CAN PROBES RELIABLY GUIDE AGENT DECISIONS? EVIDENCE FROM CONTROLLED FILE WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a probe trained on a language model's internal state reliably guide an agent's next action? Not necessarily, even when the task state remains readable. We study a file task: test the current contents, or publish them if a successful test already applies. This setting lets us change the visible history without changing the correct decision. In a confirmation on 32 new file-roster families, testing another file reduces a fixed, feedback-trained Qwen probe's pairwise accuracy from 0.984 to 0.428, while the model's output scores retain 0.988. Both prespecified tests pass Holm correction. An exploratory centroid trained only on initial histories achieves 0.996 at the same layer and token position: the distinction remains linearly accessible there. Ranking is not enough, however: with the added test, the output score classifies only 53.9 percent of inputs correctly at its old threshold. Workflow experiments in Qwen and Gemma show why the reuse question matters: feedback training repairs repeated decisions, but the Qwen confirmation exposes a limit of that repair. Readable task state need not yield a reusable decision rule. The controlled task separates failures of ranking, thresholds, and feedback handling; it does not establish their frequency in real software projects.

1 INTRODUCTION

An agent tests `auth.py`, edits it, then restores the tested version. Does the earlier test still apply? It does after restoration, but not while the untested edit is current. Now suppose the agent tests a different file. That event should not change the answer about `auth.py`. A controller that uses a model's internal state to decide whether to test or publish needs to preserve these distinctions.

A *probe* is a small predictor trained on internal activations. Prior work shows that model states can encode task states and other structured information (Li et al., 2023; Nanda et al., 2023; Marks & Tegmark, 2024). High decoding accuracy makes a probe a plausible monitoring signal. It does not establish that the same predictor will remain useful after the history changes, a distinction central to probing controls (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022).

One question and one central result. *Can an unchanged probe lose a task distinction that the model's output scores still retain?* Yes, in our controlled file task. A Qwen probe trained on histories containing execution feedback repairs repeated decisions. Yet adding a successful TEST of another file leaves target coverage unchanged while reducing that probe's pairwise accuracy from 0.984 to 0.428 on new file rosters; the output score retains 0.988. Both endpoints were fixed before capture. An exploratory Initial-only centroid retains the distinction at the same layer and token position, establishing linear accessibility there rather than a unique cause of the probe's failure. Ranking alone is insufficient: at its old threshold, the output score's classification accuracy falls from 67.9% to 53.9% with the added test. Readable task state need not yield a reusable decision rule.

Why use a small file task? A repository benchmark asks whether an agent can solve a software problem. We ask whether its proposed monitoring signal continues to mean the same thing when irrelevant context changes. Our task makes that comparison exact: file bytes and test records establish the answer independently of the model. We can hold the target state fixed while changing the surrounding events. A deterministic tracker solves the task from visible history, confirming that the necessary evidence is available. This simplicity is the experimental control, not a claim to reproduce software engineering as a whole.

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.

How the evidence fits together. We first present the controlled confirmation (Section 3). We then explain the practical failure that motivated it: a probe can keep requesting TEST after testing succeeds, and feedback training repairs that workflow (Section 4). A small post-hoc live check at the end of that section finds no repeated TEST in the models' own generated follow-up. New histories expose the limits of the training repair (Section 5). Finally, interventions and alternative reading locations test what the probe's direction and site explain (Section 6).

2 A File Task with an Explicit State

Task at a glance
Input: the visible WRITE, TEST, and RESTORE history of four files.
Question: does a prior successful TEST cover the queried file's current bytes?
Choices: TEST the file, or propose PUBLISH; at most three choices per workflow.
Success: PUBLISH is accepted before the budget ends. A validator checks coverage.

A test record belongs to contents, not just to a filename. For file i, let c_i be its current bytes and R_i the hashes recorded by successful isolated tests. Its coverage state is

$$s_i = \mathbf{1}\{\mathrm{SHA256}(c_i) \in R_i\}, \qquad \mathbf{s} \in \{0,1\}^4. \tag{1}$$

"Covered" means a matching successful-test record exists, not that the software is correct in every respect. TEST runs in a separate Python process. The model sees public events; private byte-hash receipts supply labels and validator checks.

The distinction the model must track. An *order* history changes whether TEST comes before or after WRITE. A *restore* history returns to a specified version. The table gives the final state explicitly; r0 and r1 are different contents of the same file.

Visible history	Current	TEST covered	Covered?
`WRITE r0; WRITE r1; TEST`	r1	r1	Yes
`WRITE r0; TEST; WRITE r1`	r1	r0	No
`WRITE r0; TEST; WRITE r1; RESTORE r0`	r0	r0	Yes
`WRITE r1; TEST; WRITE r0; RESTORE r0`	r0	r1	No

Both last rows restore r0, but only one has a test of r0. Thus recognizing the word RESTORE is not enough; the tested contents must match the current contents.

Who makes the decision? A *readout* assigns a coverage score. A fixed external program compares it with a threshold to choose TEST or PUBLISH, then repeats after feedback (Figure 1). In the main readout experiments the model supplies scores, not action commands; Appendix Z separately tests generated actions. Our non-neural reference, a *deterministic state tracker*, reconstructs revisions and tests from the same visible history. The validator independently rejects PUBLISH when the actual receipt does not match.

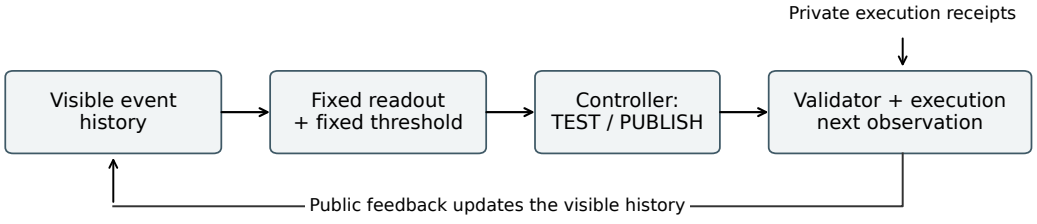


Figure 1: A score guides proposals; a receipt authorizes publication. Public history reaches the readout, while private execution receipts reach only the validator. Training the readout and handling feedback are separate controller choices.

Models, scores, and decision rules. We use native BF16 Qwen3.8-27B (Qwen Team, 2026b) and Gemma-4-31B-IT (Google, 2026), both with hybrid attention architectures. We write Lk for the activation after zero-indexed block k. Primary logistic probes read the final input position at Qwen L47 or Gemma L44, stored in float32. The *model output score* is the final yes-minus-no logit difference. A *centroid* projects activations onto a training class-mean difference (Appendix C). Learned text-only baselines are in Appendices D and T.

Training names used in the tables. *Initial-only* uses the original histories. *+Feedback* adds successful-TEST histories and denied-PUBLISH notices. *+History controls* also adds repeated-TEST histories with both coverage labels. These are training sets for the readout, not backbone fine-tuning. Training, validation, and decision families are separate; Section 4 gives their sizes.

Our *5% FNR rule* chooses the largest validation threshold whose false-negative rate on covered cases is at most 5%. Such an error requests an unnecessary TEST; the rule does not cap mistaken PUBLISH proposals. The *balanced-accuracy rule* maximizes validation balanced accuracy. Neither threshold is selected on test outcomes.

3 An Unrelated Test Separates Probe and Output Scores

Change the context without changing the answer. Take a covered and an uncovered RESTORE history with matching final contents and event counts. Append a successful TEST of another file (the *suffix*) to each. The target's bytes and receipts stay fixed, so each label stays fixed. The question is whether each score still ranks the covered member above the uncovered member.

Pairwise accuracy counts these correct orderings, with half credit for ties, first within each family and then averaged across families. It answers a different question from *classification accuracy at the old threshold*: scores may preserve their relative order while both cross the old threshold. We report both rather than treating ranking as a decision rule.

A confirmation with endpoints fixed before capture. An earlier analysis suggested a probe–output gap (Section 5). For this follow-up, a local plan fixed 32 new file-roster families, 2,048 histories, and two primary tests before any new model forward; this was not public preregistration. Every history passes independent byte-level execution and isolated TEST checks. The Qwen model, feedback-trained probe, training-fitted standardization, threshold, and history-to-text template remain unchanged. A family supplies filenames and event interleaving; the new families use fresh file rosters and bytes, not new language templates.

	Pairwise accuracy		With other-file TEST
Readout	No suffix	Other-file TEST	Old-threshold accuracy (%)
Fixed +Feedback probe (L47)	0.984	0.428	49.8
Model output score	1.000	0.988	53.9
Initial-only centroid (L47)*	1.000	0.996	77.1
Deterministic state tracker	1.000	1.000	100.0

Table 1: Another file's TEST disrupts the feedback probe's ranking, not every readout at its site; retained ranking does not ensure accurate old-threshold decisions. New Qwen families: 512 pairs per suffix, with 512 covered and 512 uncovered inputs. The first two rows enter the prespecified tests. *Exploratory Initial-only centroid at the same layer *and* position. Old thresholds use the 5% FNR rule. The tracker is a score reference, not a new policy run. Full rows: Table 7.

The probe loses a distinction retained by the output scores. Its suffix-induced ranking change is -0.557 [95% CI -0.604, -0.510]. Under the suffix, the output score exceeds it by 0.561 [0.514, 0.605]. Both prespecified family-level sign-flip tests pass Holm correction ($p_{\mathrm{adj}} = 0.00002$, 100,000 draws). Table 1 shows that this is not simply a loss of the distinction at the model's output.

The information is also accessible at the probe's own site. An exploratory CPU analysis applies an already fitted Initial-only centroid to the same L47 final-position activations. It retains 0.996

pairwise accuracy. The +Feedback and +History-controls centroids retain 0.711 and 0.914 (Appendix P). These results establish linear accessibility at this site, not that all projections are robust. The strongest comparison changes both readout form and training data, so it cannot isolate either as the cause. All three centroid results are exploratory, separate from the two confirmed endpoints.

Retaining a distinction still does not make a good action rule. The model output score retains high pairwise accuracy, but at its old threshold its classification accuracy falls from 67.9% to 53.9%. Of 512 uncovered inputs, 472 would trigger mistaken PUBLISH proposals, which the validator would reject. The Initial-only centroid falls from 93.8% to 77.1%. The confirmed ranking gap and these descriptive threshold results support different claims: retaining an ordering does not ensure that the old decision rule remains useful.

4 Feedback Training Repairs a Workflow

Why did we test a feedback-trained probe? It repairs a concrete controller failure. This section explains that benefit before testing its limits.

Testing succeeds, but the controller tests again. In the historical executed Qwen workflow, a probe trained on both order and restore histories chose correctly on the first step in 497/512 cases, yet completed only 365/512. Of the 147 failures, 143 began with a correct choice. After a first successful TEST, 169/267 next choices were unnecessary tests. Figure 2 shows one such case: the state changes correctly, but the probe keeps advising the wrong next action.

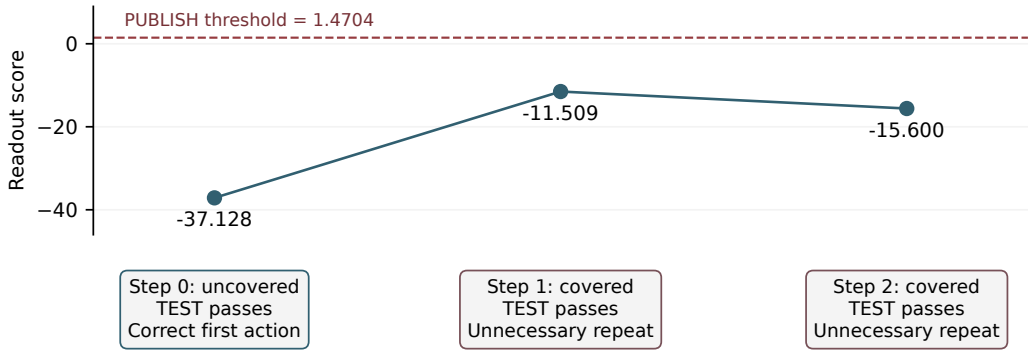


Figure 2: Successful testing need not stop a score-based controller from testing again. In this archived executed trajectory (Qwen case 0000), step 0 correctly tests an uncovered file. Subsequent scores stay below the fixed PUBLISH threshold of 1.4704, exhausting the budget. Gemma shows an analogous cached-replay failure (Appendix G.1), a separate evidence type.

Train on the histories created by feedback. The original queried histories contain one TEST per file. Successful testing and rejected publication produce additional event patterns. We fit new decision probes to three nested training sets:

Training set	Views	What is added?
Initial-only	4,096	Original queried histories.
+Feedback	14,336	Successful-TEST histories and denied-PUBLISH notices.
+History controls	22,528	Multi-TEST histories with both covered and uncovered final states.

All added views come from the 32 training families, not the 16 decision families. The final increment adds 4,096 examples per label, balanced at each added TEST count. Standardization, fitting settings, and 2,048 original validation views are shared (Appendix C). Only the readout is retrained. The feedback set changes several history features together, including both TEST outcomes and denial notices.

Compare controllers within the same new run. For both models, we execution-check the reachable histories and cache their scores. We then replay the controller's state transitions for 512 cases, each with a three-choice budget and no WRITE during the episode. This is cached replay over execution-checked histories, not new live policy execution. Its Initial-only baseline belongs to this run; it is not interchangeable with the historical 365/512 result.

Gemma also repeats successful tests. In this replay, its Initial-only probe gets 500/512 first choices right but completes only 389/512 cases under the 5% FNR rule. Of 123 failures, 115 start correctly. Following a successful step-0 TEST, 137/268 next choices repeat TEST. Appendix G.1 gives a trace and the matched Qwen counts.

Model	Training set	5% FNR rule	Balanced
Qwen	Initial-only	341/512	358/512
Qwen	+Feedback	512/512	512/512
Qwen	+History controls	512/512	512/512
Gemma	Initial-only	389/512	436/512
Gemma	+Feedback	512/512	512/512
Gemma	+History controls	511/512	512/512

Table 2: Completed workflows out of 512. Adding feedback histories repairs every case under both validation rules. Adding history controls retains one Gemma failure under the 5% FNR rule. Paired changes and all controller costs are in Appendix D.

Feedback training is useful here. It repairs the repeated-decision workflow in both models (Table 2). It is not necessary: the visible state tracker completes the task, and a *fixed recovery rule* also completes every case. The rule keeps the readout's first choice, then PUBLISHes after a successful TEST, or TESTs and PUBLISHes after a denial. Each initial false positive (FP) or false negative (FN) adds one action, giving a mean of $1.5+(FP+FN)/512$, versus 1.500 for the tracker. Training and controller design are two distinct ways to avoid repeated mistakes.

Does the model itself repeat the test? Not in a small exploratory live check. Each model generates TEST/PUBLISH actions on 64 reused cases (16 families), with task rules in its prompt and greedy decoding. The interface accepts one valid JSON action, bare or in one JSON-tagged or unlabelled Markdown fence. Earlier strict bare-JSON runs left Gemma at 41/64 completions (7/64 with a workflow instruction); after this post-hoc amendment, all arms of both models were rerun. Early and rerun results are not pooled (Appendix Z). Each model then completes 64/64 cases and never repeats TEST after success. The Initial-only probe's first choices are more accurate (Qwen 62 vs. 50 generated; Gemma 61 vs. 49), yet re-queried in cached replay it completes only 44/64 and 49/64. The fixed recovery rule completes every case; for generated actions it rescues none but issues the same follow-ups without inference, cutting calls from 110 to 64 (Qwen) and 111 to 64 (Gemma). Prompts and evidence types differ, so these comparisons are descriptive, not controlled.

5 What the Repair Does and Does Not Transfer

Does the repaired readout still recognize which contents were tested? We hold each head and threshold fixed but change the histories and which contents a receipt covers. These static two-model diagnostics motivated the later Qwen confirmation.

Matched histories expose two different failures. We construct 512 write-order pairs and 256 restore-receipt pairs, or 1,536 inputs in 16 decision families. Let r_{cur} be the target's current, uncovered revision before a restore pair. The pair appends:

Pair	Matching receipt	Nonmatching receipt
Write order	`WRITE r2; TEST`	`TEST; WRITE r2`
Restore receipt	`TEST; WRITE r2; RESTORE` r_{cur}	`WRITE r2; TEST; RESTORE` r_{cur}

Each pair has the same final contents, words, events, and TEST count, but only one member tests the final contents. Restore pairs return to r1 in 128 order-origin cases and r0 in 128 restore-origin cases. Appendix C gives a complete worked example.

Alongside pairwise accuracy and old-threshold accuracy, we measure *pooled AUC*: how well the score orders all covered versus uncovered histories, including across pairs. These are three different questions (Figure 3). For example, positive/negative score pairs $(11, 10)$ and $(-10, -11)$ are both ordered correctly, yet no single threshold classifies all four inputs correctly.

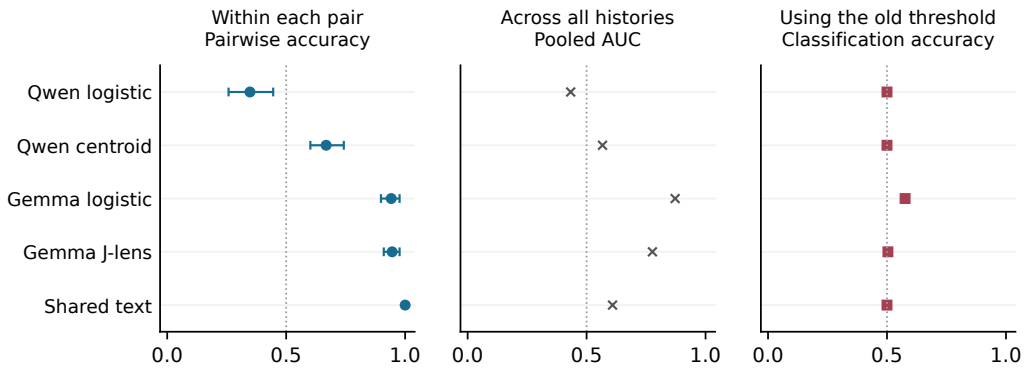


Figure 3: Restore histories: correct ordering within each pair, across all histories, and classification with the old threshold are different achievements. Only pairwise accuracy has family-bootstrap intervals here; pooled AUC and old-threshold accuracy are point estimates. Learned readouts use +History controls.

Loss of ordering is different from a bad threshold. Qwen's logistic probe reverses restore-pair order (0.348); its centroid ranks above chance (0.668). Both classify only half the inputs correctly at their old thresholds. Gemma's logistic probe instead retains strong pairwise ordering (0.941), but its threshold gives 57.62% accuracy, misclassifying 217 uncovered inputs as covered with no opposite errors. The same head reaches 96.19% on write-order pairs. These are readout- and history-specific failures, not evidence that either model lacks the state altogether.

Compare against the model output, not just chance. On write-order histories, Qwen's output score gives more accurate classifications than any of its logistic probes; Gemma shows the opposite pattern (Table 3). The model output is therefore an essential comparator, although its own old threshold fails on restore histories (50.0% Qwen and 50.4% Gemma accuracy; Table 8).

Readout	Qwen accuracy (%)	Gemma accuracy (%)
Model output score	94.24	57.62
Initial-only probe	81.05	96.68
+Feedback probe	53.52	81.74
+History-controls probe	60.45	96.19

Table 3: Model output beats the probes in Qwen; the reverse holds in Gemma. Write-order classification on 1,024 histories at old 5% FNR-rule thresholds, using the same validation inputs for all readouts. These are score-based decisions, not generated-answer accuracy.

Well-ordered scores can still sit on the wrong side of the old threshold. Qwen's +Feedback head has write-order pooled AUC 0.964 but falsely calls 476/512 uncovered histories covered. On restore histories, feedback-trained heads also lose ordering. The earlier cache cannot isolate why because revision names, length, and event roles change together (Appendix L).

Remove a specific shortcut, then add irrelevant activity. In 3,072 new histories per model, a base history ending in WRITE of identical bytes or RESTORE can have either coverage label. We

call that base history the *core*. We vary revision names and append either another file's TEST or eight events on that file, leaving target contents and receipts unchanged. A rule that merely treats the final WRITE as uncovered cannot explain the perfect pairwise ordering on the same-bytes-WRITE cores.

In the earlier discovery set, the other-file TEST reduces Qwen +Feedback restore pairwise accuracy from 0.984 to 0.531, while its output score retains 0.990. Gemma's feedback heads retain perfect pairwise ordering, although threshold errors change. Their changes are not a single uniform score shift. Qwen also makes 39 unnecessary-TEST errors on same-bytes-WRITE without a suffix and 83 after the other-file TEST. All eight suffix accuracy contrasts pass Holm correction across the 12 original tests (Appendix K). The ranking decomposition was post-hoc, prompting the independently frozen confirmation in Section 3. Full counts and intervals remain in Table 6 and Appendix M.

An earlier transfer study likewise found that a head trained only on order or restore histories need not transfer to the other mechanism. Alias balancing removes one revision-name cue without fully repairing transfer; joint-mechanism training helps. These supporting controls are kept in Appendix A, rather than treated as additional explanations of the confirmed suffix effect.

6 Reading Location and Output Sensitivity

Sections 3–5 ask what an external program can read; an intervention asks a separate question: does changing the component along that readout also change the model's own output? We keep both because predictive utility and output sensitivity are different forms of evidence.

Change one component and compare matched controls. For each model, we select 64 pairs in 16 families by salted hash rather than scores. At the final position, we replace the component along a fixed +History-controls logistic or centroid direction by its value from an opposite-label donor history. Three random shams match each perturbation within 1% in realized norm *after BF16 rounding*, without using output scores. We measure how far the output score moves toward the donor's label and subtract the sham effect, Holm-adjusting eight tests. Full activation replacement supplies a reference scale. The procedure was frozen before capture (Appendix H).

Model	Readout	Histories	Effect (logits) [95% CI]	Holm p	Flips
Qwen	Logistic	Write order	-0.004 [-0.021, 0.016]	1.000	0/64
Qwen	Logistic	Restore receipt	0.003 [-0.008, 0.014]	1.000	0/64
Qwen	Centroid	Write order	0.140 [0.116, 0.161]	$<$0.001	1/64
Qwen	Centroid	Restore receipt	0.013 [-0.005, 0.031]	0.666	0/64
Gemma	Logistic	Write order	0.336 [0.272, 0.405]	$<$0.001	0/64
Gemma	Logistic	Restore receipt	0.212 [0.165, 0.260]	$<$0.001	0/64
Gemma	Centroid	Write order	1.634 [1.192, 2.149]	$<$0.001	3/64
Gemma	Centroid	Restore receipt	0.785 [0.632, 0.951]	$<$0.001	0/64

Table 4: Direction edits can move output scores without changing yes/no choices. Each cell has 64 directed interventions from 32 pairs in 16 families. Effect is movement toward the donor's label minus three norm-matched shams. Flips count sign changes of the yes-minus-no score, not crossings of the controller's fixed threshold.

Edits move scores but rarely flip choices. Five of eight effects are positive after Holm adjustment: all four Gemma cells and Qwen centroid/write-order. All 1,536 shams meet tolerance. Logistic edits flip no choices; centroid edits flip 1/64 Qwen and 3/64 Gemma write-order choices. Each model gives the same yes/no choice across its restore inputs. Full donor replacement flips neither model's choice, yet shifts the margin by 0.223 logits in Qwen and 2.223 in Gemma. Zero flips therefore do not mean an insensitive site. Qwen logistic's null average contrast in *both* history types also prevents attributing it specifically to poor restore ranking.

A different reading location can help without fixing the threshold. The Qwen confirmation preselected six sites: blocks 31, 47, and 63, at the final token and at the target's last event. New logistic heads are fitted only on original training families. The best prespecified final-position refit is L63, the last of 64 blocks, with suffix pairwise accuracy 0.912 but old-threshold accuracy 63.9%.

