

From Internal Prediction to Measured Control

An independent reproduction dossier across frozen language models

Logan Matthew Napolitano | Proprioceptive AI | 25 September 2026

Final complete-asset independent-scientist edition. The studies below were executed separately. Their shared question is whether internal prediction can be turned into useful measured control; they were not one historical end-to-end controller. The accompanying folder supplies original experimental records, historical state banks, three exact model snapshots, independent raw-row checks, frozen protocols and explicit tests for an outside researcher.

Abstract

Can a compact internal measurement predict a declared future, and can a numerically qualified hidden-state write reduce later internal error? On a frozen Hermes-3-Llama-3.1-8B, eight explicit coordinates predict a 68-coordinate target with 68.68-76.15% lower standardized loss than a training-future mean. A separate layer-24 intervention reduces signed downstream internal error by 21.68-43.79% across eight registered cells after a prospective write-fidelity gate. Original-bank analyses have reproduced exactly; an independent full-model recapture and protected confirmation remain open. Distinct behavioral studies show 30 repairs/12 harms for Hermes M170 on 3,000 opened development items, 17/6 for a Qwen3.5-122B prefix write on 503 construction questions, and 179/228 correct for an answer-free Qwen2.5-32B Fiber reader versus 161/228 untouched. Crucial controls narrow these findings: a nearby-layer Qwen3.5 arm nearly matches the target, always using Fiber scores 180/228, and a measured Fiber-to-gateway writer scores 160/228 versus 161/228 untouched. This dossier provides the original code, rows, banks, model snapshots and frozen decision boundaries needed to reproduce each claim and its adverse control independently. It does not present the separate studies as one validated controller.

Scientist's map: claims, controls and verdicts

The scientific proposition is not that one historical agent already performs every step. It is that an explicit internal measurement, a numerically verified intervention and a decision rule can be studied as separable mechanisms and then tested together. The following map keeps the demonstrated links separate from the proposed connection.

Study map. Historical observations, their strongest included controls, and the scientist's next reproducibility step. These are different experiments, not sequential components of one historical controller.

Study	Saved observation	Closest control	External task
Hermes 8B Q0	Eight coordinates predict a 68-coordinate target; loss lower by 68.68-76.15%.	Mean, ablations, native readers; held question/family/dose.	Rerun saved-bank analysis; attempt separate full-model recapture.
Hermes 8B TCSD-CCR	Qualified layer-24 write; signed internal error lower by 21.68-43.79%.	Carrier/off, sign, source and precision gates.	Verify realized write, common support, all signed bounds.
Hermes 8B M170	30 repairs / 12 harms on 3,000 opened development items.	Fixed action budget, calibration and adverse M17P follow-up.	Reproduce model-to-event replay and separate efficacy class.
Qwen3.5 M2	Trained 17 repairs / 6 harms on 503 construction items.	Reverse 7/14; nearby layer 16/6; selected-protocol mismatch.	Rerun all four arms and audit actual hook and score.
Qwen2.5 ARC/Fiber	Answer-free selective reader 179/228 vs untouched 161/228.	Always Fiber 180/228; matched native 176/228.	Reproduce reader and report failed stronger-control gate.
Qwen2.5 gateway / bridge	Three-layer writes and changed logits; selected bridge 160/228.	Zero-dose parity; always-on 159/228; untouched 161/228.	Verify writes and report failed answer-benefit gate.

“Reproduce” has three different meanings here. Recomputing a saved table from bundled rows checks reporting. Running original analysis against the frozen captured bank checks the analysis. Capturing a model afresh checks the full apparatus using the bundled, separately hash-bound weights and state assets. A result at one level must not be reported as a pass at a higher one. Every full-model run should use a new root, retain the exact source and environment identity, and keep failed numerical gates.

The highest-value first experiment is the Hermes 8B mechanism pair: reproduce Q0's registered prediction and TCSD-CCR's qualified downstream restoration on their original banks, then attempt a full-model recapture under the frozen protocol. The Qwen studies ask different questions about answer changes, answer-free selection and realized writing. Their stronger controls and the answer-negative bridge are as important to the independent verdict as the positive rows. The scientist should return separate verdicts for packet integrity, saved-bank analysis, full-model execution, internal causal effect and answer benefit; a failure at one level must remain visible even when another level passes.

Independent reproduction: the complete-asset route

The accompanying folder contains the historical Hermes 8B response and event banks, its 12-file model snapshot, the 78-file original Qwen3.5 snapshot, and the 24-file Qwen2.5-32B snapshot. These large assets are bundled under `assets/` alongside, but distinct from, the original source and result files; every claim remains bound to its own manifest. The Qwen3.5 historical manifest named and sized its 15 weight shards without recording their SHA-256. The original snapshot's content-addressed download etags have therefore been preserved separately, and the packet verifier hashes the staged bytes against them. The Qwen2.5 snapshot is checked against the original 24-file hash list. Original Python environments are specified but not copied.

The scientist's first action is to run `verify_release.py` and `verify_cross_ip.py`, then verify each model and Hermes state bank. Those checks establish custody and recount copied outcomes; they are not model inference. `SCIENTIST_START_HERE.md` provides commands from the packet root. Next, rerun Q0 and signed TCSD-CCR from the bundled captured bank into fresh roots and compare complete results. Only then attempt full-model recapture, preserving the frozen exact predecessor gate and reporting hardware-specific differences. The original 8B method should be judged even if a new apparatus fails that exact gate.

The original TCSD full-capture program is bound to its original host inside the frozen protocol: it verifies absolute paths, Python and Torch versions, GPU identity and index, and separate filesystem device IDs. The packaged inputs are complete, but rewriting the launch command alone cannot reproduce that run on a different machine. The packet's `TCSD_HOST_BINDING_AUDIT_20260925.md` identifies each binding. A scientist should preserve the original protocol, register and hash a host-specific derivative before execution, disclose every changed environmental field, and report any failed gate. No full-model TCSD capture has yet passed from this exported folder.

The complete-asset handoff was tested on the bundled bank. Under the matched NumPy 2.2.6 runtime, Q0 reproduced every output array and result byte-for-byte. A preserved first attempt under NumPy 1.26.4 kept the same fold IDs, losses and gate verdicts but missed strict array equality by at most 1.42×10^{-14} . Signed TCSD-CCR reran from bundled canonical and replay banks with byte-equal outputs. These are analysis checks on opened states, not new model forwards or an independent question set.

The Qwen3.5 construction protocol must preserve all four paired conditions and distinguish its earlier selected candidate from its later 503-item scoring route. The Qwen2.5 reader must retain always-Fiber and matched-native comparisons; the gateway and integration must retain zero-dose, always-on and untouched controls. No scientist should tune on the opened rows and relabel the same set as fresh confirmation. The requested deliverable is a claim-by-claim report with raw outputs, actual write measurements, all failed gates and a verdict at each reproduction level. A protected test may follow only after its model, policy, comparators, scoring and harm ceiling are frozen without looking at those outcomes.

1. A compact, testable internal state

A useful account of a neural computation should do more than describe an activation after the fact. It should predict a specified future on held inputs, retain information whose destruction worsens prediction, and distinguish an intended intervention from the change actually executed by finite-precision hardware. This paper brings those questions together while keeping their experimental answers separate.

The headline result is direct: eight stored scalars predict a registered 68-coordinate internal future with substantially lower loss than both a fixed mean baseline and the native-input alternatives in a separately specified same-bank pilot, even when the evaluation question, perturbation family and dose are held out. The scalars are explicit functions of three candidate response vectors. No learned nonlinear encoder is used. A second study shows that a source-derived native-state correction can reduce later internal error when its actual numerical realization passes a prospective fidelity gate.

