SAME HEAD, DIFFERENT CALLS? AUDITING FUNCTIONAL ROLE CLAIMS IN LOOPED TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Looped Transformers apply one block across calls within a forward pass. Its weights are shared, yet its recurrent states evolve: a changing intervention effect can therefore invite a claim that a head changes function. What does such a profile establish? We specify the recipient states, donor rule, readout horizon, and response construction, pairing each claim with a diagnostic comparison. In pretrained Ouro-2.6B, a selected head shifts the answer margin at a trained early exit by 0.72 logits, yet the effect vanishes after the next call. Across six behavior-selected pointer models, target-unrelated donors reproduce 46% of the matched patch contrast and show progress-aligned structure while promoting their own states. Matched donors additionally promote the intended clean state: contrast improvement and target promotion are distinct. In separate vector analyses, fitted phase gains depend on response coordinates, and the role-slot gain weakens when the clock represents the task boundary. Sparse within-clock phase variation and head-weighting sensitivity leave phase information beyond time unresolved. In a cued positive control, scalar checks retain a known terminal effect. The evidence supports transient influence and donor-conditioned target promotion, but does not identify functional role switching.

1 INTRODUCTION

Looped Transformers deepen computation by calling the same block repeatedly within a forward pass (Yang et al., 2024; Geiping et al., 2025; Zhu et al., 2025). The block's weights are tied across calls, while its input state and downstream computation change. A head whose intervention effect peaks near completion may therefore appear to switch from intermediate processing to answer commitment. In our controlled pointer models, a preselected head's routing and ablation sensitivity change near completion (Appendix A.13). We ask what such a profile establishes about a shared component's functional role.

Such a profile can arise without a change in function. Later calls can erase an early effect, and a donor can improve a contrast by suppressing the wrong state without promoting the intended one. Repeated semantic slots can make vector responses appear phase-organized, while a phase label may encode a task boundary omitted by the clock. These alternatives bear on distinct claims—persistence, intended-target promotion, response organization, and phase information beyond time—each requiring its own evidence. Separating them matters for interpreting recurrent components and intermediate exits.

Prior work shows that patching conclusions depend on corruption and metrics (Zhang & Nanda, 2024), that recovery need not imply portable computation (Quirke, 2026), and that recurrent-state analysis depends on how states are accessed and represented (Wang & Reid, 2026). We examine these issues at the level of a repeatedly used component. Each measurement specifies its recipient, donor, readout horizon, and response construction; each functional claim is paired with a comparison that probes the relevant alternative (Figure 1). This is a claim-conditioned analysis, not a universal role detector.

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.

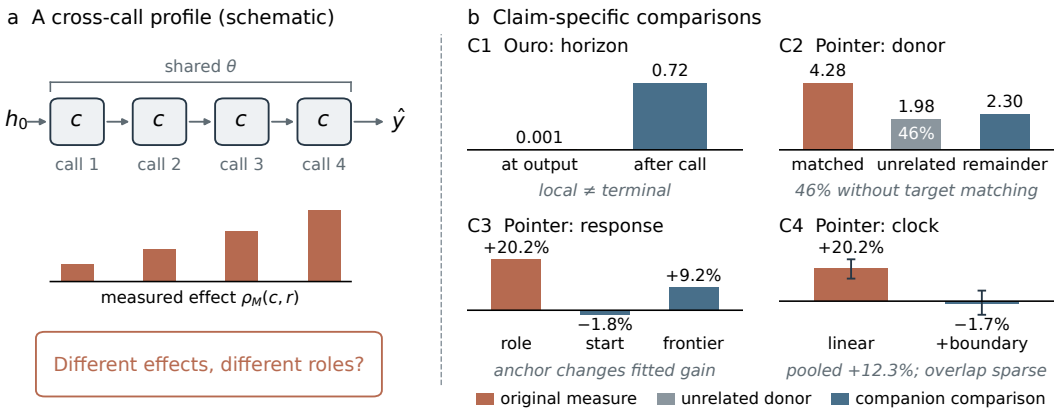

Figure 1: **From a cross-call profile to four distinct claims.** (a) Shared parameters can yield call-dependent measured effects (schematic). (b) Distinct comparisons in Ouro (C1) and controlled pointer models (C2–C4): horizon, donor content, response anchors, and task-boundary representation. C1–C2 use logits; C3–C4 show equal-head held-depth increments (%). The C3 role and C4 linear bars repeat one estimate. The panels address different claims, not successive tests of one component.

Our contributions are threefold:

- **Claim-conditioned analysis of cross-call effects:** we distinguish comparisons for persistence (C1), intended-target promotion (C2), response construction (C3), and phase information beyond a specified clock (C4; Section 3).
- **Transient influence and donor-conditioned target promotion:** a selected Ouro head affects an early trained answer readout without terminal persistence; in six pointer models, target-unrelated donors already yield 46% of the matched contrast and show progress-aligned structure, while matched donors add clean-target promotion (Sections 4–5).
- **Phase predictability is not phase identification:** with interventions held fixed, fitted gains vary with response coordinates and weaken for role slots when the clock represents the task boundary; sparse within-clock support leaves phase information beyond time unresolved (Section 6).

2 Related Work

What repeated calls compute. Looped and equilibrium models can learn iterative algorithms and benefit from additional computation (Dehghani et al., 2019; Schwarzschild et al., 2021; Bansal et al., 2022; Anil et al., 2022; Yang et al., 2024; Saunshi et al., 2025; Fan et al., 2025). Huginn and Ouro bring latent recurrence to language-model scale (Geiping et al., 2025; Zhu et al., 2025). Mechanistic studies offer different predictions for cross-call effects: attention patterns can stabilize as recurrent states approach cyclic fixed points (Blayney et al., 2026); an advancing computation frontier can link loop count to solved sequence position (Zhang et al., 2026), motivating our $r - d$ analysis; and workspace reconstruction can erase the terminal influence of an early write (Wang & Reid, 2026). The last study also relates deeper-network and repeated-program readings through the choice of coordinates, motivating our C3 comparison. Iteration-level patching has been used to study multi-hop composition (Kohli et al., 2026), but we do not attribute a component-level role-switch claim to that work.

Explicit roles versus latent roles. AIR shares a Transformer between low- and high-level updates, distinguishing them through input injection (Shen et al., 2026). This is a concrete published two-update design, but its fixed schedule makes update identity predictable from within-step position. A claim of genuine schedule-indexed specialization need not pass our beyond-clock C4 check; that check applies only to a stronger claim of phase information independent of a specified clock. Our pointer models have no recurrence embedding and provide a different, endogenous-state setting. We audit an interpretation of our own selected head, not AIR's published causal claim.

What an intervention measures. Activation and path patching localize effects, while mediation separates direct from total effects (Vig et al., 2020; Goldowsky-Dill et al., 2023; Geiger et al., 2021; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Corruption, metric, and replacement alter their conclusions, and patched hybrids may be off-distribution (Makelov et al., 2024; Grant et al., 2026; Miller et al., 2024). Same-answer controls reveal broad transfer mistaken for target specificity (Quirke, 2026); our unrelated donor tests an analogous cross-call contrast, while retaining the possibility of a meaningful necessity-like effect. Later computation can compensate for an earlier perturbation (McGrath et al., 2023; Rushing & Nanda, 2024), and an intermediate readout may expose only part of a recurrent state (Lu et al., 2025; Sharma & Vu, 2026). Because validation need not identify a unique mechanism (Méloux et al., 2025), causal interpretations require their assumptions to be explicit (Lin & Liu, 2026); causal abstraction likewise depends on an alignment between neural interventions and task variables (Geiger et al., 2025). Our contribution is not a new general warning about patching. In repeated computation, the same component can have a transient trained-exit effect, while an improved contrast can promote a donor's state rather than the intended target. These observations make the claim-to-comparison pairing operational; advancing-frontier solutions further show why phase must be compared with $r - d$.

3 Tasks, Interventions, and Readouts

From measured effects to functional claims. The shared block fixes a component's parameterized map, not the states on which it acts. We specify a finite-intervention measurement by the recipient set $\mathcal{S}$, donor rule D, readout horizon (boundary) h, and response construction T; $\rho_M(c, r)$ denotes its mean effect at component c and call r:

$$M = (\mathcal{S}, D, h, T), \qquad \rho_M(c, r) = \mathbb{E}_{x \sim \mathcal{S},\, x' \sim D(x)} \big[T_h(x; a_{c,r} \leftarrow a_{c,r}(x')) - T_h(x) \big]. \tag{1}$$

Here $a_{c,r}(x')$ is the component activation on the donor input. The index h can be a fixed absolute exit or a fixed *rule*, such as reading just after the intervened call ($h = r$). The response T_h may be scalar or vector; later sections specify signs for patch recovery and resampling damage. A difference between $\rho_M(c, r)$ and $\rho_M(c, r')$ establishes a *call-dependent measured effect*. Its functional interpretation determines the companion comparison. A *functional role claim* names a task variable and readout boundary, not a unique internal algorithm. Effect magnitude can change across calls while the task contribution remains the same; C1–C4 test alternative sources of the measured difference for the relevant functional claim.

C1 (horizon): read an effect at the claimed boundary, distinguishing target-logit influence from reliable target decoding; a local effect need not persist to output. **C2 (donor):** intended-target restoration requires comparison with a target-unrelated donor of comparable displacement; the remainder is conditional on those donor rules. **C3 (response):** phase organization should be examined under non-repeating task coordinates and alternative anchors, rather than presumed intrinsic to one role-slot construction. **C4 (clock):** a claim of phase information *beyond time* requires within-clock phase variation in $\mathcal{S}$ (Petersen et al., 2012) and comparison with a clock able to express the task boundary. These checks constrain particular claims, not every possible role: a call-scheduled role can be real without C4 (Giannou et al., 2023), and a local role need not influence the final answer. Nor do finite replacements certify that hybrid states lie on the natural-state manifold.

3.1 Controlled task and intermediate-state readout

Each pointer-chasing example contains a shuffled single-cycle map over 16 symbols, a start symbol, and requested depth $d \in \{1, \ldots, 8\}$. A tied model repeatedly queries this fixed memory by cross-attention for eight calls and predicts the symbol reached after d adaptive lookups. It has one 128-dimensional, four-head attention/MLP block (253K parameters), no recurrence embedding, and final-answer-only supervision. The primary analysis concerns six behavior-selected runs whose computation usually reaches the target near $r = d$; criteria and the full eight-run denominator are in Appendices A.3 and A.11. This gate can enrich relative-progress patterns, so results describe an advancing-frontier solution family, not every accurate model. The controlled task isolates alternative explanations; it does not estimate their prevalence in pretrained systems. For a requested depth d and call r, we call $r < d$ *Active*, $r = d$ *Commit*, and $r > d$ *Hold*. The output-derived variant and its relationship to the schedule are specified below.

The final-answer output head decodes intermediate Active states poorly. We therefore train a linear current-state probe on frozen, disjoint trajectories; Table 1 checks it on clean and cycle-preserving corrupt inputs, without refitting. This validates the readout on unpatched trajectories, but not the support of hybrid states.

Table 1: Current-state decoding accuracy (%) averaged over six models. The native output head is a final-answer readout; the independent probe supplies the intermediate-state measurement used below.

Readout and input	Active	Commit	Hold
Native output, clean	2.1	99.1	100.0
Independent probe, clean	99.2	96.8	100.0
Independent probe, cycle-preserving corrupt	99.0	96.6	100.0

3.2 Boundary-specified finite intervention responses

Let a tied recurrent model apply a shared transformation G_θ for $r = 1, \ldots, R$, producing recurrent-boundary states h_r. For component c at recurrence r, we replace or ablate its activation and read a vector of task-relevant quantities $z_1, \ldots, z_K$ at a fixed boundary:

$$\mathbf{f}_{c,r} = [\delta_{c,r}(z_1), \ldots, \delta_{c,r}(z_K)]\,. \tag{2}$$

Here $\delta_{c,r}(z) = -\rho_M(c, r)$ for $T_h = m_z^{(h)}$ and $h = r$ under resampling; it is measured after the intervened call. It is not a layer-local "direct effect": subsequent layers execute normally and may respond nonlinearly to the replacement. Model parameters, the uncorrupted input fields, and the readout horizon—not downstream activations—are held fixed. For a target symbol z, the signed logit margin at readout h is

$$m_z^{(h)} = \ell_z^{(h)} - \max_{v \neq z} \ell_v^{(h)}. \tag{3}$$

Under resampling, $\delta_{c,r}(z) = m_z(\text{base}) - m_z(\text{intervention})$, so positive values denote local margin damage. Matched patching instead reports the increase in the clean-versus-corrupt logit contrast after inserting the clean activation into the corrupt trajectory (and normalized recovery only when the clean–corrupt gap is at least one logit unit).