All six results, including failures, are in Appendix O. Earlier event positions read a prefix before the suffix and query, not the full input. The exploratory L47 centroid in Section 3 indicates that the distinction is accessible before the final block. We have not tested combining sites into one predictor.

Lenses provide another same-site comparison. A direct logit lens applies final normalization and yes/no output weights at L47. An empirical Jacobian lens (J-lens; Gurnee et al., 2026) first applies a fixed matrix fitted without task labels. Under the suffix their pairwise accuracies are 0.537 and 0.818 (Table 7). Its BF16 matrix failed our numerical check, so the J-lens is only an empirical control (Appendix F).

7 Related Work

Decoding state versus reusing a predictor. Probe-control work separates label recovery from task relevance (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022); Sahoo et al. (2026) examine format confounds in reasoning probes. Othello and truth-representation studies add transfer or intervention evidence (Li et al., 2023; Nanda et al., 2023; Marks & Tegmark, 2024). Work on reportable concepts, spatial representations, and latent computation motivates richer internal monitoring (Gurnee et al., 2026; MirroS Team, 2026; Lai et al., 2026). Our contribution is an execution-labeled perturbation that holds target state fixed while testing reuse of an unchanged readout.

Feedback, thresholds, and readout methods. DAgger addresses inputs changed by a learner's actions (Ross et al., 2011); we enumerate feedback histories offline and refit only a predictor. Reflexion uses linguistic feedback and memory to improve later attempts (Shinn et al., 2023); our backbone and controller remain fixed. Calibration and dataset-shift work motivates separating ranking from classification at an old threshold (Guo et al., 2017; Ovadia et al., 2019). Direct and tuned lenses map activations to vocabulary scores (nostalgebraist, 2020; Belrose et al., 2023), while representation engineering and causal abstraction motivate direction interventions (Zou et al., 2023; Geiger et al., 2021).

When probe signals affect behavior. Cox et al. (2026) test pre-answer probe directions with norm-matched steering controls. Collu et al. (2026) use refusal probes to guide a prompt-search loop. Such results motivate testing usefulness directly; they do not imply that a probe will transfer across histories. Here we distinguish output sensitivity under activation edits from an external controller's use of a score.

Controlled diagnostics complement agent benchmarks. SWE-bench and tool-use evaluations test broader task competence (Jimenez et al., 2024; Yao et al., 2024). Here byte-level state supplies an unambiguous target, and irrelevant file events supply a controlled perturbation. The task isolates a monitoring failure that aggregate completion alone would not identify.

External support can help without validating a probe. Reasoning–acting loops, workflow memory, feasibility-grounded skill selection, and generated control programs improve agents in their evaluated settings (Yao et al., 2023; Wang et al., 2025; Ichter et al., 2023; Liang et al., 2023), but these gains are not a safety guarantee. Safety shielding instead constrains actions against a specification (Alshiekh et al., 2018). Our validator provides enforcement; the fixed recovery rule reuses a still-valid test receipt. Neither establishes one agent framework's superiority.

8 Test Reuse, Not Just Decoding

Readable task state need not yield a reusable decision rule. The same fixed probe can lose its correct coverage ordering under irrelevant context while other readouts retain it. Feedback training can nevertheless make that probe useful in a particular workflow. These facts coexist: fitting the states reached during execution solves one problem, but does not guarantee a portable ranking or decision threshold.

For controller design, compare with model output and a visible-state baseline, evaluate old thresholds, and inspect decisions after feedback. A successful TEST should constrain later choices *while*

its receipt still matches current contents and no state change invalidates it. The fixed recovery rule can supply that follow-up after either a probe's or a model's generated first choice (Appendix Z). The validator enforces publication safety; readouts and control logic determine unnecessary work.

What recalibration could and could not repair. A new threshold could help when ranking survives, but moving a threshold cannot repair incorrect pair ordering. A common positive temperature scaling also preserves ranking. Testing online recalibration would require a separate protocol for obtaining feedback labels and updating thresholds; it is not an evaluated remedy here. The validator, not score calibration, prevents invalid publication in our task.

What this controlled task establishes. It supplies a counterexample to treating high decoding accuracy or repaired completion as sufficient evidence for reusing a readout. It does not estimate how often this occurs in real repositories. The tracker makes internal activations unnecessary in this task; its role is to establish a solvable reference, not to show that learned monitoring is useless where exact tracking is unavailable. We identify a failure of a projection and its old threshold, not a unique neural shortcut or absence of world state.

Scope. The transfer, feedback-training, receipt, intervention, and live-check comparisons cover Qwen and Gemma. Gemma's feedback probes retain ranking in the two-model suffix discovery; the confirmed ranking failure concerns Qwen. The original executed workflow, alias and joint-training/MLP follow-ups, and the new-family confirmation and six-site analysis use Qwen only.

We use one history template and recoverable three-action workflows. New readout policies use cached replay; the exploratory live check reuses 64 cases and does not test the other-file suffix. New rosters reuse templates, and same-site centroid scoring is exploratory. Fixed probe settings establish a failure case, not robustness across all regularization choices or probe classes. Single-site interventions do not test general controllability. Earlier captures have weaker identity evidence; native BF16 Jacobians remain numerically unqualified, and CPU lens arithmetic is not exact CUDA replay.

Reproducibility Statement

The anonymous supplement contains frozen scores, traces, histories, numeric heads, analysis code, and file hashes. Its CPU entry point recomputes workflow, receipt, intervention, crossed-history, and confirmation statistics. It includes 3,072 lossless discovery residuals and exact output rows for direct-lens recomputation, all confirmation texts and token records, six fitted heads, and the frozen protocol. The exploratory extension prints all 84 crossed-history cells and checks all new centroid statistics, with projection replay on eight preselected residual witnesses per head. Local checks projected all 2,048 new residuals. The full new capture archive and dense J matrix are hash-bound but outside this compact package; neither backbone capture nor six-site refitting is repeated. A separate CPU entry point verifies the live check's action transcripts, receipts, early and final interface records, and case-matched replay comparison. The capsule retains archived table names; the supporting guide (p. 13) and Appendices N–P explain the earlier evidence scopes.

Ethics Statement

Experiments use synthetic file histories in owned temporary workspaces, without human-subject data or private deployment traffic. An external validator blocks invalid proposals. General deployment safety requires controls appropriate to the actual task and environment.

AI Use Statement

OpenAI Codex, Anthropic Claude, Kimi K3, and xAI Grok assisted with research discussion, software implementation and review, interpretation checks, and manuscript drafting and editing. AI-generated interpretations are not experimental evidence. Executable checks, artifact hashes, statistical recomputation, and visual inspection were used to validate the work. Responsibility for the final claims and artifacts is not delegated to these tools.

REFERENCES

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11797. URL https://ojs.aaai.org/index.php/AAAI/article/view/11797.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112.

Matteo Gioele Collu, Riccardo Conte, Alberto Giaretta, Denis Kleyko, Mauro Conti, Matteo Zavatteri, and Roberto Confalonieri. Refusal before decoding: Detecting and exploiting refusal signals in intermediate LLM activations, 2026. URL https://arxiv.org/abs/2605.28553v2. Version 2.

Kyle Cox, Darius Kianersi, and Adrià Garriga-Alonso. Post-hoc reasoning in chain of thought: Decoding and steering pre-committed answers, 2026. URL https://arxiv.org/abs/2603.01437v2. Version 2.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Google. Gemma-4-31B-IT: Frozen model configuration, 2026. URL https://huggingface.co/google/gemma-4-31B-it/blob/842da3794eaa0b77d5f08bae87a17459d91ff475/config.json. Revision 842da3794eaa0b77d5f08bae87a17459d91ff475.

Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed September 25, 2026. Instruction-tuned 31B benchmark results.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, July 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Matteo He, William F. Shen, Xinchi Qiu, and Nicholas D. Lane. Sparse readout prism: Explaining logit-lens scores in features instead of tokens, 2026. URL https://arxiv.org/abs/2609.01936.

Lovis Heindrich, Philip Torr, Fazl Barez, and Veronika Thost. Do sparse autoencoders generalize? a case study of answerability, 2025. URL https://arxiv.org/abs/2502.19964.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Jeffrey Lai, Anthony Bao, John Quinn, and William Gilpin. Fractal basins trap latent reasoning, 2026. URL https://arxiv.org/abs/2609.04963.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.13382.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023. URL https://arxiv.org/abs/2209.07753v4. Version 4.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

MirroS Team. S-Space: Exploring spatial workspace in multimodal models, September 2026. URL https://mirros.ai/blog/s-space. Research blog post.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models, 2023. URL https://arxiv.org/abs/2309.00941.

nostalgebraist. Interpreting GPT: The logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Research blog post.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Qwen Team. Qwen3.8-27B: Official model card and benchmark results, 2026a. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed September 25, 2026. Vendor-reported SWE-bench Pro result.

Qwen Team. Qwen3.8-27B: Frozen model configuration, 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B/blob/1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0/config.json. Revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.05806.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of EACL*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states, 2026. URL https://arxiv.org/abs/2606.02907v2. Version 2.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25bx.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Mario Zechner and contributors. pi-agent-core, 2026. URL https://github.com/earendil-works/pi/tree/main/packages/agent. Software, version 0.87.1, MIT license.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

GUIDE TO THE SUPPORTING EVIDENCE

The main paper follows one question rather than the chronological order of experiments. The appendices retain the complete measurements. For the central confirmation, start with Appendix O; for the exploratory Initial-only centroid, use Appendix P. For the training repair and receipt tests, use Appendices C and D. For interventions, see Appendix H; for the live check, Appendix Z. Older transfer, lens, and executed-workflow records are supporting evidence, not additional confirmatory samples.

Evidence	What it directly tests	What not to infer
New-family confirmation	Two ranking endpoints fixed before capture	New templates or production failure frequency
Initial-only centroid	Exploratory fixed projection of the same activations	A third prespecified primary test
Workflow replay	Controller transitions over execution-checked histories	Thousands of newly executed policies
Static receipt controls	Scores and old thresholds on changed histories	Successful open-ended software engineering
Direction interventions	Output-margin sensitivity to single-site edits	General controllability or deployment safety
Live check	Generated first choices and follow-ups with the amended parser	Controlled probe-versus-generation comparison or suffix robustness

Statistics and evidence levels. The original transfer corpus has 64 families split into 32 training, 16 validation, and 16 test families, each containing all 16 coverage states. Workflow, receipt, and intervention comparisons use 16 separate decision families, with 2,000 paired family-bootstrap draws for their pointwise 95% intervals. The new confirmation uses 32 new file rosters and its own frozen two-endpoint test and 5,000-draw bootstrap. Complete families, not repeated histories, are the resampling units. Local frozen plans are not public preregistration; post-hoc analyses retain their exploratory status. The individual appendices state the multiplicity correction for each frozen test family.

Names in the archived tables. "Native score" denotes the model output score; "pair ranking" is pairwise accuracy; "frozen accuracy" uses an unchanged validation threshold. "Low false-denial" denotes the 5% FNR rule on covered cases, while "selector" denotes the score-threshold controller. "Margin" denotes the yes-minus-no output score, not the controller threshold. The capsule preserves these archived names and table identifiers; the paper changes exposition, not their definitions.

Official model benchmarks are context, not our experiments. The Qwen3.8-27B model card reports SWE-bench Pro 61.7 using Claude Code, temperature 1.0, top-p 0.95, and a 256K context window; it also notes corrections to problematic benchmark tasks (Qwen Team, 2026a). The Gemma model card reports LiveCodeBench v6 80.0% for instruction-tuned 31B (Google DeepMind, 2026). These are vendor reports on different benchmarks, not scores reproduced here or bound by us to the exact snapshots used in our captures. They motivate studying models with reported coding ability; they do not validate our readout tests or live check. We do not substitute a third-party SWE-bench score for Gemma's documented coding result.

A SUPPORTING STUDY: TRANSFER ACROSS HISTORY MECHANISMS

This is the historical transfer study, called Study 1 in the archived reports. It is distinct from the mixed-mechanism decision heads in the main paper. For each model and source mechanism, we fit one head per file slot and score both test mechanisms. The transfer metric measures ranking within held-out family f and slot i:

$$A_{m\to n} = \frac{1}{4|\mathcal{F}|}\sum_{f\in\mathcal{F}}\sum_{i=0}^{3}\mathrm{AUC}(\mathbf{s}_{f,n,i}, \widehat{\mathbf{p}}^{(m)}_{f,n,i}). \tag{2}$$

The arguments are 16-element vectors: labels for eight covered and eight uncovered worlds, and scores from the head trained on mechanism m. Mechanism n supplies the test histories.

Train on → test on	Qwen AUC [95% CI]	Gemma AUC [95% CI]
Order to order	0.992 [0.987, 0.997]	0.883 [0.861, 0.905]
Order to restore	0.338 [0.307, 0.372]	0.398 [0.368, 0.428]
Restore to order	0.337 [0.312, 0.365]	0.448 [0.411, 0.484]
Restore to restore	0.981 [0.972, 0.989]	0.892 [0.867, 0.916]

Table 5: Historical mechanism transfer. Each cell uses 256 histories without a target-file question, from 16 test families at the primary layer. All original gate decisions and secondary endpoints remain in the historical appendices.

An alias control removes one candidate cue. Let z_i indicate that TEST used r1. The original construction has $s_i = z_i$ in order histories and $s_i = 1 - z_i$ in restore histories. In a Qwen follow-up, revision aliases are balanced across labels, the heads are refitted, and the transfer test is repeated. Cross-mechanism AUC improves to 0.440 and 0.457 but remains below chance; within-mechanism AUC remains 0.991 and 0.970. Both alias views enter each family-slot AUC together, rather than averaging separately scored views. This removes the specific revision-name association while leaving event-role explanations open (Appendix X).

Joint training can represent both mechanisms. Joint-mechanism training reaches AUCs of 0.978 and 0.967 on new Qwen families. A small MLP or difference-in-means readout trained on only one mechanism does not repair transfer. These results establish neither a unique shortcut nor a mechanism for the later suffix effect; they motivate separating learnability from reuse of a fixed head.

B Discovery and Confirmation: Full Summary Tables

The main paper's first table focuses on the confirmed comparison and its most informative exploratory control. The tables below retain every row of the previous main summaries, including unsuccessful readouts, lenses without calibrated thresholds, and the initial two-model discovery.

Readout	Suffix	Pairwise acc. [95% CI]	Pooled AUC	FP	FN
Qwen +Feedback	None	0.984 [0.965, 1.000]	0.949	3	58
	Other-file TEST	0.531 [0.449, 0.625]	0.539	81	162
	Eight events	0.766 [0.684, 0.836]	0.713	139	51
Qwen model output	None	0.996 [0.988, 1.000]	0.998	168	0
	Other-file TEST	0.990 [0.977, 1.000]	0.922	249	0
	Eight events	0.980 [0.965, 0.992]	0.963	188	0
Gemma +Feedback	None	1.000 [1.000, 1.000]	1.000	21	0
	Other-file TEST	1.000 [1.000, 1.000]	0.999	0	12
	Eight events	1.000 [1.000, 1.000]	1.000	3	4

Table 6: Discovery RESTORE controls: each row pools four revision assignments, giving 256 pairs in 16 families. FP/FN use the old 5% FNR-rule threshold, with 256 negatives/positives. Intervals are post-hoc family-bootstrap diagnostics; the original Holm-adjusted accuracy tests are unchanged. Appendix M gives the complete decomposition and WRITE results.

The probe loses ordering that output scores retain. The Qwen +Feedback probe loses both matched and pooled ordering on RESTORE cores (Table 6). Mean scores rise by 9.77 on uncovered inputs and fall by 7.44 on covered inputs: this is not a uniform offset. +History-controls pairwise accuracy also drops (0.973 to 0.551). Output-score pairwise accuracy remains 0.990, but pooled AUC and classification accuracy decrease. Strong pairwise ordering does not mean correct answers at the old threshold.

Preserved ordering can still leave threshold errors. Gemma's feedback heads retain perfect pairwise accuracy under all three suffixes, while classification errors change. Their score changes are not a single additive shift. Qwen also ranks every same-bytes-WRITE pair correctly, yet +Feedback

misses 39 covered cases without a suffix and 83 after the other-file TEST. The final-operation-only rule is excluded, not every shortcut. All eight suffix accuracy effects pass the original Holm correction; this ranking decomposition is post-hoc.

	Pairwise accuracy		Other-file TEST		
Readout	No suffix	Other TEST	Old-thr. acc. (%)	FP	FN
Fixed +Feedback probe (L47)	0.984	0.428	49.8	197	317
Direct lens (L47)	0.842	0.537	–	–	–
Empirical J-lens (L47)	0.969	0.818	–	–	–
Model output score	1.000	0.988	53.9	472	0
Refitted feedback probe (L31)	0.129	0.242	47.0	499	44
Refitted feedback probe (L47)	0.988	0.469	47.6	317	220
Refitted feedback probe (L63)	1.000	0.912	63.9	312	58
Initial-only centroid (L47)*	1.000	0.996	77.1	217	17
Deterministic state tracker	1.000	1.000	100.0	0	0

Table 7: Retained ranking need not preserve accurate decisions at the old threshold. Confirmation summary: 512 pairs per suffix, with 512 positives and 512 negatives. Lenses have no calibrated threshold. Refits are training-only secondary controls; L63 is the final block. *Exploratory centroid scoring. The tracker follows visible history and fixed revision bytes, not a newly executed policy. All six sites and all centroid training sets appear in Appendices O and P.

C New Histories, Matching, and Frozen Evaluation

The successor diagnostic starts from the 512 decision cases, not the original transfer test worlds. Every case contributes one write-order pair. The 256 initially uncovered cases also contribute one restore-receipt pair. The resulting 1,536 rendered inputs are all new to the preceding capture cache. They yield 768 pairs but only 16 family clusters, not 1,536 independent tasks.

One actual target, two suffixes. For `lens-use-000-order-00`, the target is `family000_slot0_qstciakh.py`. Its initial relevant history is `WRITE r0; TEST; WRITE r1`: it is at r1 without a receipt covering r1. The full input includes the other three files and the same interleaving. The following suffixes are drawn from the frozen case file; `f` abbreviates that exact target filename only for typography.

Pair	Covered suffix	Uncovered suffix
Write order	`WRITE f r2; TEST f`	`TEST f; WRITE f r2`
Restore receipt	`TEST f; WRITE f r2; RESTORE f r1`	`WRITE f r2; TEST f; RESTORE f r1`

Both members have identical operation multisets and TEST counts. Their final contents match: r2 in write order, r1 in restore receipt. In the second pair the final RESTORE position is also identical; the preceding order determines whether TEST covered r1 or r2. The renderer adds: "Query for slot 0: does a prior successful TEST cover the current contents of family000_slot0_qstciakh.py? Answer yes or no." Full positive and negative strings, labels, observation hashes, and all other families are in `data/receipt/cases.jsonl`.

The design file's `execution_verified=false` is a pre-execution design field, not a later failure status. The separate execution report checks all 1,536 histories with real byte writes and isolated tests. The analysis independently reconstructs their labels and checks matching. These checks have different roles and are retained separately.

Feedback-trained logistic heads. All three groups use the StandardScaler fitted only on the 4,096 initial training views and logistic regularization $C = 1$. Sample weights sum to 2,048 for each label, keeping total training weight at 4,096 as histories are added. The same 2,048 initial validation views determine each head's low-false-denial and balanced thresholds. Neither the language-model weights nor the decision-family inputs enter fitting or calibration.

Feedback-trained centroid readouts. For each model, we reuse the feature-wise StandardScaler fitted only on the 4,096 initial training views. Let $z(h)$ be its transform of residual h. For each training group $g \in \{I, F, FH\}$, we compute arithmetic class means $m_{g,1}$ and $m_{g,0}$ of $z(h)$ over that group's covered and uncovered training views. The groups pool both mechanisms and all queried slots from the same 32 training families, with 4,096, 14,336, and 22,528 views, respectively. The score is

$$f_g(h) = \left[z(h) - \frac{m_{g,1} + m_{g,0}}{2} \right]^\top (m_{g,1} - m_{g,0}).$$

The implementation retains the scaler's float32 transform and accumulates the means in float64. It applies no unit-length normalization, sigmoid, or separately fitted intercept. For each model and group, the two thresholds are selected separately on the same 2,048 initial validation views from 16 validation families, using the rules in Section 2. The scaler, class means, and thresholds then remain fixed for workflow and receipt evaluation. This is the implementation in `score_residuals` and `calibrate` in the archived `reanalyze.py`, not the historical transfer study's single-mechanism difference-in-means control (Appendix A).

Metrics. For pair j, with covered and uncovered scores a_j, b_j, its contribution is $\mathbf{1}[a_j > b_j] + \frac{1}{2}\mathbf{1}[a_j = b_j]$. Pair ranking averages contributions within each family and then across families. Pooled AUC instead compares every positive with every negative across families. Accuracy compares each score to the original validation threshold. For example, scores $(11, 10)$ in one positive/negative pair and $(-10, -11)$ in another give perfect pair ranking but admit no perfectly separating global threshold. No threshold is fitted on the receipt set.

Bootstrap replicates draw 16 families with replacement, keeping all their rows and pairs. Differences use the same draw on both sides. The new audit uses 2,000 draws with fixed integer seed 2026091902. It reports pointwise 95% percentile intervals; the exploratory collection is not a familywise hypothesis test. Tests of the historical four-cell transfer retain their original nested-null and multiplicity protocol.

D Full Successor Results

Readout	Pairwise acc. [95% CI]	Pooled AUC	Accuracy (%)
Qwen model output	0.564 [0.461, 0.682]	0.561	50.0
Qwen logistic probe	0.348 [0.258, 0.445]	0.433	50.0
Qwen centroid probe	0.668 [0.602, 0.742]	0.567	50.0
Gemma model output	0.928 [0.889, 0.965]	0.739	50.4
Gemma logistic probe	0.941 [0.898, 0.977]	0.872	57.6
Gemma centroid probe	0.957 [0.926, 0.984]	0.802	50.4
Gemma J-lens	0.945 [0.910, 0.977]	0.777	50.4
Shared text probe	1.000 [1.000, 1.000]	0.609	50.0

Table 8: Strong pair ordering can coexist with poor classification at the old threshold. Restore-receipt results on 256 pairs; learned heads use +History controls, and thresholds use the 5% FNR rule. Figure 3 plots a subset of these rows, including the empirical Gemma J-lens.

How the main-text names map to the archive. The appendix tables and supplementary data retain their original labels. *Model output score* is `native`, the final yes-minus-no logit difference; *deterministic state tracker* is `ledger`. *Pairwise accuracy* is `pair_ranking`, with half credit for ties and family-macro averaging. *Accuracy at the original threshold* is the earlier "frozen accuracy." The *5% FNR rule* is "low false-denial," FD, or `false_denial`; its cap applies to validation false negatives, not test-time errors. Balanced accuracy is BA. Initial-only is I, +Feedback is F, and +History controls is FH. A residual head is the logistic probe. These are presentation changes only: scores, metric definitions, thresholds, and identifiers are unchanged. "Native BF16" still denotes arithmetic, not an output-score method. The construction checks 22,448 unique execution histories, not 22,448 new policy episodes.

Model	Training comparison	Rule	Completion difference (pp) [95% CI]
Qwen	F minus I	FD	33.40 [28.71, 37.90]
Qwen	F minus I	BA	30.08 [24.61, 35.55]
Qwen	FH minus I	FD	33.40 [28.71, 37.90]
Qwen	FH minus I	BA	30.08 [24.61, 35.55]
Gemma	F minus I	FD	24.02 [21.48, 26.37]
Gemma	F minus I	BA	14.84 [10.94, 18.36]
Gemma	FH minus I	FD	23.83 [21.29, 26.17]
Gemma	FH minus I	BA	14.84 [10.94, 18.36]

Table 9: Paired-family changes over the new I baseline. These are cache replay outcomes, not the historical executed 365/512 Qwen run.