The historical prediction and restoration findings are distinct. Restoration used a separately fitted low-rank map of native source displacements. A small-model Qwen pilot in Section 11 explicitly connects the eight-coordinate packet to a fitted native correction decoder; its internal error falls while behavior worsens. Sections 13-15 then test separate Qwen3.5 and Qwen2.5 systems for answer changes, answer-free selection and measured writing. These studies cannot be joined retroactively into one historical controller.

1.1 Contributions

First, we specify an eight-coordinate representation that separates relative candidate geometry, common and relational magnitude, and signed coupling. Second, we report a question-disjoint test against geometric ablations, gain ablation, identity disruption and a specified native-summary augmentation. Third, we describe a conditional restoration procedure in which the realized write is qualified before downstream outcomes are accessed. Fourth, we use retained signed trajectories to distinguish restoration of the paired mean from restoration of the full polarity pair. Finally, we test stronger readers, a differently defined native target, fresh-model method transfer, and an explicit compact-to-native decoder. Behavioral negatives distinguish predictive representation quality from useful control.

1.2 Relationship to existing methods

Predictive state representations describe state through predictions of action-conditioned observations [1]. Our finite future registry follows that motivation, but does not demonstrate a complete predictive state representation or a closed state-update law. Causal abstraction and distributed alignment search test correspondences between high-level variables and neural computations using interventions [2,3]. Here, predictive adequacy of a fixed descriptor and efficacy of a separate native correction are tested; a faithful causal abstraction does not follow automatically.

Inference-time intervention changes activations to influence language-model behavior [4]. Our restoration endpoint instead measures downstream internal displacement and explicitly conditions on numerical write fidelity. No experiment here compares task utility against that method, so these results do not establish superiority over it. The contribution is the concrete representation, held-input evidence and prospective realization procedure, rather than priority over the general ideas of probing, predictive state or activation intervention.

2. Experimental setting and state construction

The studies use a frozen Hermes-3-Llama-3.1-8B line [5] and a 400-question PubMedQA mechanism set [6]. The Q0 protocol and results are preserved in [7]. This set was already opened for research. Holding questions out of a fitted predictor prevents direct fitting on those questions; it does not turn an opened research set into an independently protected confirmation set. The model weights were not trained during the reported analyses.

2.1 Stored response bank

The Q0 analysis reads a bank with shape [400, 52, 49, 4096], storing BF16 bit patterns. The 49 blocks are registered capture blocks, not 49 transformer layers. The file occupies 8,349,286,528 bytes including its array header. Q0 used a read-only memory map, made no full-state copies and ran no new model passes. The bank's SHA-256 is recorded in the reproducibility appendix.

Each field is a three-row matrix, with one 4096-dimensional row per candidate. Candidate order, question identity, depth, token statistic, intervention condition and reference trajectory are part of the measurement definition. Final-token and token-mean statistics are separate configurations. Acquiring the fields requires counterfactual response and reference trajectories; eight stored coordinates do not imply a free measurement from one ordinary forward pass.

2.2 Paired fields

Let $H+$ and $H-$ denote the two perturbation polarities, C the matched center, and N the no-op reference. With and without the registered later conditioning, define:

$$E0 = H+off + H-off - 2 \text{ Coff}; EL = H+on + H-on - 2 \text{ Con}$$

$$I = EL - E0; S = \text{Con} - N.$$

The identity $EL = E0 + I$ is algebraic and is not evidence by itself of useful causal cancellation. Every occurrence of a field in the analysis carries its condition and reference identities.

2.3 Eight-coordinate representation

For V with rows $V0, V1, V2$, set $\mu = (V0 + V1 + V2)/3$ and $Zc = Vc - \mu$. Use the ordered differences $D0 = Z1 - Z0$, $D1 = Z2 - Z0$ and $D2 = Z2 - Z1$. With $\epsilon = 10^{-20}$ and float64 accumulation:

$$ej = \|Dj\|^2; sj = \log(ej + \epsilon) - \text{mean}_j \log(ej + \epsilon)$$

$$gR = \frac{1}{2} \log(e0 + e1 + e2 + \epsilon); gC = \log(\|\mu\| + \epsilon)$$

$$kj = \langle Dj, \mu \rangle / (\sqrt{ej} \|\mu\| + \epsilon)$$

$$z(V) = [s0, s1, s2, gR, gC, k0, k1, k2].$$

The six shape/coupling entries occupy indices 0, 1, 2, 5, 6 and 7; the two gains occupy 3 and 4. The centered log-energy entries sum to zero up to numerical error, so eight stored numbers are not eight independent degrees of freedom. Shared orthogonal changes of hidden basis preserve the descriptor. Shared positive scaling changes the gains while approximately preserving shape and coupling away from the numerical floor. Candidate reordering is not erased: it changes the ordered coordinates and is tested as a destructive control.

3. Registered prediction experiment

3.1 Target and held evaluation

The target concatenates $z(E0)$, $z(I)$, $z(EL)$ and $z(S)$ at depths 24 and 32: $4 \times 8 \times 2 = 64$ coordinates. Four further depth-32 coordinates measure relational and common E0-I alignment and relational and common cancellation, giving 68 targets. Alignment uses normalized inner products. Cancellation in either mode is $1 - \|EL\|/(\|E0\| + \|I\| + \epsilon)$, with the corresponding common or zero-sum component used consistently.

Four source configurations are frozen: E0 at depth 16 and EL at depth 18, each using final-token and token-mean statistics. Five immutable question folds are assigned by the registered SHA-256 round-robin rule. Training uses Jacobian and random-angle perturbation families at doses 0.25, 0.5 and 0.75. Evaluation uses the severe anti-margin family at dose 1.0 on held questions. All rows belonging to a held question are excluded from fitting for that fold.

3.2 Estimator and loss

Each input and target coordinate is standardized using its training-fold mean and sample standard deviation. Required standard deviations must exceed 10^{-12} . Let Φ be the centered standardized training design and Y the standardized target matrix. The affine ridge reader uses an unpenalized intercept:

$$\lambda = 10^{-3} \text{tr}(\Phi^T \Phi) / p; B = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T (Y - \bar{Y}).$$

At evaluation, training normalization and the training design mean are reused. The loss is mean squared error across the 68 training-standardized targets. These losses need not be below one because the evaluation distribution is a severe family/dose holdout. The reported percentage reduction is $100 \times (1 - L_{\text{chart}+\text{gain}}/L_{\text{future-mean}})$; it is neither answer accuracy nor a speedup.

3.3 Ablations, controls and residual information

Readers use six shape/coupling inputs, two gains, all eight inputs, or all eight plus 15 native-state summaries. The 15 summaries are three reference cosines and two coordinate-shape statistics per candidate. References are the matching no-op candidate, matching canonical candidate and no-op prefix. Shape statistics are normalized participation ratio and mean absolute value over RMS. Thus the 23-input reader is a specific augmentation, not an unrestricted full-state competitor.

Fixed candidate cycling recomputes the descriptor after disturbing candidate identity. Within-fold question breaking reassigns predictions without fixed points. Targets retain their original identities. The remaining control predicts the training-future mean. No polynomial expansion, feature selection, nonlinear tuning or target-dependent choice among estimators was allowed in Q0.