Let $\tau_{c,r}(z)$ denote the same response with $h = R$. Comparing δ with τ reveals attenuation or amplification, but we avoid a ratio when the local denominator is small. Crucially, our replacements are finite donor interventions, not estimates of a local derivative: donor direction and magnitude vary with the realized trajectory, and the hybrid recipient may cross nonlinear regions. We therefore call $\mathbf{f}_{c,r}$ a *finite response fingerprint*. It is conditional on task, donor rule, readout representation, aggregation, and horizon; it does not identify a component's complete operation.

3.3 Computational phase, clock baselines, and interventions

Operational phase labels. Computational phase is a sample-level label derived only from the unintervened answer logits and the gold final answer y; it uses neither attention nor a learned classifier. For example x, define the answer margin and first stable-margin call together:

$$\begin{aligned} m_r(y; x) &= \ell_r(y; x) - \max_{v \neq y} \ell_r(v; x), \\ r^*(x) &= \min\{r \in \{1, \ldots, R\} : m_s(y; x) > 1 \text{ for every } s \geq r\}. \end{aligned} \tag{4}$$

Recurrences before r^* are Active, r^* is Commit, and later recurrences are Hold. If no such stable-margin recurrence exists, the trajectory is Unresolved and excluded from phase-conditioned causal comparisons. Thus different examples at the same recurrence can receive different phases. Labels are frozen before intervention; matched patching additionally requires clean and corrupt trajectories to share the same non-Unresolved phase. Because the condition includes $s = R$, it already entails a correct final prediction; this is equivalent to the implementation's explicit final-correctness guard. These phases are post-hoc operational labels for analysis, not an assumed internal controller or a deployable phase detector. Our output-independent schedule instead labels call r as Active for $r < d$, Commit for $r = d$, and Hold for $r > d$, where d is requested pointer depth. Because this label is a

deterministic coarse function of relative time, it tests schedule alignment and output independence, not information beyond the clock. On the intervention cohort, model-derived and task-schedule labels agree on 98.7% of all states ($[98.1, 99.3]$ across seed means; seed range 97.5–99.8%) and on 98.7% of resolved states. Both are therefore near-coarsenings of $r - d$ in these models. We use "phase" as a compact grouping label for position relative to the requested depth, not as a variable identified separately from relative time.

We compare causal fingerprints using a grouping contrast

$$C_{\text{group}} = D(\text{same call, different phase}) - D(\text{same phase, different call}), \tag{5}$$

where D is cosine distance for resample fingerprints and absolute distance for scalar matched-patch effects. A positive value means effects differ more between phases at one call than between calls within a phase: we call this *progress-aligned* structure, not target recovery. For scalar effects, the unit is a head–call–phase cell, averaged over eligible examples before pairwise distances and then over heads within each seed. Exact weighting and the related phase/call predictor are in Appendix D.

Two intervention estimands. We use two complementary, non-interchangeable interventions. In clean-to-corrupt patching, the corrupt pointer map remains one 16-symbol cycle; inserting a clean activation yields the local change in the clean-minus-corrupt current-state logit contrast. A matched donor carries the intended clean state. A same-depth, same-call unrelated donor carries a different current and final target. The reported donor approximates the matched donor's L_2 displacement; this rule was added retrospectively and then checked against a first-eligible donor rule. Neither rule matches direction or proves on-manifold support. We also report effect divided by the clean–corrupt gap where that gap is at least one logit.

In phase-matched resampling, the donor instead comes from another example at the same depth and phase; its multi-coordinate end-of-pass response tests geometry, not target restoration. The reported scalar comparison is the raw matched-minus-unrelated logit contrast under the stated donor rule; it was not prospectively registered as a unique primary endpoint. Gap-normalized recovery checks its scale, while vector coordinate and clock comparisons answer distinct representation questions. We aggregate within models; seed-bootstrap intervals describe only the six trained solutions.

3.4 Scope and supporting systems

Ouro-2.6B has four calls, 48 shared layers, and trained per-call LM-head exits (Zhu et al., 2025). Its selected-head case study and the six-model controlled result do not estimate prevalence across architectures. Appendices A.2 and A.3 report their configurations and denominators; Appendix B reports failed pretrained gates.

4 C1: Readout Horizon Changes a Head's Apparent Temporal Profile

A terminal readout can miss a head's earlier influence if subsequent calls erase the measured effect. Ouro answers three-link letter chains; clean and corrupt prompts differ at one link. On 600 held-out pairs, we patch the clean layer-42 head into corrupt calls 1–4 and read the answer margin immediately or at output. Per-call LM-loss makes the intermediate exit a trained answer readout (Zhu et al., 2025). Head 11 was selected on discovery data for *source-call specificity* at call 4: matched clean call-4 versus clean call-1 activations, both inserted at the call-4 site and read terminally. Call-1 interventions were not used in selection. A source-call preference at call 4 does not establish an exclusively late role; Figure 2 also shows all layer-42 heads at call 1.

The selected head has early influence that is almost invisible at output. The matched call-1 patch shifts the immediate answer margin by 0.721 logits, but nearly zero remains at call 2 or output (0.001; Figure 2). On these unpatched pairs, call-1 answer top-1 accuracy is only 37.5%/34.2% for clean/corrupt prompts, despite a 1.35-logit natural margin gap (Appendix A.2). The patch affects a trained answer readout, not reliable early answer decoding. Patching every token instead of only the final query token preserves this local-to-terminal attenuation (Appendix A.2). The path establishes loss of measured influence after one more call, but does not identify whether later computation actively compensates for the patch.

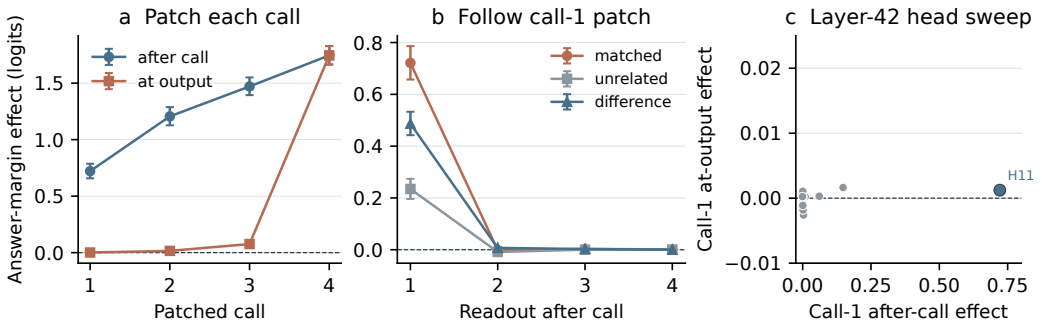

Figure 2: **A selected head's early answer-logit influence is missed at output.** (a) Patching H11 at each call and reading immediately or at output. (b) Following a fixed call-1 patch through all four readouts: a target-unrelated cross-example donor explains part, but not all, of the local effect. (c) Call-1 local and output effects for all 16 layer-42 heads; H11 is discovery-selected. Intervals bootstrap paired examples; all effects are clean-minus-corrupt answer-margin changes in logits.

On the same recipients and H11 site, a same-stage, different-target natural donor from another prompt produces 0.235 logits locally; the matched-minus-unrelated remainder is 0.487 ($[0.442, 0.533]$), and a first-eligible donor also leaves a positive remainder (0.401). Both donor effects vanish by call 2. The unrelated donor naturally predicts its own target and approximately matches activation displacement, but differs in prompt identity and direction; this is a *cross-example donor sensitivity*, not a pure target-information effect (Appendix A.2). The remainder is concentrated in stage-3 link corruptions, so it cannot establish multi-hop progress (Appendix A.2).

H11 has the largest call-1 attenuation among all 16 heads in its layer; the median is 0.001 logits (Figure 2c). Its call-4-specific terminal contrast remains valid, but that contrast alone does not justify an exclusively late functional interpretation. This selected case is not a layer-wide prevalence estimate. The horizon test leaves open what a donor contributes to an improved contrast. The controlled task supplies a validated current-state readout at each call, letting us separate contrast improvement from promotion of the intended state.

5 C2: Contrast Improvement Need Not Restore the Intended Target

An improved clean–corrupt contrast does not reveal which state a patch promotes. We test this distinction immediately after the intervened call. Every corrupt map remains a single 16-symbol cycle, and the independent probe decodes its unpatched states accurately (Table 1). A target-unrelated donor can expose contrast improvement without the intended clean state; this checks target promotion without verifying the support of patched hybrid states.

A clean–corrupt contrast can improve by raising the clean answer, suppressing the corrupt answer, or both. With mean-centered probe logits $\bar{\ell}_v = \ell_v - \frac{1}{16}\sum_u \ell_u$,

$$\Delta^{\text{patch}} = \Delta\bar{\ell}_{z_r} - \Delta\bar{\ell}_{\tilde{z}_r}, \qquad \Delta^{\text{rem}} = \Delta^{\text{patch}}_{\text{matched}} - \Delta^{\text{patch}}_{\text{unrelated}}. \tag{6}$$

Here $z_r, \tilde{z}_r$ are clean and corrupt current-state symbols, distinct from final-answer y before completion. Matched and target-unrelated donors improve the contrast in different ways. The matched donor improves it by 4.28 logits: it raises the clean state by 2.22 and lowers the corrupt state by 2.07 (Figure 3a). A same-depth, same-call target-unrelated donor produces 1.98 logits (46%) by lowering the corrupt state by 2.04 while leaving the clean state nearly unchanged (-0.06). On the 94.1% of eligible observations where its own state differs from both recipient states, that donor instead raises its own centered state logit by 2.26 (Appendix A.7). Thus the same clean–corrupt contrast can improve while a different state is promoted. The matched-minus-unrelated remainder is 2.30 logits ($[1.87, 2.55]$; 6/6 seeds), predominantly additional clean-target promotion under these donor rules. On the same donor-distinct subset, the independent probe ranks the clean state first after 10.3% of matched patches but only 0.1% of unrelated patches; unrelated patches instead rank the donor's own

state first in 9.2% of cases (Appendix A.7). These are readout destinations, not native task-recovery rates: most patched states still rank corrupt first.

Matched and unrelated displacements are close in mean L_2 norm (3.96 versus 4.04); first-eligible donors give a similar split. Noise, batch-mean and cross-depth replacements constrain simpler perturbation accounts (Appendix A.7), but neither donor direction nor hybrid-state support is matched. The remainder is therefore an operational donor-conditioned difference, not a pure target-semantic causal effect; the near-equal corrupt suppression is an aggregate, not a cellwise identity (Appendix A.7).

The unrelated donor alone has a positive progress-aligned grouping contrast of 2.58 logits ($[2.02, 3.15]$; 6/6 seeds), versus 4.92 for the matched donor; norm-matched noise yields only 0.43 (Appendix A.7). Progress-aligned structure therefore appears without intended-target promotion. Figure 3b–c shows that both the unrelated effect and the matched-minus-unrelated remainder have progress-aligned structure. The remainder's larger cells form a band near $r = d - 1$ and $r = d$; this pattern persists under coarse balancing by time since the corrupted edge (Appendix A.7). The behavior gate favors near-d completion, however, and eligibility changes as clean and corrupt paths diverge. The band describes this cohort rather than identifying a latent phase-specific operation. Gap-normalized recovery likewise preserves the donor distinction without yielding a constant remainder fraction (Appendix A.7).

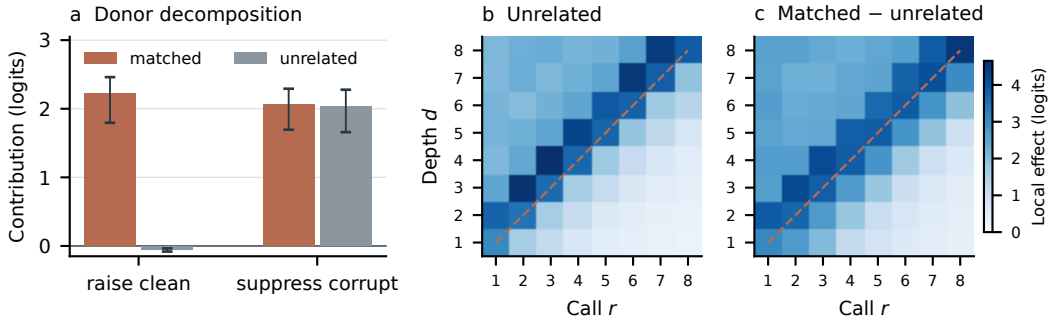

Figure 3: **Donor content changes the meaning of contrast improvement.** (a) Mean-centered clean-target promotion and corrupt-target suppression for target-matched and L_2-matched target-unrelated donors; bars are six-seed means and intervals resample seeds. (b) The target-unrelated local effect and (c) matched-minus-unrelated local remainder by requested depth and call, averaged equally across seeds after pooling heads within each cell. Both heatmaps use the same color scale; the dashed diagonal is $r = d$. Corrupt maps remain single cycles; all panels read effects after the intervened call.

The selected seed-137 head further illustrates why target and horizon must be specified together: its Active remainder is 7.25 logits for the current state after the call but 0.15 for the final answer at output. The target changes, so this is not an attenuation estimate; fixed-target checks are in Appendix A.7.