The following four tables distinguish pair ranking, pooled AUC, and the two frozen thresholds. Confusion counts are TP/TN/FP/FN, with "covered" positive. Write order has 512 examples per class, restore receipt 256. The exact ledger is shared. Text controls are shared fitted models; identical entries under both neural-model reports do not represent independent replications. The complete machine-readable output additionally includes all centroid variants and the TEST-count control.

Readout	Pair ranking [95% CI]	Pooled AUC	FD acc. (%)	TP/TN/FP/FN
Qwen native	1.000 [1.000, 1.000]	0.993	94.24	497/468/44/15
Qwen residual I	0.922 [0.863, 0.971]	0.878	81.05	331/499/13/181
Qwen residual F	0.994 [0.984, 1.000]	0.964	53.52	512/36/476/0
Qwen residual FH	0.988 [0.973, 1.000]	0.947	60.45	512/107/405/0
Qwen centroid FH	0.980 [0.953, 0.998]	0.928	74.51	488/275/237/24
Qwen J-lens	0.990 [0.980, 0.998]	0.883	59.96	510/104/408/2
Qwen direct	0.939 [0.879, 0.984]	0.785	61.91	459/175/337/53
Gemma native	0.992 [0.986, 0.998]	0.984	57.62	78/512/0/434
Gemma residual I	0.998 [0.994, 1.000]	0.993	96.68	501/489/23/11
Gemma residual F	0.998 [0.994, 1.000]	0.998	81.74	512/325/187/0
Gemma residual FH	0.998 [0.994, 1.000]	0.998	96.19	511/474/38/1
Gemma centroid FH	0.998 [0.994, 1.000]	0.992	60.84	111/512/0/401
Gemma J-lens	0.998 [0.994, 1.000]	0.993	53.52	36/512/0/476
Gemma direct	0.998 [0.994, 1.000]	0.996	62.79	131/512/0/381
Shared text I	1.000 [1.000, 1.000]	0.812	50.00	512/0/512/0
Shared text F	1.000 [1.000, 1.000]	1.000	50.00	512/0/512/0
Shared text FH	1.000 [1.000, 1.000]	1.000	50.00	512/0/512/0
Shared ledger	1.000 [1.000, 1.000]	1.000	100.00	512/512/0/0

Table 10: Write-order pairs, original false-denial thresholds.

Readout	Pair ranking [95% CI]	Pooled AUC	BA acc. (%)	TP/TN/FP/FN
Qwen native	1.000 [1.000, 1.000]	0.993	91.89	429/512/0/83
Qwen residual I	0.922 [0.863, 0.971]	0.878	81.64	348/488/24/164
Qwen residual F	0.994 [0.984, 1.000]	0.964	52.44	512/25/487/0
Qwen residual FH	0.988 [0.973, 1.000]	0.947	59.38	512/96/416/0
Qwen centroid FH	0.980 [0.953, 0.998]	0.928	85.06	433/438/74/79
Qwen J-lens	0.990 [0.980, 0.998]	0.883	79.00	369/440/72/143
Qwen direct	0.939 [0.879, 0.984]	0.785	70.51	240/482/30/272
Gemma native	0.992 [0.986, 0.998]	0.984	50.98	10/512/0/502
Gemma residual I	0.998 [0.994, 1.000]	0.993	89.06	509/403/109/3
Gemma residual F	0.998 [0.994, 1.000]	0.998	67.29	512/177/335/0
Gemma residual FH	0.998 [0.994, 1.000]	0.998	90.04	512/410/102/0
Gemma centroid FH	0.998 [0.994, 1.000]	0.992	50.68	7/512/0/505
Gemma J-lens	0.998 [0.994, 1.000]	0.993	75.59	262/512/0/250
Gemma direct	0.998 [0.994, 1.000]	0.996	52.54	26/512/0/486
Shared text I	1.000 [1.000, 1.000]	0.812	75.00	384/384/128/128
Shared text F	1.000 [1.000, 1.000]	1.000	50.00	512/0/512/0
Shared text FH	1.000 [1.000, 1.000]	1.000	50.00	512/0/512/0
Shared ledger	1.000 [1.000, 1.000]	1.000	100.00	512/512/0/0

Table 11: Write-order pairs, original balanced-accuracy thresholds.

Readout	Pair ranking [95% CI]	Pooled AUC	FD acc. (%)	TP/TN/FP/FN
Qwen native	0.564 [0.461, 0.682]	0.561	50.00	256/0/256/0
Qwen residual I	0.484 [0.387, 0.578]	0.489	46.88	181/59/197/75
Qwen residual F	0.414 [0.316, 0.512]	0.458	50.00	256/0/256/0
Qwen residual FH	0.348 [0.258, 0.445]	0.433	50.00	256/0/256/0
Qwen centroid FH	0.668 [0.602, 0.742]	0.567	50.00	256/0/256/0
Qwen J-lens	0.592 [0.520, 0.668]	0.527	51.17	236/26/230/20
Qwen direct	0.357 [0.270, 0.449]	0.443	47.85	206/39/217/50
Gemma native	0.928 [0.889, 0.965]	0.739	50.39	256/2/254/0
Gemma residual I	0.898 [0.852, 0.945]	0.718	61.33	217/97/159/39
Gemma residual F	0.930 [0.895, 0.965]	0.846	51.56	256/8/248/0
Gemma residual FH	0.941 [0.898, 0.977]	0.872	57.62	256/39/217/0
Gemma centroid FH	0.957 [0.926, 0.984]	0.802	50.39	256/2/254/0
Gemma J-lens	0.945 [0.910, 0.977]	0.777	50.39	256/2/254/0
Gemma direct	0.887 [0.828, 0.936]	0.705	52.54	256/13/243/0
Shared text I	0.172 [0.109, 0.219]	0.426	50.00	256/0/256/0
Shared text F	1.000 [1.000, 1.000]	0.644	50.00	256/0/256/0
Shared text FH	1.000 [1.000, 1.000]	0.609	50.00	256/0/256/0
Shared ledger	1.000 [1.000, 1.000]	1.000	100.00	256/256/0/0

Table 12: Restore-receipt pairs, original false-denial thresholds.

Readout	Pair ranking [95% CI]	Pooled AUC	BA acc. (%)	TP/TN/FP/FN
Qwen native	0.564 [0.461, 0.682]	0.561	53.52	127/147/109/129
Qwen residual I	0.484 [0.387, 0.578]	0.489	48.63	198/51/205/58
Qwen residual F	0.414 [0.316, 0.512]	0.458	50.00	256/0/256/0
Qwen residual FH	0.348 [0.258, 0.445]	0.433	50.00	256/0/256/0
Qwen centroid FH	0.668 [0.602, 0.742]	0.567	50.78	254/6/250/2
Qwen J-lens	0.592 [0.520, 0.668]	0.527	51.56	56/208/48/200
Qwen direct	0.357 [0.270, 0.449]	0.443	46.48	28/210/46/228
Gemma native	0.928 [0.889, 0.965]	0.739	61.72	247/69/187/9
Gemma residual I	0.898 [0.852, 0.945]	0.718	55.66	250/35/221/6
Gemma residual F	0.930 [0.895, 0.965]	0.846	50.00	256/0/256/0
Gemma residual FH	0.941 [0.898, 0.977]	0.872	51.56	256/8/248/0
Gemma centroid FH	0.957 [0.926, 0.984]	0.802	59.57	255/50/206/1
Gemma J-lens	0.945 [0.910, 0.977]	0.777	50.00	256/0/256/0
Gemma direct	0.887 [0.828, 0.936]	0.705	62.30	221/98/158/35
Shared text I	0.172 [0.109, 0.219]	0.426	39.06	0/200/56/256
Shared text F	1.000 [1.000, 1.000]	0.644	50.00	128/128/128/128
Shared text FH	1.000 [1.000, 1.000]	0.609	64.06	128/200/56/128
Shared ledger	1.000 [1.000, 1.000]	1.000	100.00	256/256/0/0

Table 13: Restore-receipt pairs, original balanced-accuracy thresholds.

Paired comparison	Restore-pair ranking difference [95% CI]
Gemma native: J lens minus logit	0.059 [0.006, 0.119]
Gemma: F minus I	0.031 [-0.016, 0.082]
Gemma: FH minus F	0.012 [-0.012, 0.039]
Gemma: FH minus I	0.043 [0.008, 0.082]
Qwen FP32 reference: J lens minus logit	0.209 [0.115, 0.309]
Qwen native: J lens minus logit	0.234 [0.145, 0.328]
Qwen: F minus I	-0.070 [-0.148, 0.000]
Qwen: FH minus F	-0.066 [-0.129, -0.004]
Qwen: FH minus I	-0.137 [-0.199, -0.078]
Qwen: native minus fp32	0.025 [0.006, 0.045]

Table 14: Direct differences, not comparisons of two marginal intervals. Qwen native versus FP32 is small but its pointwise paired interval excludes zero. This does not establish native numerical qualification.

Model	Rule	J complete	Direct complete	Difference (pp) [95% CI]
Qwen	Low false-denial	278	278	0.00 [-3.71, 2.54]
Qwen	Balanced	336	330	1.17 [-7.42, 8.98]
Gemma	Low false-denial	508	506	0.39 [-0.59, 1.56]
Gemma	Balanced	393	435	-8.20 [-13.67, -2.34]

Table 15: Native BF16 empirical J-lens and direct logit-lens readouts on 512 cached workflows per row. Differences resample paired families. The original validation thresholds are unchanged.

Readout	Rule	Controller	Completed	Invalid/case	Extra/case	Actions/case
Gemma J	BA	recovery	512	0.248	0.006	1.754
Gemma J	BA	repeat	393	0.713	0.008	1.756
Gemma J	FD	recovery	512	0.244	0.039	1.783
Gemma J	FD	repeat	508	0.244	0.086	1.822
Gemma direct	BA	recovery	512	0.113	0.129	1.742
Gemma direct	BA	repeat	435	0.113	0.520	1.982
Gemma direct	FD	recovery	512	0.242	0.023	1.766
Gemma direct	FD	repeat	506	0.242	0.062	1.793
Qwen ref. J	BA	recovery	512	0.162	0.139	1.801
Qwen ref. J	BA	repeat	337	0.486	0.518	2.000
Qwen ref. J	FD	recovery	512	0.449	0.012	1.961
Qwen ref. J	FD	repeat	276	1.348	0.035	1.973
Qwen ref. direct	BA	recovery	512	0.111	0.189	1.801
Qwen ref. direct	BA	repeat	330	0.307	0.723	2.076
Qwen ref. direct	FD	recovery	512	0.453	0.023	1.977
Qwen ref. direct	FD	repeat	278	1.355	0.037	1.984
Qwen J	BA	recovery	512	0.158	0.139	1.797
Qwen J	BA	repeat	336	0.475	0.527	2.000
Qwen J	FD	recovery	512	0.443	0.014	1.957
Qwen J	FD	repeat	278	1.330	0.041	1.971
Qwen direct	BA	recovery	512	0.111	0.189	1.801
Qwen direct	BA	repeat	330	0.307	0.723	2.076
Qwen direct	FD	recovery	512	0.453	0.023	1.977
Qwen direct	FD	repeat	278	1.355	0.037	1.984

Table 16: All new lens controller cells, 512 cases per row. Native rows are empirical controls after failed finite-difference diagnostics; Qwen reference uses the separately qualified FP32 tail. Every recovery cell completes; repeat-selector action means are not matched-success efficiency measures.

E Archived Error Decomposition and Native Comparators

This descriptive audit joins the saved per-input predictions to the 1,536 frozen receipt histories. It reproduces 56 metric cells (two models, seven readouts, two thresholds, two history types) exactly, without new model calls, head fits, or threshold selection. The seven readouts are native and the initial-only, feedback, and history-control logistic and centroid heads. Counts use the same 16 decision families; the strata below are not additional independent experiments or multiplicity-adjusted discoveries.

Model	Readout	Write FP	Write FN	Restore FP	Restore FN
Qwen	Native	44/512	15/512	256/256	0/256
Qwen	Initial	13/512	181/512	197/256	75/256
Qwen	+Feedback	476/512	0/512	256/256	0/256
Qwen	+History controls	405/512	0/512	256/256	0/256
Gemma	Native	0/512	434/512	254/256	0/256
Gemma	Initial	23/512	11/512	159/256	39/256
Gemma	+Feedback	187/512	0/512	248/256	0/256
Gemma	+History controls	38/512	1/512	217/256	0/256

Table 17: Conditional error counts at the frozen low-false-denial threshold. FP means an uncovered input is classified covered; FN means a covered input is classified uncovered. Each denominator is the corresponding true-label count, not the number of all inputs. Native uses its validation-calibrated final-logit threshold.

Saturation is not a single failure mechanism. For Qwen's feedback head, the write-order negative and positive score medians are 15.64 and 39.33, both above its threshold 0.756. Its AUC is still 0.964, although 476/512 negatives become false positives. With history controls, the medians are 12.16 and 33.17 against threshold 1.198, and AUC is 0.947. Restore histories differ: every Qwen F/FH score exceeds its threshold, while FH pair ranking is 0.348 and pooled AUC 0.433. Thus the write-order result retains strong rank discrimination at a misplaced operating threshold; the restore result additionally loses the desired ordering. Scores from separately fitted heads are not treated as one invariant scale.

What the event strata can identify. The audit groups predictions by last target event, target TEST count, the last TEST's target-event index, events since that TEST, whether it follows the last content-changing event, query slot, source mechanism, and family. Within every one of the 768 pairs the target TEST counts match. In write-order pairs the last event is TEST for positives and WRITE for negatives, so this split confounds event type with label. In restore pairs the last event is always RESTORE. The full score summaries, confusion counts, and I-to-F-to-FH prediction changes are retained in the supplement, without computing AUC for single-label strata. These observations constrain simple explanations, but do not isolate the feature causing an error or establish that the model uses a count shortcut.

F Additional J-Lens Methods and Results

At the final input position T, the task-label-free matrix is

$$\bar{J} = \frac{1}{4}\sum_{p=1}^{4} \frac{\partial r^{(p)}_{\text{final,pre-norm},T}}{\partial h^{(p)}_{\ell,T}}. \tag{3}$$

We apply $\bar{J}h$, the final normalization once, and the original output head, retaining Gemma's final softcap. The score is the yes-minus-no logit contrast; the direct logit lens omits $\bar{J}$. The matrix uses no task labels, but its action threshold uses labeled validation data. This same-position heuristic differs from the future-summed J-space construction (Gurnee et al., 2026).

Native BF16 matrices fail the specified finite-difference checks. A post-hoc amendment evaluates them as empirical readouts while preserving that failure. The separately qualified Qwen FP32-tail reference remains distinct (Appendix I).

Gemma J-lens completes 508/512 workflows under the low-false-denial rule, but the direct lens already completes 506; their paired difference is 0.39 percentage points with an interval spanning zero. Under the balanced rule, J-lens completes 393 versus 435 for direct, a paired difference of -8.20 [-13.67, -2.34] points. Qwen's two differences (278 versus 278 and 336 versus 330) have intervals spanning zero (Table 15). On restore pairs, J-minus-direct pair ranking increases by 0.234 [0.145, 0.328] in native Qwen and 0.059 [0.006, 0.119] in Gemma; the FP32 Qwen reference also improves.

Higher first-action accuracy need not mean higher completion. Gemma J-lens makes 367 correct first choices under low false-denial and 382 under the balanced rule, yet completion drops from 508 to 393. Under the first rule, 144 cases recover from an initially wrong choice. Under the second, all 119 failures propose PUBLISH on uncovered contents three times: the threshold keeps denial feedback on the PUBLISH side. This complements the repeated-TEST failure in Figure 2; exact joins follow.

G Exact First-Choice and Outcome Joins

G.1 Repeated Testing by Logistic Probes in Both Models

These post-hoc descriptive joins use the same archived Qwen and Gemma cached replays as Table 2. They are not new model calls or live controller executions. The first-choice/completion intersection is counted by case identity, not inferred by subtracting totals. Repeated TEST here means the second choice after a successful TEST at step 0; later transitions are not counted again in that denominator. The added CPU audit checks all 6,144 traces against the archived scores, thresholds, and state transitions, and reports all three training sets under both rules.

Under the 5% FNR rule, the new Qwen Initial-only probe is initially correct in 499/512 cases but completes 341/512; 165/171 failures start correctly, and 184/261 post-TEST choices repeat TEST. Gemma is initially correct in 500/512 but completes 389/512; its corresponding counts are 115/123 and 137/268. These two-model replays remain separate from the historical executed Qwen result (497/512, 365/512, 143/147, and 169/267; Appendix Y). Feedback training completes all 512 cases under either rule in both models, but does not eliminate every unnecessary TEST. For example, the

Model	Training	Rule	First correct	Completed	Correct-first failures	Repeated TEST
Qwen	Initial-only	FD	499	341	165/171	184/261
Qwen	Initial-only	BA	494	358	146/154	166/252
Qwen	+Feedback	FD	493	512	–	2/267
Qwen	+Feedback	BA	499	512	–	1/257
Qwen	+History controls	FD	498	512	–	5/264
Qwen	+History controls	BA	498	512	–	4/254
Gemma	Initial-only	FD	500	389	115/123	137/268
Gemma	Initial-only	BA	509	436	75/76	114/257
Gemma	+Feedback	FD	503	512	–	1/265
Gemma	+Feedback	BA	507	512	–	1/257
Gemma	+History controls	FD	507	511	1/1	3/261
Gemma	+History controls	BA	508	512	–	1/258

Table 18: A correct first choice can lead to repeated testing in both models. First-correct and completed counts are out of 512. Correct-first failures divide the exact intersection by all failures; a dash means no failures. Repeated TEST divides repeated second actions by choices following a successful step-0 TEST. FD is the 5% FNR rule; BA is balanced accuracy. These are descriptive counts on the same 16 families, not independent new experiments.

+Feedback heads retain 2/267 (Qwen) and 1/265 (Gemma) repeated next choices under FD while still finishing within budget. The Gemma +History-controls FD head retains one failure.

Step	State before action	Choice	Score	Fixed threshold	Outcome
0	Uncovered	TEST	-22.414	5.4593	Success
1	Covered	TEST	-10.053	5.4593	Success
2	Covered	TEST	3.180	5.4593	Success

Table 19: Gemma cached replay, case `lens-use-000-restore-00`, Initial-only probe and 5% FNR threshold. TEST succeeds and makes the current contents covered, but both subsequent scores remain below the fixed PUBLISH threshold. The controller exhausts its three choices. This illustrative trace is the first case by identifier among correctly initialized three-TEST failures, not a randomly selected or confirmatory sample.

This is a probe-controller result, not a claim that Gemma's generated policy repeats TEST. In the separate live check, both models generate PUBLISH after every observed successful TEST (Appendix Z). That check uses different prompts, 64 reused cases, and a post-hoc parser amendment.

G.2 J-Lens First-Choice and Outcome Joins

Readout	Rule	Initial FP/FN	Wrong/success	Wrong/fail	Correct/success	Correct/fail
qwen native	FD	227/7	3	231	275	3
qwen native	BA	81/71	30	122	306	54
gemma native	FD	125/20	144	1	364	3
gemma native	BA	127/3	11	119	382	0
qwen fp32 reference	FD	230/6	3	233	273	3
qwen fp32 reference	BA	83/71	32	122	305	53

Table 20: Direct joins on case identity for the new J-lens repeat selector. "Wrong" and "correct" describe the first TEST/PUBLISH choice, not the final outcome. Every row sums to 512 across its last four columns.

Gemma/FD has 144 initially incorrect cases that finish and only one initially incorrect case that fails. Its successful patterns include 124 `PUBLISH denied; TEST success; PUBLISH accepted` traces. Three correctly initialized cases instead spend all three decisions testing. Under Gemma/BA, 119 cases repeatedly propose an uncovered publication and never dispatch a TEST. All are initial false positives. Of the 118 cases completed by FD but not BA, the improved static

operating point therefore fails to preserve the favorable feedback trajectory. Three cases complete under BA but not FD; one fails under both. Net counts alone would conceal these discordances.

The audit reconstructs 59,392 cached episodes across all saved residual, centroid, text, native, ledger, and lens settings, checking every score comparison, prefix transition, action, terminal outcome, and count. This is a computational consistency check on cached policies, not 59,392 independent tasks or fresh live executions. Stored traces and exact examples are included in the supplement.

H Readout-Direction Interchange Diagnostic

The original post-hoc diagnostic and the later norm-matching correction each froze their code and analysis before their respective model forwards. The correction was motivated by the original realized-norm imbalance, not designed as an independent replication. Both use the exact revisions of Appendix I on one RTX PRO 6000 96 GB GPU with native BF16 weights, a full outer-model forward, eager attention, and no quantization or offload. The runtime is PyTorch 2.8.0+cu128, Transformers 5.8.1, Accelerate 1.12.0, and NumPy 2.3.4, distinct from the earlier three-GPU captures. The chat template and no-thinking settings are unchanged; the endpoint is the native lowercase single-token yes-minus-no logit margin. No token is generated, no head or threshold is refitted, and no Jacobian is used.

Selection and directions. From the frozen receipt corpus of Appendix C, one pair is drawn per decision family in each of the four pair-type by source-mechanism cells, choosing the lowest salted SHA-256 of the pair identifier. Each model therefore has 64 pairs and 128 histories; both members of every pair are kept. For a score $w^\top((h-\mu)/\sigma)+b$, the raw-residual gradient is w/σ, normalized to a unit vector u. The +History-controls logistic head and the centroid readout of Appendix C each supply one u per model; the centroid direction is its score gradient, not the unstandardized class-mean displacement. Centers and intercepts remain part of score reporting but do not enter the intervention.

Conditions. Each history has ten conditions: baseline; the logistic interchange $h' = h + u\,u^\top(h_d - h)$ with matched opposite-label donor h_d; its three random shams; the centroid interchange and its three shams; and full final-position donor replacement. The original run matched requested norms. The correction retains all selected inputs and active interventions but matches the shams' delivered norms. Each run comprises 256 baseline and 2,304 intervention forwards across both models.

Matching the actual BF16 perturbation. The hook already performs addition in float32 before casting back to BF16. Therefore merely promoting the addition would not fix the imbalance. For requested change d, define the delivered change as $q(h,d) = \mathrm{BF16}(\mathrm{FP32}(h) + \mathrm{FP32}(d)) - h$. Each active perturbation fixes the target $\|q(h, d_{\mathrm{active}})\|_2$. A sham's nonnegative scalar is chosen from arithmetic alone: bracket the target by doubling an initial upper scalar of 1 up to at most 64, run 60 bisections, and retain the closest candidate, breaking ties toward the smaller scalar. The relative norm error must be at most 1%.

Quantization plateaus make a target unattainable on some rays. We try the original ray first, then at most 15 deterministic isotropic Gaussian proposals, keeping the first feasible one. Their seed is derived from the base residual and original requested vector, without logits, labels beyond the already fixed active intervention, or outcome tests. Six of 768 shams per model require a replacement; maximum proposal indices are 4 for Qwen and 12 for Gemma. All inputs remain in the analysis. Proposals are isotropic before matching; the accepted, BF16-delivered controls are conditioned on feasibility and are not claimed to remain isotropic. Scalars, replacement vectors, and predicted patch hashes were fixed before the new forwards.