3.4 Inference and decision rule

A question-cluster bootstrap with 5,000 draws and seed 202608160301 constructs two-sided simultaneous maximum-t intervals across 24 effects, conditional on the fitted cross-fold readers. All 20 gain, geometry, identity and mean-baseline effects must have positive simultaneous lower bounds and positive means in every fold. For each of four augmentation effects, the simultaneous upper bound of $L_{\text{chart}+\text{gain}} - L_{\text{augmented}}$ must be at most $0.05 \times L_{\text{augmented}}$. This is an operational approximate-sufficiency test relative to the specified comparator and margin, not a theorem about all information in a hidden state.

4. Prediction results

Table 1. Held-future loss under question/family/dose holdout. MSE over 68 training-standardized targets; n = 400 questions.

Source	Mean baseline	Eight coordinates	Reduction
E0 final	13.96799	3.78977	72.87%
E0 mean	17.07891	5.34923	68.68%
EL final	13.96799	3.33141	76.15%
EL mean	17.07891	4.87397	71.46%

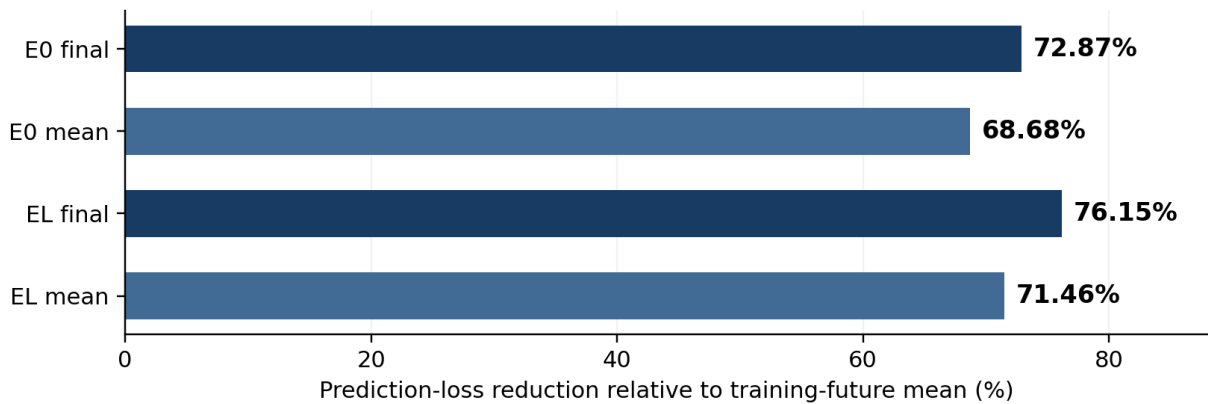


Figure 1. Point estimates from Q0. Bars do not represent accuracy or hardware speed.

Across the four source configurations, eight-coordinate prediction reduces loss by 68.68-76.15% relative to the training-future mean. The percentages in Table 1 are calculated from the full-precision result file, rather than reconstructed from rounded printed losses. The four configurations are correlated measurements on the same research set, not four independent replications.

All 20 specificity effects pass the simultaneous lower-bound criterion, and all 100 associated fold-specific checks are positive. The smallest fold margin is 0.544386 and the maximum-t critical value is 2.882920. Removing gain increases loss by 8.44676-10.88785; removing shape and coupling while retaining gains increases loss by 0.71016-2.37860. The result therefore uses both geometric and magnitude information.

Candidate cycling increases loss by 1.85396-8.00843; breaking question identity increases it by 1.22186-3.20933. The representation carries candidate-specific and question-specific information about the registered future. These controls do not identify a unique semantic variable, but they rule out the tested identity-free explanations.

5. What the compact representation retains

Table 2. Eight-coordinate loss minus augmented-reader loss. Positive values favor augmentation; all upper bounds are below the frozen margin.

Source	Loss difference	Simultaneous interval	5% margin
E0 final	-0.09868	[-0.11696, -0.08040]	0.19442
E0 mean	+0.08393	[+0.05689, +0.11097]	0.26327
EL final	-0.03243	[-0.04817, -0.01670]	0.16819
EL mean	-0.01514	[-0.04100, +0.01071]	0.24446

Adding native-state summaries worsens prediction in two source configurations by the simultaneous interval criterion, offers a small detectable improvement in E0 mean, and has an interval spanning zero in EL mean. Every upper bound remains below its frozen practical margin. In the one clearly improving configuration, the improvement is 1.59% of augmented-reader loss, below the 5% rule. We therefore describe the result as approximate sufficiency for the registered future against the registered augmentation.

This conclusion depends on the comparator. The original Q0 experiment does not test an unrestricted native-state reader. The later comparator pilot in Section 9 addresses several such alternatives; neither experiment exhausts nonlinear probes, causal abstractions or modern representation-learning methods. A stronger reader may extract information missed by the registered summaries. Any resulting noninferiority conclusion remains relative to the compared estimator family and future registry.

5.1 Relation to the later neighborhood analysis

Q1 attempted a source-only neighborhood stress test but failed a geometry-validity requirement. Its favorable stored contrasts are not used as evidence here. A later Q1B redesign used conditional donor groups and found supportive finite-neighborhood results. Because Q1B was designed after that invalid attempt on an opened set, it is supporting exploratory evidence, not independent confirmation. The central quantitative claim of this manuscript rests on Q0.

5.2 A compact predictor is not a closed dynamical law

The preceding R2A experiment rejected the conjunction that one fixed learned transition map predicted and composed across depth. Q0 asks a different question: whether a fixed source descriptor predicts a finite downstream target. Its success does not restore the rejected global-composition claim. A predictive descriptor may be useful even when transitions depend on state, conditioning or the chosen observation registry.

5.3 Replay and chronology

Q0's canonical and cold-replay result, future atlas and audit were reported byte-identical. Both analyses consumed the same recorded state bank. This establishes computational reproducibility of that analysis, not a second biological dataset or an independent model capture. Peak resident memory was approximately 7.16 GB; the evidence root was 130,124 bytes per run.

The canonical runner printed its classification token before the replay audit was complete. The source record states that the protocol, source, thresholds and interpretation were not changed afterward, and replay followed without an adaptive modification. We preserve this result-order deviation and do not describe Q0 as human-blind through replay. An external replication should suppress scientific output until its frozen replay comparison closes.

A later independent run of the original frozen analysis against the retained 8.35 GB BF16 state bank reproduced its seven atlas arrays and final result byte-for-byte in a new root [14]. This is an independent bank-to-analysis computation, not a new model capture. Two separate attempts to capture the 8B model anew on a different GPU stopped at the unchanged predecessor exact-replay gate after 226 of 20,800 conditions: the first 208 historical score/state rows matched and the next 18 differed, starting with a fifth-question no-op reference [14]. The numerical cause remains unresolved; the failed runs are part of the reproducibility record rather than a negative verdict on the original experiment.

6. Precision-qualified native-state restoration

The second study [8] asks whether a correction predicted from a source displacement can reduce a later internal error. It uses a native low-rank transition predictor, not Q0's eight-coordinate ridge reader. Its two source regimes are E0 depth-16 final and EL depth-18 final. Four registered branches are relational gain (GR), common gain (GC), joint gain (GJ) and whole-chart write (W). Mean-token restoration was excluded before outcome access because the predecessor live census left insufficient support for the same rank-16 design.

6.1 Source-derived correction

The predecessor transition study fits polarity-paired lesion displacements at the source to lesion displacements at depth 24. Candidate-common and candidate-relational components are fitted separately. Through-origin maps use rank 16, sketch width 24 and a ridge scale of 10^{-6} with immutable question cross-fitting. In reduced source and target bases B_x and B_y , the map has the form:

$$K = (Zx^T Zx + \lambda I)^{-1} Zx^T Zy; \hat{Y} = (X B_x) K B_y^T; u = -\hat{Y}.$$

The instance-specific held intact destination vector is excluded from correction generation. Matched intact trajectories are used afterward for evaluation. Source displacement measurements themselves require the registered reference/counterfactual apparatus, so this is not yet an endogenous error sensor running inside an ordinary inference.