6 C3–C4: Phase Predictability Depends on Response and Clock Design

Beyond scalar target promotion, we ask what makes a vector response predictable from phase. This separate analysis uses phase-matched cross-example resampling. We first vary the coordinates while holding the interventions fixed, then vary the clock while holding each response construction fixed.

For this vector question, we predict mean responses in head–call–depth–phase cells while leaving one requested depth out. The *linear clock* uses $(r - d)/R$ and r/d; the boundary clock adds indicators for $r = d$ and $r > d$. Both include answer-margin confidence features.

$$\Delta_{\text{phase}} = 1 - \frac{\text{MSE(clock+confidence+phase)}}{\text{MSE(clock+confidence)}}. \tag{7}$$

We report raw and cellwise L_2-normalized responses. Our default summary takes this ratio within each head before averaging heads and seeds equally, asking how a typical head's fitted gain changes. Pooling head errors before taking the ratio weights heads with larger baseline errors more heavily (Appendix D).

6.1 C3: Response alignment changes the fitted phase gain

The role-slot response aligns effects with four semantic slots (predecessor, current, next, final). Near completion, several slots can denote the same symbol, so an apparent phase gain may reflect the response itself. When all four slots denote answer y,

$$\mathbf{f}_{\text{role}} = \delta_{c,r}(y)\mathbf{1}_4, \qquad \mathbf{f}_{\text{deweighted}} = \tfrac{1}{2}\delta_{c,r}(y)\mathbf{1}_4. \tag{8}$$

We compare six constructions on the same interventions: role slots, deweighted slots, nine unique path symbols, and all 16 symbols aligned by fixed identity, start position, or current frontier (Appendix A.6). With a *linear* relative clock plus confidence, adding model phase reduces held-depth error by 20.2% in role coordinates (6/6 seeds; 12.9% after normalization). Deweighting retains a 19.3% raw gain, so duplicate weight alone does not explain the result. The start-aligned non-repeating response instead gives -1.8%, while frontier alignment retains $+9.2\%$.

These comparisons concern accessibility to the fitted predictor, not simply information content. Start and frontier responses differ by a known within-cell permutation: applying it to both a response and its prediction preserves squared error, but refitting a restricted predictor need not. Role selection and repetition make further changes beyond that permutation. The positive frontier result is therefore partial, not construction-general. Since role slots begin to coincide at $r = d$, we next ask whether the linear clock missed that same task boundary.

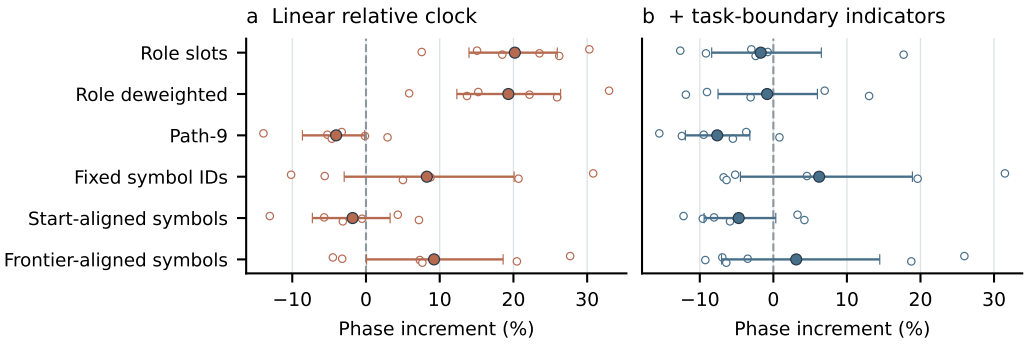

Figure 4: **Phase gain depends on response construction and the task boundary.** Adding model phase to (a) a linear relative-clock baseline or (b) the same clock augmented only with two task-boundary indicators. Small points are individual seeds; large points and bars are six-seed means and bootstrap 95% intervals of raw independent-probe increments, with heads averaged equally within seed. The dashed line is zero. C3's construction comparison uses (a); (b) tests how much the boundary changes that comparison. Pooled-head and flexible-clock sensitivities are in Appendix D.

6.2 C4: Task-boundary controls limit beyond-clock interpretation

Task-schedule phase is exactly a coarsening of $q = r - d$: Active for $q < 0$, Commit for $q = 0$, and Hold for $q > 0$. Consequently, for any response Y,

$$\mathbb{E}[Y \mid q, \phi_{\text{task}}] = \mathbb{E}[Y \mid q], \qquad \phi_{\text{task}} = g(q). \tag{9}$$

This identity is definitional, not an empirical null or a rejection of schedule-indexed roles. A linear clock may still miss that boundary.

A boundary-only replay adds just two task-boundary indicators to the linear clock (Figure 4b). For raw role slots, adding them reduces the phase increment from $+20.2\%$ to -1.7% under equal-head averaging ($[-8.4, 6.5]$ for the latter). The same comparison under pooled-head MSE falls from $+31.0\%$ to $+12.3\%$ ($[-1.2, 25.3]$ for the latter). The role/start gap also narrows from 22.0 to 3.0 points under equal-head averaging. Thus representing the task boundary attenuates the *role-slot* linear-clock gain under either aggregation. The residual depends on head weighting, however, and some raw all-symbol pooled gains remain positive (Appendix D).

Identifying information beyond time also requires phase variation at fixed clock values. Model-derived phase agrees with the task schedule on 98.7% of states; in the vector cohort, only 0–16 of 256

fitted head–depth–call strata per seed have at least 20 examples in each of two phases (Appendix D). These are descriptive support counts, not a fitting filter. Sparse overlap prevents either a positive or negative phase-beyond-time inference from the residual gain.

With a more flexible clock (categorical $r - d$, quadratic r/d), none of the six current response constructions has a positive mean model-phase increment (Appendix A.6); an earlier output-head construction has the opposite sign (Appendix D). Both are protocol-specific predictor comparisons, not tests of the definitional identity in Equation 9.

7 Discussion

What changes in a component analysis. An improved patch contrast and a progress-aligned profile do not alone establish target restoration. In C2, an unrelated donor improves the recipient contrast while promoting its own state; the matched donor additionally promotes the intended state. Reporting clean-, corrupt-, and donor-state logits at the same readout and on a stated cohort makes the difference visible; tracing one patch through trained exits separates transient from terminal influence. These are claim-specific results across systems, not a threshold for declaring a role.

A compatible account, not an identified mechanism. In the selected advancing-frontier solutions, repeated lookup on changing states could produce a completion-peaked effect without changing a head's operation. Before completion, the current state and final answer differ; at $r = d$ they coincide, and after completion the state is held. This accounts for a current-state effect missing a final-answer readout and for role-slot collinearity near the boundary. The selected seed-137 head changes routing with progress, but its Active readouts change both target and horizon (Section 5; Appendix A.13). The data do not isolate state-dependent reuse from a phase-specific operation.

Positive-control calibration. In an explicitly cued control, every matched patch reaches the oracle switched final target, and the terminal donor-conditioned remainder is positive in all three models (mean 15.22 logits; Appendix F). Unrelated donors still reproduce 56% of the matched contrast. That share is not a role score: the scalar checks retain a known effect, while spontaneous specialization and the C3/C4 vector claims remain uncalibrated.

Inferential scope. The six behavior-selected seeds, not heads or intervention rows, are inferential units; their bootstrap intervals are descriptive. Seed 71 and a failed untied gate limit the generality of this solution family (Appendices A.3 and A.13); Ouro contributes one selected head, and other pretrained gates failed (Appendix B). Donor direction and hybrid-state support remain uncontrolled; sparse within-clock support and aggregation sensitivity leave beyond-time claims unresolved.

Next identifying design. A test of phase *beyond time* needs exogenous variation in progress at fixed relative time, with a frozen site, donor, response, aggregation, and untouched examples. Schedule-indexed roles need not pass that test: AIR's L/H updates are identified by within-step position (Shen et al., 2026).

8 Conclusion

Cross-call intervention profiles reveal changes in measured influence, but do not by themselves identify changes in a head's function. A selected Ouro head has transient influence on a trained early answer readout. In six pointer models, matched donors add intended-target promotion beyond contrast improvement from target-unrelated donors, which also promote their own states. For vector responses, phase gains depend on coordinate construction; representing the task boundary attenuates the role-slot gain without resolving phase information beyond time. An explicitly cued control retains a known terminal effect, so the scalar comparisons can preserve a functional conclusion rather than merely weaken one. In practice, the comparison must match the claim: trace persistence, distinguish intended-target from alternative-state promotion, and assess phase gains against both response and clock choices. Separating state-dependent reuse from phase-specific operation remains a distinct identification problem.

Reproducibility Statement

The appendix records model and data-generation protocols, component selection, interventions, readouts, aggregation rules, seed-level results, and failed eligibility gates. Discovery and validation splits were separated before component search; additional model runs are distinguished from retrospective measurements on frozen checkpoints. Model, evaluation, and resampling seeds are fixed in the experiment configurations and analysis scripts.

AI Use Statement

Generative AI tools assisted literature search and discovery; research ideation, experimental planning and execution; and drafting and polishing portions of the manuscript. The synthetic task examples were sampled by the documented generator rather than generated directly by AI, and AI was not used to prove mathematical propositions. All AI-assisted work was reviewed, and numerical claims were checked against stored artifacts. The authors take responsibility for the final text, claims, and artifacts.

References

Cem Anil, Ashwini Pokle, Kaiqu Liang, Johannes Treutlein, Yuhuai Wu, Shaojie Bai, J. Zico Kolter, and Roger Grosse. Path independent equilibrium models can better exploit test-time computation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Extrapolation without overthinking. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael M. Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal Transformers. In *International Conference on Learning Representations*, 2019.

Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. Looped transformers for length generalization. In *International Conference on Learning Representations*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380.

Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11398–11442, 2023. URL `https://proceedings.mlr.press/v202/giannou23a.html`.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. *arXiv preprint arXiv:2604.07822*, 2026.

Zezheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims. *arXiv preprint arXiv:2605.08012*, 2026. URL https://arxiv.org/abs/2605.08012.

Wenquan Lu, Yuechuan Yang, Kyle Lee, Yanshu Li, and Enqi Liu. Latent chain-of-thought? Decoding the depth-recurrent transformer. *arXiv preprint arXiv:2507.02199*, 2025.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.20914.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling*, 2024.

Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012. doi: 10.1177/0962280210386207.

Philip Quirke. Necessary, decodable and reversible, yet not transferable: A stress test for attention-head role claims. *arXiv preprint arXiv:2606.08292*, 2026. version 2, 4 August 2026.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855, 2024.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, 2025.

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. Can you learn an algorithm? Generalizing from easy to hard problems with recurrent networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Mark Shapiro. Retrofitting recurrent depth into a pretrained language model: Installation, extrapolation, transfer, and retention at two parameter budgets. *arXiv preprint arXiv:2608.11233*, 2026.

Rituraj Sharma and Tu Vu. Dense supervision is not enough: The readout blind spot in looped language models. *arXiv preprint arXiv:2606.24898*, 2026.

Jucheng Shen, Barbara Su, and Anastasios Kyrillidis. One model, two roles: Emergent specialization in a shared recurrent transformer. *arXiv preprint arXiv:2605.17811*, 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Wenlong Wang and Fergal Reid. Looped transformers under the Jacobian lens: Does the global workspace survive recurrence? *arXiv preprint arXiv:2609.01924*, 2026.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Tong Zhang, Junhao Hu, Yun Peng, and Tao Xie. When does recurrence become an algorithm? Convergence selection in weight-tied looped transformers. *arXiv preprint arXiv:2607.20594*, 2026.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

A Additional Experimental Details

The appendix follows the main claims: model and evidence provenance precede the Ouro horizon audit, controlled-model readout and donor checks, vector response and clock sensitivities, and calibration/scope controls. Replays on the same six frozen checkpoints are measurement sensitivities, not additional model replications; separate training cohorts are identified where used.

Controlled architecture. The shared block contains pre-normalized four-head cross-attention from a single query state to fixed task memory, followed by a residual connection and a pre-normalized two-layer GELU MLP with expansion ratio four and a second residual connection. Attention and MLP dropout are zero. A final layer normalization and bias-free linear classifier read every recurrent-boundary state. Tied models reuse one block across passes, without a loop-index embedding.

Canonical pointer protocol. Unless an experiment is explicitly marked otherwise, the controlled models use 16 symbols, maximum task depth eight, eight recurrent passes, width 128, and four attention heads. They are trained for 12,000 steps with batch size 512, AdamW, learning rate 3×10^{-4}, weight decay 0.01, and cosine decay after 500 warmup steps to a minimum learning rate of 3×10^{-5}. The objective predicts only the final answer. Table 2 centralizes the frozen-state and intervention sample sizes used by the principal six-model analyses; all splits use distinct random seeds and the base checkpoints remain frozen.

Table 2: Shared configuration for the six-model probe and causal replays. Counts are per task depth and per model. Representation variants share the same resampled activations.