Endpoint and inference. For each cell, the effect is the family-macro mean of the donor-signed native-margin change, with donor sign $2y_{\mathrm{donor}} - 1$, minus the mean donor-signed change of the three shams. Bootstrap intervals use 2,000 paired family-cluster replicates with seed 2026092202. Exact two-sided sign-flip tests on the 16 family effects are Holm-adjusted across the eight primary cells. Secondary outputs are requested and realized BF16 perturbation norms, native yes/no mass, full-vocabulary KL from baseline, and choice flips at the native zero threshold.

Model	Readout	Histories	Sham/active norms	Max error (%)	Active mean KL
Qwen	Logistic	Write order	0.9983–1.0021	0.214	0.000
Qwen	Logistic	Restore receipt	0.9918–1.0070	0.819	0.000
Qwen	Centroid	Write order	0.9938–1.0084	0.844	0.001
Qwen	Centroid	Restore receipt	0.9955–1.0056	0.556	0.000
Gemma	Logistic	Write order	0.9971–1.0083	0.826	0.022
Gemma	Logistic	Restore receipt	0.9906–1.0098	0.979	0.002
Gemma	Centroid	Write order	0.9981–1.0020	0.196	0.502
Gemma	Centroid	Restore receipt	0.9925–1.0087	0.868	0.047

Table 21: Per-intervention realized-norm ratios in the correction, not ratios of cell averages. All 1,536 shams satisfy the 1% tolerance. KL describes the active perturbation and is unchanged from the original run.

Model	Readout	Histories	Readout norm ratio	Sham norm ratios	Mean KL	Flips
Qwen	Logistic	Write order	1.014	1.016–1.018	0.000	0/64
Qwen	Logistic	Restore receipt	0.942	0.917–0.924	0.000	0/64
Qwen	Centroid	Write order	1.001	1.001–1.001	0.001	1/64
Qwen	Centroid	Restore receipt	1.012	1.010–1.011	0.000	0/64
Gemma	Logistic	Write order	1.000	0.506–0.525	0.022	0/64
Gemma	Logistic	Restore receipt	1.001	0.303–0.314	0.002	0/64
Gemma	Centroid	Write order	1.000	0.934–0.937	0.502	3/64
Gemma	Centroid	Restore receipt	1.000	0.572–0.575	0.047	0/64

Table 22: Historical requested-norm controls, retained unchanged: realized-to-requested L_2 ratios (ratios of cell means), active mean KL, and flips. Gemma's original shams were smaller. The corrected main results use Table 21's realized-norm matching instead.

Model	Histories	Baseline zero-threshold acc. (%)	Full-state mean KL	Full-state flips
Qwen	Write order	56.2	0.240	4/64
Qwen	Restore receipt	50.0	0.003	0/64
Gemma	Write order	65.6	2.914	10/64
Gemma	Restore receipt	50.0	0.051	0/64

Table 23: Reference condition: full final-position donor replacement on the same 64 histories per cell. Baseline accuracy uses the native zero threshold on the selected subset. Both models make a constant native restore choice at baseline. Empirically, no tested restore intervention, including full replacement, changes a choice; baseline constancy alone would not imply this.

Model	Readout	Histories	Original effect [95% CI]	Holm p	Flips
Qwen	Logistic	Write order	-0.001 [-0.016, 0.018]	1.000	0/64
Qwen	Logistic	Restore receipt	-0.003 [-0.018, 0.012]	1.000	0/64
Qwen	Centroid	Write order	0.137 [0.114, 0.159]	$<$0.001	1/64
Qwen	Centroid	Restore receipt	0.015 [-0.003, 0.033]	0.480	0/64
Gemma	Logistic	Write order	0.296 [0.226, 0.372]	$<$0.001	0/64
Gemma	Logistic	Restore receipt	0.221 [0.170, 0.275]	$<$0.001	0/64
Gemma	Centroid	Write order	1.619 [1.168, 2.137]	$<$0.001	3/64
Gemma	Centroid	Restore receipt	0.770 [0.602, 0.953]	$<$0.001	0/64

Table 24: Original requested-norm results, before the arithmetic correction. They remain part of the record and are not pooled with the corrected contrasts as independent data.

Reading the tables. The same five contrasts remain positive after realized-norm matching. All 256 baseline states/logits and 768 active/full-state interventions are byte-identical across runs, so this comparison specifically changes the random controls. Gemma centroid/write has mean KL 0.502 and maximum 7.088; its small relative residual change need not be a small output disruption. On the selected restore subsets, external logistic probes at their frozen thresholds classify 50.0% (Qwen) and 56.25% (Gemma) correctly, separately from native zero-threshold choices. Qwen logistic changes individual native margins in 24/64 write-order and 16/64 restore histories despite null mean contrasts. Null sensitivity is not specific to the reversed-ranking setting and cannot by itself diagnose a shortcut. Constant restore choices and unsuccessful full-state flips also limit localization claims at this one site. No outcome-selected position/layer scan was run.

Custody and recomputation. Snapshot manifests were freshly checked before both runs (Qwen 32 files, 55,586,114,863 bytes; Gemma 12 files, 62,578,686,256 bytes). Two fresh processes per model reproduced eight baseline witnesses exactly in tokens, residual bytes, and full-logit hashes; a zero-delta hook left full logits unchanged. Local CPU verification reproduces all 2,304 corrected patched states, all norm matches, and all eight effects, intervals, and p values. The compact supplement reconstructs every frozen float64 perturbation from stored scalars, shared directions, and the few replacement vectors, checking its hash before replay. Full-vocabulary KL is retained from the GPU run, not independently recomputed from missing full logit arrays. The original evidence archive has SHA-256 `f8b9405e988fe6d273db67bf6250eafcfd54f7fbf28347c76b4fdd7ac8d45d88`; the correction (3,445,754 bytes, 332 files) has SHA-256 `40a82cffec93f01edf518c2ade4ba4e418fc7da0e56f76a87d6a91b9eb51a0dd`. The original norm mismatch is preserved, not retroactively reclassified as matched.

I Successor Numerical Identity and Evidence Scope

New receipt captures and lens fits use three RTX 5090 32 GB GPUs, PyTorch 2.6.0a0+ecf3bae40a.nv25.01 and Transformers 5.17.0. Full outer-model forwards preserve architecture-specific preprocessing, final normalization, and Gemma's logit softcap. Loaded weights are native BF16; saved residuals are float32. No quantization or CPU/disk model offload is used. Matrix dimensions are 5120^2 for Qwen L47 and 5376^2 for Gemma L44. Four neutral prompts are averaged, without task labels; no affine intercept is fitted, and the action threshold is supervised.

The exact model revisions are Qwen `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0` and Gemma `842da3794eaa0b77d5f08bae87a17459d91ff475`. Current snapshot verification covers Qwen's 32 files (55,586,114,863 bytes) and Gemma's 12 files (62,578,686,256 bytes). For each model, two fresh processes recapture the same 32 witnesses and match residuals, tokenization, and logits exactly. Main receipt capture then records all 1,536 inputs. This is not a full independent replay of all new inputs. Earlier coverage-v1 captures retain their original provenance limits; manifest verification now cannot retroactively certify their complete historical runtime.

Qualification and amendment. The original native-BF16 finite-difference check stops at its first failing tested direction; it does not establish that every direction fails. Its original dense fits remain Not Run. The new empirical-control amendment makes the finite-difference checks nonblocking, keeps the diagnostics and failure logs, and completes native dense fits and scoring in both models. The separate Qwen FP32-tail reference passes its specified checks in two fresh processes. It promotes downstream blocks and normalization while preserving the loaded BF16 values and original head. It is not a replacement verdict for native arithmetic.

Precision sensitivity. On 4,784 unique cached inputs, Qwen native and FP32-tail J-lens scores have Spearman correlation 0.999545. The mean matrices have cosine 0.999946 and relative Frobenius difference 0.010518. These descriptive agreements do not imply equal decisions near a threshold. On the new 512 restore rows, separately validation-calibrated decisions disagree once under FD and nine times under BA (99.80% and 98.24% agreement). On 1,024 write-order rows they disagree seven and six times. The direct native-minus-reference restore ranking difference is 0.0254 [0.0059, 0.0449] (Table 14); its positive paired interval is not inferred from whether the two marginal intervals cross chance. Both arithmetic versions still have poor frozen-threshold receipt accuracy.

Supplement boundary. The minimal anonymous capsule packages the actual scores, traces, new receipt residuals, numeric probe/scaler parameters, input histories, execution-report extracts, model file hashes, and the deterministic CPU verifier. It checks the new primary statistics and saved-head scores without loading pickled estimators. Dense matrices, model weights, and all older raw GPU captures are omitted for size; hashes bind the retained larger archives. The capsule cannot be used to claim an independent numerical requalification or a new GPU reproduction. Historical appendices below retain the earlier experimental identities and verdicts.

J Additional Historical Controls

The original difference-in-means and shallow-text controls are retained here because they concern Study 1 rather than the feedback-trained decision heads.

Coordinates	O→O	O→R	R→O	R→R
Raw	0.752 [0.736, 0.770]	0.442 [0.427, 0.458]	0.433 [0.420, 0.448]	0.764 [0.750, 0.780]
Train-standardized	0.762 [0.747, 0.779]	0.437 [0.419, 0.455]	0.407 [0.391, 0.424]	0.772 [0.760, 0.785]

Table 25: Historical Qwen L47 difference-in-means AUCs with family intervals.

Readout	O→O	O→R	R→O	R→R
Unigram TF–IDF	0.500	0.500	0.500	0.500
Ordered 1–3-gram TF–IDF	0.696	0.375	0.383	0.625
Residual probe (L47)	0.992	0.338	0.337	0.981
Last relevant operation	1.000	0.500	1.000	0.500
Exact visible ledger	1.000	1.000	1.000	1.000

Table 26: Historical Qwen unqueried shallow-text and deterministic controls.

K HISTORIES THAT SEPARATE COVERAGE FROM SURFACE EVENTS

Additional representation context. SVCCA and CKA compare representation spaces (Raghu et al., 2017; Kornblith et al., 2019); sparse-feature evaluations examine decomposition quality and generalization (Heindrich et al., 2025; He et al., 2026). Our unit of evaluation is instead a fitted task score and its decision threshold. S-space and latent-dynamics studies examine spatial workspaces and recurrent latent computation (MirroS Team, 2026; Lai et al., 2026); the controls below test a narrower property, invariance of a coverage judgment to specified changes in visible history.

This review-motivated follow-up freezes its protocol, inputs, fitted heads, and thresholds before any new forward pass. It uses the existing 16 held-out decision-family file rosters and all four query slots. No training or validation family is added. The 64 family–slot units are crossed with two target-history constructions, four revision assignments, three suffixes, and two labels, giving 3,072 histories per model and 1,536 matched pairs. Families, not histories, are the 16 resampling units.

A WRITE need not invalidate a test. Initialize all four files at r0. For the queried file, let a be the final contents and b different contents. The following cores end in the same operation on both sides and have matching word and event multisets:

Core	Covered	Uncovered
Same-bytes WRITE	`WRITE b; WRITE a; TEST; WRITE a`	`WRITE b; TEST; WRITE a; WRITE a`
RESTORE	`WRITE a; TEST; WRITE b; RESTORE a`	`WRITE a; WRITE b; TEST; RESTORE a`

The last WRITE changes no bytes on either side. The positive side has tested a, while the negative side has only tested b. A rule that invalidates every WRITE would therefore fail on positives. Each core has exactly one target-file TEST. We use $(a, b) = (r0, r1), (r1, r0), (r0, r2), (r2, r0)$, so revision identity does not fix the label.

Changes that leave target coverage fixed. Each history receives one of three suffixes: none, one TEST of the next file in the roster, or four WRITE-r0/TEST cycles on that other file (eight events). Neither suffix changes the queried file or its receipts. The suffixes test whether irrelevant activity changes a frozen decision. They are not a pure manipulation of token count: adding events also changes the visible event sequence. The r2 contrast substitutes r2 for r1 in corresponding histories while retaining target coverage. The original renderer and its fixed revision legend are unchanged.

All 3,072 histories pass actual byte-hash checks and isolated TEST execution before capture, in addition to a symbolic reconstruction. These are static readout tests, not new controller rollouts. The exact ledger remains correct throughout.

Frozen capture and scoring. We use the same Qwen and Gemma revisions, native BF16 eager forward passes, and primary layers as in the preceding study. The full outer model supplies the final yes/no logits; the selected last-position residual is saved in float32. No tokens are generated. Initial-only, +Feedback, and +History-controls logistic and centroid heads reuse their original initial-only scaler and both original validation thresholds. Numeric exports of the logistic heads reproduce the saved classifiers exactly on archived residual witnesses. Native scoring uses the original validation-calibrated yes-minus-no contrast.

Two fresh processes per model repeat eight hash-selected inputs. Their token, residual, and full-logit hashes must match, and the main capture must reproduce those inputs. Snapshot manifests are checked file by file. These identity checks concern this capture; they do not strengthen the identity evidence of older captures retroactively.

Estimands and multiplicity. For each cell, we report matched-pair ranking, pooled AUC, frozen accuracy, and conditional error counts. Family bootstrap draws retain all slots, revision assignments, labels, and suffixes (2,000 draws, integer seed 2026092307). For each model and the two feedback-trained logistic heads, the three primary accuracy contrasts are other-file TEST minus no suffix, eight irrelevant events minus no suffix, and r2 substitution minus its r1 counterpart. These 12 paired

family-sign-flip tests receive one Holm adjustment. Score changes, decision disagreements, other readouts, and cellwise intervals are descriptive. The transformations test sensitivity to these inputs, not an exclusive neural feature or mediation mechanism.

Model	Target core	Readout	Pair rank	No suffix	Other TEST	Eight events
Qwen	Same-bytes WRITE	Native	1.000	84.96	65.82	82.42
Qwen	Same-bytes WRITE	Initial	1.000	90.23	70.51	96.68
Qwen	Same-bytes WRITE	+Feedback	1.000	92.38	83.79	99.80
Qwen	Same-bytes WRITE	+History controls	1.000	87.30	86.13	99.41
Qwen	Same-bytes WRITE	Centroid +controls	1.000	98.63	94.73	88.67
Qwen	RESTORE	Native	0.996	67.19	51.37	63.28
Qwen	RESTORE	Initial	0.984	87.11	52.93	64.84
Qwen	RESTORE	+Feedback	0.984	88.09	52.54	62.89
Qwen	RESTORE	+History controls	0.973	87.89	52.54	60.74
Qwen	RESTORE	Centroid +controls	0.996	88.67	57.42	58.98
Gemma	Same-bytes WRITE	Native	1.000	86.72	79.30	81.25
Gemma	Same-bytes WRITE	Initial	1.000	87.50	99.80	99.80
Gemma	Same-bytes WRITE	+Feedback	1.000	83.01	96.48	99.80
Gemma	Same-bytes WRITE	+History controls	1.000	85.16	98.24	100.00
Gemma	Same-bytes WRITE	Centroid +controls	1.000	93.95	87.50	90.62
Gemma	RESTORE	Native	1.000	92.38	96.68	93.36
Gemma	RESTORE	Initial	1.000	96.68	96.09	99.02
Gemma	RESTORE	+Feedback	1.000	95.90	97.66	98.63
Gemma	RESTORE	+History controls	1.000	94.73	98.44	97.66
Gemma	RESTORE	Centroid +controls	1.000	98.24	98.63	98.24

Table 27: Frozen low-false-denial accuracy (percent) across all four revision assignments; 512 histories per model/core/suffix cell. Pair rank is for the no-suffix cell only. Native is validation-calibrated. The complete 672 cells, including both threshold rules and all seven readouts, are in the supplement.

Model	Readout	Change	Accuracy change (pp) [95% CI]	Holm p	Decisions changed (%)
Qwen	+Feedback	Other-file TEST	-22.07 [-24.41, -19.53]	0.0004	24.61
Qwen	+Feedback	Eight events	-8.89 [-12.89, -4.20]	0.0160	23.14
Qwen	+Feedback	r2 for r1	0.85 [-0.00, 1.76]	0.5054	8.40
Qwen	+History controls	Other-file TEST	-18.26 [-21.00, -15.62]	0.0004	25.10
Qwen	+History controls	Eight events	-7.52 [-10.35, -4.20]	0.0086	25.49
Qwen	+History controls	r2 for r1	-0.85 [-2.28, 0.59]	0.6470	9.70
Gemma	+Feedback	Other-file TEST	7.62 [5.37, 9.86]	0.0013	11.13
Gemma	+Feedback	Eight events	9.77 [8.20, 11.13]	0.0004	10.94
Gemma	+Feedback	r2 for r1	-0.39 [-1.30, 0.46]	0.6470	3.12
Gemma	+History controls	Other-file TEST	8.40 [7.32, 9.47]	0.0004	10.35
Gemma	+History controls	Eight events	8.89 [7.62, 10.16]	0.0004	11.23
Gemma	+History controls	r2 for r1	-0.65 [-1.50, 0.13]	0.5801	3.12

Table 28: All 12 primary paired accuracy contrasts. Suffix contrasts pair 1,024 target-identical views; revision contrasts pair 1,536 views. Intervals are pointwise paired-family intervals; Holm correction spans all 12 exact family-sign-flip tests. Decision disagreement can exceed net accuracy change because opposite errors can cancel.

Reading the controlled effects. Both feedback-trained heads have perfect pair ranking on same-bytes-WRITE histories without a suffix, despite WRITE being the last target event for both labels. On the new restore cores their rankings are 0.984/0.973 for Qwen F/FH and 1.000/1.000 for Gemma. These shorter cores differ from the original receipt construction; their strong rankings do not erase the earlier failure. They show that neither the last operation nor target TEST count alone determines the scores across these inputs.

Adding one other-file TEST changes Qwen F/FH decisions on 24.61%/25.10% of paired views and Gemma decisions on 11.13%/10.35%. The accuracy effect is strongly concentrated in Qwen's restore cores: F drops from 88.09% to 52.54%, and FH from 87.89% to 52.54%. Gemma improves in both cores. More irrelevant events do not simply amplify the one-event accuracy effect: Qwen's same-bytes-WRITE accuracy increases with the eight-event suffix even while restore accuracy decreases. A single monotone length explanation is therefore insufficient for the observed cell pattern.

All eight suffix accuracy contrasts pass the 12-test Holm correction; none of the four revision-substitution contrasts does. The latter is not invariance: r2 substitution changes 8.40%/9.70% of Qwen F/FH decisions and 3.12% of Gemma decisions, with opposite errors partly cancelling in

accuracy. The score shifts and all cells remain available in the machine-readable results. These controlled input effects establish context sensitivity at fixed target coverage, not the unique cause of the original receipt-set score drift.

L What the Existing Cache Can Identify

We join each receipt-history score to the initial history from which it was constructed. All initial histories omit r2 in their events; all receipt histories include it. Receipt construction adds exactly 70 characters for write-order and 109 for restore histories. The joint design matrix with intercept, receipt-panel indicator, r2-in-events indicator, event count, and character count has rank three out of five. A regression on these columns cannot identify an independent r2 or length effect. The paired score changes are therefore combined history shifts, not estimates of either factor's causal contribution. The new crossed histories in Appendix K address this limitation with controlled input changes.

Whole-state margin shifts give the intervention an output scale. For the original 64 directed interventions per model/history cell, let $d = 2s_{\mathrm{donor}} - 1$ and m be the native yes-minus-no margin. We summarize $d(m_{\mathrm{patched}} - m_{\mathrm{base}})$ for full donor replacement and for each one-dimensional direction before sham subtraction. Ratios divide the mean direction shift by the mean full-state shift on the same families. They are not ratios of the sham-adjusted effects in Table 4, nor fractions of computation explained. The denominator is a signed mean and cancellation can make a ratio unstable; it is omitted if its bootstrap interval includes zero.

Model	Histories	Full-state shift [95% CI]	Logistic/full	Centroid/full
Gemma	Restore	2.223 [1.791, 2.671]	0.095	0.328
Gemma	Write order	12.106 [9.813, 14.457]	0.019	0.108
Qwen	Restore	0.223 [0.053, 0.385]	0.000	0.061
Qwen	Write order	3.033 [2.574, 3.467]	-0.003	0.053

Table 29: Descriptive output scale at the original fixed layer and last-token site. Shifts are donor-signed logits; ratios use raw active shifts, with paired-family intervals in the machine-readable supplement. Zero restore-choice flips coexist with nonzero full-state margin shifts.

M SEPARATING RANKING FROM FROZEN-THRESHOLD ERRORS

This post-hoc analysis reuses every frozen score from the crossed-history capture. It adds no model calls, fitted heads, threshold choices, or confirmatory tests. The original 12 Holm-adjusted accuracy tests in Table 28 remain unchanged. The decomposition was motivated by inspecting their different accuracy effects, so its intervals are descriptive rather than selection-adjusted discoveries.

Aggregation and dependent views. For each model, readout, target core, and suffix, we pool all four revision assignments. Each cell contains 512 histories (256 of each label) in 256 matched pairs from 16 families. Pair ranking averages positive-minus-negative comparisons with half credit for ties. Every family contributes the same number of pairs. FP and FN are counts over the 256 uncovered and 256 covered histories, respectively. Pooled AUC compares all covered scores with all uncovered scores after pooling assignments; it is not the mean of four assignment-specific AUCs.

We draw 2,000 whole-family bootstrap samples, using integer seed 2026092307. Each draw keeps the four query slots, assignments, labels, and paired histories together. The same family draws are used across suffixes and readouts. Pair ranking, FP/FN, accuracy, and pooled AUC are recomputed within each draw, including cross-family positive-negative comparisons for pooled AUC. Tables 30–32 give pointwise percentile intervals. All seven readouts under both original threshold rules are printed in Appendix P (168 rule-specific records). Machine-readable results additionally retain score distributions, count intervals, and 24 paired native-minus-probe comparisons.

Readout	Suffix	Pair ranking [95% CI]	Pooled AUC [95% CI]	FP	FN
Same-bytes WRITE					
+Feedback	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0	39
	Other-file TEST	1.000 [1.000, 1.000]	0.997 [0.994, 1.000]	0	83
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1	0
+History controls	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0	65
	Other-file TEST	1.000 [1.000, 1.000]	0.994 [0.990, 0.998]	0	71
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0	3
Native score	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	77	0
	Other-file TEST	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	175	0
	Eight events	1.000 [1.000, 1.000]	1.000 [0.999, 1.000]	90	0
RESTORE					
+Feedback	None	0.984 [0.965, 1.000]	0.949 [0.934, 0.967]	3	58
	Other-file TEST	0.531 [0.449, 0.625]	0.539 [0.505, 0.575]	81	162
	Eight events	0.766 [0.684, 0.836]	0.713 [0.665, 0.760]	139	51
+History controls	None	0.973 [0.953, 0.988]	0.949 [0.928, 0.969]	19	43
	Other-file TEST	0.551 [0.473, 0.629]	0.544 [0.512, 0.579]	135	108
	Eight events	0.730 [0.660, 0.793]	0.660 [0.611, 0.706]	160	41
Native score	None	0.996 [0.988, 1.000]	0.998 [0.996, 1.000]	168	0
	Other-file TEST	0.990 [0.977, 1.000]	0.922 [0.907, 0.937]	249	0
	Eight events	0.980 [0.965, 0.992]	0.963 [0.949, 0.977]	188	0

Table 30: Qwen, all suffixes and both target cores at the unchanged low-false-denial thresholds. Each row contains 512 histories; 256 is the conditional denominator for either FP or FN. Native uses the original calibrated yes-minus-no score, not newly generated answers.