6.2 Requested writes and realized writes

A predicted vector is not automatically the vector executed by the numerical carrier. The restoration path runs a BF16 prefix through depth 24, constructs the requested target in FP32, casts once to FP16 and runs layers 25-32 and the output suffix in FP16. The intact, lesion-only, treatment and control paths share the carrier. If h is the pre-write state and Q the target cast, define the realized write relative to the carrier-only baseline:

$$\tilde{u} = \text{FP32}(Q(\text{FP32}(h) + u)) - \text{FP32}(Q(\text{FP32}(h))).$$

The prospective gate requires finite positive action, relative realization error ≤ 0.05 , direction cosine ≥ 0.995 , relative norm error ≤ 0.05 , action-to-state norm ratio ≤ 0.25 , and preservation of non-target prefix and padding positions under the registered carrier comparison. Candidate-specific valid completion positions are used instead of a common padded index. A failed gate propagates the carrier-only lesion baseline; it is not clipped into an unregistered partial correction.

Applicability is established before downstream treatment/control states are opened. The downstream files are committed by hash at the seal. For paired endpoints, both polarities must pass; treatment-control comparisons require common physical support for the treatment and the particular control. Precision, branch and locus are therefore part of the experimental treatment, rather than incidental implementation choices.

7. Signed restoration results

The original restoration experiment retains 2,048 source proposals and includes 19,168 treatment/control rows per capture. The prospective gate admits 18,743 rows (97.7828%) and 9,242 of 9,584 paired opportunities (96.4316%). Each of the canonical and cold-replay captures runs 25,664 model passes. The source report records ten byte-identical scientific members across these full captures.

An independent invocation of the frozen signed analysis against the preserved canonical and replay state/event banks recomputed 9,584 pairs and 122 effects; both newly written result files matched the corresponding originals byte-for-byte [14]. This strengthens bank-to-analysis reproduction without adding new model forwards.

The original endpoint measures the paired-even field. A later frozen retrace [9] uses both retained polarities. For downstream errors $e+$ and $e-$ relative to matching intact trajectories, define:

$$E = (e+ + e-)/2; O = (e+ - e-)/2$$

$$\|E\|^2 + \|O\|^2 = (\|e+\|^2 + \|e-\|^2)/2.$$

Branch-specific projections are applied consistently. Restoration gain is one minus treated error energy divided by lesion-only energy. Signed-total gain uses the sum of channel energies, not an average of separately normalized gains. The signed analysis includes 1,974 supported treatment pairs with nondegenerate denominators and uses no new model calls.

Table 3. Signed-field restoration. Parentheses show adjusted one-sided lower bounds, not standard deviations.

Branch / source	Even gain (lower)	Odd gain (lower)	Total gain (lower)
GR / E0 d16	0.2763 (0.2459)	0.0638 (0.0468)	0.2168 (0.1911)
GR / EL d18	0.3464 (0.3125)	0.1498 (0.1267)	0.3059 (0.2746)
GC / E0 d16	0.4102 (0.3860)	0.1385 (0.1173)	0.3545 (0.3329)
GC / EL d18	0.4623 (0.4368)	0.2884 (0.2569)	0.4379 (0.4124)
GJ / E0 d16	0.3165 (0.2979)	0.1039 (0.0909)	0.2542 (0.2381)
GJ / EL d18	0.4040 (0.3827)	0.2328 (0.2102)	0.3666 (0.3470)
W / E0 d16	0.2921 (0.2752)	0.0947 (0.0834)	0.2330 (0.2182)
W / EL d18	0.3708 (0.3508)	0.2013 (0.1826)	0.3332 (0.3145)

Each estimate is a question-equal-weighted fractional error-energy reduction; parentheses contain the multiplicity-adjusted one-sided lower bound from the signed analysis. GR is the relational mode, GC the common mode, and GJ/W the joint field. These are internal-energy effects, not percentage-point changes in accuracy.

All 24 efficacy and all 90 registered specificity endpoints pass their familywise-adjusted criteria and all-five-fold positivity conditions. Controls include sign reversal, question breaking, matched random writes and candidate cycling where defined; cycling is degenerate in the pure common branch. Twenty thousand question-cluster bootstrap draws are used, with familywise alpha 0.05 separately across each complete endpoint family. All eight even-minus-odd contrasts have positive adjusted lower bounds.

8. A stronger internal result, a bounded behavioral claim

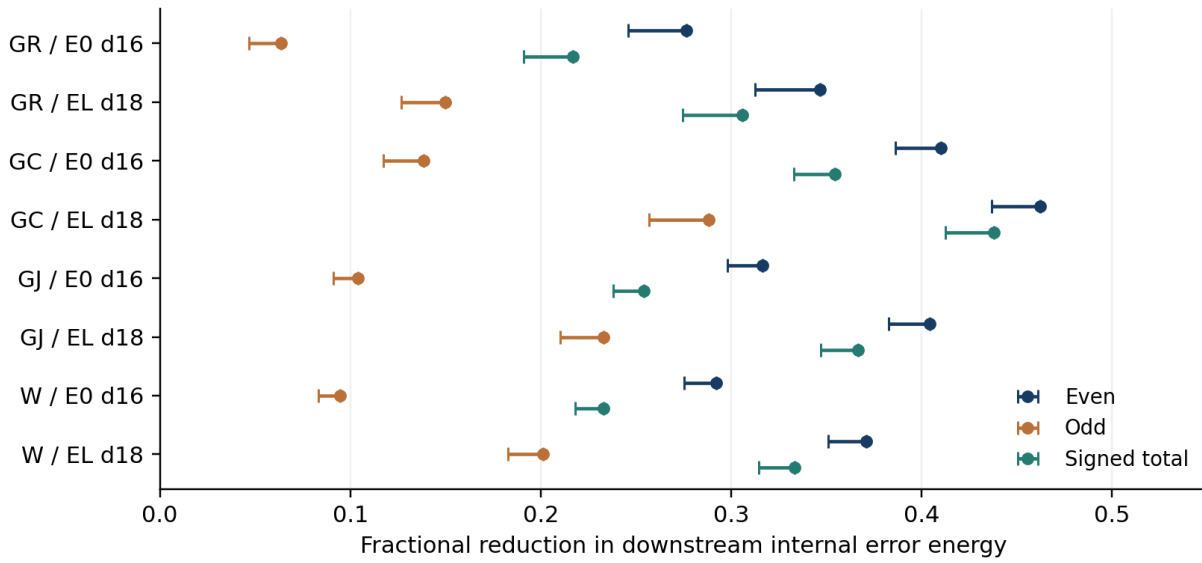


Figure 2. Signed restoration estimates with left whiskers extending to one-sided familywise-adjusted lower bounds. Whiskers are not two-sided confidence intervals.

The signed retrace changes the interpretation of the original paired-mean finding. A same-sign correction derived from the paired source displacement reduces complete signed error energy, including a smaller but positive odd component. Complete paired error-energy reductions range from 21.68% to 43.79%; odd reductions range from 6.38% to 28.84%. This cannot be summarized solely as repairing an averaged field while ignoring the two signed trajectories.

The historical restoration predictor consumes native source-displacement information; its results cannot be attributed to an eight-coordinate correction generator. Section 11 separately tests a compact decoder. Neither the historical result nor that new pilot establishes useful improvement in natural language behavior or separately optimal repair of each polarity.