Stage	Examples	Seed / optimization
Linear-probe train	512	907; 400 steps, batch 4096
Linear-probe test	512	919
Phase-matched resample	256	947; batch 64
Two-edge patch sensitivity	128	953; off-support cohort
Cycle-preserving patch	128	1000; separate fixed cohort
Seed bootstrap	6 models	10,000 resamples

A.1 Evidence status and analysis timing

Table 3 separates new training runs from re-analyses on the same six frozen models. "Fixed before" names the relevant stage of this study, not external preregistration.

A.2 Ouro component localization

Model and readout. We use the released Ouro-2.6B checkpoint with 48 shared decoder layers and four recurrent calls. Early exit is disabled for the intervention sweep, so every example completes all

Table 3: Evidence status for the controlled results. Only new training runs add models; re-analyses check measurement sensitivity on frozen checkpoints.

Analysis	Denominator	Fixed before	Evidential role
New training runs			
Pointer discovery	4 canonical seeds	confirmation training	discovery
Pointer confirmation	2 new seeds	model training	seed replication
Wider replication	3 new seeds	model training	width transfer
Operator factorial	6 paired seeds	causal runs; 3 before training	construction comparison
Re-analyses of six frozen checkpoints			
Probe/task replay	same 6 models	replay outcomes	measurement sensitivity
Cycle-preserving patch	same 6 models	corruption/probe outcomes; not L_2 match	validity audit
Coordinate decomposition	same 6 models	decomposition outcomes	measurement sensitivity
Boundary-only clock	same 6 models	retrospective	clock sensitivity

four calls. We read the trained LM-head exit after each call, and patch one-based layer 42, head 11 at the pre-output-projection site unless stated otherwise.

Exit-readout diagnostic. On the same frozen 600 confirmation pairs used below, we replayed the unpatched clean and corrupt prompts through all four trained LM-head exits. Table 4 reports answer top-1 accuracy and the clean-minus-corrupt gap in their answer-pair margins. The call-1 gap is positive despite low top-1 accuracy: an intervention can change answer logits at that trained exit without showing that the model already solves the three-link task there. The cohort contains 100, 230, and 270 pairs with corruption at links 1, 2, and 3, respectively. The diagnostic uses no new component selection or examples.

Table 4: Unpatched answer readout on the frozen Ouro confirmation pairs. Top-1 accuracy is for each trajectory's own answer; the gap is the clean minus corrupt mean margin for the same clean-versus-corrupt answer pair.

Exit after call	Clean top-1 (%)	Corrupt top-1 (%)	Margin gap (logits)
1	37.5	34.2	1.35
2	62.2	62.5	3.22
3	86.3	85.5	5.14
4	100.0	100.0	6.09

The Ouro component was selected only on the discovery split. We first compared attention and MLP outputs across recurrent passes and found the first reproducible family-level path in attention during pass 4. Non-overlapping eight-layer windows then identified layers 41–48; refinement to layers 41–42 and a single-layer comparison attributed the effect to layer 42 rather than layer 41. Only layer 42 was subsequently opened into its 16 pre-output-projection attention heads.

The selection statistic fixes the recipient intervention at call 4 and compares terminal recovery when the inserted clean activation comes from call 4 versus call 1 of the same example. It does not compare interventions performed at calls 4 and 1. Head 11 had discovery specificity 0.115 with a 95% interval of $[0.099, 0.130]$; head 13 was a distant second at 0.018, and the remaining heads were near zero. Head numbers in the paper are one-based, so head 11 corresponds to array index 10. We froze this site before generating the confirmation cohort. Among 600 fresh eligible pairs, its specificity was 0.099 with a 95% interval of $[0.088, 0.109]$. No component was reselected on confirmation data.

The horizon follow-up uses the same 600 confirmation pairs. At the pre-o_{proj} site we either replace H11 only at the final query token or at every aligned token position, then read the answer contrast after the patched pass and after R4. The all-position R1 estimates are 0.730 ($[0.666, 0.796]$) locally and 0.000 ($[-0.004, 0.005]$) terminally. A separate R1 sweep applies the final-token patch to all 16 layer-42 heads. H11 ranks first in attenuation; 13/16 point estimates are positive, but the median is 0.001 and the second largest is 0.146 logits. Thus the attenuation control survives broader token scope, while the selected head's local magnitude is explicitly atypical.

Joint call-1 horizon and donor replay. Before this follow-up, we fixed H11, the final-query-token patch at call 1, the same 600 recipients, and the clean-minus-corrupt answer margin at all four loop ends. The call-1 and terminal matched means exactly reproduce the earlier horizon run. For each recipient, target-unrelated donors are natural clean prompts from other rows with the same corruption stage and a final answer distinct from both recipient answers. Among eligible donors, we choose the H11 vector whose L_2 distance from the corrupt recipient most closely matches the paired clean vector's distance; a deterministic first-eligible donor is a sensitivity check. Neither rule inspects the patched response. All 600 recipients have eligible donors, the nearest donor's own final target is predicted correctly in all 600 cases, and its median unrelated/matched displacement ratio is 1.139 (90th percentile 2.35). The two prompts still differ in more than target semantics, and vector directions and patched-state support remain uncontrolled. This donor replay reuses the confirmation cohort; it is not an untouched test.

Table 5: **One H11 call-1 patch, four readout boundaries.** Entries are mean clean-minus-corrupt answer-margin effects (logits) across 600 recipients. Donors are cross-example natural prompts; the matched-minus-nearest call-1 paired-bootstrap 95% interval is $[0.442, 0.533]$.

Donor / contrast	After call 1	After call 2	After call 3	At output
Matched clean	0.721	-0.002	0.003	0.001
Target-unrelated, nearest L_2	0.235	-0.009	0.000	0.001
Matched $-$ nearest unrelated	0.487	0.007	0.003	0.000
Matched $-$ first eligible	0.401	0.010	-0.002	0.000

The matched-minus-nearest call-1 remainder is positive at each corruption stage (0.072, 0.193, and 0.890 logits for stages 1–3), though most of the aggregate effect comes from stage 3. The all-position control above tests patch scope; the donor replay tests target relation. They are different controls and should not be combined into a single target-specific estimate.

A.3 Pointer denominator and phase assignment

Table 6 reports the complete denominator of eight primary final-answer-only pointer runs used to define the canonical subgroup. The behavior-only canonical rule requires mean final accuracy $\geq 95\%$, minimum per-depth final accuracy $\geq 90\%$, and mean final-answer accuracy at the requested-depth exit $\geq 85\%$. It was fixed before untouched seeds 137 and 223 were trained; the earlier development runs remain exploratory. No causally measured fingerprint enters this classification.

Table 6: Complete primary pointer-run denominator. All accuracies are percentages; "canonical" means that all three behavior-only criteria are satisfied.

Seed	Selected step	Mean final	Min. depth	Depth-exit acc.	Classification
7	4,000	100.0	100.0	100.0	canonical
42	8,500	97.9	83.4	60.3	behavior gate fail
314	3,000	100.0	100.0	99.7	canonical
11	3,000	100.0	100.0	99.7	canonical
29	5,000	100.0	99.8	99.2	canonical
71	9,000	100.0	100.0	22.8	noncanonical boundary
137	4,500	100.0	100.0	99.6	canonical confirmation
223	6,500	100.0	100.0	95.2	canonical confirmation

Phase labels implement Equation 4 directly with threshold 1.0. They use the gold final symbol and baseline answer logits, are recomputed for each example, and contain no learned phase classifier. Unresolved rows do not enter phase-conditioned donor pools or causal contrasts. Table 7 reports the prespecified intervention cohorts. The small excluded fraction includes Unresolved states and rare samples without an eligible same-phase donor with distinct targets. This coverage audit used 256 examples per depth for development seeds 7 and 314 and 128 for the other four models, accounting for their approximately twofold count difference. It is a donor-availability diagnostic, not the common 256-example-per-depth vector-resampling denominator in Table 2.

Table 7: Phase and donor coverage before head replication. Every recurrence has at least two phases in every seed. Counts are among eligible states.

Seed	Active	Commit	Hold	Eligible	Excluded
7	7,157	2,042	7,173	16,372	0.073%
314	7,055	2,037	7,268	16,360	0.146%
11	3,572	1,009	3,580	8,161	0.378%
29	3,560	1,011	3,595	8,166	0.317%
137	3,538	1,020	3,622	8,180	0.146%
223	3,600	1,019	3,563	8,182	0.122%
Total	28,482	8,138	28,801	65,421	0.175%

Using the same evaluation seed and 256 examples per requested depth, we directly compare model-derived labels with the task schedule before intervention. Agreement averages 98.7% over all 98,304 seed–example–recurrence states ($[98.1, 99.3]$ across seed means) and 98.7% after excluding Unresolved model states; seed-level agreement ranges from 97.5% to 99.8%. Similar marginal phase counts therefore reflect row-level agreement, not just similar phase proportions.

The stored complete progress cohorts for four pre-confirmation seeds also permit a label-only threshold audit. Relative to the primary threshold 1.0, lowering the stable-margin threshold to 0 preserves 94.3–98.0% of sample–recurrence labels; raising it to 2 preserves 97.8–99.8%. This checks label stability only: donors and interventions were not reselected, so we do not present it as an independent causal replication.

A.4 Output-independent phase sensitivity

Because the primary phase label uses the baseline answer margin, we perform a retrospective sensitivity analysis on the already frozen causal rows. Pointer depth supplies a model-output-independent schedule: recurrence r is Active when $r < d$, Commit when $r = d$, and Hold when $r > d$. We replace the primary label with this task-defined label and repeat the same leave-one-depth-out phase-versus-loop prediction. Gold-schedule phase has lower MSE than absolute loop for every head in all six canonical models (24/24 heads). The mean phase/loop ratio across seed means is 0.373, with a seed-bootstrap 95% interval of $[0.321, 0.429]$.

This analysis does not use answer logits, attention, or the tested component to assign phase. A separate head-independent check corrupts individual input edges and measures when each path edge becomes causally available. In fresh canonical seeds 11 and 29, the half-maximum frontier differs from one-edge-per-pass progress by 0.000 and 0.078 edges on average, respectively. These checks support the canonical advancing schedule without turning it into a claim that Active, Commit, and Hold form a universal latent-state ontology; seed 71 remains a counterexample.

A.5 Independent readout and component-site sensitivity

This subsection isolates readout, donor-definition, and component-site changes. The following subsection then holds each resampling intervention fixed and performs the coordinate decomposition used for the main interpretation.

The earlier model-favored coordinates use symbol margins from the model's own output projection. To distinguish that representation from a validated intermediate-state decoder, we audit its unintervened argmax against the exact gold trajectory state at every boundary. Table 20 reports cell-balanced means under the output-independent schedule. Across the six canonical seeds, accuracy is only 2.0% in Active, but 98.9% at Commit and 100.0% in Hold on this earlier output-head audit cohort; Table 1 uses the separate probe-test cohort (2.1% Active). The model therefore does not simply emit each intermediate answer during Active, and this output-head fingerprint alone does not establish a decoded latent frontier.

We next fit a frozen-state linear readout. For each canonical checkpoint, one multinomial probe is trained to predict the exact gold trajectory state from the output-normalized recurrent state. Probe-training, held-out probe-test, resample, and matched-patch examples use four disjoint seeds; the

base model remains frozen. The held-out probe reaches 99.2%/96.8%/100.0% mean accuracy in Active/Commit/Hold, versus 2.1%/99.1%/100.0% for the original head on the same cohort (Table 1). Across three fixed optimization initializations per model, mean Active accuracy remains 99.2%, the largest within-model range is 0.08 percentage points, and held-out predictions agree in 99.83% of cases on average. This checks optimization stability, not independent base-model replication.

We replay every recurrence and attention head under both interventions. Unlike the earlier output-head/model-favored coordinates, this sensitivity fixes predecessor, current, next, and final symbols directly from the gold trajectory, so neither readout defines the tested coordinates. This creates an explicit decomposition. Moving from the earlier output-head/model-favored analysis to output-head/gold coordinates retains a 23.6% raw incremental phase improvement, versus 39.9% in the earlier analysis. Holding the gold coordinates fixed and replacing the output head with the probe retains a 20.2% improvement. Thus coordinate choice changes the effect size, whereas the readout swap itself changes the estimate only modestly.

The prespecified endpoint for this replay was the raw attention-head incremental phase improvement; normalized fingerprints and matched patching are secondary checks, and the whole-MLP site is an extension. Table 8 summarizes the replay whose phase and donor selection use model-derived phase and prediction-matched donors. With the independent probe, phase beats absolute loop for raw and normalized attention fingerprints in all six seeds. Against relative clock plus contemporaneous confidence, adding phase reduces held-out error by 20.2% (seed-bootstrap 95% CI $[13.9, 25.8]$; 6/6 seeds) for raw fingerprints and 12.9% ($[6.0, 19.7]$; 5/6) after normalization. Phase alone is only approximately tied with that combined baseline after normalization (mean ratio 0.983; 3/6), so the incremental rather than dominance claim is the appropriate strong-baseline conclusion. Matched-patch contrasts remain positive in 6/6 seeds and 23/24 heads.