A frozen readout and a native score need not lose the same contrast. For Qwen +Feedback on restore cores, the other-file TEST changes pair ranking by -0.453 $[-0.535, -0.363]$ and pooled AUC by -0.410 $[-0.445, -0.376]$. The +History-controls ranking change is -0.422 $[-0.496, -0.348]$. For native scores the pair-ranking change is only -0.006 $[-0.021, 0.008]$, but pooled AUC decreases by -0.076 $[-0.091, -0.062]$. Thus native within-pair discrimination is mostly preserved, not every native metric. Its accuracy at the frozen calibrated threshold falls from 67.19% to 51.37%; we do not call the native answers correct on these histories.

Readout	Suffix	Pair ranking [95% CI]	Pooled AUC [95% CI]	FP	FN
Same-bytes WRITE					
+Feedback	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	87	0
	Other-file TEST	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	18	0
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1	0
+History controls	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	76	0
	Other-file TEST	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	9	0
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0	0
Native score	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	68	0
	Other-file TEST	0.996 [0.988, 1.000]	1.000 [0.999, 1.000]	106	0
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	96	0
RESTORE					
+Feedback	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	21	0
	Other-file TEST	1.000 [1.000, 1.000]	0.999 [0.998, 1.000]	0	12
	Eight events	1.000 [1.000, 1.000]	1.000 [0.999, 1.000]	3	4
+History controls	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	27	0
	Other-file TEST	1.000 [1.000, 1.000]	0.999 [0.998, 1.000]	0	8
	Eight events	1.000 [1.000, 1.000]	0.999 [0.996, 1.000]	2	10
Native score	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	39	0
	Other-file TEST	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	17	0
	Eight events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	34	0

Table 31: Gemma under the same aggregation and fixed thresholds. Perfect pair ranking in the two feedback heads coexists with changing threshold errors and small changes in pooled AUC.

Model	Readout	Added suffix	Pair-ranking change [95% CI]	Pooled-AUC change [95% CI]
Qwen	+Feedback	Other-file TEST	-0.453 [-0.535, -0.363]	-0.410 [-0.445, -0.376]
Qwen	+Feedback	Eight events	-0.219 [-0.301, -0.145]	-0.236 [-0.287, -0.188]
Qwen	+History controls	Other-file TEST	-0.422 [-0.496, -0.348]	-0.406 [-0.436, -0.376]
Qwen	+History controls	Eight events	-0.242 [-0.312, -0.180]	-0.290 [-0.340, -0.243]
Qwen	Native score	Other-file TEST	-0.006 [-0.021, 0.008]	-0.076 [-0.091, -0.062]
Qwen	Native score	Eight events	-0.016 [-0.027, -0.004]	-0.035 [-0.049, -0.022]
Gemma	+Feedback	Other-file TEST	0.000 [0.000, 0.000]	-0.001 [-0.002, -0.000]
Gemma	+Feedback	Eight events	0.000 [0.000, 0.000]	-0.000 [-0.001, -0.000]
Gemma	+History controls	Other-file TEST	0.000 [0.000, 0.000]	-0.001 [-0.002, 0.000]
Gemma	+History controls	Eight events	0.000 [0.000, 0.000]	-0.001 [-0.004, 0.000]
Gemma	Native score	Other-file TEST	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Gemma	Native score	Eight events	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Table 32: Suffix minus no suffix on RESTORE cores, retaining the same families and input matching in every draw. These are post-hoc effect summaries, not additional multiplicity-adjusted hypothesis tests.

Neither "only a shift" nor "the model lost coverage information." On Qwen +Feedback restore histories, the mean other-file-TEST score change is $+9.77$ for uncovered inputs and -7.44 for covered inputs; the per-input shift has standard deviation 11.32. This is a compression and reordering of the two score groups, not a common offset. Gemma +Feedback preserves all matched orderings but shifts scores nonuniformly: corresponding label means are -4.89 and -8.29, with shift standard deviation 6.80. Its improved accuracy describes threshold crossings on these inputs, not a demonstrated increase in general robustness. The mean and range of every paired score shift are in the supplement.

Qwen's no-suffix and one-other-TEST WRITE cores both have pair ranking 1.000, while +Feedback FN counts rise from 39 to 83. Eight irrelevant events instead reduce these FN to zero while restore ranking remains degraded. This supports separate evaluation of ranking and operating points and rules out a single monotone length account of the observed accuracy pattern.

The dissociation concerns the existing frozen projection at the last position of L47, not all possible readouts from that layer. The earlier +History-controls intervention tests another cohort at the same fixed site; a null donor-aligned mean effect is compatible with, but does not demonstrate, a context-sensitive projection that differs from the native scoring route. This discovery capture did not scan sites. Appendix O reports the later six-site readout comparison; neither identifies a causal binding feature. Appendix P prints all 84 model/core/suffix/readout combinations, including all centroid heads, under both original thresholds.

N POST-HOC SAME-SITE READOUTS OF THE QWEN RESIDUAL CACHE

Question and scope. The probe/native comparison changes both the readout and the computation following its site. This CPU follow-up instead holds the observed residual fixed: all 3,072 Qwen block-47 outputs at the final input position from Appendix K. We test both cores, all three suffixes, four revision assignments, and all 16 decision families. The analysis scope was recorded before computing these new scores, but the question was motivated by observed results. It is therefore a post-hoc diagnostic, not confirmation on new families. No backbone forwards, model-token generation, head fitting, threshold selection, or site search were performed.

Exact weights and normalization. The model revision is unchanged. We downloaded the two shards containing the output normalization and unembedding, then recomputed their complete SHA-256 hashes against the pre-existing snapshot manifest. The retained asset contains the 5,120 normalization weights and the two 5,120-dimensional rows for tokens `yes` (9405) and `no` (2083). Qwen's zero-centered RMS normalization is

$$N(h) = \frac{h}{\sqrt{\mathrm{mean}(h^2) + 10^{-6}}} \odot (1 + w), \qquad g_D(h) = u_{\mathrm{yes}}^{\top} N(h) - u_{\mathrm{no}}^{\top} N(h). \tag{4}$$

There is no Qwen output softcap. In particular, omitting the $1+w$ offset would implement a different readout. The primary CPU calculation uses float32 on the exact BF16-stored weights and residual values. A sensitivity calculation rounds the input, normalized vector, and each output logit to BF16 with round-to-nearest-even; products accumulate in float32. This is an arithmetic sensitivity check, not a claim of exact equivalence to the archived CUDA vocabulary projection.

Secondary empirical matrix. The second readout is $g_J(h) = g_D(\bar{J}h)$, using the already archived native-BF16 mean matrix, with output coordinates as rows. Its fit used four neutral prompts and no task labels. We neither refit it nor use the distinct FP32-tail reference matrix. Its original finite-difference qualification failed and remains failed; the new result describes an empirical fixed transformation, not a validated derivative. The matrix hash and its original fit record are preserved. For the BF16 sensitivity calculation, the transported vector is rounded before normalization.

Scoring and intervals. All scores have the fixed yes-minus-no direction. We pool the four revision assignments before computing AUC. Pair ranking gives half credit to ties. The same 2,000 whole-family bootstrap draws (seed identifier 2026092307) are shared across readouts and suffixes. Intervals and contrasts are pointwise descriptive diagnostics. No new frozen-threshold accuracy is reported: these CPU score conventions were not separately calibrated, and a ranking comparison does not require inventing an operating point. Tables 33–35 report all cells and the targeted RESTORE contrast.

What this comparison establishes. On RESTORE cores after another file's TEST, the direct lens also deteriorates, so a simple vocabulary projection does not remove this probe/native dissociation. The empirical J-lens nevertheless retains substantial ranking at the identical site. The frozen probe is therefore not an exhaustive test of information available there. This does not identify the native computation's causal feature, prove complete state semantics, or exclude improvements at later layers or other positions. A failed lens alone would not prove absence of information either. The later fresh-roster confirmation and six-site comparison are reported separately in Appendix O; they do not turn this earlier diagnostic into a prespecified analysis.

CPU audit boundary. All 3,072 residual hashes are checked against the original capture records before lossless BF16 packing. The compact supplement contains those arrays, token/residual hash witnesses, the two output rows and normalization weights, and CPU code. Its clean audit recomputes the direct-lens scores within declared numerical tolerance and their ranking statistics. For the secondary J-lens it recomputes statistics from the archived per-input scores, but does not repeat the dense matrix multiplication: the approximately 100 MiB matrix is retained in the source archive rather than this compact submission package. A separate local run performed that multiplication with the hash-verified matrix. No audit here reruns the backbone or turns the original numerical gate into a pass.

Readout	Suffix	Pair ranking [95% CI]	Pooled AUC [95% CI]
+Feedback probe	None	0.984 [0.965, 1.000]	0.949 [0.934, 0.967]
	Other-file TEST	0.531 [0.449, 0.625]	0.539 [0.505, 0.575]
	Eight other-file events	0.766 [0.684, 0.836]	0.713 [0.665, 0.760]
Native score	None	0.996 [0.988, 1.000]	0.998 [0.996, 1.000]
	Other-file TEST	0.990 [0.977, 1.000]	0.922 [0.907, 0.937]
	Eight other-file events	0.980 [0.965, 0.992]	0.963 [0.949, 0.977]
Direct lens (FP32)	None	0.883 [0.836, 0.926]	0.832 [0.795, 0.873]
	Other-file TEST	0.563 [0.500, 0.621]	0.566 [0.529, 0.600]
	Eight other-file events	0.699 [0.645, 0.750]	0.637 [0.609, 0.670]
Direct lens (BF16 rounding)	None	0.873 [0.820, 0.922]	0.832 [0.794, 0.873]
	Other-file TEST	0.563 [0.496, 0.625]	0.566 [0.528, 0.601]
	Eight other-file events	0.689 [0.635, 0.742]	0.636 [0.608, 0.669]
Empirical J-lens (FP32)	None	0.977 [0.957, 0.992]	0.914 [0.898, 0.935]
	Other-file TEST	0.844 [0.793, 0.891]	0.727 [0.689, 0.772]
	Eight other-file events	0.934 [0.887, 0.973]	0.757 [0.724, 0.798]
Empirical J-lens (BF16 rounding)	None	0.977 [0.957, 0.992]	0.914 [0.898, 0.934]
	Other-file TEST	0.840 [0.781, 0.891]	0.726 [0.689, 0.772]
	Eight other-file events	0.941 [0.895, 0.977]	0.757 [0.724, 0.798]

Table 33: Same Qwen L47 final-position activations, different readouts on RESTORE controls. Each row has 512 histories and 256 matched pairs. Native scores are a later-computation comparator, not another readout at L47. The J-lens matrix is numerically unqualified and retained only as an empirical control.

Readout	Suffix	Pair ranking [95% CI]	Pooled AUC [95% CI]
+Feedback probe	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Other-file TEST	1.000 [1.000, 1.000]	0.997 [0.994, 1.000]
	Eight other-file events	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
Native score	None	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Other-file TEST	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Eight other-file events	1.000 [1.000, 1.000]	1.000 [0.999, 1.000]
Direct lens (FP32)	None	0.492 [0.430, 0.555]	0.525 [0.486, 0.568]
	Other-file TEST	0.582 [0.535, 0.637]	0.607 [0.573, 0.647]
	Eight other-file events	0.836 [0.770, 0.898]	0.728 [0.695, 0.763]
Direct lens (BF16 rounding)	None	0.496 [0.435, 0.563]	0.526 [0.486, 0.569]
	Other-file TEST	0.584 [0.539, 0.637]	0.607 [0.573, 0.647]
	Eight other-file events	0.832 [0.768, 0.893]	0.728 [0.695, 0.763]
Empirical J-lens (FP32)	None	0.996 [0.988, 1.000]	0.954 [0.942, 0.968]
	Other-file TEST	0.977 [0.957, 0.992]	0.899 [0.880, 0.920]
	Eight other-file events	1.000 [1.000, 1.000]	0.958 [0.945, 0.971]
Empirical J-lens (BF16 rounding)	None	0.996 [0.988, 1.000]	0.953 [0.941, 0.968]
	Other-file TEST	0.977 [0.957, 0.992]	0.899 [0.879, 0.921]
	Eight other-file events	1.000 [1.000, 1.000]	0.958 [0.945, 0.971]

Table 34: All same-site readouts on same-bytes-WRITE controls, using the same inputs, aggregation, and arithmetic conventions as Table 33.

Readout	Pair-ranking change [95% CI]	Pooled-AUC change [95% CI]
+Feedback probe	-0.453 [-0.535, -0.363]	-0.410 [-0.445, -0.376]
Native score	-0.006 [-0.021, 0.008]	-0.076 [-0.091, -0.062]
Direct lens (FP32)	-0.320 [-0.363, -0.277]	-0.266 [-0.312, -0.227]
Direct lens (BF16 rounding)	-0.311 [-0.357, -0.264]	-0.266 [-0.312, -0.227]
Empirical J-lens (FP32)	-0.133 [-0.176, -0.094]	-0.188 [-0.222, -0.149]
Empirical J-lens (BF16 rounding)	-0.137 [-0.184, -0.094]	-0.188 [-0.223, -0.148]

Table 35: Paired RESTORE changes after one other-file TEST, relative to no suffix. All readouts use shared family draws. These are post-hoc contrasts, separate from the original Holm-adjusted accuracy tests.

O FROZEN RANKING CONFIRMATION AND SIX PRESELECTED READOUT SITES

What is new, and what is held fixed. This follow-up is motivated by the earlier post-hoc decomposition but fixes its sample, ranking endpoints, and six sites before any new model forward. It uses Qwen3.8-27B at the same exact revision, native BF16, eager attention, and batch size one on a single PRO 6000 96 GB. There is no quantization, model offload, or token generation. The 32 new families have fresh file rosters and byte contents; they do not introduce new natural-language templates. Texts and filenames are disjoint from all earlier training, validation, decision, and crossed-history inputs. The old feedback head, scaler, and low-false-denial threshold supply the primary confirmation; none is refitted.

Construction and execution. The design crosses 32 families, four queried files, four revision assignments, two labels, and two suffixes: 2,048 histories. All files first receive WRITE r0. For distinct revisions a, b, the covered core is `WRITE a; TEST; WRITE b; RESTORE a`; the uncovered core is `WRITE a; WRITE b; TEST; RESTORE a`. The revision assignments are $(r0, r1)$, $(r1, r0)$, $(r0, r2)$, and $(r2, r0)$. The suffix is absent or one successful TEST of the next roster file. It never changes the target's current bytes or receipts. All 2,048 histories pass independent exact-byte execution and isolated TEST checks. There are 512 covered/uncovered pairs per suffix, 512 paired contrasts between their ranking outcomes, and 1,024 same-label history matches across suffixes.

Exactly two primary tests. Within each family, pair ranking averages all four slots and four revision assignments, with half credit for an exact tie. We then average the 32 family means. Endpoint 1 is the old probe's suffix-minus-no-suffix ranking change, predicted negative. Endpoint 2 is native-minus-old-probe ranking under the suffix, predicted positive. Each receives a two-sided family-level sign-flip Monte Carlo test with 100,000 draws and plus-one correction; Holm adjustment spans these two tests only. Confirmation requires both adjusted $p < 0.05$ and both predicted directions. Shared percentile intervals use 5,000 whole-family bootstrap draws. Numeric seed identifiers are 2026092311 for materials, 2026092312 for sign flips, and 2026092313 for bootstrap; they are arbitrary fixed integers, not claims about future execution dates.

Primary results. The frozen probe's ranking change is -0.55664 [-0.60352, -0.50977]. Native-minus-probe ranking under the suffix is 0.56055 [0.51367, 0.60547]. Both raw Monte Carlo p values equal $1/100001$; both Holm-adjusted values equal $2/100001$. These are finite Monte Carlo resolutions, not exact tail probabilities of zero. Both declared confirmation conditions pass. The native score nevertheless retains only 53.91% accuracy at its old threshold, with 472 false positives out of 512 uncovered inputs. Thus high native pair ranking remains distinct from a successful action rule.

Readout	Suffix	Pair ranking [95% CI]	Pooled AUC [95% CI]	Acc. (%)
Frozen probe	None	0.984 [0.975, 0.994]	0.966 [0.953, 0.977]	88.87
Frozen probe	Other TEST	0.428 [0.381, 0.475]	0.477 [0.450, 0.501]	49.80
Native score	None	1.000 [1.000, 1.000]	1.000 [0.999, 1.000]	67.87
Native score	Other TEST	0.988 [0.979, 0.996]	0.900 [0.886, 0.914]	53.91
Direct / FP32	None	0.842 [0.795, 0.885]	0.810 [0.779, 0.841]	–
Direct / FP32	Other TEST	0.537 [0.508, 0.566]	0.569 [0.552, 0.586]	–
Direct / BF16	None	0.846 [0.800, 0.889]	0.810 [0.779, 0.841]	–
Direct / BF16	Other TEST	0.537 [0.509, 0.566]	0.569 [0.552, 0.586]	–
J-lens / FP32	None	0.969 [0.951, 0.984]	0.899 [0.878, 0.921]	–
J-lens / FP32	Other TEST	0.818 [0.779, 0.854]	0.717 [0.689, 0.746]	–
J-lens / BF16	None	0.969 [0.951, 0.984]	0.900 [0.877, 0.921]	–
J-lens / BF16	Other TEST	0.814 [0.775, 0.850]	0.716 [0.689, 0.745]	–

Table 36: All fixed confirmation readouts, 1,024 histories per row, with 512 positives and 512 negatives. Pooled AUC combines observations before scoring. Lens controls have no newly fitted threshold; FP32 scoring is primary for those controls and BF16-rounded scoring is a sensitivity calculation. The J matrix remains an empirical, numerically unqualified control.

Six sites fixed before capture. We capture outputs of zero-based blocks 31, 47, and 63, all full-attention blocks in the pinned hybrid architecture. L63 is the last of its 64 blocks, directly upstream of final normalization and unembedding. Each block is read at the final input token and the final token overlapping the queried file's last event line. The latter is found from the renderer's character span and tokenizer offsets, never scores. We recapture all 14,336 original F training views and all 2,048 original initial validation views at these sites. Each head gets a site-specific scaler fitted to the 4,096 initial training views, followed by logistic regression with $C = 1$, lbfgs, and maximum 2,000 iterations. Training weights total 2,048 per class. Both action thresholds are chosen on the original validation set. All six heads are frozen before loading test scores, and all six cells are reported. The new L47 head is a separate recapture/refit control, not a replacement for the old primary head.

Site	Suffix	Pair ranking [95% CI]	AUC	Low-denial acc.	Balanced acc.
L31 / final	None	0.129 [0.102, 0.158]	0.279	35.16	43.16
L31 / final	Other TEST	0.242 [0.209, 0.275]	0.348	46.97	40.04
L31 / event	None	1.000 [1.000, 1.000]	0.718	52.64	72.07
L31 / event	Other TEST	1.000 [1.000, 1.000]	0.718	52.64	71.48
L47 / final	None	0.988 [0.979, 0.996]	0.977	93.07	92.48
L47 / final	Other TEST	0.469 [0.412, 0.529]	0.480	47.56	46.97
L47 / event	None	0.563 [0.547, 0.580]	0.527	50.00	51.95
L47 / event	Other TEST	0.568 [0.555, 0.584]	0.527	50.00	51.86
L63 / final	None	1.000 [1.000, 1.000]	0.995	89.75	88.67
L63 / final	Other TEST	0.912 [0.881, 0.939]	0.784	63.87	62.89
L63 / event	None	0.695 [0.666, 0.725]	0.556	50.39	60.25
L63 / event	Other TEST	0.693 [0.664, 0.723]	0.553	50.00	59.77

Table 37: Every preselected site, with accuracy in percent and ranking intervals clustered by the 32 new families. These are secondary comparisons, not a selection of the highest-scoring site. Full pooled-AUC intervals and paired ranking changes are in the machine-readable supplement. Event positions precede the later suffix and query.

Reading the location comparison. The final-position L47 refit loses ranking under the suffix, much like the old frozen head. The final-block L63 refit retains 0.912 ranking, but its frozen low-denial threshold is only 63.87% accurate: 312 false positives and 58 false negatives. L31 final-position ranking remains below chance. No score inversion or test-set threshold selection is performed. The empirical J-lens retains 0.818 ranking at L47, so the L63 result does not establish that the information first appears after L47. The matrix is the same native-BF16 empirical matrix used in Appendix N, not the distinct qualified FP32-tail reference; its numerical FAIL remains unchanged. The exploratory centroid control in Appendix P further narrows the same-site interpretation.

Earlier positions and finite-precision nonidentity. The target-event controls precede the suffix and query; their position is selected using the evaluator's queried-file identity. The paired token prefixes through the anchor are identical in all 1,024 matches. This is a different observation boundary, not an autonomous target locator. Although the architecture is causal, full-sequence BF16 captures are not byte-identical across sequence lengths: zero pairs have identical vectors at any site. At the earlier positions, median relative L_2 differences are 0.00835, 0.01217, and 0.01797 for L31, L47, and L63; maxima are 0.01656, 0.03397, and 0.04640. We report this numerical nonidentity rather than claiming exact suffix invariance or backward information flow. Its numerical source is not isolated here. The two-process replay gate compares the same complete inputs and does pass exactly; it does not promise cross-length identity. In particular, L31/event's 1.000 ranking does not demonstrate suffix robustness at the final-input boundary.

Identity, replay, and compact audit. All 32 snapshot files (55,586,114,863 bytes) were rehashed before capture. Runtime, source, chat template, fixed token IDs, protocol, package, and inputs are bound by hashes. Two fresh processes reproduce all-site residuals and full final-logit hashes exactly on the same 16 score-blind witnesses; the main capture reproduces them again. Native BF16 vectors are saved losslessly as 16-bit storage and bound by hashes of their decoded float32 values. The full evidence archive contains 18,496 verified files and is 928,131,152 bytes; all members were verified against its manifest.

The compact submission contains all new confirmation texts, token records, per-input scores, six fitted heads, fit summaries, and a privacy-minimized protocol copy with the original hash. The CPU entry point independently reconstructs labels and prefix matches, checks head hashes, and recomputes all 24 ranking/AUC cells, operating-point counts, paired changes, and both primary tests. The full new residual arrays and training captures remain in the hash-bound source archive; the compact audit does not refit these heads or rerun the backbone. Separately, local validation replays all seven probe projections and the direct readout from the archived new residuals. The previous appendix's direct-lens residual audit remains part of the same submitted package.

Site	Validation AUC	Low-denial threshold	Balanced threshold	Iterations
L31 / final	0.849584	-7.8370	0.2158	242
L31 / event	1.000000	6.6917	0.3385	9
L47 / final	0.995885	0.1085	-0.5787	48
L47 / event	1.000000	7.4369	0.7590	11
L63 / final	0.995684	-0.4208	-0.7871	66
L63 / event	0.999999	8.7496	-0.5082	16

Table 38: Validation-only selection and convergence of all six new heads. The low-false-denial cap is conditional on covered validation cases, not a false-publication cap.

Readout/site	Suffix	Threshold	TP	TN	FP	FN
Frozen probe	None	0.7562	409	501	11	103
Frozen probe	Other TEST	0.7562	195	315	197	317
Native score	None	-8.1250	512	183	329	0
Native score	Other TEST	-8.1250	512	40	472	0
L31 / final	None	-7.8370	161	199	313	351
L31 / final	Other TEST	-7.8370	468	13	499	44
L31 / event	None	6.6917	256	283	229	256
L31 / event	Other TEST	6.6917	256	283	229	256
L47 / final	None	0.1085	466	487	25	46
L47 / final	Other TEST	0.1085	292	195	317	220
L47 / event	None	7.4369	256	256	256	256
L47 / event	Other TEST	7.4369	256	256	256	256
L63 / final	None	-0.4208	512	407	105	0
L63 / final	Other TEST	-0.4208	454	200	312	58
L63 / event	None	8.7496	236	280	232	276
L63 / event	Other TEST	8.7496	233	279	233	279

Table 39: Frozen low-false-denial operating points for both primary readouts and all six site refits. Each row has 512 positive and 512 negative histories. Balanced-rule errors are also retained in the supplement.