8.1 Behavioral follow-up

BHV1 was invalid for its intended causal timing claim. BHV2 [10] repaired the temporal apparatus and evaluated a fixed future yes/no self-verifier. It passed the apparatus and cold-replay checks, with 4,001 of 4,096 inverse events physically admitted and 4,000 treatment/sign-reversal events on common support. Nonetheless, all 16 registered mediation and sign-specificity endpoints failed both the adjusted-lower-bound and all-fold-positive predicates. The largest mediation point estimate was 0.000547.

This is a valid negative for universal mediation of that fixed verifier across the registered branches and regimes. It does not erase the measured internal restoration. It shows that the proposed bridge from that restoration to this behavioral readout was not established. The conclusion does not license a claim of improved truthfulness, answer accuracy, safety or general reasoning.

9. A stronger native-input comparator pilot

A new protocol [11] was frozen before new comparator fitting on September 18, 2026. All 400 original questions, all four source configurations and all five original outer folds were retained. The full native input is the 12,288 entries of the same three candidate source-response vectors used to construct the compact descriptor; it is not all model layers. No new model inference was required.

For each outer fold, the next question fold supplied inner validation and the remaining three folds supplied fitting data. Hyperparameter selection used only Jacobian/random-angle conditions at doses 0.25, 0.5 and 0.75. Final testing retained anti-margin dose 1.0 on outer held questions. All feature/target scaling, PCA and kernel centering were fitted solely on the corresponding training split.

Comparators were native linear ridge, eight-component PCA plus ridge, a fixed seeded Gaussian eight-coordinate projection plus ridge, and native RBF kernel ridge. A tuned descriptor reader was included alongside the original fixed reader. The normalized linear-kernel regularizer was n times α for α in $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}$. RBF used $\exp(-\gamma \times \text{squared distance} / \text{median training squared distance})$, with γ in $\{0.25, 1, 4\}$. PCA used seven randomized power iterations and seed 20260918. A single inner fold made this a bounded pilot rather than exhaustive nested optimization.

Table 4. Stronger same-bank comparator pilot. Held-out standardized MSE; lower is better. All four source configurations retained.

Source	Eight	Native linear	PCA8	Random8	RBF
E0 final	3.7898	8.3244	12.4959	9.9736	8.9578
E0 mean	5.3492	13.2589	19.9682	13.2211	9.9038
EL final	3.3314	10.1157	15.4033	10.2777	8.8418
EL mean	4.8740	14.3037	20.6618	12.3598	9.4819

The original reader reproduced archived per-question losses within 4.70×10^{-7} . It reduced loss by 45.99-62.32% relative to the best tested native-input comparator in each configuration. All four native alternatives were worse in each configuration and each held fold. Simultaneous two-sided maximum-t intervals over all 20 paired effects, including tuned-descriptor comparisons, excluded zero in favor of the original descriptor. These intervals condition on the fitted models and do not account for historical study selection.

Native linear ridge selected an interior α in every split. RBF bandwidth selection reached the lowest registered γ in eight of 20 splits, and PCA regularization reached its lowest α in ten. The original descriptor outperformed its tuned variant in aggregate in every configuration: selection on the training intervention distribution did not guarantee improved extrapolation. These observations motivate a broader separately declared challenge rather than retuning the completed result.

The compact map is nonlinear in the source field. Consequently a linear native reader is not a strict superset of the descriptor reader. The native RBF alternative addresses one nonlinear family but does not establish optimality against all possible readers. Feature count also does not measure acquisition cost: the descriptor still requires the registered source-response capture apparatus.

9.1 Broader readers and a direction-sensitive target

A second frozen challenge [12] retained the two original final-token source settings and all five outer folds. It added PCA with 64 components, two further fixed Gaussian projection seeds, and a neural-network reader of training-fitted PCA-64 features. The MLP used one tanh hidden layer of width 32 or 128, Adam at learning rate 0.001, L2 coefficient 0.001 and at most 250 epochs; the same inner held-question fold selected width. Scaling and PCA used training data only. Both inner-width fits reached their iteration limit in all 20 inner comparisons; 19 of 20 selected outer fits also emitted convergence warnings. This is a bounded MLP challenger, not an optimized neural-network ceiling.

The original 68-coordinate future is the secondary target. The primary new target consists of 32 fixed Gaussian projections of the actual candidate-stacked E0 or EL field at depth 32, with projection seed 2026091804. This target preserves some native-direction information and is not constructed from the eight-coordinate formula. Its numerical loss scale is different from that of the original future, so losses must be compared within target rows only.

Table 5. Broader-reader challenge. Compare losses within each target only; target standardizations differ.

Target / source	Eight	Native linear	PCA64	MLP/PCA64
original68 / E0-final	3.7898	8.3244	15.6991	10.2259
original68 / EL-final	3.3314	10.1157	19.5386	10.3158
native_projection32 / E0-final	59.7009	52.7802	52.7805	57.2465
native_projection32 / EL-final	58.6811	50.8348	51.8838	56.7221

On the original future, every added comparator remains worse in both final-token source settings, with all simultaneous intervals favoring the original eight-coordinate reader. On the direction-sensitive target, native linear readers have losses 52.78018 and 50.83481 versus 59.70088 and 58.68107 for the original eight-coordinate reader. The simultaneous intervals for native-minus-eight are [-7.80452, -6.03690] and [-8.85844, -6.83408]. Native readers therefore win this deliberately different task. PCA-64 also improves on the compact reader for this target.

This is an informative boundary. Shared orthogonal rotations preserve the compact packet while rotating native-vector projections; exact recovery of arbitrary direction cannot follow from an invariant packet alone. The new negative does not erase predictive value for the registered geometric future, but rules out presenting the packet as a generally sufficient replacement for the entire source field. Each target family uses its own 5,000-draw simultaneous maximum-t comparison family over 18 effects; no target was changed after observing outcomes.

10. Fresh-model method transfer

The transfer protocol [12] fixed Qwen2.5-1.5B-Instruct revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306 [13], FP32 execution, the coordinate construction and the new test definitions before capture. One hundred twenty unique three-choice arithmetic prompts were generated with seed 2026091803 from addition, subtraction and remainder templates. Hash-ranked prompt identities assigned 60 questions to fitting, 20 to development and 40 to final evaluation. This is a fresh experimental task set, not a claim that arithmetic was absent from pretraining.

Three teacher-forced A/B/C continuations define candidate rows. A prefix-token perturbation after block 7 uses per-question dense Gaussian or 10%-support Gaussian directions at doses 0.05, 0.10 and 0.15 during fitting and development. Final evaluation uses held Rademacher directions at dose 0.25. Each dose scales a unit direction by the current prefix norm. A crossed block-11 perturbation uses a fixed seeded direction at relative magnitude 0.03. Plus/minus pairs, no-op and later-only references define E0, EL, I and S. Candidate states after block 14 supply the unchanged eight-coordinate formula; blocks 21 and 28 supply the unchanged 68-coordinate target construction.

The same comparator families and tuning grids as Section 9 were fitted on the new model. Development questions under training intervention conditions selected parameters; refitting then used 80 fit-plus-development questions. Only the 40 final questions under the held family and dose determine Table 6. This tests transfer of the representation and fitting procedure; the 8B predictor weights were not transported to Qwen, so it is not zero-shot cross-model prediction. Intervention families, layer mapping and tasks also change, making this a method-transfer pilot rather than an exact Q0 replication.

Table 6. Fresh Qwen method transfer on 40 held questions. Standardized prediction loss; lower is better.