The same replay replaces the complete output of the shared MLP at one recurrence. Under the independent probe, raw and normalized phase/relative-confidence MSE ratios are 0.715 and 0.776, both in the predicted direction for 6/6 seeds. Adding phase reduces error by 31.9% ($[23.2, 40.9]$) and 12.9% ($[7.8, 21.1]$), respectively; matched-patch contrasts are positive in 6/6 seeds.

Because that replay still uses model-derived phase labels and output predictions to match donors, we add a retrospective sensitivity in which $r < d$, $r = d$, and $r > d$ define phase and donors differ in gold current and final states. No output-head quantity then enters the phase label, donor match, or fingerprint coordinates. Under the probe, attention raw/normalized phase-to-absolute-loop ratios in the role-aligned representation are $0.240/0.301$; adding phase to relative clock plus confidence reduces error by 41.4% ($[37.0, 46.1]$) and 34.4% ($[25.6, 43.3]$), respectively, with all six seed directions and all 24 matched-patch head contrasts positive. The whole-MLP raw result also persists (53.5% incremental reduction, 6/6), as does its patch contrast (6/6). Its normalized incremental reduction is only 5.3%, however, with a seed-bootstrap interval crossing zero ($[-27.1, 29.5]$; 5/6), so we treat the MLP result as a limited component-boundary extension rather than evidence for an independent MLP mechanism or neuron-level specialization. Because task-schedule phase is a function of $r - d$, these gains over a linear clock do not identify phase information beyond relative time.

Table 8: Readout, coordinate, and component-site sensitivity across six canonical seeds. P/A is phase/absolute-loop MSE; ΔRC is the fractional error reduction from adding phase to relative-clock plus confidence. Raw/Norm denote unnormalized and unit-normalized fingerprints. Patch reports positive seed-level contrasts. "Model" uses model-derived phase and prediction-matched donors; "Task" uses task-schedule phase and gold-state-matched donors.

Protocol	Site	Readout	Raw P/A	Norm P/A	Raw ΔRC	Norm ΔRC	Patch
Model-derived	Attention heads	Output head	0.634	0.748	0.236	0.134	6/6
Model-derived	Attention heads	Linear probe	0.640	0.630	0.202	0.129	6/6
Model-derived	Whole MLP	Output head	0.530	0.627	0.172	0.217	6/6
Model-derived	Whole MLP	Linear probe	0.566	0.761	0.319	0.129	6/6
Task-schedule	Attention heads	Linear probe	0.240	0.301	0.414	0.344	6/6
Task-schedule	Whole MLP	Linear probe	0.179	0.610	0.535	0.053	6/6

The seed is the inferential unit throughout; head counts and the multiple related raw/normalized endpoints are descriptive robustness checks, not independent replications. With only six seeds, bootstrap intervals describe variation among these selected solutions, not population uncertainty.

A.6 Representation decomposition and coordinate sensitivities

The coordinate audit computes six margin representations from each identical finite resample before any cell aggregation. "Role" uses the four semantic slots; "deweighted" divides each occurrence of a symbol repeated m times by $\sqrt{m}$, so that symbol has unit total squared weight. "Path-9" orders nine unique positions from the start. Fixed-ID all-symbol retains vocabulary order, whereas canonical all-symbol orders all 16 symbols by cycle position relative to the start. The sixth representation rotates this non-repeating vector so coordinate zero is the gold current state. Because its rotation is fixed within each (d, r) cell, rotating the cell mean is algebraically identical to rotating examples before averaging. If predictions are rotated by the same cell-specific permutation $P_{d,r}$, $\|P_{d,r}\mathbf{f} - P_{d,r}\widehat{\mathbf{f}}\|_2^2 = \|\mathbf{f} - \widehat{\mathbf{f}}\|_2^2$. Refit restricted predictors can nevertheless have different errors in the two alignments. Thus role/deweighted tests duplicate weight, fixed/canonical tests cross-example alignment, and canonical/frontier tests anchoring without changing coincidence.

Table 9 gives the paired model-phase replay. Deweighting barely changes the capacity-matched increment, whereas Path-9 and start-anchored canonical all-symbol remove it; frontier anchoring recovers a smaller gain. Fixed-ID all-symbol has substantially lower aggregation retention, $\|\mathbb{E}u\|/\mathbb{E}\|u\|$, than the canonical order. Under the tail-pooled flexible clock no representation has a positive mean increment. The decomposition rules out duplicate weight alone and shows that anchoring, slot coincidence, semantic selection, and clock parameterization cannot be collapsed into one explanation.

Table 9: Paired coordinate-audit attention results with model-derived phase and the independent probe. ΔRC and ΔFC add phase to capacity-matched and flexible relative-clock/confidence models. Ret. is aggregation retention. Counts are seed directions; all rows use the same resamples.

Representation	ΔRC	ΔFC	Ret.
Role	0.202 (6/6)	-0.194 (2/6)	0.856
Role, deweighted	0.193 (6/6)	-0.185 (1/6)	0.844
Path-9	-0.040 (1/6)	-0.036 (1/6)	0.739
All-symbol, fixed ID	0.083 (4/6)	-0.152 (1/6)	0.464
All-symbol, start	-0.018 (2/6)	-0.036 (1/6)	0.701
All-symbol, frontier	0.092 (4/6)	-0.134 (2/6)	0.701

The role-aligned representation above changes its coordinate multiplicity with phase. At Commit, current, next, and final can denote the same symbol; during Hold, all four slots can coincide. Repeating one symbol changes both its weight and the direction of the fingerprint. We therefore define a fixed alternative coordinate set from the start symbol and the next eight states on the underlying 16-symbol single cycle. These nine symbols are unique for every example, and the set does not depend on requested depth, recurrence, phase, or either readout. We rerun all six models rather than attempting to transform the aggregated four-coordinate artifacts.

Table 10 shows a sharp estimand boundary. Under model-derived phase, the output-head raw phase/absolute ratio is near parity (0.983), and adding phase to relative clock plus confidence changes error by -2.3% (1/6 positive). In the unique-path replay, the independent probe gives 1.091 and -4.0% ($[-8.5, -0.2]$; 1/6), respectively. Its mean agrees with the independent coordinate-decomposition run in Table 9 to six decimal places; both round to -0.040 in the tables. Under task-schedule phase the probe result is still weaker: a raw ratio of 1.154 and a -13.1% increment ($[-24.7, -6.1]$; 0/6). Normalization does not restore the attention result. The role-aligned phase advantage is therefore not invariant to a fixed non-repeating task-coordinate set.

The earlier all-symbol audit uses the same target-versus-best-alternative margin for all 16 symbols, so role, path, and full-space results differ in represented symbols rather than outcome transformation. Every symbol appears once within an example, but fixed symbol IDs do not align task roles across examples. Under both phase definitions, phase beats absolute loop for the raw probe vector in 6/6 seeds (P/A 0.806 model-derived and 0.584 task-defined). Its increments over relative clock plus confidence are small and uncertain: 8.3% ($[-2.8, 19.8]$; 4/6) and 6.2% ($[-1.7, 15.0]$; 3/6). A separate centered-logit sensitivity gives negative mean increments (-7.6% and -8.8%), but is not used to isolate representation because it also changes the outcome transformation.

For the earlier centered-logit all-symbol response (not the target-versus-best- alternative margins in Table 9), replacing the linear clock with categorical $(r - d)$ and quadratic r/d, then adding model-derived phase, changes raw probe error by -12.4% ($[-27.0, 1.6]$; 2/6 positive) and normalized error by -7.7% ($[-14.8, -0.7]$; 2/6). We do not treat the task-schedule analogue as added phase information because its labels are algebraically determined by $(r - d)$. This is a deliberately flexible sensitivity analysis, not a new primary model or a replacement for the capacity-matched comparison.

This diagnostic does not make the role-aligned estimand invalid: predecessor, current, next, and final are meaningful computational roles. It instead narrows the claim to that alignment. Matched patching is unchanged by the vector-coordinate replacement and remains positive in 6/6 seed means and 24/24 attention heads under the task schedule; the routing and ablation case study likewise uses direct edge attention and a single current-state contrast. These results support clean-directed influence, grouping, and margin sensitivity, but do not identify a complete causal operator.

Table 10: Attention fingerprints in fixed non-repeating readouts. "Path" is the nine-symbol unique-path coordinate set; "All" is the all-symbol margin response. P/A and ΔRC follow Table 8; Patch is unchanged by the vector representation and distinguishes the estimands.

Basis	Phase/donor	Readout	Raw P/A	Norm P/A	Raw ΔRC	Norm ΔRC	Patch
Path	Model	Output	0.983	0.945	−0.023	0.010	6/6
Path	Model	Probe	1.091	1.174	−0.040	−0.028	6/6
Path	Task	Output	1.040	0.972	−0.127	−0.037	6/6
Path	Task	Probe	1.154	1.254	−0.131	−0.088	6/6
All	Model	Output	0.608	0.643	0.084	0.102	6/6
All	Model	Probe	0.806	0.866	0.083	0.005	6/6
All	Task	Output	0.377	0.505	0.261	0.175	6/6
All	Task	Probe	0.584	0.774	0.062	−0.059	6/6

A.7 Cycle-preserving corruption and donor-control audit

The original corruption exchanged two outgoing targets. For a single directed cycle, cutting two edges and reconnecting them crosswise necessarily produces two cycles. We therefore repeat the complete six-model scalar audit with a three-edge cycle-ordered rotation: one cut source lies on the requested path, two lie off it, and their targets are rotated in the order in which the sources occur on the cycle. An automated invariant check verifies one cycle for all 6,144 generated corrupt maps. Unit tests also verify that the former two-edge operation produces two cycles and that the replacement preserves the source and target permutations.

The probe is trained only on disjoint clean trajectories and is not refit for this audit. On unpatched corrupt trajectories its mean current-state accuracy is 99.0% in Active ($[97.0, 100.0]$), 96.6% at Commit ($[95.3, 97.7]$), and 100.0% in Hold ($[99.9, 100.0]$); final-state accuracy is 99.7%. Intervals resample the six training seeds. These checks validate task support and readout decodability before intervention, but do not certify that patched hybrids are on-manifold.

Table 11 reports the primary effect and the target-unrelated control. For each head, the control donor comes from another example whose current and final symbols differ from the recipient's clean targets. Hence it lacks the intended target match. We select the eligible donor whose L_2 displacement from the corrupt recipient is nearest to the displacement of the matched donor. Overall mean distances are 4.04 and 3.96 (ratio 1.020); corresponding cosine distances are 0.731 and 0.693, so the match controls magnitude closely but not direction. The two target-logit columns show why the control is essential: both donors suppress the corrupt target, whereas only the matched donor raises the clean target. A paired reanalysis of the stored head–depth–call cells makes this qualification precise. The six-seed mean matched-minus-unrelated difference decomposes into 2.274 logits of extra clean-target promotion ($[1.834, 2.529]$) and 0.028 logits of extra corrupt-target suppression ($[0.012, 0.043]$). The six-seed means of clean-promotion and total matched-minus-unrelated differences, formed after pooling heads within each seed, are positive in all 64 depth–call cells; these correlated cells are descriptive, not 64 independent replications. Cellwise suppression differences have both signs (range $[-0.072, 0.135]$ logits), so "equal suppression" is only a useful aggregate approximation. The L_2 control also does not match donor direction or certify equal hybrid support. The matched-minus-

unrelated pairing retains a positive effect in every seed. For grouping, matched, unrelated, and their paired difference are respectively 4.92 ($[3.89, 6.13]$), 2.58 ($[2.02, 3.15]$), and 2.35 ($[1.89, 2.97]$) logits, each positive in 6/6 seeds. Using normalized recovery instead of logits gives matched and unrelated grouping contrasts of 0.199 and 0.101; their paired difference is 0.097 ($[0.085, 0.114]$), positive in all six seeds. The matched-specific fraction $(M-U)/M$, computed within each seed before aggregation, is 49.5%, 51.9%, and 60.2% in Active, Commit, and Hold. The Hold–Active difference is 10.7 points ($[6.2, 14.7]$; 6/6 seeds) on the raw scale and 10.4 points ($[6.2, 13.9]$; 6/6) after gap normalization. This rejects a strictly phase-constant share within the selected family, but Hold has the smallest absolute effects and the ratios do not identify a clock-independent phase. The unrelated-donor concept was specified before this audit, but nearest-L_2 selection was added after inspecting the initial control distances; it is a targeted retrospective sensitivity rather than a confirmatory experiment.

Table 11: Single-cycle scalar audit, averaged over six seeds. Effects and centered target contributions are logits; M/U recovery divides by eligible clean–corrupt gaps and is shown as a fraction.