P Complete Readout Inventory and Exploratory Centroid Follow-Up

Scope and chronology. The complete inventory below reprints existing crossed-history results. The centroid follow-up was planned after reading the confirmation results and reviewer feedback. It adds no GPU forwards, sample selection, fitted heads, thresholds, or hypothesis tests. It is not a third confirmatory endpoint. All three already-frozen centroid heads are scored, rather than choosing the best head on the new families. Decimal half-up rounding is used for the new inventory and confirmation tables; exact machine-readable values are retained.

Frozen scoring, including scale and offset. For each of the 2,048 confirmation residuals at L47/final, we apply the original initial-only scaler in float32 and the archived float64 centroid direction and center (Appendix C). Thus $g(h) = (z(h) - c)^\top u$ retains its original offset and scale; it is neither unit-normalized nor recalibrated. Initial-only, +Feedback, and +History-controls use their respective original training data and both original validation thresholds. The same 5,000 whole-family bootstrap draws used in the confirmation give descriptive ranking intervals and paired suffix changes. Pooled AUC combines all assignments before scoring.

				Low false-denial		Balanced	
Centroid training	Suffix	Pair ranking [95% CI]	AUC	Acc.	FP/FN	Acc.	FP/FN
Initial-only	None	1.000 [1.000, 1.000]	1.000	93.8	64 / 0	98.9	1 / 10
Initial-only	Other TEST	0.996 [0.988, 1.000]	0.951	77.1	217 / 17	76.7	2 / 237
+Feedback	None	0.969 [0.951, 0.984]	0.823	60.4	381 / 24	55.6	0 / 455
+Feedback	Other TEST	0.711 [0.664, 0.762]	0.577	50.4	500 / 8	54.5	66 / 400
+History controls	None	0.998 [0.994, 1.000]	0.957	89.2	53 / 58	80.5	3 / 197
+History controls	Other TEST	0.914 [0.893, 0.936]	0.705	58.7	369 / 54	63.6	149 / 224

Table 40: Exploratory frozen-centroid results on 32 new file-roster families. Each row has 1,024 histories, including 512 negatives and 512 positives. Accuracy is in percent; pair-ranking intervals are pointwise, while AUC and operating-point columns are point estimates. No operating point was chosen using these results.

Information is present; the old decision rule can still fail. The Initial-only centroid retains 0.996 pair ranking under the suffix, with a change of -0.004 $[-0.012, 0.000]$. Its pooled AUC falls from approximately 1.000 to 0.951, and low-denial accuracy from 93.8% to 77.1%. The balanced rule changes the error mix rather than resolving the transfer problem: it produces only 2 false positives but 237 false negatives, versus 217 and 17 under the low-denial rule. +Feedback and +History-controls centroid ranking changes are -0.258 $[-0.309, -0.205]$ and -0.084 $[-0.105, -0.063]$. The +Feedback centroid is less transferable than the Initial-only centroid under this suffix, and its low-denial accuracy is already 60.4% without a suffix. The training expansion that repaired the logistic workflow does not uniformly improve centroid transfer. These exploratory results qualify any claim that all readouts at L47 lose the distinction; they do not show that native computation uses the centroid direction.

The non-neural reference. We independently replay the visible WRITE/TEST/RESTORE events as a deterministic interpreter, tracking current revisions and successful-test records per file. It agrees with all 3,072 execution-checked discovery labels and all 2,048 confirmation labels, giving static accuracy and pair ranking 1.000 with zero FP/FN. Revision identity denotes fixed bytes within these histories. This is an explicit-state reference, not a new set of live policy executions and not evidence that natural deployment logs are always complete.

Numerical comparisons have different meanings. The event-site changes in Table 41 coexist with the cross-length BF16 vector differences described in Appendix O. Their numerical source is not isolated, and their small ranking changes are not a global noise bound for the final-position probe. The latter compares genuinely different visible inputs. The supplement therefore separates `CAUSAL_PREFIX_NUMERICS.json` from `FINAL_INPUT_DIFFERENCES.json`, while preserving the original combined object unchanged. Same-complete-input fresh-process replay remains an independent exact check.

Readout	Pair-ranking change [95% CI]
Native score	-0.012 [-0.021, -0.004]
L31 / event	0.000 [0.000, 0.000]
L47 / event	0.006 [-0.002, 0.014]
L63 / event	-0.002 [-0.016, 0.010]

Table 41: Other-file TEST minus no suffix in the original confirmation scores. Native ranking changes slightly despite remaining high. Event-site changes compare the same causal token prefix at different full-input lengths; these observations do not bound numerical error at other sites or on different inputs.

Recomputation and provenance. The local follow-up checks every archived residual's file and decoded-value hash before applying all three heads. The compact capsule carries all resulting scores and eight preselected confirmation residual witnesses per head; it reprojects those witnesses and recomputes all six centroid cells, the 84-cell inventory, and both ledger checks. It does not reproject all new residuals or rerun the backbone. Its manifest binds the exact included scores, token records, redacted design and protocol, original design/protocol hashes, heads, and analysis scripts. Hashes establish identity and local continuity, not an externally attested freeze timestamp. Original primary-test files are unchanged.

All crossed-history cells. Tables 42–45 enumerate 2 models $\times 2$ cores $\times 3$ suffixes $\times 7$ readouts: 84 overlapping cells, not 84 independent experiments. Both threshold rules share the same ranking and AUC; the two operating points give 168 rule-specific records. Every row pools four revision assignments into 512 histories, with 256 examples per label, 256 pairs, and 16 families. The printed values are point estimates; family-bootstrap intervals for every ranking, AUC, and accuracy are in the exact expected-output JSON. This inventory includes the strong Qwen Initial-only centroid result as well as the feedback heads highlighted in the main text.

Across the 84 overlapping cells, 74 have pair ranking at least 0.95; 28 of those have frozen low-false-denial accuracy below 90%, and 30 have at least 25 false positives out of 256 uncovered histories. These review-motivated cutoffs are descriptive, not additional tests; the cells share histories and families.

Readout	Suffix	Pair rank	AUC	Low false-denial Acc.	Low false-denial FP/FN	Balanced Acc.	Balanced FP/FN
Native	None	1.000	1.000	85.0	77 / 0	99.8	1 / 0
Native	Other TEST	1.000	1.000	65.8	175 / 0	99.6	2 / 0
Native	8 events	1.000	1.000	82.4	90 / 0	94.5	0 / 28
Logistic: Initial	None	1.000	1.000	90.2	0 / 50	93.0	0 / 36
Logistic: Initial	Other TEST	1.000	0.999	70.5	0 / 151	74.4	0 / 131
Logistic: Initial	8 events	1.000	1.000	96.7	0 / 17	97.3	0 / 14
Logistic: +Feedback	None	1.000	1.000	92.4	0 / 39	93.6	0 / 33
Logistic: +Feedback	Other TEST	1.000	0.997	83.8	0 / 83	86.3	0 / 70
Logistic: +Feedback	8 events	1.000	1.000	99.8	1 / 0	99.6	2 / 0
Logistic: +History controls	None	1.000	1.000	87.3	0 / 65	89.6	0 / 53
Logistic: +History controls	Other TEST	1.000	0.994	86.1	0 / 71	88.5	0 / 59
Logistic: +History controls	8 events	1.000	1.000	99.4	0 / 3	99.8	0 / 1
Centroid: Initial	None	1.000	1.000	97.7	12 / 0	100.0	0 / 0
Centroid: Initial	Other TEST	1.000	1.000	96.1	20 / 0	99.4	0 / 3
Centroid: Initial	8 events	1.000	1.000	99.6	2 / 0	99.4	0 / 3
Centroid: +Feedback	None	1.000	0.997	96.9	15 / 1	55.7	0 / 227
Centroid: +Feedback	Other TEST	1.000	0.974	68.6	161 / 0	73.8	0 / 134
Centroid: +Feedback	8 events	1.000	0.937	55.1	230 / 0	77.7	9 / 105
Centroid: +History controls	None	1.000	1.000	98.6	0 / 7	89.8	0 / 52
Centroid: +History controls	Other TEST	1.000	0.997	94.7	27 / 0	95.9	4 / 17
Centroid: +History controls	8 events	1.000	0.992	88.7	55 / 3	95.5	6 / 17

Table 42: Complete Qwen same-bytes-WRITE inventory. Accuracies are percentages at the two original validation thresholds.

Readout	Suffix	Pair rank	AUC	Low false-denial Acc.	Low false-denial FP/FN	Balanced Acc.	Balanced FP/FN
Native	None	0.996	0.998	67.2	168 / 0	98.6	0 / 7
Native	Other TEST	0.990	0.922	51.4	249 / 0	83.6	13 / 71
Native	8 events	0.980	0.963	63.3	188 / 0	67.4	0 / 167
Logistic: Initial	None	0.984	0.954	87.1	2 / 64	88.5	5 / 54
Logistic: Initial	Other TEST	0.570	0.572	52.9	20 / 221	54.9	26 / 205
Logistic: Initial	8 events	0.797	0.763	64.8	30 / 150	67.4	44 / 123
Logistic: +Feedback	None	0.984	0.949	88.1	3 / 58	89.1	7 / 49
Logistic: +Feedback	Other TEST	0.531	0.539	52.5	81 / 162	53.7	96 / 141
Logistic: +Feedback	8 events	0.766	0.713	62.9	139 / 51	60.5	160 / 42
Logistic: +History controls	None	0.973	0.949	87.9	19 / 43	87.7	24 / 39
Logistic: +History controls	Other TEST	0.551	0.544	52.5	135 / 108	52.1	150 / 95
Logistic: +History controls	8 events	0.730	0.660	60.7	160 / 41	60.0	170 / 35
Centroid: Initial	None	0.996	1.000	95.9	20 / 1	98.2	0 / 9
Centroid: Initial	Other TEST	0.996	0.968	81.3	88 / 8	72.5	0 / 141
Centroid: Initial	8 events	0.988	0.990	94.7	10 / 17	57.4	0 / 218
Centroid: +Feedback	None	0.988	0.825	58.0	207 / 8	57.8	0 / 216
Centroid: +Feedback	Other TEST	0.730	0.622	50.2	255 / 0	57.4	37 / 181
Centroid: +Feedback	8 events	0.715	0.605	50.4	254 / 0	58.4	77 / 136
Centroid: +History controls	None	0.996	0.953	88.7	36 / 22	82.4	0 / 90
Centroid: +History controls	Other TEST	0.930	0.755	57.4	203 / 15	69.3	80 / 77
Centroid: +History controls	8 events	0.902	0.763	59.0	200 / 10	69.7	82 / 73

Table 43: Complete Qwen RESTORE inventory. The Initial-only centroid preserves high ranking across suffixes; the frozen projection's failure is not shared by every same-site head.

Readout	Suffix	Pair rank	AUC	Low false-denial Acc.	Low false-denial FP/FN	Balanced Acc.	Balanced FP/FN
Native	None	1.000	1.000	86.7	68 / 0	95.1	25 / 0
Native	Other TEST	0.996	1.000	79.3	106 / 0	93.6	33 / 0
Native	8 events	1.000	1.000	81.3	96 / 0	95.5	23 / 0
Logistic: Initial	None	1.000	1.000	87.5	64 / 0	80.7	99 / 0
Logistic: Initial	Other TEST	1.000	1.000	99.8	1 / 0	93.9	31 / 0
Logistic: Initial	8 events	1.000	1.000	99.8	1 / 0	97.5	13 / 0
Logistic: +Feedback	None	1.000	1.000	83.0	87 / 0	79.1	107 / 0
Logistic: +Feedback	Other TEST	1.000	1.000	96.5	18 / 0	91.2	45 / 0
Logistic: +Feedback	8 events	1.000	1.000	99.8	1 / 0	96.7	17 / 0
Logistic: +History controls	None	1.000	1.000	85.2	76 / 0	78.7	109 / 0
Logistic: +History controls	Other TEST	1.000	1.000	98.2	9 / 0	94.9	26 / 0
Logistic: +History controls	8 events	1.000	1.000	100.0	0 / 0	100.0	0 / 0
Centroid: Initial	None	1.000	1.000	91.8	42 / 0	99.0	5 / 0
Centroid: Initial	Other TEST	1.000	1.000	82.0	92 / 0	98.4	8 / 0
Centroid: Initial	8 events	1.000	1.000	88.7	58 / 0	99.6	2 / 0
Centroid: +Feedback	None	1.000	1.000	97.5	13 / 0	100.0	0 / 0
Centroid: +Feedback	Other TEST	1.000	1.000	93.9	31 / 0	98.2	0 / 9
Centroid: +Feedback	8 events	1.000	1.000	92.4	39 / 0	100.0	0 / 0
Centroid: +History controls	None	1.000	1.000	93.9	31 / 0	100.0	0 / 0
Centroid: +History controls	Other TEST	1.000	1.000	87.5	64 / 0	99.8	1 / 0
Centroid: +History controls	8 events	1.000	1.000	90.6	48 / 0	100.0	0 / 0

Table 44: Complete Gemma same-bytes-WRITE inventory. Ranking retention and threshold accuracy remain separate measurements.

Readout	Suffix	Pair rank	AUC	Low false-denial Acc.	Low false-denial FP/FN	Balanced Acc.	Balanced FP/FN
Native	None	1.000	1.000	92.4	39 / 0	99.8	1 / 0
Native	Other TEST	1.000	1.000	96.7	17 / 0	99.0	0 / 5
Native	8 events	1.000	1.000	93.4	34 / 0	99.8	1 / 0
Logistic: Initial	None	1.000	1.000	96.7	17 / 0	86.5	69 / 0
Logistic: Initial	Other TEST	1.000	0.999	96.1	0 / 20	98.6	4 / 3
Logistic: Initial	8 events	1.000	1.000	99.0	1 / 4	96.3	19 / 0
Logistic: +Feedback	None	1.000	1.000	95.9	21 / 0	88.3	60 / 0
Logistic: +Feedback	Other TEST	1.000	0.999	97.7	0 / 12	98.4	6 / 2
Logistic: +Feedback	8 events	1.000	1.000	98.6	3 / 4	97.1	15 / 0
Logistic: +History controls	None	1.000	1.000	94.7	27 / 0	85.4	75 / 0
Logistic: +History controls	Other TEST	1.000	0.999	98.4	0 / 8	98.0	7 / 3
Logistic: +History controls	8 events	1.000	0.999	97.7	2 / 10	98.6	7 / 0
Centroid: Initial	None	1.000	1.000	96.9	16 / 0	100.0	0 / 0
Centroid: Initial	Other TEST	1.000	1.000	97.9	11 / 0	99.0	0 / 5
Centroid: Initial	8 events	1.000	1.000	97.1	15 / 0	99.8	1 / 0
Centroid: +Feedback	None	1.000	1.000	99.4	3 / 0	100.0	0 / 0
Centroid: +Feedback	Other TEST	1.000	1.000	99.2	1 / 3	94.9	0 / 26
Centroid: +Feedback	8 events	1.000	1.000	99.2	4 / 0	99.8	0 / 1
Centroid: +History controls	None	1.000	1.000	98.2	9 / 0	100.0	0 / 0
Centroid: +History controls	Other TEST	1.000	1.000	98.6	4 / 3	98.0	0 / 10
Centroid: +History controls	8 events	1.000	1.000	98.2	9 / 0	100.0	0 / 0

Table 45: Complete Gemma RESTORE inventory. All rows use unchanged heads and original operating points.

Q EXECUTION IDENTITY AND EVIDENCE INVENTORY

The local source records for the transfer and original native-lens experiments were committed on 12 September 2026; the numerical-reference and reference-decision records on 13 September 2026. These are local provenance timestamps, not independent public registration certificates. Date-like seeds remain unchanged integers and are not these timestamps.

The model identifier is `Qwen/Qwen3.8-27B`. The frozen revision is

`1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`.

The snapshot contains 32 files totaling 55,586,114,863 bytes. The stored snapshot-manifest SHA-256 is

`b419d0bfff0100bbd1b2754fb53bb8e37a5e444025fc7daf0718e8d0f49957c9`.

The state-transfer source commit is `bfa28d5fc3a1dba67a913053991ac25345dc578c`; the original native-lens commit is `0679053ef92abc6ff90d13a9825f31fecad317f5`. The numerical-reference commit is `be2406b368224b6b1da7e1e8d10d787576864ffd`, and the reference-lens agent commit is `cc294d11973033c597b619782751aee523dd35f8`. These identities keep the failed original experiment separate from its successors. The model has 64 layers and residual width 5,120. Capture uses PyTorch 2.10 with ROCm 7.2.4 and Transformers 5.16.1. Analysis uses scikit-learn 1.6.1. Runtime records bind the tokenizer, chat template, dtype, device assignment, and no-thinking configuration.

The complete capture took 8,080.7 seconds and its full fresh replay took 8,093.2 seconds. The replay matches the residual and action-logit payloads exactly; it is not a second independent sample. The transfer archive is 1,095,501,034 bytes; the original native-lens failure archive is 2,924,193 bytes. The separate reference-lens custody archive is 592,683,689 bytes. All are locally retained and SHA-256 verified. The smaller result extracts link figures to these records without embedding machine paths in the PDF.

Stage	Observed status
Executed corpus and independent ledger replay	PASS
Full residual capture and fresh replay	Complete; exact-byte match
Frozen four-cell state-transfer analysis	Complete; bidirectional gate FAIL
Local-lens CPU preparation	PASS
Original native BF16 derivative qualification	FAIL on first direction
Original second process, fit, and five agent arms	NOT RUN
Separate FP32-tail reference qualification	PASS; two exact-replay processes
Reference mean dense fit	Complete; 160 chunks reconstruct matrix
Reference five-selector executed comparison	Complete; 512 cases, 2,560 episodes
Independent action and archive audit	PASS; 28,920 lifecycle records

Table 46: Qwen evidence accounting by experiment identity. NOT RUN in the original native BF16 protocol is not a zero-utility measurement. The reference executes its separate complete comparison. Appendix V records the later Gemma transfer replication.

The original downstream comparison was not executed after its prerequisite failed. The reference protocol froze its distinct arithmetic before fitting or agent evaluation, reused the previously unevaluated case/input package without changing its hashes, and then completed every arm. The 1,184 named parameter tensors in the fresh agent process are all native BF16 and match the original model identity. Complete phase archives and independent per-file backups agree. Saved decisions have been independently checked against the frozen selectors and thresholds, without new model calls.

R SECONDARY ENDPOINTS AND CONTROLS

Table 47 gives the queried condition and its paired difference. The paired unit remains a family; the queried observations are not a second independent corpus. No primary gate is redefined using these endpoints. All results in this appendix concern Qwen. The corresponding Gemma diagnostics appear in Appendix V.

Train to test	Queried AUC [95% CI]	Exact match	AUC change [95% CI]
Order to order	1.000 [1.000, 1.000]	1.000	+0.008 [0.003, 0.013]
Order to restore	0.793 [0.768, 0.820]	0.094	+0.455 [0.412, 0.499]
Restore to order	0.988 [0.978, 0.996]	0.387	+0.651 [0.624, 0.677]
Restore to restore	0.998 [0.997, 1.000]	0.715	+0.017 [0.009, 0.026]

Table 47: Secondary L47 target-queried results and queried-minus-unqueried differences. Heads are fitted separately by query slot and condition. Paired family bootstrap intervals do not identify whether the query retrieves, transforms, or computes the state.

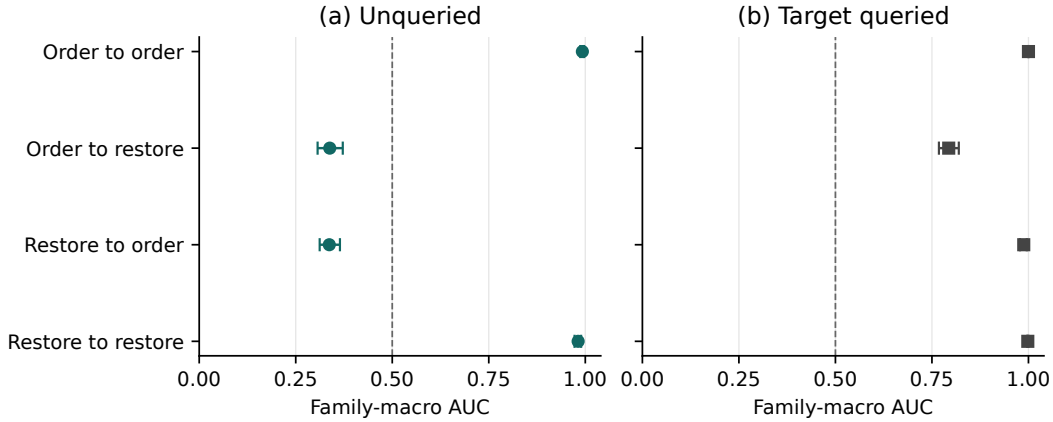


Figure 4: Preserved Qwen query comparison. Separately fitted heads rank states more accurately at target-queried prefixes than at the unqueried boundary. Error bars are family-cluster 95% intervals. Changed inputs and refitted heads prevent a causal extraction interpretation.

For the static native endpoint, the probability is normalized over the selected yes and no tokens. The reported slot decisions use the fixed threshold 0.5, without post-hoc calibration. This table measures forced-endpoint statistics, not the outcomes of the historical executed study in Appendix T.

Mechanism	AUC [95% CI]	Mean slot accuracy	Exact match	Brier
Order	0.997 [0.995, 0.999]	0.504	0.062	0.423
Restore	0.760 [0.745, 0.776]	0.505	0.062	0.468

Table 48: Native post-query two-token endpoint. Ranking and decision accuracy differ markedly.

The hidden-update controls hold the complete observation fixed while changing one external coverage bit. A saved-payload audit confirms identical input-token hashes for all 16 pairs, exact-byte residual equality at all five captured layers, and identical native yes/no logits. The maximum absolute residual difference is zero. All four stored heads from each training mechanism also tie at L47 on every pair. These controls are excluded from the ordinary matrix: they test the observation boundary, not average performance on an unobservable target. No hidden-pair J-lens measurement is inferred from this audit.

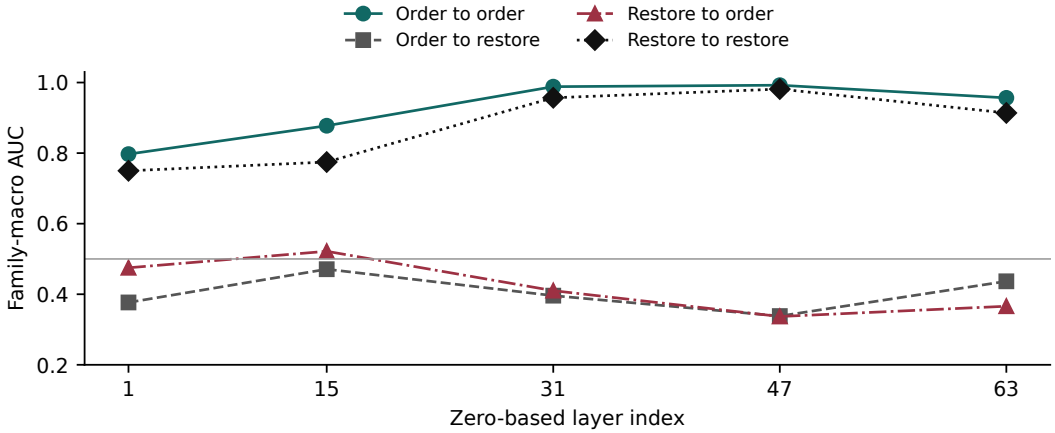


Figure 5: Secondary unqueried layer diagnostics. Marker shapes and line styles distinguish all four cells in grayscale. L47 remains the frozen primary endpoint; the plot is not a search for a replacement gate.

