PAIR IDENTITY

This algebraic observation adds no fitted model or empirical experiment. Let $K \geq 2$ cells each contain m attempts, and write $p_c = n_c^+/m$, $\bar{p} = K^{-1}\sum_c p_c$, and $v = K^{-1}\sum_c (p_c - \bar{p})^2$ for their empirical pass-rate mean and population-normalized variance. Assume $0 < \bar{p} < 1$ so pooled AUROC is defined. Substitution into Eq. (1) gives

$$\begin{aligned}\lambda &= \frac{m^2 \sum_c p_c(1-p_c)}{(Km)^2 \bar{p}(1-\bar{p})} \\ &= \frac{1}{K}\frac{\bar{p}(1-\bar{p}) - v}{\bar{p}(1-\bar{p})} = \frac{1}{K}\left(1 - \frac{v}{\bar{p}(1-\bar{p})}\right). \end{aligned} \tag{4}$$

Thus $0 \leq \lambda \leq 1/K$, with upper equality if all empirical pass rates coincide. When within-cell pairs exist, both components are defined and their AUROCs lie in $[0, 1]$, so

$$|A_{\text{pooled}} - A_{\text{between}}| = \lambda |A_{\text{within}} - A_{\text{between}}| \leq \lambda \leq 1/K.$$

If no within-cell opposite-outcome pairs exist, $\lambda = 0$ and pooled and between-cell AUROC coincide directly; within AUROC is undefined, not assigned an arbitrary value. The deterministic bound gives an $O(1/K)$ discrepancy along equally sampled panels, not a claim that the discrepancy is always of order exactly $1/K$. It concerns the empirical rates, requires no independence assumption, and does not attribute between-cell discrimination solely to difficulty.

Equal cell size is essential to this particular bound. For example, one cell with K^2 positives and K^2 negatives and $K - 1$ cells with one positive each has $\lambda = K^2/(K^2 + K - 1) \to 1$, not zero. The count of cells alone therefore does not control the within-pair fraction for unbalanced panels.

F.5 TOKEN-PREFIX IDENTITY AND WITHIN-PROBLEM RANKING

This post-hoc diagnostic uses W and the unchanged union-trained OOF scores from Appendix D.12. It asks whether repeated early token prefixes account for the weak measured within-problem ranking; it neither fits a new probe nor measures information absent from these scores. We retokenize the complete public solution strings with the locally cached tokenizer, using the extraction convention (special tokens for the prompt, none for the solution). All 5,632 prompt and stored-window lengths agree with the extraction index. The original extraction did not pin a tokenizer revision, so length agreement is an alignment check, not proof of historical token-ID identity. The supplement records the tokenizer and source hashes.

For each anchor t, we group attempts by problem and exact equality of their first t generated token IDs. Among eligible opposite-outcome pairs within a problem, let q_t be the fraction sharing a prefix. Partitioning the pair average gives

$$A_{\text{within}}(t) = q_t A_{\text{same prefix}}(t) + (1 - q_t) A_{\text{different prefix}}(t),$$

where components are reported only when their pair sets are nonempty. Scores are never refitted, rounded, or forced to tie. In particular, identical reconstructed prefixes need not have bit-identical stored scores: numerical effects can remain in the replay/fitting pipeline. At $t = 4$, the maximum score spread within a same-prefix group is 4.77×10^{-7} and its pair-weighted AUROC is 0.4991.

There are 192,776 within-problem opposite-outcome pairs at $t = 4$, of which 93,507 (48.51%) share their token prefix. Across problems, the 5,632 attempts occupy only 55 problem-specific prefix groups. Different-prefix pairs give AUROC 0.5310, compared with 0.5155 overall. The same-prefix fraction falls to 1.1781% at $t = 64$ and 0.0088% at $t = 128$; no identical-prefix pairs remain at the four later anchors. Table 19 retains all anchors and eligible sample counts. Thus prefix repetition contributes to early ties but does not alone account for weak measured ranking on W, including after textual divergence. This does not establish representational indistinguishability or a null for alternative probes. Removing pairs changes problem weights; these conditional AUROCs are descriptive and have no newly estimated confidence intervals or multiplicity-adjusted tests. Later anchors also change the surviving sample population.

Table 19: Post-hoc token-prefix diagnostic on W using unchanged OOF scores. Same-prefix fractions count within-problem opposite-outcome pairs; AUROCs are descriptive point estimates without new confidence intervals. Later anchors retain only eligible trajectories.