Source	Eight	Native linear	PCA8	Random8	RBF
E0 final	0.80995	2.63591	3.24284	2.62196	1.11297
EL final	0.81075	2.63470	3.26572	2.64137	1.11330

The fixed eight-coordinate reader has 27.23% and 27.18% lower prediction loss than the best tested native-input alternative, RBF, for E0 and EL respectively. Loss reductions against the mean baseline are 48.93% and 48.88%. Simultaneous 95% intervals for native-RBF-minus-eight loss are [0.19776, 0.40829] and [0.19711, 0.40800]; all eight native-comparator/source contrasts favor the compact reader. The tuned-eight and original-eight contrast intervals include zero. Inference uses 5,000 question bootstrap draws over the 40 held questions and a simultaneous maximum-t family containing all ten comparator/source effects.

The capture evaluates 10,800 candidate sequences in 360 batched model calls. The stored model, prompt registry, state arrays and source hashes bind these results. The two source settings share questions, a model and closely related trajectories; they are not independent replications. The small synthetic arithmetic domain is materially narrower than general language understanding.

11. A compact-to-native correction test

The fresh capture also supports a declared integration test [12]. An affine eight-coordinate E0 reader predicts the negative plus-polarity prefix displacement at block 21. A native-field linear decoder is fitted as a comparison. Decoder regularization is selected solely by development native-vector error, then both decoders are refitted on the 80 development-plus-fitting questions. The 40 current test intact destination vectors are used only for evaluation, not correction generation. The paired candidate probes needed for sensing remain an explicit cost.

Each held question receives the block-7 Rademacher dose-0.25 lesion. At block 21, the system attempts one of four actions: compact correction, native correction, sign-reversed compact correction, or a seeded same-norm random correction. A prospective gate requires finite positive action, cosine at least 0.995, relative realization and norm errors at most 0.05, and requested/realized action-to-state norm at most 0.25. Only the final prefix token is written; rejected actions propagate the unchanged baseline. All 160 attempted actions pass in this FP32 experiment. This is a new compact decoder, not the historical rank-16 correction map.

The primary behavioral outcome is constrained three-choice negative log likelihood (NLL), evaluated over A/B/C logits before the candidate continuation. Positive lesion-minus-treatment NLL would indicate benefit. Secondary outcomes are three-choice accuracy, repaired/harmed questions and final prefix error relative to the clean trajectory. Candidate continuations were checked not to alter earlier prefix states or logits within the fixed numerical tolerance.

Table 7. Fresh compact-decoder intervention. NLL benefit is lesion minus treatment; negative means worse. Internal error is mean squared-vector distance.

Arm	NLL	Accuracy	Internal error	NLL benefit interval
Clean	3.03343	17/40	0	--
Lesion	2.98287	17/40	56.920	--
eight	2.99352	17/40	54.538	[-0.01529, -0.00600]
native	2.99588	17/40	53.951	[-0.01860, -0.00742]
reverse	2.97217	17/40	60.056	[+0.00604, +0.01537]
random	2.98568	17/40	57.473	[-0.00861, +0.00301]

The compact decoder reduces mean internal squared error by 4.18%, but worsens NLL by 0.01065: the simultaneous interval for lesion-minus-treatment NLL is [-0.01529, -0.00600]. The native decoder likewise modestly reduces internal error but worsens NLL. Sign reversal worsens internal error and improves NLL. Accuracy remains 17/40 (42.5%) in clean, lesion and all treatment arms; there are zero repaired and zero harmed classifications. The perturbation itself has no demonstrated behavioral damage: lesion-minus-clean NLL is -0.05055 with interval [-0.12530, 0.02420].

This is a negative result for useful compact correction in this assay. A weak behavioral damage assay limits conclusions about recovery, but cannot turn the observed NLL worsening into a benefit. The opposing internal and NLL directions demonstrate that reducing displacement toward a clean trajectory need not improve the measured task outcome. Forty extra batched calls, evaluating 480 candidate sequences, completed the intervention stage. No task-specific intervention was selected after test outcomes were observed.

12. A separately reproduced action-selection result

M17O [15] tests a behavioral and control question distinct from Q0 prediction and TCSD-CCR internal restoration. Its original frozen Hermes/Llama-family 8B pipeline made 18,000 model forward passes across 3,000 opened development questions and six passive/action conditions, retaining full 4,096-coordinate states. A registered equal-domain low-rank adapter fitted action-specific benefit and harm heads using leave-one-domain-out data. At a fixed top-5% action budget, the selected policy produced 30 wrong-to-right repairs and 12 right-to-wrong harms, net +18. The three held-domain tallies were FEVER 16/3, CLIMATE-FEVER 5/4 and SNLI 9/5. The archived cold replay reports 14 of 14 pipeline checks passing from model capture through event identity, adapter fitting and per-event prediction export [15].

This is an executed model-to-action pipeline and a bounded development-set answer effect; it is not the same experiment as Q0 or TCSD-CCR. Q0's eight-coordinate reader was not the M17O action controller, and TCSD-CCR's restorative write was not established as the cause of M17O's repairs. The M17O result has an important weakness: predicted selected benefit was 84.899% versus 20.000% observed, although harm calibration improved relative to its predecessor. The paper therefore presents M17O as reproduced historical behavior, not as fresh deployment validation.

M17P [16] later reallocated actions using corrected prior semantics on the already opened action bank. Its exact arithmetic replay passed, but the frozen primary P2 arm made 61 repairs and 28 harms and failed its harm and calibration gates. A stronger +58 net P1 result was a noneligible diagnostic. This adverse follow-up is retained so the M17O behavior result is not mistaken for a generally validated controller. An independent scientist should reproduce the M17O forward-to-ledger chain and the M17P primary gate separately.

13. Qwen3.5 M2: answer changes under a source-defined prefix write

The Qwen3.5-122B-A10B-int4-AutoRound M2 construction run tests a distinct question from TCSD: does a trained state write change which of three complete candidate answers scores highest? Its frozen evaluator scores the log probability of each candidate continuation after a question-and-answer prefix while adding trained vectors to the prompt-prefix states at layers 11, 23 and 35. Each of 503 questions receives four paired conditions in order: no-op, trained write, reversed sign and a layer-shift-plus-one control. The 503 consist of 403 construction-training rows followed by 100 construction-selection rows. This is not an untouched 503-question confirmation test, free-form autoregressive generation or a prospective answer-free controller.

The copied original raw file has 2,012 rows, with its full SHA-256 bound in Module B. An independent block-and-ID recount of those rows gives 316/503 correct at no-op, 327/503 under the trained write, 309/503 under sign reversal, and 326/503 under the adjacent-layer control. Against no-op, trained write makes 17 repairs and 6 harms; reversal makes 7 repairs and 14 harms; the adjacent-layer arm makes 16 repairs and 6 harms. The post-run exact paired test reports $p=0.03469$ for trained versus no-op, $p=0.000912$ for trained versus reversed sign, and $p=1.0$ for trained versus the adjacent-layer control. The trained polarity has a bounded answer effect in this construction apparatus; the layer-specific interpretation does not survive the nearby-layer comparison.