Layer	O→O	O→R	R→O	R→R
1	0.797	0.377	0.475	0.750
15	0.878	0.471	0.522	0.775
31	0.988	0.396	0.410	0.957
47	0.992	0.338	0.337	0.981
63	0.957	0.437	0.366	0.914

Table 49: All captured unqueried layer AUCs, including the primary L47 row.

Train to test	Median	Central 95% across projections	Maximum
Order to order	0.559	[0.498, 0.632]	0.639
Order to restore	0.505	[0.416, 0.584]	0.603
Restore to order	0.492	[0.430, 0.574]	0.603
Restore to restore	0.563	[0.508, 0.649]	0.663

Table 50: Secondary random-projection reference distribution. Each row summarizes 100 directions with validation-only sign and threshold selection. The central range is not a confidence interval on the population of tasks.

S HISTORICAL NUMERICAL QUALIFICATION AND REPRODUCTION DETAILS

Let F map the injected last-position post-block residual to the final pre-normalization residual, with the rest of the model input fixed. Qualification chooses 16 evenly spaced output coordinates I and computes the corresponding autograd rows J_I. For a seeded float32 unit direction v and scale α, it compares

$$a = J_I v, \qquad b_\alpha = \frac{F(h + \epsilon v)_I - F(h - \epsilon v)_I}{2\epsilon}, \qquad \epsilon = \alpha \|h\|_2. \tag{5}$$

The comparison uses $\alpha \in \{0.01, 0.03, 0.10\}$. Its two criteria are

$$\frac{a^\top b_\alpha}{\|a\|_2 \|b_\alpha\|_2} \geq 0.85, \qquad \frac{\|b_\alpha - a\|_2}{\max(\|a\|_2, 10^{-12})} \leq 0.35. \tag{6}$$

At least two scales must satisfy both criteria for each direction. Two directions are planned for each of the first two neutral prompts, with seed 20260914 and a per-prompt offset. The native forward computation, including BF16 rounding of injected states, is retained. Storing the derivative rows in float32 does not make this an FP32 forward model.

The four prompts used to average the reference matrix, with frozen rendered token identities, are: "Describe the appearance of a quiet public garden in three sentences."; "Explain how a pendulum moves without using equations."; "List the ingredients commonly used to prepare vegetable soup."; "Describe the differences between a triangle and a rectangle." The first numerical-qualification prompt, which is not one of those four, asks for a common desk item. The original artifacts include the 5,120-dimensional source and target residuals, two selected logits, and a $16 \times 5{,}120$ row matrix. Its first direction fails the scale-count rule (Table 51), so later checks in that original run are unobserved. The error and arrays are retained. Four analytical CPU tests check orientation, normalization placement, hooks, and finite differences on known functions. They do not override the failed runtime qualification.

Step / residual norm	Cosine	Relative error	Scale result
0.01	0.747	1.669	FAIL
0.03	0.929	0.423	FAIL
0.10	0.980	0.198	PASS

Table 51: Original native BF16 qualification, first direction. Only one of three scales passes; two are required. The FAIL is unchanged by later reference measurements.

Reference experiment. The separate numerical protocol captures native prefix contexts and promotes the downstream tail to FP32 while retaining the exact values loaded from the BF16 snapshot. It checks sixteen neutral prompts in two fresh processes. All 356 saved arrays per process agree byte for byte. Finite differences use two directions on each of the first two prompts. Table 52 summarizes those four directions in one process; the second process reproduces all twelve measurements exactly, yielding 24 recorded checks. Native-tail replay agrees with the checked native full-model outputs and derivative rows. This reduces a path-wiring explanation, but does not by itself isolate precision as the sole cause of the original discrepancy. In particular, the reference and native diagnostic directions use different frozen seed streams. The reference does not retroactively qualify the native derivative.

Step / residual norm	Minimum cosine	Maximum relative error	Passing directions
0.01	1.000000	0.000810	4/4
0.03	0.999999	0.001492	4/4
0.10	0.999883	0.016110	4/4

Table 52: Separate FP32-tail reference, extrema over four checked directions per scale in one process. The same cosine/error criteria are used. These checks concern local numerical sensitivity, not semantic or agent accuracy.

The reproduction sequence preserves each protocol's scope: construct and audit the executed corpus; capture and replay; analyze the frozen transfer matrix; preserve the failed native numerical

prerequisite; qualify the separately frozen reference; fit that reference; and run its fresh native BF16 decision process. Hash-checked closed evidence is copied before advancing. Numerical thresholds, case membership, and action thresholds are not revised using outcomes.

T Reference Readout and Executed Decision Details

Readout and calibration. For the reference lens, the coverage score is the yes-minus-no logit contrast after $\bar{J}h$, one native final normalization, and original unembedding. The decision residual and text classifiers are separate from Study 1's four-head, source-mechanism-specific probes. Each pools both mechanisms and all four target-query slots, using the queried slot's executed bit as the binary label. There are 4,096 training rows from 32 families and 2,048 validation rows from 16 families. The residual classifier uses L47, a training-only StandardScaler, and logistic $C = 1$; text uses word 1–2-gram TF–IDF and $C = 10$. These values maximize pooled validation AUC over the frozen grid, with ties choosing smaller C. Neither classifier is selected using the decision test families.

The largest threshold satisfying the frozen 5% false-denial cap is 3.2421875 for J-lens, 1.47038565 for residual, and -1.55730406 for text. Its denominator contains only truly covered validation rows: predicting uncovered on these rows incurs an unnecessary TEST. It does not cap PUBLISH on uncovered states. Scores below threshold select TEST; all others select PUBLISH. Native shadow uses yes logit $>$ no logit with no validation calibration. Its later calibrated diagnostic instead uses yes-minus-no ≥ -8.125. Table 53 separates model-derived scores, supervised heads, and visible rules.

Selector	Input and score	Fitting and action boundary
Native shadow	Target-queried final yes/no logits	No fitted head; yes $>$ no means PUBLISH.
Reference J-lens	Target-queried L47; mean local derivative then yes/no contrast	Matrix from four neutral prompts; task-label validation threshold.
Residual	Target-queried L47; one binary classifier	Both mechanisms and all query slots; 4,096 training rows; validation threshold.
Text	Same target-queried visible history; one binary TF–IDF classifier	Same training/validation rows as residual; no activation input.
Ledger	Same visible roster and events	Exact coverage interpreter; no fitted head or threshold.
Calibrated native	Same native contrast	Post-hoc validation threshold; exact-cache reenactment.

Table 53: Decision readout specifications. The first five are the original arms. Supervised decision heads pool mechanisms rather than choosing a mechanism-specific head at test time.

The static queried-boundary analysis (Table 54) reuses the earlier hash-bound capture, not the subsequent decision episodes. For each family it pools the four query slots before computing AUC, then averages families. This differs from the average of slot-specific AUCs in Equation 2 and the earlier static native table. Native scores here are recorded yes-minus-no logit contrasts. These comparisons do not revise the unqueried transfer gate. Higher static ranking than a text classifier does not ensure a better thresholded policy under a limited action budget.

Readout	Order AUC [95% CI]	Restore AUC [95% CI]
Reference J-lens	0.843 [0.824, 0.862]	0.705 [0.682, 0.730]
Direct logit lens	0.830 [0.798, 0.860]	0.699 [0.681, 0.716]
Native final two-token readout	0.997 [0.994, 0.999]	0.739 [0.726, 0.752]
Residual linear probe	1.000 [1.000, 1.000]	0.995 [0.991, 0.998]
Visible-text classifier	0.688 [0.688, 0.688]	0.625 [0.625, 0.625]

Table 54: Secondary queried-state ranking on the preserved capture. Each mechanism has 1,024 queried observations from 16 held-out families; intervals resample families. AUC pools query slots within a family and is a separate endpoint from executed completion.

Action and outcome definitions. Each case reconstructs its actual initial files and receipt set. The target slot is family index modulo four. After TEST, the current contents become covered and a TEST event is appended to the visible history. An untested PUBLISH is denied before dispatch, increments the invalid-proposal count, and returns the same public instruction to obtain a fresh TEST. A covered PUBLISH writes a publication receipt binding the case, target, and content hash, then completes the episode. It does not copy files to an external service or mutate a repository. Testing already-covered contents increments the unnecessary-test count. The three-decision budget includes denied actions. These definitions distinguish a proposal error, a safe rejection, and failure to complete.

Comparison	Completion change (pp)	Invalid-proposal change	Extra-test change
J minus shadow	+45.12 [40.23, 49.80]	+1.365 [1.260, 1.453]	-2.326 [-2.387, -2.258]
J minus text	-3.32 [-5.47, -1.37]	-0.002 [-0.096, 0.078]	+0.043 [0.000, 0.094]
J minus residual	-17.97 [-22.46, -13.28]	+1.354 [1.248, 1.441]	-0.596 [-0.686, -0.502]
J minus ledger	-46.68 [-48.63, -44.34]	+1.365 [1.260, 1.453]	+0.043 [0.000, 0.094]

Table 55: All frozen paired contrasts, with 95% family-bootstrap intervals. Count differences are per episode; completion differences are percentage points. No comparison is dropped based on its sign.

Mechanism	Selector	Completion (%)	Invalid proposals	Extra tests
Order	Native final-readout shadow	13.67	0.000	2.301
Order	Reference J-lens	53.12	1.406	0.000
Order	Residual linear probe	85.94	0.012	0.344
Order	Visible-text classifier	63.28	1.234	0.000
Order	Exact visible-history ledger	100.00	0.000	0.000
Restore	Native final-readout shadow	2.73	0.000	2.438
Restore	Reference J-lens	53.52	1.324	0.086
Restore	Residual linear probe	56.64	0.012	0.934
Restore	Visible-text classifier	50.00	1.500	0.000
Restore	Exact visible-history ledger	100.00	0.000	0.000

Table 56: Both executed mechanisms, 256 cases and 16 families each. Invalid proposals and extra tests are means per episode. Aggregate comparisons need not have the same sign within every mechanism. The complete result extract retains mechanism-specific paired intervals.

Cost accounting. Every decision computes every readout, even when only one selects the action. Table 57 gives the measured component times. Hence the exact-ledger arm also incurs the instrumented model forward; its measured episode time is not the cost of deploying a ledger alone. Fit cost is reported separately and is not amortized using an assumed deployment volume.

Component	Mean milliseconds per call
Native forward and capture	789.45264
Reference J-lens	4.50597
Residual linear probe	0.52604
Visible-text classifier	0.70088
Exact visible-history ledger	0.01445

Table 57: Recorded component times over 5,272 calls per component. These are not separately benchmarked end-to-end deployments. Reference fitting takes 5,599.32 s; native BF16 decision execution takes 4,574.21 s.

U REVIEWER-MOTIVATED POST-HOC DIAGNOSTICS

The following analyses were designed after observing the original results. Their settings were hash-frozen before the new computations, but they are post-hoc diagnostics rather than a new confirmatory gate. The original matrices, calibration rules, and five-arm table remain unchanged.

Head alignment. We express each stored logistic head in the original residual coordinates by dividing its coefficient vector by its training-scaler standard deviation, with the corresponding intercept correction. Table 58 compares same-slot order and restore heads. Their cosines are far from -1, so below-chance transfer is not explained by the two fitted coefficient vectors being global negatives of one another. Same-input probability correlations apply both heads to the same pooled held-out residual rows. These descriptive quantities identify neither semantic axes nor a causal mechanism.

Slot	Raw-coordinate weight cosine	Same-input probability correlation
0	-0.142	-0.220
1	-0.054	0.350
2	-0.118	-0.253
3	-0.098	-0.437

Table 58: Stored L47 head alignment. Correlations pool both test mechanisms on identical input support for the two heads; they are not correlations between different inputs sharing a label.

Fixed nonlinear and mixed-training diagnostics. We fit a 32-unit ReLU MLP with four sigmoid outputs, Adam, regularization 0.001, batch size 64, initial learning rate 0.001, and at most 200 iterations. Stopping uses training loss only (tolerance 10^{-4}, 20 non-improving iterations); validation and test data do not select the architecture or a seed. All three integer seeds, 91531–91533, are reported. Each source-trained scaler and head is applied unchanged to both test mechanisms. Mixed training uses both mechanisms from the training families, not the held-out families; its linear heads select regularization using validation AUC equally averaged across the two mechanisms. Linear controls refitted on a single mechanism reproduce all four original point estimates.

Training	Test	Linear AUC [95% CI]	MLP AUC [95% CI]
order	order	0.992 [0.988, 0.997]	0.983 [0.976, 0.991]
order	restore	0.338 [0.306, 0.372]	0.320 [0.298, 0.344]
restore	order	0.337 [0.310, 0.362]	0.378 [0.356, 0.401]
restore	restore	0.981 [0.972, 0.990]	0.955 [0.935, 0.972]
mixed	order	0.978 [0.970, 0.986]	0.987 [0.980, 0.992]
mixed	restore	0.967 [0.955, 0.978]	0.958 [0.944, 0.971]

Table 59: Post-hoc L47 unqueried diagnostics using Equation 2. MLP estimates average the three seed-specific AUCs, not ensemble predictions. Intervals resample the same 16 test families jointly across seeds; seeds are not independent task samples. Mixed training evaluates new families of two known mechanisms, not a third, unseen mechanism.

The small MLP does not rescue cross-mechanism transfer, while mixed linear training gives high held-out ranking on both known mechanisms. This narrows a decoder-capacity explanation without excluding other nonlinear decoders, and motivates testing training coverage separately from state decodability.

Calibrated native selector. We select the largest native yes-minus-no threshold that satisfies the same 5% false-denial cap on the original 2,048 validation rows. The resulting threshold is -8.125, with validation false denial 4.004%. No decision-case labels enter threshold selection. All 512 original cases are reenacted in fresh local workspaces with the unchanged three-decision harness and common validator. A model output is reused only when its complete visible input matches an archived input exactly and duplicate input IDs and logits agree. Any missing prefix would make the policy comparison incomplete; none occurs.

MLP training	Seed	Order test AUC	Restore test AUC
order	91531	0.985	0.335
order	91532	0.979	0.315
order	91533	0.986	0.309
restore	91531	0.409	0.952
restore	91532	0.384	0.955
restore	91533	0.340	0.958
mixed	91531	0.990	0.969
mixed	91532	0.985	0.953
mixed	91533	0.985	0.952

Table 60: Every fixed MLP seed, without best-seed selection. A separate saved-weight audit recomputes predictions, AUCs, decision accuracies, and confidence intervals for all 36 model–seed–test cells.

The preserved archive contains 5,272 original decisions and 1,767 distinct prefixes. The new selector makes 970 decisions, all covered by that cache. An independent saved-action audit checks every input hash, thresholded action, authorization, lifecycle hash chain, outcome, and publication receipt. This is a cache-backed policy reenactment with new local TEST operations, not a fresh model capture or a new GPU latency measurement.

Calibrated native outcome	Estimate [95% CI]
Completion (%)	61.133 [57.227, 65.234]
Invalid proposals	1.166 [1.043, 1.283]
Extra tests	0.006 [0.000, 0.012]
Unsafe dispatches	0.000 [0.000, 0.000]
Completion minus shadow (pp)	52.930 [48.242, 57.227]
Completion minus jlens (pp)	7.812 [3.320, 12.305]
Completion minus text (pp)	4.492 [0.386, 8.403]
Completion minus residual (pp)	-10.156 [-15.625, -5.273]
Completion minus ledger (pp)	-38.867 [-42.773, -34.766]

Table 61: Post-hoc calibrated native selector, 512 cases in 16 families. Intervals resample complete families. Count metrics are per episode. Completion exceeds the original reference lens despite using the native readout, but invalid proposals remain a limitation and the validator still determines safety.

V GEMMA CROSS-FAMILY REPLICATION

Question and frozen scope. The Gemma design was fixed after the Qwen results but before Gemma capture or scientific evaluation. It reuses the executed corpus, 32/16/16 family split, logistic selection, 2,000 family-bootstrap draws, and 100 complete-vector null fits per training mechanism. Reusing observations adds no independent families. Heads are fitted separately by model; Holm correction covers each protocol's four cells. Table 62 complements the primary AUC intervals with every gate and thresholded metric.

Model	Train to test	Min. slot AUC	Exact match	Brier	Holm p	Gate
Qwen	Order to order	0.989	0.781	0.052	0.0396	PASS
Qwen	Order to restore	0.140	0.031	0.516	1.0000	FAIL
Qwen	Restore to order	0.244	0.062	0.518	1.0000	FAIL
Qwen	Restore to restore	0.962	0.602	0.107	0.0396	PASS
Gemma	Order to order	0.834	0.391	0.188	0.0396	PASS
Gemma	Order to restore	0.324	0.047	0.487	1.0000	FAIL
Gemma	Restore to order	0.331	0.059	0.485	1.0000	FAIL
Gemma	Restore to restore	0.812	0.402	0.190	0.0396	PASS

Table 62: All primary unqueried cells, 256 worlds from 16 families each. Exact match requires all four validation-thresholded predictions to be correct. Min. slot AUC and corrected p are the other numerical gate components; mean AUC and its interval appear in Table 5. No score is inverted after evaluation.

Model identity and architecture. The model is `google/gemma-4-31B-it`, revision

`842da3794eaa0b77d5f08bae87a17459d91ff475`.

The snapshot has 12 files totaling 62,578,686,256 bytes. We retain the full native conditional-generation model, including components unused by these text-only inputs. Its text configuration has 60 blocks, width 5,376, full attention at indices $5 \bmod 6$, sliding attention elsewhere with window 1,024, per-layer embedding input width zero, zero KV-shared layers, and no MoE blocks. Native K=V attention semantics and the final logit softcap of 30 are preserved (Google, 2026).

The layer mapping, fixed before capture, is

$$\ell_{\text{Gemma}} = \text{round}_{\text{half-up}}\left((\ell_{\text{Qwen}} + 1)\frac{60}{64} - 1\right). \tag{7}$$

Qwen L47 maps to Gemma L44, a sliding-attention block. Full-attention neighbors L41 and L47 are equally distant; the protocol chooses lower L41 as a secondary diagnostic. Captured indices are 1, 14, 29, 41, 44, and 59. This depth mapping does not claim block-type equivalence. Table 63 includes all sampled layers, not just the favorable endpoints. L41 cross-mechanism AUCs are 0.410 [0.376, 0.442] and 0.438 [0.395, 0.480], consistent with failure at primary L44 without replacing it.

Integrity and interpretation. Technical execution and independent audit pass; the bidirectional scientific transfer gate fails. The final closed analysis archive SHA-256 is

`0f3b296695627be9003dd9e6828ccc2a899c7c72cb4a2fcbf32064d7d12a4321`.

The manifest binds design, capture identity, signed predictions, heads, null records, and verified GPU archives. Transfer failure does not imply absent state information. This historical replication tested Gemma state transfer, not a Gemma J-lens or policy loop. The later Gemma coverage and interchange experiments in Sections 4 and 6, and the lens experiments in Appendix F, are separate successors, with their own captures and provenance. Mixed-training/MLP diagnostics of the original transfer heads remain Qwen-only.

Gemma unqueried readout	O→O	O→R	R→O	R→R
L1 (sliding)	0.750	0.521	0.395	0.599
L14 (sliding)	0.806	0.531	0.559	0.759
L29 (full)	0.943	0.429	0.440	0.928
L41 (full), neighbor	0.894	0.410	0.438	0.907
L44 (sliding), primary	0.883	0.398	0.448	0.892
L59 (full)	0.791	0.450	0.394	0.834
Unigram TF–IDF	0.500	0.500	0.500	0.500
Ordered 1–3-gram TF–IDF	0.696	0.375	0.383	0.625
Last relevant operation	1.000	0.500	1.000	0.500
Exact visible ledger	1.000	1.000	1.000	1.000

Table 63: All Gemma unqueried layer and visible-control AUCs. Only L44 supplies the confirmatory four-cell gate. Unigram counts tie exactly; ordered text is a shallow baseline, while the ledger recovers ordinary state exactly. Deterministic controls repeat across source columns only for alignment.

Native execution and replay. All parameters remain BF16 on one RTX PRO 6000 Blackwell 96GB GPU. PyTorch 2.8.0+cu128, CUDA 12.8, and Transformers 5.7.0 run eager attention at batch size one; residuals are stored in float32. Native rendering, token IDs, tokenizer/chat-template hashes, and no-thinking settings are bound, without truncation at the 2,048-token limit. No sampling, quantization, or offload is used. Two fresh 64-row technical passes precede two full 10,272-row passes. Capture/replay take 1,440.56/1,453.42 seconds, with peak allocated memory 62,747,952,128 bytes; these are not controlled ROCm speed comparisons.

Full replay matches every residual/action-logit byte and all 10,272 full-vocabulary logit hashes. All 16 hidden-update pairs also tie exactly; they are boundary controls, not extra ordinary worlds. CPU analysis uses scikit-learn 1.6.1. Independent closeout recomputes 50 endpoints from 12,800 prediction records, reproduces saved-head predictions with zero maximum discrepancy, and recomputes p values/Holm correction from saved nulls without refitting them.

Queried and native endpoints. Table 64 reports queried heads and paired changes; Table 65 reports native two-token probabilities. Strong order ranking contrasts with much lower restore joint accuracy. Neither table measures closed-loop completion.

Train to test	Queried AUC [95% CI]	Exact match	AUC change [95% CI]
Order to order	1.000 [1.000, 1.000]	0.980	+0.117 [0.095, 0.139]
Order to restore	0.760 [0.744, 0.776]	0.062	+0.362 [0.327, 0.397]
Restore to order	0.391 [0.360, 0.423]	0.121	-0.057 [-0.110, 0.000]
Restore to restore	1.000 [0.999, 1.000]	0.949	+0.107 [0.084, 0.133]

Table 64: Secondary Gemma L44 target-queried endpoints. Differences compare separately fitted queried and unqueried heads on paired families, not one head before and after a query. No queried endpoint changes the primary gate.

Mechanism	Native AUC [95% CI]	Slot accuracy	Exact match	Brier
Order	0.999 [0.999, 1.000]	0.984	0.941	0.015
Restore	0.788 [0.773, 0.801]	0.523	0.066	0.468

Table 65: Gemma native post-query two-token endpoint, threshold 0.5. All four slot predictions must be correct for joint exact match; these rates are not agent completion rates.

W Policy Recovery and Revision-Alias Diagnostics

These reviewer-motivated diagnostics were designed after the primary results. Their source and inputs were locally hash-frozen before execution. They preserve all original captures, thresholds, task definitions, and arm results. No GPU or model call was made.

Appendix Z separately tests model-generated actions and the fixed recovery rule in a later live check; its cases and evidence are not pooled with these historical results.

Executed elementary strategies and common recovery. Each of the original 512 cases is reconstructed in a fresh workspace for each of eight new arms. Two policies ignore model outputs: TEST then PUBLISH, and PUBLISH followed by TEST and PUBLISH only if denied. The remaining six keep the first choice of the original shadow, reference lens, residual, text, ledger, or calibrated native selector. They then use identical recovery: a successful TEST is followed by PUBLISH; a denied PUBLISH by TEST then PUBLISH. The unchanged harness raises an error if the test receipt does not match current bytes. No WRITE occurs during these decision episodes. The external validator still authorizes every publication; recovery does not bypass it.

The archive supplies 5,272 original decision records and 1,767 unique full-prefix hashes. All 512 initial prefixes are covered; duplicate inputs have identical readouts. Every recovered original-arm first choice matches its archived action. The calibrated native threshold remains -8.125, without refitting. Later recovery steps use public action consequences rather than additional model predictions. This is an executed controller change with cached initial scores, not a fresh GPU model run.