Phase	Matched	Unrelated	Difference	M clean up	M corrupt down	M recovery	U recovery
Active	6.08	3.08	3.00	2.91	3.17	0.247	0.123
Commit	7.45	3.61	3.84	3.74	3.71	0.246	0.118
Hold	2.47	0.97	1.51	1.43	1.04	0.052	0.020
Overall	4.28	1.98	2.30	2.22	2.07	0.138	0.065

Table 12: Seed-level overall scalar patch effects (logits) under the cycle-preserving corruption. The primary within-seed contrast is matched minus target-unrelated. Seeds 137 and 223 are the two behavior-gated confirmation models; nearest-L_2 donor selection is a retrospective measurement sensitivity on all six frozen models.

Seed	Matched	Unrelated	Matched − unrelated
7	4.519	2.023	2.497
11	4.671	2.165	2.506
29	4.551	2.104	2.447
137	4.764	2.265	2.499
223	2.344	1.096	1.248
314	4.861	2.247	2.614

The mean within-seed contrast is 2.302 logits (six-seed standard deviation 0.519). A two-sided Student-t interval on these six seeds is $[1.757, 2.847]$; the exact sign test gives $p = 0.03125$. The former assumes approximately normal seed-level variation, which cannot be checked well with $n = 6$; both summaries describe this selected solution family, not a population of architectures.

Donor-selection sensitivity. We replayed the identical checkpoints, generated examples, probe training, and local contrast with the first eligible donor whose current and final targets differ from the clean targets, instead of choosing the eligible donor closest in L_2 displacement. The matched cells agree with the original run to 10^{-5}; all unrelated-cell identities and counts agree exactly. This deterministic rule is an unmatched donor sensitivity, not a uniform random-donor test or a new model replication. Table 14 shows that the donor-conditioned remainder and progress-aligned grouping persist, while the mean donor displacement rises from 4.04 to 4.46. Unrelated-donor grouping is positive in 6/6 seeds for each rule; the first-eligible mean is 2.46 logits with a seed-bootstrap interval $[1.90, 3.01]$. The control thus does not owe its qualitative sign to nearest-L_2 selection, but its magnitude and hybrid-state support remain donor-rule-dependent.

What the unrelated donor writes. We retrospectively replayed the original nearest-L_2 donor on the same recipients and frozen probes, retaining rows where the donor's own current state differs from both recipient states (94.1% of eligible observations). The donor's unpatched own state is probe top-1 in 98.6% of Active, 96.4% of Commit, and over 99.9% of Hold observations on this subset. After patching, its own-state mean-centered logit rises by 2.26 (seed-bootstrap interval $[1.78, 2.54]$), while the recipient corrupt-state logit falls by 2.17 and the clean-state change is -0.06.

The positive clean–corrupt contrast therefore combines corrupt-state suppression with promotion of a donor-associated state; it does not imply the unrelated donor carries the recipient's clean answer. The donor-conditioned remainder remains an operational contrast under the specified donor rule, not a proof that the head changes its operation.

Where the probe points after patching. On the same donor-distinct recipient subset, we classify the independent probe's immediate top-1 state as recipient clean, recipient corrupt, donor-own, or another symbol. The three conditions in Table 13 share identical recipients and frozen probes. Every original full-cohort C2 cell count and mean is reproduced within 5.2×10^{-4}. Matched patches increase clean top-1 from nearly zero to 10.3%, whereas unrelated patches increase donor-own top-1 to 9.2%. The corrupt state remains top-1 in more than 86% of patched observations. Thus the logit decomposition has a corresponding, but limited, categorical readout signature. This retrospective replay neither validates hybrid-state support nor identifies a head operation.

Table 13: Immediate probe top-1 destination (%) on the common recipient subset where the unrelated donor's own current state differs from both recipient states (94.1% of eligible observations). Equal-seed means over six models; columns are mutually exclusive. Seed-bootstrap 95% intervals are $[9.3, 11.3]$ for matched clean and $[7.9, 10.6]$ for unrelated donor-own. This retrospective replay is not independent model replication or native task-success measurement.

Recipient condition	Clean	Corrupt	Donor-own	Other
Unpatched corrupt	0.0	99.1	0.1	0.8
Matched patch	10.3	86.6	0.2	2.9
Unrelated patch	0.1	87.1	9.2	3.6

The heatmap's completion-adjacent band is also sensitive to which corrupted edge has already been traversed. Using the original eligible rows, a coarse stratification by calls since that edge (0, 1, or 2+) retains 60 comparable head–depth–lag strata per seed. The near-completion minus earlier-Active remainder is 1.47 logits after equal-stratum and equal-seed averaging ($[1.23, 1.68]$), versus 1.38 without stratification. The band therefore persists under this coarse lag balancing. Neither lag nor proximity to completion is randomized, so this does not identify a phase-specific head function.

Table 14: Retrospective donor-rule replay on the same six models. Patch effects, grouping contrasts, and L_2 displacements are six-seed means. The matched effect is 4.28 logits under both rules; percentages are unrelated/matched.

Unrelated donor	Effect	Share	Remainder	Grouping	Mean L_2
Nearest-L_2	1.98	46.3%	2.30	2.58	4.04
First eligible	1.94	45.4%	2.35	2.46	4.46

Perturbation and cross-context donor controls. To test whether the unrelated-donor result merely reflects patch magnitude or same-schedule donor selection, we retrospectively replayed the same six frozen models, probes, cycle-preserving recipients, heads, and end-of-call contrast. Matched and original unrelated cell counts agree exactly with the prior run; the largest numerical difference in a local cell mean is below 0.0006 logits. Each new control uses the same eligible recipient cells. Norm-matched noise adds an isotropic Gaussian direction to the corrupt head vector at the paired clean–corrupt L_2 displacement (one fixed draw). Mean ablation inserts the clean-activation mean for the depth, call, head, and 64-example batch. The cross-context control uses a clean natural donor from the next depth in cyclic order and a different call, with both current and final targets different from the recipient clean targets; among eligible candidates, it matches the paired displacement most closely. Candidate eligibility is complete in all six seeds.

Table 15: **Perturbation and cross-context donor controls.** Equal-seed means of local independent-probe effects and grouping contrasts (logits), centered target-logit changes, and donor–recipient L_2 displacement. "Suppress" is the negative corrupt-target change. The three new rows are retrospective sensitivities, not independent confirmation.

Replacement	Effect	Grouping	Clean promote	Corrupt suppress	L_2
Matched clean	4.28	4.92	2.22	2.07	3.96
Same-context unrelated	1.98	2.58	-0.06	2.04	4.04
Norm-matched noise	0.36	0.43	0.01	0.35	3.96
Within-batch mean	1.92	2.68	0.11	1.81	3.18
Cross-depth/call unrelated	2.03	2.70	-0.01	2.04	4.03

Across seeds, the noise effect is 0.36 logits ($[0.28, 0.41]$), versus 2.03 ($[1.65, 2.26]$) for the cross-context donor; their grouping intervals are $[0.35, 0.50]$ and $[2.10, 3.34]$, respectively. This rules out matched displacement magnitude alone as an explanation of the natural-donor result. The mean replacement has a shorter displacement and includes a 1/64 share of the recipient clean vector, so it is not a semantic negative control. Changing depth and call also changes donor context; similarity of its aggregate effect does not establish on-manifold hybrid support or a phase-independent function.

Frozen-head horizon–donor audit. The seed-137 paper head 2 is index 1 in the stored activation tensor. It was identified by a prior role-slot match, not by the following horizon–donor values (Appendix A.13). We reaggregate only that head's cycle-preserving scalar patch cells from the same evaluation cohort used for the six-model C2 result. Matched and target-unrelated cells have identical depth–call–phase keys and eligible counts, with 64 paired cells in total. Both horizons use the same independent probe and contrast *form*, but not always the same target symbols: the after-call margin compares clean and corrupt current states $z_r, \tilde{z}_r$, while the output margin compares their final answers $y, \tilde{y}$. The patch, recipients, donors, and eligible rows are fixed. Thus this paired table describes two task readouts, not an isolated change of horizon. The donor-conditioned remainder subtracts unrelated from matched within each readout. Table 16 makes the local/terminal distinction on one site and donor contrast explicit. The paired rows are correlated examples and cells in one selected model, not independent replications. The separate routing and zero-ablation illustration uses another cohort and intervention; its numbers are not treated as stages of this scalar patch.

Table 16: Descriptive C1/C2 audit of one frozen pointer head (seed 137, paper head 2). Effects are logits. "Remainder" is matched minus target-unrelated; after-call and output values use the same patch and recipient cohort, with current-state and final-answer target symbols respectively.

Phase	Paired rows	Matched after	Matched output	Rem. after	Rem. output
Active	1,805	13.63	2.88	7.25	0.15
Commit	1,024	31.60	36.12	16.51	19.39
Hold	3,584	8.57	6.16	5.72	4.59

Fixed-target horizon replay. We replayed both donors on the same 1,805 Active rows, fixing the target-symbol pair across readout boundaries without refitting the probe. Table 17 shows that the current-state pair is decoded only after the call, and the final-answer pair only at output. Logit shifts at the other boundaries remain measurable, but do not by themselves establish a decoded task contribution or isolate a change in that contribution. At Commit and Hold, the two targets coincide, so Table 16 does hold target identity fixed.

Table 17: Selected seed-137 head, Active rows: fixed target symbols at both horizons. Top-1 accuracy is for the clean/corrupt target on the corresponding unpatched trajectory. The gap is the clean-minus-corrupt margin difference; the remainder is matched minus unrelated patch effect (logits). Rows with near-zero target accuracy measure logit sensitivity, not successful target readout at that boundary.

Fixed target	Readout	Top-1 clean/corrupt (%)	Gap	Remainder
Current state	After call	99.9 / 99.9	26.39	7.25
Current state	Output	0.06 / 0.11	−2.40	4.88
Final answer	After call	0.00 / 0.00	0.80	−0.35
Final answer	Output	99.8 / 99.5	52.16	0.15

A.8 Two-edge corruption sensitivity

This earlier analysis is retained for auditability but is not the primary scalar evidence because its two-edge corruption splits the map into two cycles. The absolute-distance grouping contrast discards sign, so a signed sensitivity separately aggregates the signed local effect already stored in every eligible patch cell. Positive means that the clean activation increases the current clean-state versus corrupt-state contrast in the corrupt recipient. Within seed and phase, cell means are weighted by eligible example count and pooled over heads, depths, and recurrences; the six training seeds remain the replication units. Table 18 shows that every seed mean is positive in every phase. The overall gated normalized recovery is 13.7% ($[13.2, 14.0]$; 6/6), but its phase pattern is not substituted for the separately defined grouping statistic.

Table 18: Signed local patch effects in logits. Recovery uses only rows with an absolute clean–corrupt gap of at least one.

Seed	Active	Commit	Hold	Overall recovery
7	7.250	7.416	2.312	0.141
11	6.022	8.721	2.835	0.141
29	6.144	6.724	2.876	0.142
137	5.400	6.504	2.858	0.134
223	3.348	4.743	1.014	0.128
314	5.419	6.998	3.024	0.135

As a sensitivity check on grouping rather than sign, restricting Equation 5 to Active and Commit yields 6.30 logits ($[4.53, 8.55]$; 6/6 seeds). We also permute phase labels among depth cells within each head and recurrence, preserving each stratum's cell count and weights before recomputing the statistic. Every observed seed mean exceeds its permutation 97.5th percentile. This is a descriptive calibration—task phase is not exchangeable with depth under the data-generating process—and is not used as a causal test of a latent phase variable.

A.9 Scalar matched-patch overlap audit

Across the six model-derived patch cohorts there are 49,152 candidate sample–recurrence rows; 38,478 have different clean/corrupt local and final states, and 33,363 also share a resolved phase (8,153 Active, 4,887 Commit, 20,323 Hold). For a scalar clock adjustment we aggregate raw independent-probe effects into (head, depth, recurrence, phase) cells, record mean contemporaneous output-head answer margin and mean absolute clean–corrupt gap, and require at least ten rows. We then ask whether each (head, depth, recurrence) stratum contains at least two phases, so a phase coefficient is supported rather than extrapolated. The supported fractions for seeds 7, 11, 29, 137, 223, and 314 are respectively 0, 0, 1.6, 0, 13.1, and 0%. Our predeclared 50%-per-seed overlap rule therefore fails and no adjusted phase coefficient is interpreted. This audit concerns the model-derived scalar patch analysis; task-schedule phase is deterministically defined by $(r - d)$ and cannot independently identify a phase effect either.

A.10 WIDER-MODEL REPLICATION FIXED BEFORE TRAINING

To test transfer beyond the original model capacity, we fixed three new tied pointer models before training, doubling width from 128 to 256 and head count from four to eight. Parameter count rises from approximately 253K to 998K; all other task, recurrence, objective, optimizer, and behavior-gate choices match the final-only canonical setup. All three seeds pass the gate: their mean final accuracy is 99.99%, mean requested-depth Commit accuracy is 99.87%, and the worst per-depth final accuracy is at least 99.80%.