The construction run must not be mistaken for confirmation of the separately selected M2 candidate. The preserved vector-training record selected dose 0.25 on 100 selection questions using mean (length-normalized) token log probability. The later exact-503 evaluator runs dose 1.0, sums candidate-token log probabilities, and uses a different prompt/tokenization route: the canonical trainer includes special tokens and a leading space on the completion, while the later evaluator omits added special tokens and inserts a line break after the answer prefix. Its selection-subset no-op count is 58/100, whereas the earlier training selection record reports 71/100. This is a protocol mismatch whose causes and effect on the result require an external rerun; the 17/6 count is a separate construction diagnostic, not the frozen selected-dose result. The original evaluator serializes the final two active blocks with the same trained_write action label as the actual trained block. The block order in the source and the frozen input identity, not that field, distinguishes reversal and layer shift. The original summary's total_questions of 403 also reflects only the training subset; the raw file and four exact 503-ID blocks establish the actual denominator. Neither defect has been silently corrected in the historical files. The external task is to rerun the copied source with the hashed checkpoint, two construction inputs and vector artifact, and to audit realized hooks and paired scores. The protected excluded and confirmation sets must remain separate until an external protocol is sealed.

14. Qwen2.5 ARC/Fiber: answer-free reader versus strong readers

A separate frozen Qwen2.5-32B-Instruct apparatus measures a candidate-conditioned internal feature and chooses before access to the answer key. Four MMLU test items from each of 57 subjects supplied 228 questions, each in original and deterministically permuted choice order. The no-write baseline, matched-64 Fiber reader, native-state reader, random-projection reader and direct verification controls were all saved per item. The selective policy replaces the baseline choice only when Fiber disagrees and its normalized score margin is at least 0.075, fixed before the new cohort. Capturing an answer-free decision is real control over a choice rule; it is not yet hidden-state actuation.

The copied 456-row analysis ledger independently recounts original-order accuracy as 161/228 baseline, 180/228 always Fiber, 176/228 matched native, 178/228 random projection and 179/228 selective Fiber. The selective rule triggers 23 original-order questions, with 19 repairs and 1 harm relative to baseline. The separate permuted presentations yield 152 baseline, 176 always Fiber and 178 selective correct out of the same 228 questions. Because the prespecified primary gate required superiority to both always Fiber and matched native, the reader did not pass its stronger-control gate: always Fiber scores one more correct on original order. The two presentations are not 456 independent questions. This was newly evaluated relative to specified estate captures; unknown pretraining exposure and the opened outcome now prevent its reuse as fresh confirmation for a revised rule.

15. The archived writer and the actual reader-to-writer bridge

The archived production_v2 gateway was separately exercised on a real Qwen2.5-32B model under two fixed synthetic prompts. It made nonzero hidden-state writes at layers 16, 32 and 48, changed downstream logits and preserved exact zero-dose parity. The short generated answers and tokens did not change in that mechanical gate. This is evidence of causal actuation, not of useful answers.

An integration sweep then used the same frozen Fiber reader and 0.075 threshold to select that gateway write on all 228 already opened original-order items. Its runner retained baseline, exact-zero, always-on, selected and sign-reversed arms; observer hooks recorded nonzero writes at all three layers when active, and the 228 raw rows recorded downstream logit changes for the always-on arm. Twenty-three items triggered the selected branch. The copied per-row ledger and independent recount give 161/228 untouched, 159/228 always-on (1 repair, 3 harms), and 160/228 selected gateway (0 repairs, 1 harm). The direct Fiber answer reader remains 179/228. Thus a real answer-free sensor was connected to a real writer, but this frozen pairing did not improve answer accuracy. This is a development integration result on an opened cohort, not a new confirmation test. A post-hoc opposite-sign diagnostic must not be called a prospective success.

16. What the connected evidence establishes

The three model lines support different rungs. Q0 provides a specified internal prediction under held questions/family/dose on saved Hermes states; TCSD-CCR provides conditional, precision-gated causal restoration of a downstream internal field; M170 provides a replayed model-to-action outcome ledger on opened data. Qwen3.5 M2 demonstrates prefix-actuation changes in candidate scoring under a construction protocol, with an adjacent-layer control too close to claim precise layer localization. Qwen2.5 ARC/Fiber demonstrates answer-free readout and an answer-changing selection rule, while a stronger always-Fiber rule defeats its primary superiority claim. The gateway demonstrates realized hidden-state writes and downstream logits; integrating that reader with that writer was mechanically successful and answer-negative. No single study in this paper demonstrates all of compact prediction, a beneficial hidden-state correction, answer-free selection and superiority to strong simpler methods on fresh protected questions.

That separation is the central research lesson. The IP contains concrete, runnable measurement and intervention components, not merely retrospective prose. It also contains direct evidence that internal restoration and measured writing do not automatically improve the answer endpoint. The external scientist should report an evidence-level verdict for each component and for the cross-component hypothesis, rather than collapsing every positive internal metric into a product claim.

17. Limitations and decisive independent replication

The original Q0/TCSD capture was on one 8B research apparatus; saved-bank arithmetic is not an independent new capture. Two later full-model Q0 recapture attempts stopped at the unchanged exact predecessor gate after 226/20,800 conditions, with cause unresolved. M170 and the M2 construction inputs are opened development evidence. The ARC/Fiber selection cohort was new to specified estate captures, but it is now opened, and the gateway integration reused it. No represented result is an external scientist's independent validation.

17.1 Integrity and independent recount, before model loading

The reproduction package separates three model-specific modules. Module A stages the original Hermes source and small frozen inputs; its 14.82 GB historical state/event assets and 16.07 GB checkpoint are bundled under assets/ and separately hash-identified. Module B contains the original Qwen3.5 run, 503 construction inputs, evaluator, vector and controlled-swap record, with the exact model snapshot under assets/qwen35_model/. Module C contains Qwen2.5 reader captures, probes, writer gate and integration ledger, with the exact model snapshot under assets/qwen25_32b_model/. Start in a clean copy: run `verify_release.py`, `verify_cross_ip.py` and Module A's source-packet verifier, then issue an independent per-item recount. Record any discrepancy before model execution. These checks validate the packet and saved arithmetic, not the model.

17.2 Original-bank and full-model reproduction

After hashing the bundled large assets, rerun Q0 and signed TCSD-CCR analyses from the original banks into new roots. Compare their complete outputs, not only headline percentages. Then use the frozen Hermes capture route and M170 model-to-event pipeline. Keep the original Q0 exact-predecessor gate: the earlier different-GPU attempts matched 208 rows and diverged on 18 beginning at condition 209. Report that point, environment and hardware even if a new run fails. A change of GPU or numerical backend may require a separately declared equivalence analysis, but it must not silently replace the historical exact gate.

On Qwen3.5, treat the selected 0.25-dose normalized-score candidate and the later 1.0-dose summed-score 503-question evaluator as two protocols. Verify prompt encoding, vector identity, actual layer/token writes, paired choice scores and all four controls. On Qwen2.5,

rerun the reader with its original/permutated option presentations and matched strong readers; then separately test gateway zero-dose parity, three-layer writes and downstream logits before rerunning the opened-cohort bridge. Every claim must retain its original negative comparator. The package README and module instructions give the precise commands, required hashes and safe path relocation steps.

17.3 Evidence report and the new protected test

For each module, return the source and input hashes, model and environment identity, exact command, fresh output root, complete per-item ledger, independent recount, actual write measurements, comparator results and every failed gate. Classify saved-row recount, original-bank analysis, full-model capture, downstream causal effect and answer benefit separately. The researcher should produce a short discrepancy log and a manuscript correction list, including negative outcomes. A bank-to-analysis pass does not cure a failed model recapture; a measured write does not establish an answer benefit; and a baseline improvement does not satisfy a gate against always-on or matched simpler controls.

A decisive next test would freeze the target model, split, readout, intervention policy, dose, comparator costs and harm ceiling before touching a genuinely new question set. Measure acquisition cost and realized writes; compare no-op, sign, nearby layer, always-on and matched simpler readers under paired scoring. Report behavior separately from internal-state restoration. A positive result on one rung does not by itself validate the full measurement-guided control claim.