Anchor	Rows	Same-prefix pairs (%)	Within AUROC	Different-prefix AUROC
4	5,632	48.5055	0.5155	0.5310
8	5,632	27.9034	0.4963	0.4949
16	5,632	12.8123	0.5095	0.5111
32	5,632	6.4972	0.5265	0.5283
64	5,632	1.1781	0.4781	0.4778
128	5,632	0.0088	0.5088	0.5088
192	5,631	0.0000	0.4987	0.4987
256	5,630	0.0000	0.5089	0.5089
384	5,472	0.0000	0.5130	0.5130
512	5,322	0.0000	0.4988	0.4988

The numeric supplement supplies problem/attempt/anchor-specific prefix-group IDs and the existing OOF scores, permitting exact pair-count replay without redistributing raw solutions or token sequences. The preparation entry point can reconstruct group IDs from separately supplied public YAML files and a tokenizer. Group-ID replay does not independently verify that upstream tokenization; no new generation, activation extraction, or probe fitting is part of this check.

F.6 SENSITIVITY TO COARSENING THE OBSERVED ANCHOR GRID

We recompute labels from the existing continuation counts and unchanged restart panels. The full-record reconstruction matches all 178 saved regime labels, T_F anchors, and primary T_V anchors/event statuses exactly. All subsequent comparisons retain the recorded four/eight-attempt counts, restart interpolation, and confirmed terminal annotations; only the set of anchor measurements entering the label rule changes. They are post-hoc fixed-record sensitivities, not fresh experiments or simulations of an alternative adaptive sampling path. Unmeasured anchors are not imputed.

Initial grid only. All three cohort manifests specify $64, 128, \ldots, 8{,}192$ as the initial grid. Keeping only these recorded anchors removes 138 measurements from 69 cells, including adaptively added 16/32-token measurements. No T_F event status changes, although 54 T_F anchors and 62 brackets change. All 43 instant labels necessarily disappear because the smallest retained anchor is 64: four instant-prefix-advantaged cells become prefix-limited. All 43 instant cells have only the 1,024-token restart measurement, not a measured 4,096-token rate. For the 39 instant-scratch-solvable cells, the unchanged classifier maps them to no-crossing-at-primary-delta after removing their instant status. That mapping is a consequence of the missing higher-budget measurement and precedence, not evidence of a new budget regime. These are the only regime changes. The prefix-limited count consequently rises from one to five, not because four new advantages were measured, but because early resolution and the instant precedence were removed. The non-instant primary T_V count rises from three to six as membership in that group also changes; it is not a comparison on a fixed non-instant subset.

Available dyadic subset. To separate early-resolution loss from interior refinement, a second sensitivity retains all recorded power-of-two anchors, including 16/32 where available. This is not a uniform prespecified 16–8,192 grid: the early measurements were acquired adaptively. Removing the 44 remaining intermediate measurements from 24 cells leaves all 178 regime labels and all T_F event statuses unchanged. Ten T_F anchors and 16 brackets change. Primary T_V anchors/event statuses change in three cells, including one event disappearance; non-instant event counts at $\delta = 0.25, 0.5, 0.75$ change from $7, 3, 2$ to $6, 2, 1$.

Table 20: Fixed-record grid sensitivities over 178 cells. Changes are relative to all recorded anchors. The dyadic subset retains available 16/32-token measurements; it was not a uniformly sampled initial grid.

Measure	Full	Initial grid only	Dyadic subset
Removed anchor measurements	0	138	44
Cells with removals	0	69	24
Changed regime labels	0	43	0
Changed T_F event status	0	0	0
Changed T_F anchor	0	54	10
Changed T_F bracket	0	62	16
Changed primary T_V anchor/status	0	7	3
Changed primary T_V event status	0	1	1
Non-instant T_V events, $\delta = 0.25$	7	11	6
Non-instant T_V events, $\delta = 0.5$	3	6	2
Non-instant T_V events, $\delta = 0.75$	2	2	1

For the sole full-record prefix-limited Gemma-4 cell, the T_F bracket widens from $(768, 896]$ to $(512, 1{,}024]$, and the primary T_V anchor shifts from 896 to 1,024, retaining the regime. For Ministral-3 on the same problem, the primary T_V event at 768 disappears, while the budget-limited label persists because of taxonomy precedence. Another budget-limited Ministral-3 cell's primary T_V shifts from 320 to 512. The supplement identifies every cell and retains all margins and transitions.

Broad-label stability under interior coarsening must therefore not be read as stability of every advantage or exact boundary. The label hierarchy can hide changes in T_V. Neither comparison quantifies sampling uncertainty, repairs sparse restart interpolation, nor validates bisection on noisy non-monotone rates. These results favor reporting operational brackets and grid sensitivity alongside the frozen labels, rather than interpreting a refined anchor as a statistically identified transition time.