We then apply the task-schedule phase, gold-current donor, role-aligned gold-coordinate, and independent-probe protocol above without changing its primary endpoint. The raw attention phase increment over relative clock plus confidence is 31.8% ($[27.9, 36.9]$; 3/3 directions); after unit normalization it is 36.5% ($[31.6, 43.5]$; 3/3). Matched-patch seed means are positive in 3/3 models, with 20/24 individual heads positive. Table 19 compares this fixed-before-training extension with the original six models. Because $n = 3$ permits a minimum two-sided sign-test value of only 0.25, these results are directional architecture transfer evidence for the specified role-aligned measurement, not phase information beyond time or a new population-level confirmation.

Table 19: Task-defined, role-aligned attention-head sensitivity by controlled-model scale. P/A and ΔRC follow Table 8; parentheses in Patch report positive head counts.

Model	Params	Gate	Active probe	Raw P/A	Norm P/A	Raw/Norm ΔRC	Patch
Width 128, 4 heads	253K	6/6	99.2%	0.240	0.301	0.414/0.344	6/6 (24/24)
Width 256, 8 heads	998K	3/3	99.9%	0.304	0.336	0.318/0.365	3/3 (20/24)

A.11 COMPLETE-DENOMINATOR SENSITIVITY

The same audit exposes all eight primary runs rather than analyzing only the canonical subgroup. Requested-depth final-answer accuracy and mean-head phase advantage vary continuously together (Pearson $r = 0.949$, exact two-sided permutation $p = 0.012$; Spearman $\rho = 0.719$, $p = 0.052$). This relationship is descriptive: schedule adherence and causal organization are properties of the same learned solution, and this is a retrospective sensitivity rather than a prespecified test.

Table 20: Output-head readout and phase organization for the complete pointer denominator. A/C/H are gold-trajectory argmax accuracies in task-defined Active/Commit/Hold cells; Adv. is mean-head phase advantage. Asterisks mark the six behavior-defined canonical solutions.

Seed	Active	Commit	Hold	Adv.
7*	2.8	100.0	100.0	0.692
42	3.6	60.3	93.7	0.084
314*	0.2	99.7	100.0	0.494
11*	3.5	99.7	100.0	0.559
29*	4.1	99.2	100.0	0.401
71	17.2	22.8	48.3	−0.230
137*	0.1	99.6	100.0	0.541
223*	1.6	95.2	100.0	0.659

Table 21: Held-out behavior and role-coordinate fit in controlled pointer models. Phase-favoring heads and role matching are descriptive in the stated coordinates, not evidence of a phase variable beyond the clock; lower matching cost is better.

Seed	Final acc.	Depth-exit acc.	Phase-favoring heads	Match / null
137	100.0	99.6	4/4	0.069 / 0.522
223	100.0	95.2	4/4	0.130 / 0.499

A.12 SEED-LEVEL FACTORIAL EFFECTS

This supporting 2×2 comparison crosses cyclic-translation versus dihedral-affine operator banks with absolute-position versus current-state table selection. It asks which task constructions covary with the measured statistics; it does not identify noncommutativity or selection as a fixed-weight cause. The paragraphs following the tables give the competence gates and timing of the two three-seed cohorts.

Table 22: Paired factorial main effects by training seed. The operator effect is positive for both phase advantage and matched-patch contrast in all 6/6 seeds (exact two-sided sign test $p = 0.03125$ for each). The heterogeneous selection effect on patching is shown for comparison ($5/6$, $p = 0.21875$).

Seed	Operator: phase adv.	Operator: patch	Selection: patch
107	0.022	0.344	0.822
653	0.114	1.436	0.622
809	0.082	0.439	1.146
557	0.267	0.942	0.844
683	0.375	2.175	−0.588
941	0.068	0.901	0.506

Table 23: Raw cell means for the six-seed 2×2 factorial. Rows vary the operator family and columns vary operator selection. All metrics use higher values for stronger relative-progress organization under the stated measurement.

Operator family	Selection	Phase adv.	Resample	Matched patch
Cyclic translation	Position-selected	0.193	0.065	0.143
Cyclic translation	State-selected	0.286	0.083	0.885
Dihedral affine	Position-selected	0.416	0.122	1.365
Dihedral affine	State-selected	0.372	0.087	1.741

Table 24: Controlled algorithm comparison. Phase advantage is averaged over heads; Resample and Patch are descriptive positive-CI head counts. Higher values favor role alignment; seed-level effects carry family-level inference.

Condition	Seeds	Final acc.	Phase adv.	Resample	Patch
Cyclic composition	3	$\approx 100\%$	0.052	1/12	0/12
Translation	4	$\geq 99.9\%$	0.205	9/16	4/16
Adaptive random table	4	$\geq 99.9\%$	0.439	15/16	16/16

A.13 SUPPORTING ROLE-COORDINATE ANALYSES AND THEIR LIMITS

The figures in this subsection retain their originally specified role-aligned coordinates. They describe those measurements, not a role variable identified independently of the coordinate and clock checks in Section 6.

The direct seed-137 case was selected without using the routing or ablation values shown in Figure 5. We first learned four role slots from the earlier seed-7 normalized resample and matched-patch profiles. For each later model, we exhaustively evaluated the $4!$ assignments and froze the minimum-cost permutation; seed-137 head 2 occupied the seed-7 head-2 slot. Only then were its edge attention and zero-ablation responses summarized. This ordering makes the case a held-out mechanistic illustration of a pre-existing role match, while the six-seed patch analysis—not this one head—carries the replication claim.

The original role-aligned population analysis remains useful within its stated representation. In the two originally held-out confirmation models, all 8/8 heads favor phase over absolute loop and frozen role-set matching is substantially better than permutation controls. Along one seed-11 training trajectory, the phase/loop ratio is near one before task acquisition, falls to 0.522 when accuracy reaches 100%, and remains low thereafter. This is a temporal association from one trajectory, not a general learning-dynamics result and not evidence beyond the coordinate audits.

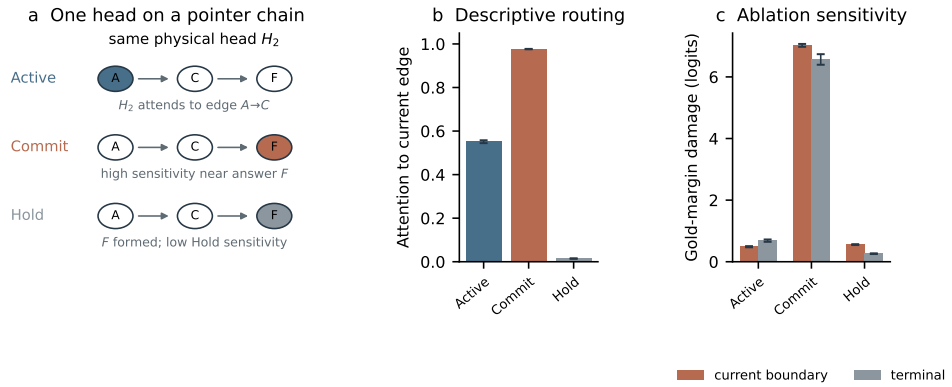

Figure 5: Preselected seed-137 head-2 illustration. Attention routing and zero-ablation sensitivity concentrate at the Commit boundary, while the same head is comparatively inactive in Hold. Selection used the frozen role match, not the values displayed here; the six-model matched-patch analysis supplies replication evidence within the selected solution family.

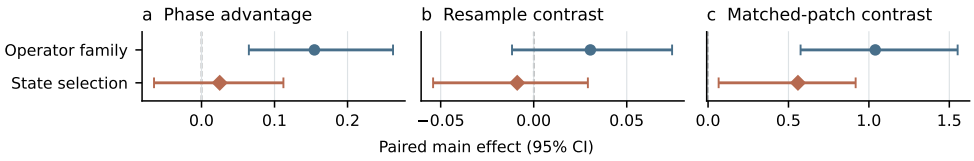

Figure 6: Paired main effects from the controlled 2×2 operator-family by selection extension across six training seeds. Positive values indicate stronger role-aligned phase predictability or patch-effect grouping. This is evidence for a construction-level association, not primary identification.

All controlled tasks use symbols sampled independently of semantic identity and uniform task depths. Pointer maps are single cycles, preventing short accidental orbits. The matched-family study uses 8 symbols, 4 complete shuffled tables, four recurrent passes, width 128, four heads, two development seeds (157 and 263), and two untouched confirmation seeds (401 and 509). Training uses 16K steps, batch size 256, cosine decay after 500 warmup steps, and a final-answer-only curriculum that raises the maximum sampled depth every 3K steps. Evaluation and causal cohorts use seeds 277 and 283; within-model bootstrap uses seed 293 and the paired seed-level bootstrap uses 317.

Matched-family competence gates. The final comparison followed two behavior-only calibrations whose causal branches remained closed. An initial 16-symbol protocol presented eight complete tables (128 facts): repeated lookup reached 100% accuracy, but translation averaged 17.9% and independent permutation composition 6.2% (6.25% chance). A uniformly reduced calibration then used 8 symbols, 4 tables, and 4 recurrences. Both state-selected lookup seeds reached 100%, whereas translation was seed-sensitive (100% versus 77.5%) and permutation stopped near a two-step frontier. Neither calibration met the common behavioral gate, so neither entered causal analysis. The final depth curriculum was fixed uniformly across conditions before training and was the first protocol eligible for the matched comparison reported in Figure 8.

Competence-gated untied comparison. To test whether independently parameterized depth spatializes roles that a tied model must virtualize, we froze a same-width, final-answer-only comparison using the untouched pointer seeds 137 and 223, the tied models' 12K-step optimization contract, and a 95% mean / 90% minimum-depth behavioral gate. The untied models reach only 74.1% and 78.1% mean final accuracy, with minimum-depth accuracies of 35.7% and 23.6%. Neither model meets the gate and they are not analyzed causally. We do not replace these models with an earlier behavior-matched untied pilot trained with intermediate-state supervision, because doing so would conflate

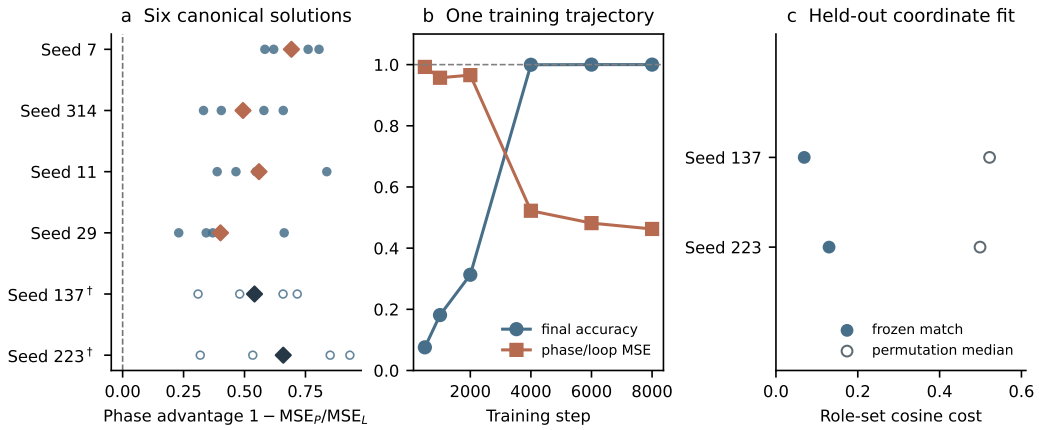

Figure 7: Supporting role-aligned analyses, conditional on these coordinates. (a) Per-head phase advantage and seed means; † marks originally held-out models (the plotted replay is retrospective). (b) Accuracy and role-aligned phase/loop prediction error along one training trajectory. (c) Frozen permutation-invariant role-set matching in those models.

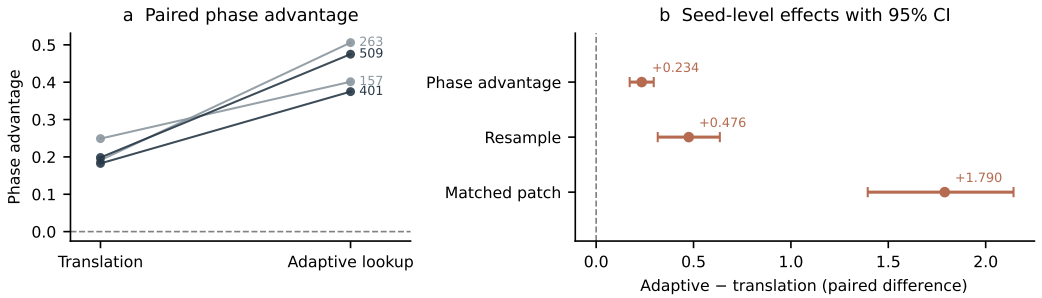

Figure 8: Prospective binary matched-family measurement after the prespecified behavior gate. (a) Phase advantage for four paired training seeds. (b) Paired adaptive-minus-translation effects with seed-bootstrap 95% intervals. Both conditions use the same model, input size, final-only curriculum, and recurrent budget; training seed is the replication unit.

architecture with the learning objective. This outcome limits the desired spatial-versus-temporal comparison; it is not evidence that untied models cannot spatially specialize.