18. Conclusion

The most distinctive result is the combination of a compact, specified internal-future predictor with a separate intervention whose realized numerical write is checked before its downstream effect is scored. That is a serious mechanism claim for independent reproduction. The wider programme also contains bounded answer changes, answer-free choice selection and causal Qwen hidden-state writing across different models. Their first measured reader-to-writer combination executes but fails the answer-benefit test. The supplied packet lets a scientist reproduce each result and its adverse control without treating them as one finished controller. A successful independent recapture would strengthen the mechanism paper; a later, frozen protected test would decide whether the connected policy improves answers.

Appendix A. Reproduction contract and provenance

The accompanying three-module evidence packet includes original records, stronger-comparator pilot, follow-up protocols, source, per-question outcomes, Qwen M2 construction rows, ARC/Fiber capture, the gateway integration ledger, the original 8B state banks and three model snapshots. The model and state assets live in the packet's assets/ folder and have their own verification commands. Rebuilding tables verifies reporting; rerunning capture code reproduces the experimental apparatus. Neither is an external independent replication by itself.

Table A1. Artifact identifiers binding the reported results to the accompanying records.

Evidence object	SHA-256
Q0 result	fae1ff72d6e5b0d6bd12dc054e54e481 a8442128f2987f64886578aa79f7e7a1
Original restoration result	b16c815bed444f1d893c465ae882a44c e1cafd9d2cc3f7a51ccb9ee20de7116
Signed restoration result	fe22fc4ac139105a7cf7f93c8717b8d1 f0bebb0a46cc1fb290280d02068ba246
Original response bank	7182606ecb741f9ad00feca0107526ad7 98c28f6a6b9abdbcc0337c55bb9b38a

Q0 reproduction requires the exact BF16 bank/index and frozen source/target order, normalization, folds, holdouts, effect ordering, bootstrap and apparatus rules.

Restoration reproduction requires matching model/tokenizer, proposal identities, polarities, token positions, precision and cast order; seal applicability before outcomes and verify common support, abstention and the signed-energy identity.

A.1 Deviations retained in this manuscript

Q0 emitted an early classification token before cold replay, without adaptive protocol changes. Q1's invalid instrument is excluded; Q1B is exploratory. The signed decomposition rejected mismatched polarity indices before endpoints; its corrected schema preserved estimands. A metadata hash labeled as a result identified a report, while executable inputs were separately bound. BHV1 is invalid and BHV2 is a valid negative. Full deviations remain in the underlying records.

A.2 Disclosure and filing

The author has a commercial interest through Proprioceptive AI. USPTO acknowledgment records provisional 64/157,514, received 18 September 2026, 11:28:20 AM ET, confirmation 2289. The 39-page specification's local SHA-512 matches the receipt. This records a provisional submission, not an issued patent, validity finding or scientific validation.

A.3 Data availability

The supplied research package retains protocols, source, results, audits, historical 8B banks and the three model snapshots. The full asset folder is much larger than the PDF and should be transferred intact with independent hash verification against its MANIFEST.json. No public archival repository or completed external replication is claimed. The model captures and bank reanalyses described here were performed within the research workflow; an outside scientist must establish their own reproduction level and disclose deviations. A protected confirmation set is not bundled with the opened development data.

References

- [1] Littman, M. L., Sutton, R. S., and Singh, S. Predictive Representations of State. Advances in Neural Information Processing Systems 14, 2001 conference.
<https://proceedings.neurips.cc/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf>
- [2] Geiger, A., Lu, H., Icard, T., and Potts, C. Causal Abstractions of Neural Networks. NeurIPS 2021.
<https://arxiv.org/abs/2106.02997>
- [3] Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. CLeaR 2024; arXiv first submitted 2023.
<https://arxiv.org/abs/2303.02536>

- [4] Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv:2306.03341, 2023; revised 2024. <https://arxiv.org/abs/2306.03341>
- [5] Nous Research. Hermes-3-Llama-3.1-8B model card. <https://huggingface.co/NousResearch/Hermes-3-Llama-3.1-8B>
- [6] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. EMNLP-IJCNLP 2019. <https://aclanthology.org/D19-1259/>
- [7] Napolitano research record. COG-202R2Q Q0 Future Sufficiency Protocol and Post-Replay Report. 16 August 2026. Accompanying evidence: COG202R2Q_Q0_FUTURE_SUFFICIENCY_PROTOCOL.json; COG202R2Q_Q0_RESULT.json; COG202R2Q_Q0_POST_REPLAY_REPORT.md.
- [8] Napolitano, L. M. The Proprioceptive Transformer: Causal Observability, Precision-Conditioned Reachability, and Replay-Closed Conditional Causal Restoration in a Frozen Language Model. Research manuscript, 17 August 2026. Accompanying file: PROPRIOCEPTIVE_TRANSFORMER_TCSD_CCR_20260817.md.
- [9] Napolitano research record. TCSD-CCR Signed-Field Decomposition: Post-Replay Report and Machine-Readable Result. 20 August 2026. Accompanying files: TCSD_CCR_SIGNED_FIELD_DECOMPOSITION_POSTREPLAY_REPORT_20260820.md; TCSD_CCR_SIGNED_FIELD_DECOMPOSITION_RESULT.json.
- [10] Napolitano research record. TCSD-CCR BHV2 Post-Closure Falsification and Localization Report and Independent Post-Closure Audit. 17 August 2026. Accompanying evidence files retain these complete titles in their filenames.
- [11] Napolitano research record. Q0 Strong-Comparator Pilot. 18 September 2026. Accompanying frozen protocol, source, per-question losses, nested selections and result JSON.
- [12] Napolitano research record. Three Follow-Up Experiments: Fresh-Model Transfer, Representation Challenge and Useful Control. 18 September 2026. Accompanying PROTOCOL.md, freeze records, capture identities, source and result archives.
- [13] Qwen Team. Qwen2.5-1.5B-Instruct model card. <https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>. Experimental snapshot 989aa7980e4cf806f80c7fef2b1adb7bc71aa306.
- The external references establish methodological context and model/dataset provenance. They do not independently validate the numerical findings in this manuscript. References [7]-[12] identify the author's underlying experimental records.
- [14] Codex independent reconstruction record. Focused Q0/TCSD Validation Register, original capture-route audit, Q0 raw-bank rederivation receipt, signed TCSD raw-bank rederivation receipt and Q0 failed full-model recapture audit. 24 September 2026. Supplied in the Kolabtree handoff.
- [15] Napolitano research record. M17O Full-Coordinate Causal Workspace and Low-Rank Adapter Result; M17O End-to-End Replay Audit and Reproducibility Runbook. 14 August 2026. Supplied in the Kolabtree handoff.
- [16] Napolitano research record. M17P Action Validity Result and Exact Cold Replay Audit. 14 August 2026. Supplied in the Kolabtree handoff.
- [17] Napolitano research record. Qwen3.5 M2 exact-503 construction run 20260922T045613Z-436071, original 2,012-row JSONL, frozen evaluator, controlled swap, vector artifact and post-run verification. Included in Module B.
- [18] Napolitano research record. Answer-free ARC/Fiber Qwen2.5-32B 228-question/456-presentation run, frozen protocol, capture, analysis and independent raw recount. Included in Module C.
- [19] Napolitano research record. Archived production-v2 gateway real-model gate and full 228-item opened-development reader-to-writer integration, original source, raw ledgers, controls, independent audit and recovery. Included in Module C.