Initial rule + recovery	Invalid proposals [95% CI]	Extra TESTs [95% CI]	Actions [95% CI]
TEST-first	0.000 [0.000, 0.000]	0.500 [0.500, 0.500]	2.000 [2.000, 2.000]
PUBLISH-first	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	2.000 [2.000, 2.000]
Native shadow	0.000 [0.000, 0.000]	0.492 [0.482, 0.500]	1.992 [1.982, 2.000]
Reference J-lens	0.455 [0.422, 0.482]	0.020 [0.000, 0.043]	1.975 [1.955, 1.990]
Residual probe	0.004 [0.000, 0.010]	0.025 [0.012, 0.043]	1.529 [1.516, 1.545]
Visible text	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	2.000 [2.000, 2.000]
Calibrated native	0.389 [0.346, 0.426]	0.004 [0.000, 0.010]	1.893 [1.853, 1.930]
Exact ledger	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	1.500 [1.500, 1.500]

Table 66: Post-hoc common-recovery costs, 512 episodes per arm. Every arm completes 100% with zero unsafe dispatch. Intervals use 2,000 paired family-cluster bootstrap draws (seed 91561); the 16 families, not 4,096 episodes, are the resampling units.

With 256 initially covered targets, each unnecessary test is an initial false negative and each invalid proposal an initial false positive. Mean action count equals $1.5 + (\mathrm{FP} + \mathrm{FN})/512$ under this particular recovery rule. This identity checks observed logs; it is not an assumed monetary exchange rate. The residual probe's 15 initial errors incur 1.529 actions per episode, near the ledger's 1.500; the reference lens has 243 errors and incurs 1.975. Completion is saturated, so the evidence concerns initial-choice costs under a known recovery policy. It does not establish improvement over the exact ledger or unrestricted agent utility.

There are also 256 initially uncovered targets. The lens's 233 initial false positives therefore imply a conditional false-positive rate of $233/256 = 91.0\%$, whereas its 0.455 invalid proposals per episode uses all 512 cases as denominator. Both measures are correct; neither is the validation false-denial rate constrained during calibration.

Recovery minus original	Completion change (pp)	Invalid-proposal change	Extra-test change
Native shadow	+91.797 [87.891, 95.117]	+0.000 [0.000, 0.000]	-1.877 [-1.930, -1.814]
Reference J-lens	+46.680 [44.336, 48.633]	-0.910 [-0.965, -0.844]	-0.023 [-0.062, 0.000]
Residual probe	+28.711 [24.414, 33.008]	-0.008 [-0.020, 0.000]	-0.613 [-0.697, -0.529]
Visible text	+43.359 [41.797, 44.922]	-0.867 [-0.898, -0.836]	+0.000 [0.000, 0.000]
Calibrated native	+38.867 [34.570, 42.578]	-0.777 [-0.852, -0.691]	-0.002 [-0.006, 0.000]
Exact ledger	+0.000 [0.000, 0.000]	+0.000 [0.000, 0.000]	+0.000 [0.000, 0.000]

Table 67: Paired changes after retaining each initial choice and replacing only subsequent selection with common recovery. All entries include 95% family-cluster intervals. Calibrated native's original comparator is its prior post-hoc cache reenactment.

An independent audit verifies 4,096 publication receipts, 42,211 chained lifecycle records, all action transitions and source hashes, and recomputes reported costs. The elementary rules' theoretical

100% completion and 0.5 error-cost means agree with execution. Their degenerate intervals follow exact within-family balance and deterministic rules, not uncertainty-free generalization to other tasks. No case is excluded or cache entry imputed.

The available revision shortcut. We replay visible WRITE/RESTORE/TEST events for all 2,048 ordinary histories, recovering 8,192 slot labels that agree with executed receipts. Each file has exactly one historical TEST. Let z_i indicate that its tested revision is r1. In every audited order world $s_i = z_i$; in every restore world $s_i = 1 - z_i$. A rule with the sign learned in either source mechanism therefore has matrix $(1, 0; 0, 1)$ for within/cross-mechanism AUC. This construction-level explanation is compatible with, but not established by, the models' signed transfer results. Non-opposite fitted weight vectors and matched unigram counts do not exclude such a relation-sensitive shortcut.

We prepare both bijective display maps r0↦vA/r1↦vB and r0↦vB/r1↦vA for each original world, keeping r2↦vC and a fixed legend. The resulting 4,096 unqueried views preserve content, receipts, event order, labels, and family splits. Tested-alias and final-alias identities are exactly balanced by split, mechanism, state vector, and slot; unigram multisets match within each family/mechanism/view. The tested-alias rule becomes AUC 0.5 in all four cells. These are paired views of existing worlds, not independent new tasks. The subsequent model evaluation is reported in Appendix X; this construction audit alone is not a model result.

The fixed legend contains no reverse mapping to the original revision identities. The following is a verbatim excerpt from an actual balanced input, followed by the three selected event lines for slot 1 (other roster/event lines omitted):

```
Each named file has exact revision contents vA, vB, and vC.
WRITE replaces a file with the named exact revision
contents.
TEST runs the named file in an isolated interpreter and
records an exact-byte receipt.
Every TEST listed in this history completed successfully.
RESTORE replaces a file with the referenced revision
contents.
WRITE family000_slot1_bkfodysm.py vA
TEST family000_slot1_bkfodysm.py
WRITE family000_slot1_bkfodysm.py vB
```

Only the display names, not the hidden map metadata, enter the model input. This makes the tested-name association uninformative while retaining event roles and positions.

X Fresh Alias-Balanced Transfer

This post-hoc follow-up freezes its inputs and analysis before new model forwards. It reuses the exact Qwen revision in Appendix Q, with native BF16 weights, float32 last-position post-block L47 residuals, eager attention, batch one, no generation, and no quantization or offload. The runtime is a single RTX PRO 6000, PyTorch 2.8.0+cu128, and Transformers 5.8.1. It is a separate CUDA measurement, not an assertion of byte identity with the historical ROCm study.

The 6,176 inputs comprise 4,096 balanced views, 2,048 original-name unqueried histories for a same-CUDA bridge, and 32 hidden-update observations. Two fresh 16-row qualification processes pass exact-byte checks before full capture. Full capture and a new-process replay each process all 6,176 inputs; residual and action-logit arrays match exactly. All 16 hidden pairs also tie. Capture takes 2,219.6 seconds and replay 1,982.9 seconds, with peak allocated memory 59,799,156,224 bytes. An earlier instrumentation attempt stopped before any forward because an optional device-map attribute was absent. Its failure archive is retained; a separately hashed compatibility correction changes device-identity reporting, not scientific inputs or analysis.

For each source mechanism we fit four new logistic slot heads. Both aliases enter training and validation together: 1,024 training and 512 validation views per mechanism, representing 512 and 256 unique worlds. Every test cell has 512 views of 256 worlds in the original 16 held-out families.

We retain the original C grid, validation-only selection and thresholds, family bootstrap, and four-cell gate. The 100 nested null fits per source mechanism permute complete state vectors within families and assign both aliases of a world the same permuted label. The original-name CUDA bridge has independently fitted heads and a separate four-cell correction. No old head is merely tested on renamed inputs.

For the balanced AUC in Table 68, we pool both alias views within each family and slot before computing AUC, then average the four slots and 16 test families as in Equation 2. Each family-slot AUC therefore ranks 32 observations (16 worlds, two views each), including comparisons across views. It is not the average of two separately computed view AUCs. Bootstrap resampling keeps each complete family, including both views of every world, together. The earlier full audit also computed separate-view AUCs as a diagnostic, without replacing the pooled estimand. The current minimal capsule includes the saved predictions and historical table extracts; its clean entry point focuses on the successor experiments.

Train to test	Balanced AUC (95% CI)	Original-name CUDA AUC (95% CI)	Paired difference (95% CI)
Order→Order	0.991 [0.985, 0.996]	0.994 [0.990, 0.998]	-0.004 [-0.006, -0.001]
Order→Restore	0.440 [0.426, 0.455]	0.330 [0.301, 0.357]	+0.110 [0.079, 0.142]
Restore→Order	0.457 [0.431, 0.487]	0.361 [0.334, 0.390]	+0.096 [0.067, 0.131]
Restore→Restore	0.970 [0.961, 0.978]	0.980 [0.971, 0.987]	-0.010 [-0.018, -0.002]

Table 68: Fresh Qwen unqueried L47 transfer. Differences subtract separately fitted original-name CUDA estimates from balanced estimates using paired family bootstrap. Both rows for a world's aliases remain together; they do not double the independent sample size. Historical ROCm results in Table 5 are not substituted for this bridge. The retained full alias audit additionally records slot AUCs, thresholds, and gate statistics.

Both balanced within-mechanism cells pass, with Holm-adjusted $p = 0.0396$; both transfer cells fail, with adjusted $p = 1.000$. The minimum slot AUCs are 0.988/0.950 within mechanisms and 0.412/0.380 across them. The original-name CUDA bridge has bidirectional gate FAIL; its separately fitted four-cell estimates above keep the runtime comparison distinct from the historical ROCm experiment. Balancing and refitting increase cross-mechanism AUC by +0.110 [0.079, 0.142] and +0.096 [0.067, 0.131], respectively. Thus the reversed-transfer pattern is attenuated, not unchanged, but neither interval for the resulting AUC reaches chance.

Train to test	Slot agreement [95% CI]	View 0 accuracy	View 1 accuracy	Both correct
Order→Order	0.932 [0.918, 0.944]	0.939	0.924	0.897
Order→Restore	0.837 [0.804, 0.869]	0.472	0.486	0.397
Restore→Order	0.824 [0.793, 0.856]	0.452	0.511	0.394
Restore→Restore	0.839 [0.810, 0.865]	0.841	0.826	0.753

Table 69: Alias-paired predictions on the same 256 worlds per cell. Accuracy and agreement average slots and families; "both correct" requires the slot prediction to be correct in both views. Agreement alone is not evidence of correct state decoding. The retained full alias audit additionally records intervals, joint agreement, and joint exact match.

Within-mechanism decoding remains strong, but cross-mechanism scores remain below chance after removing the tested-name association and refitting. The remaining failure cannot be attributed solely to that association in the new balanced dataset. It neither reveals the original heads' causal feature nor rules out event-position, role, or renderer-specific explanations. Paired agreement is imperfect, so these data also do not establish an alias-invariant representation.

Y Archived Feedback Error Decomposition

This post-hoc analysis reads the original 2,560 episodes and 5,272 actually visited prefixes. It verifies the archive, each used member, every lifecycle hash chain, input/residual telemetry, frozen thresholds, and the exact initial-input link to common recovery. No model forward, missing-prefix

imputation, or new policy execution occurs. Initial correctness uses pre-action coverage; coverage after TEST must not be substituted for that label.

Selector	First correct	Completed	Failures with correct first	Next TEST after first TEST
Reference J-lens	269	273	3/239	6/33
Exact ledger	512	512	0/0	0/256
Residual probe	497	365	143/147	169/267
Native shadow	260	42	242/470	491/508
Text	256	290	0/222	–

Table 70: Archived initial decisions and feedback, 512 cases per original arm. The fourth column is an exact intersection divided by that arm's noncompletions; 0/0 denotes no failures. The last column conditions on a successful first-step TEST with a subsequent readout. Text never selects TEST initially. These arm-selected transition populations differ and are not a paired cross-arm accuracy comparison.

The residual initial confusion counts are TP=243, TN=254, FP=2, FN=13. Accuracy is $497/512 = 97.07\%$ with a 16-family bootstrap interval [95.51,98.44]%. Of 147 original failures, 143 (97.28%) start correctly; four start incorrectly. This exact trajectory intersection says when errors occur, not which representation feature causes them.

After a first-step successful TEST, 169/267 residual next actions are unnecessary TESTs. The rate differs by mechanism: order 52/128 (40.63%, CI [26.77,53.91]) and restore 117/139 (84.17%, CI [75.35,92.42]). Across all successive TEST-to-readout transitions, 314/436 are wrong (72.02%, CI [64.99,78.09]). Another 145 final-step TESTs have no subsequent readout; they are excluded from transition denominators, not treated as correct. Intervals use 2,000 family-cluster resamples, retaining dependent decisions within each family. The mean residual score rises by 16.603 after the first TEST (median 12.834), yet frequently remains below the frozen PUBLISH threshold 1.47038565.

Case	Step	Covered before	Residual score	Proposal	Execution
0000	0	0	-37.128	TEST	TEST succeeds
0000	1	1	-11.509	TEST	TEST succeeds
0000	2	1	-15.600	TEST	TEST succeeds
0100	0	0	2.744	PUBLISH	denied
0100	1	0	14.502	PUBLISH	denied
0100	2	0	14.502	PUBLISH	denied

Table 71: Two exact archived residual trajectories. Case 0000 correctly TESTs initially, then repeats TEST despite coverage. Case 0100 repeatedly proposes PUBLISH on uncovered content; the validator denies each proposal. Case identifiers are diagnostic examples, not independent effect estimates.

A separate support audit checks all 4,096 training and 2,048 validation queried rows: each target history contains exactly one prior TEST per file. It verifies that every one of the 1,764 observed successful TEST-to-next-readout transitions across arms adds a TEST, placing the target's count at two or more. This is a concrete input-support shift. Because no matched intervention varies only that count, it remains a candidate explanation rather than a causal attribution. The evidence supports checking feedback-conditioned readout validity and using execution receipts while they remain matched to current contents, rather than treating any past successful TEST as permanently sufficient.

Z LIVE GENERATED ACTIONS AND THE FIXED RECOVERY RULE

Question and status. Does the repeated-TEST failure also occur when a model generates its own actions? This small exploratory live check was designed after the main results. It uses 64 existing workflow cases, not new families or the unrelated-file perturbation. It separately tests a workflow instruction and the fixed recovery rule. The following comparison helps distinguish accurate initial state prediction from successful reuse after feedback; it is not a new agent-architecture contribution.

Z.1 SETUP AND INFORMATION ACCESS

We select one case in each of 16 decision families $\times$ two mechanisms $\times$ two initial labels. Within each cell, the smallest salted SHA-256 of the case identifier determines selection, without model scores. The salt, selected identifiers, histories, and labels are in the capsule. Models use the same exact Qwen and Gemma revisions as the main study, native BF16 on one RTX PRO 6000 96GB, eager attention, and no quantization or offloading. Decoding is greedy with at most 64 output tokens and a 4,096-token input cap; each case is one deterministic sample per arm. Eight fixed, non-benchmark action-format checks precede each model's task run.

The execution loop uses `pi-agent-core` 0.87.1, restricted to TEST and PUBLISH (Zechner & contributors, 2026). JSON text is translated to tool calls; this is not the full coding agent. Each episode receives a fresh real-file workspace, the unchanged execution harness, and the same publication validator. A proposal is not a publication: a matching test receipt must authorize dispatch. The model sees public event history and feedback, never the private labels or byte-hash receipts. There is no WRITE during these three-attempt episodes. Runtime-issued actions also consume attempts.

The four arms cross the workflow instruction (absent/present) with recovery (off/on). Without recovery, the model chooses each action. With recovery, a successful TEST is followed by runtime PUBLISH; a rejected PUBLISH is followed by TEST and, after success, PUBLISH. A malformed output consumes an attempt and does not activate recovery. A failed TEST stops rather than establishing coverage. This rule guarantees recovery only under the stated valid-action, successful-tool, unchanged-content conditions; the existence of a valid strategy does not force a generated policy to follow it.

Prompt differences matter. The archived probe answers a yes/no coverage query. The live task preserves the event history but replaces that query with the target filename, the goal of publication, and the remaining action budget. It also adds the system instruction below and accumulates action/observation turns. Thus matched case identifiers do not mean identical model inputs. The base instruction is:

> Complete the controlled file-history task by publishing the target file within three action attempts. You have exactly two tools, TEST and PUBLISH, both acting on the target file shown in the task. TEST tests the target's current contents. PUBLISH requests publication; the common validator permits it only when a successful TEST covers the current contents. The visible history lists successful tests. No file is edited during this decision episode. Respond with exactly one JSON object: {"action":"TEST"} or {"action":"PUBLISH"}. Do not include explanations, Markdown, multiple actions, or other keys. Tool results will report the next observation. A rejected or malformed action consumes one attempt.

The added workflow text is supplied directly, including the following guidance:

> Use the visible history to decide whether a successful TEST has covered the target file's current contents. A test of another file does not cover the target. Restoring previously tested identical content can restore coverage.
>
> If current target content is known to be covered, choose PUBLISH. If it is not covered or coverage is uncertain, choose TEST. After a successful target-file TEST, choose PUBLISH next, provided the content has not changed. Do not keep retesting the same unchanged content. After PUBLISH is denied for missing coverage, choose TEST, then PUBLISH after success, within the remaining action budget. Do not claim completion unless PUBLISH actually succeeds.

The complete verbatim `SKILL.md`, including its header and validation caveat, is included. This is an instruction-content comparison: it provides restore semantics as well as workflow advice. It does not test automatic retrieval, a skill packaging format, or one framework against another.

Z.2 Interface Amendment and Live Results

The early interface (V1) accepted only a bare JSON object. Both models passed their eight simple format checks, but Gemma wrapped a valid single action in one Markdown fence in 160 of 256 task episodes: 23 in each arm without the instruction and 57 in each instruction arm. V1 rejected all 480 such outputs (3×160), leaving Gemma at 41/64 completions without and 7/64 with the instruction. None of these outputs hit the output-length cap. Qwen completed all cases. Because fencing was more common with the instruction, V1's apparent instruction penalty includes format incompatibility, not just differences in action choice.

We preserved V1 and then specified the final interface (V2), which also accepts exactly one complete JSON or unlabelled Markdown fence around one valid action object. Fencing violates the prompt's format instruction; V2 accepts it only when the fence encloses one valid action object. It does not infer an action, repair semantics, change the prompt, or add retries. Both models and all arms were rerun live. The first raw response is identical between versions in all 256 episodes per model. V2 normalized 160 Gemma outputs and no Qwen outputs. This is a post-hoc interface amendment, not an unchanged preregistered result; versions are not pooled. The failure in V1 was real interface incompatibility, but should not be interpreted as a failure of the enclosed action choices. Both versions' records are retained.

Model	Arm	Done	First correct	FP	FN	Restore correct
Qwen	Minimal	64	50	0	14	18
Qwen	Instruction	64	52	2	10	20
Qwen	Recovery	64	50	0	14	18
Qwen	Instruction + recovery	64	52	2	10	20
Gemma	Minimal	64	49	0	15	17
Gemma	Instruction	64	58	4	2	26
Gemma	Recovery	64	49	0	15	17
Gemma	Instruction + recovery	64	58	4	2	26

Table 72: Every V2 arm completes 64/64 cases; the instruction changes first choices, not completion. First-choice counts are out of 64 and restore counts out of 32. Order histories are 32/32 in every cell. FP is an invalid PUBLISH proposal; FN is an unnecessary TEST. All invalid proposals are blocked.

Model	Arm	Actions/case	Model calls	Input tokens	Model seconds
Qwen	Minimal	1.719	110	81,270	65.0
Qwen	Instruction	1.688	108	102,995	70.3
Qwen	Recovery	1.719	64	35,346	33.4
Qwen	Instruction + recovery	1.688	64	48,914	37.3
Gemma	Minimal	1.734	111	91,012	74.9
Gemma	Instruction	1.594	102	104,479	86.5
Gemma	Recovery	1.734	64	38,838	41.2
Gemma	Instruction + recovery	1.594	64	52,534	56.3

Table 73: Recovery saves calls and tokens without changing the paired action sequence. Counts exclude eight format checks per model. Seconds sum recorded model-generation time, not total experiment time or a monetary estimate. The instruction itself adds input tokens.

In the minimal live arm, the model publishes after every successful TEST: 46/46 Qwen and 47/47 Gemma transitions. No episode required recovery to complete; recovery replaces choices the models already made correctly. It cuts model calls by 41.8% and 42.3%, and input tokens by 56.5% and 57.3%, respectively. These gains concern avoided inference, not better task completion. Both first outputs and action sequences match between each recovery arm and its corresponding model-follow-up arm.

Instruction gains occur only on restore histories. Gemma improves from 17 to 26 correct restore first choices, trading FN 15 to 2 against FP 0 to 4. Qwen improves from 18 to 20, with FN 14 to 10 and FP 0 to 2. All erroneous PUBLISH proposals are blocked by the common validator. The paired instruction-minus-minimal mean-action differences are -0.031 [-0.094, 0.031] for Qwen and -0.141 [-0.203, -0.078] for Gemma. With these successful observed trajectories, mean actions equal $1.5+(FP+FN)/64$; action savings and first-choice improvements are not independent findings. Recovery-minus-minimal model-call differences per case are -0.719 [-0.750, -0.672] and -0.734 [-0.750, -0.703]. Intervals use 10,000 paired family-bootstrap resamples with seed 20260925. They describe variation across these reused families, not decoding randomness or confirmatory discoveries.

Z.3 Same Cases, Different Decision Sources

First choice	Later choices	Qwen First / done / actions	Gemma First / done / actions
Generated action (live)	Model	50 / 64 / 1.719	49 / 64 / 1.734
Generated action (live)	Fixed recovery	50 / 64 / 1.719	49 / 64 / 1.734
Initial-only probe (replay)	Same probe	62 / 44 / 1.859	61 / 49 / 1.828
Initial-only probe (replay)	Fixed recovery	62 / 64 / 1.531	61 / 64 / 1.547
+Feedback probe (replay)	Same probe	61 / 64 / 1.547	62 / 64 / 1.531
Output score (replay)	Same score	40 / 40 / 1.875	48 / 56 / 1.766

Table 74: Better first choices need not survive repeated use. Counts are out of the same 64 cases; actions are means. Probe and output-score rows are archived cached replay with the old 5% FNR-rule thresholds. Generated rows are live V2 runs. Prompts, evidence types, and runs differ: this is a descriptive juxtaposition, not a controlled comparison of policy sources.

The Initial-only probe is initially correct in 62/64 Qwen and 61/64 Gemma cases, compared with 50/64 and 49/64 generated choices. Nevertheless, it completes only 44/64 and 49/64 when repeatedly queried. On covered states immediately following successful TEST, it requests TEST again on 41/54 and 33/53 available transitions; generated minimal policies do so on 0/46 and 0/47. These are transition counts, not independent episodes. The fixed recovery rule keeps the probe's more accurate first choices and completes every case with 1.531/1.547 actions, versus 1.719/1.734 for generated first choices. +Feedback probes also complete all 64 cases in replay.

The descriptive probe-minus-generated first-choice difference is 0.188 [0.141, 0.234] for Qwen and 0.188 [0.125, 0.234] for Gemma. Generated-minus-repeated-probe completion is 0.313 [0.234, 0.391] and 0.234 [0.188, 0.281]. The same 16-family resampling scheme is used. These comparisons illustrate that the repeated-TEST behavior belongs to this repeated score-based rule, not to all ways of eliciting behavior from the model. They do not establish a causal advantage of probes over generation, robustness to the unrelated-file suffix, or a neural mechanism.

Z.4 Evidence and Reproduction

The added `live_audit` directory contains both versions' raw action transcripts, model-request records, lifecycle journals, publication receipts, resulting file bytes, run code, and the existing replay rows used in the comparison. Its CPU audit checks proposal/action correspondence, lifecycle hash chains, receipt hashes, model identities, per-version totals, all 64-case summaries, and descriptive intervals. It verifies 512 V2 publications and also retains the 352 V1 publications and 160 unsuccessful V1 episodes. Raw scientific records are kept separate from an anonymized deployment-metadata projection; original evidence archive hashes remain provenance, not hashes of the anonymized package. No model calls, new fitting, or new significance tests are performed by this audit. The software license is included. The experiment demonstrates a bounded use of execution feedback, not general agent reliability.