The graded-adaptivity experiment uses three paired seeds and four nested doses $k \in \{0, 1, 2, 3\}$. All conditions use cyclic translation tables. The initial operator is fixed; among the remaining three choices, the first k use the newly computed state modulo four to select the next table, while the rest use their absolute step. The behavior gate and all causal and slope statistics were prespecified before training any dose.

The operator-by-selection factorial first uses paired seeds 107, 653, and 809; before further training, we designate paired seeds 557, 683, and 941 as untouched replications under the identical 16K-step final-only curriculum, behavior gate, causal cohort, and analysis. All 24 selected-checkpoint cells pass the prespecified behavioral gate. Because the common validation rule can select 12K or 16K checkpoints, we also evaluate the 16K milestone as a prespecified sensitivity analysis for the untouched seeds. All 12 fixed-step cells pass the same gate. The operator-family association survives both the untouched replication and fixed-step analysis; the state-selection patch effect does not exclude zero in either three-seed sensitivity. The factorial therefore associates learned solutions under a controlled training contract; it does not intervene on abstract algebra at fixed weights.

For resampling, donors share task depth and model-derived phase but have different current predictions and final states. For matched patching, we apply a task-native operator corruption, require clean

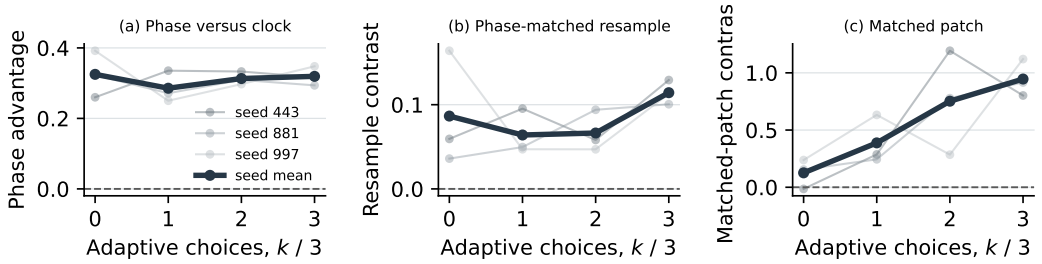

Figure 9: Frozen graded-adaptivity test in the stated response coordinates. Panel (a) displays phase advantage, $1 - \mathrm{MSE}_{\mathrm{phase}}/\mathrm{MSE}_{\mathrm{loop}}$. Thin colored lines are paired training seeds and black is their mean. All panels use higher values for stronger effects under the stated phase-indexed measurements. Only matched patching exhibits the prespecified direction in all seeds; the phase/loop and resample metrics are inconsistent with a common monotone dose trend in these three seeds.

and corrupt trajectories to share phase, and replace the corrupt pre-output-projection head value with its clean counterpart. Every reported local effect is read after the remainder of the current recurrent pass. Bootstrap intervals resample examples rather than individual recurrence rows.

The recurrent-Qwen audit uses the public rank-16 adapter of Shapiro (2026) with a Qwen2.5-0.5B-Instruct base model. It uses the release-order first 32 examples at each depth, same-depth donors with different chain states at every observed recurrence, and last-token replacements at attention/MLP outputs of recurrent layers 6, 9, 13, and 17. Its Coda readout is fixed before intervention.

B Pretrained Audit Gates

Ouro variable-hop feasibility. We reuse frozen three-link confirmation graphs, retaining graph edges and examples while changing only the final request to one, two, or three links. Stage-one divergence yields 396 complete triplets. At four recurrent passes, clean accuracy is 95.2%, 53.5%, and 62.1% by requested depth; only 31 graphs have both clean and corrupt answers correct at all three depths. Allowing stage-one or stage-two divergence yields 792 paired two-/three-link graphs. Clean accuracy is 57.8% and 69.9%, and 134 graphs satisfy joint clean/corrupt correctness and a contrast gap above one at both depths. On this R4-selected subset, the joint correct rate is only 4.5% at R2 for two-link requests, versus 67.9% at R3 for three-link requests. Thus this prompt family lacks a reliable requested-depth completion boundary, especially at two links. We do not use its selected subset to claim Ouro effects organize by $r - d$; a distinct behavior-valid elicitation is needed before such an intervention comparison is interpretable.

Huginn's mean held-out z_1/z_2 accuracy at R8/R10 is 30.9% for a recurrent-boundary linear probe, 33.3% after the Coda, and 32.5% for a 256-unit nonlinear probe. The best method improves by only 2.4 points and fails the prespecified 50% absolute and 10-point improvement thresholds. Its preceding four-block by attention/MLP resample pilot is therefore treated as underpowered.

For recurrent Qwen, the prespecified coarse audit covers attention and MLP outputs at recurrent layers 6, 9, 13, and 17. No site passes both the effect-size and positive-interval gates. Layer-13 attention is locally causal (effect norm 0.589), but its state-over-clock contrast is 0.062 with 95% CI $[-0.115, 0.160]$. Since intermediate-state decoding is nearly perfect, this is an observable coarse boundary rather than a readout failure; it is not a population-level null.

C State-Local Differential Diagnostic

For seed-137 head 2, we differentiate the previous, current, next, and final task-coordinate margins with respect to the pre-output-projection head activation on an independent 256-example-per-depth cohort. The recurrent-boundary logits are reconstructed exactly before differentiation. A leave-one-depth-out nested model gives a phase/loop MSE ratio of 0.420. After phase is known, adding loop reduces error by only 0.45%; adding phase after loop reduces it by 58.2%. Differential sensitivity and the independently measured finite-resample effect agree across 272 eligible cell–coordinate

pairs (Pearson $r = 0.831$; Spearman $\rho = 0.678$). The prespecified direct distance contrast using per-coordinate gradient norms is -0.00082 with 95% CI $[-0.00125, 0.00006]$ and is therefore inconclusive. A secondary $JJ^\top$ geometry has the predicted sign, but we do not promote it over the failed primary contrast. This one-head diagnostic compares a true local derivative with the finite response, but does not turn donor replacements elsewhere in the paper into Jacobian estimates.

D Relative-Clock, Confidence, and Magnitude Controls

For scalar grouping, we pool eligible rows over requested depths into an example-count-weighted mean $e_{h,r,p}$ for head h, recurrence r, and phase p. The first term of Equation 5 equally averages all available $|e_{h,r,p} - e_{h,r,p'}|$ pairs; the second equally averages all $|e_{h,r,p} - e_{h,r',p}|$ pairs. Heads are then averaged within seed. Thus task-schedule comparisons at the same recurrence draw phases from different requested depths; the contrast is unchanged if every effect reverses sign. For vector responses we additionally report the phase advantage $1 - \mathrm{MSE}_{\mathrm{phase}}/\mathrm{MSE}_{\mathrm{loop}}$ and nested depth-only, depth-plus-loop, depth-plus-phase, and combined predictors under leave-one-depth-out cross-validation. Numerical head identities are not matched across seeds; role sets use a fixed permutation-invariant assignment. Head-level positive-interval counts are descriptive rather than independent replications.

The capacity-matched relative clock adds $(r-d)/R$ and r/d to depth; confidence adds linear and quadratic asinh answer margin. The flexible clock instead uses categorical $(r-d)$ and linear/quadratic r/d. To avoid unestimated dummy levels under leave-one-depth-out evaluation, offsets $\pm(R-1)$ are pooled with $\pm(R-2)$. These are predictor comparisons under a stated response protocol, not exhaustive controls for temporal structure. For each head, an ordinary least-squares multi-output fit uses all other requested depths to predict the held depth. There is no learned regularizer or fold-fitted scaler: numeric features are defined by the formulas above, and categorical levels by the fixed eight-call domain. Each eligible response cell has equal weight; MSE averages squared errors over held cells and response coordinates, then heads are averaged within seed before seed-level summaries. Model phase enters as two indicators for Commit and Hold, with Active omitted.

Boundary-only clock sensitivity. To isolate the task boundary from the flexible clock's other features, a retrospective replay keeps the linear clock, confidence features, folds, cells, and least-squares fit fixed, adding only $\mathbf{1}[r-d=0]$ and $\mathbf{1}[r-d>0]$. The response cells and all six model checkpoints are the same as in the representation decomposition. Table 25 reports head-averaged increments summarized across seeds. The raw role-slot mean is -1.7% ($[-8.4, 6.5]$; 1/6 positive), compared with $+20.2\%$ over the linear clock. The fixed-ID and frontier-aligned raw means remain positive, but their seed intervals cross zero and their normalized means are negative. Sparse within-clock support still precludes phase identification.

Table 25: Retrospective boundary-only clock sensitivity on the frozen vector cohort. Entries are mean model-phase increments (%) with six-seed bootstrap 95% intervals, after adding only two task-boundary indicators to the linear clock. Ratios are averaged equally across heads within seed; positive seed counts for raw responses are shown in parentheses.

Response	Raw	Normalized
Role slots	-1.7 $[-8.4, 6.5]$ (1/6)	$+3.0$ $[-8.0, 15.6]$
Role deweighted	-0.8 $[-7.5, 6.0]$ (2/6)	$+2.7$ $[-7.4, 14.5]$
Path-9	-7.6 $[-12.0, -3.2]$ (1/6)	-5.5 $[-8.9, -2.1]$
Fixed symbol IDs	$+6.2$ $[-4.5, 18.9]$ (3/6)	-3.2 $[-10.0, 4.1]$
Start-aligned	-4.7 $[-9.4, 0.3]$ (2/6)	-4.9 $[-8.4, -1.4]$
Frontier-aligned	$+3.1$ $[-7.0, 14.5]$ (2/6)	-1.2 $[-8.8, 7.4]$

Pooling head MSEs *before* taking their ratio changes the increment (Table 26). The table includes the pooled *linear* baseline so the two clock comparisons use the same aggregation. Raw fixed-ID and frontier-aligned pooled seed-bootstrap intervals are positive, unlike their equal-head counterparts; normalization changes the pattern again. These are gains over a finite predictor, not identified phase information beyond time: within-clock model phase overlap is sparse, and this clock is not exhaustive.

For one seed, let $E_h > 0$ and $E_h^{+\phi}$ be head h's held-depth MSE before and after adding model phase. The two summaries are distinct estimands:

$$\Delta_{\text{equal}} = \frac{1}{H}\sum_h \left(1 - \frac{E_h^{+\phi}}{E_h}\right), \qquad \Delta_{\text{pooled}} = 1 - \frac{\sum_h E_h^{+\phi}}{\sum_h E_h} = \sum_h \frac{E_h}{\sum_j E_j}\left(1 - \frac{E_h^{+\phi}}{E_h}\right). \quad (10)$$

Pooling therefore weights larger-baseline-error heads more heavily; different signs are consistent with the distinct estimands.

Table 26: Alternative aggregation of the frozen clock fits: pool head MSEs within each seed before taking $1 - \overline{\text{MSE}}_{+\text{phase}}/\overline{\text{MSE}}_{\text{clock}}$. Entries are six-seed means (%) and seed-bootstrap 95% intervals. The linear and boundary columns use identical raw responses and weighting.

Response	Raw linear	Raw boundary	Normalized boundary
Role slots	$+31.0$ $[24.8, 39.8]$	$+12.3$ $[-1.2, 25.3]$	-2.7 $[-23.6, 14.5]$
Role deweighted	$+29.3$ $[21.1, 40.3]$	$+15.3$ $[-2.3, 33.0]$	-2.7 $[-24.7, 14.3]$
Path-9	-3.1 $[-5.9, -0.1]$	-2.6 $[-7.1, 1.3]$	-5.5 $[-9.1, -2.2]$
Fixed symbol IDs	$+12.5$ $[2.0, 24.8]$	$+19.3$ $[1.3, 37.1]$	-0.5 $[-5.3, 4.8]$
Start-aligned	$+0.2$ $[-4.8, 5.5]$	$+2.7$ $[-5.0, 10.3]$	-5.0 $[-8.8, -1.3]$
Frontier-aligned	$+16.5$ $[7.1, 28.2]$	$+18.5$ $[1.8, 35.2]$	$+0.8$ $[-6.6, 8.7]$

Table 27 exposes the raw role-slot fit by seed. MSE is averaged over heads, whereas each equal-head increment averages the head-specific MSE ratios. The sign flips in seeds 7, 11, and 29 when head errors are pooled first, because larger-error heads then receive more weight.

Table 27: Raw role-slot boundary-only comparison by training seed. The two MSE columns are for the boundary-augmented clock before and after adding model phase. The last two columns show distinct aggregations of the same fits.

Seed	Boundary MSE	+Phase MSE	Equal-head Δ (%)	Pooled Δ (%)
7	5.09	3.97	-3.0	$+22.0$
11	8.98	8.61	-12.7	$+4.1$
29	15.79	10.61	-0.8	$+32.8$
137	4.22	4.61	-9.2	-9.3
223	3.13	2.18	$+17.7$	$+30.1$
314	5.74	6.10	-2.4	-6.2

The clean trajectories drawn with the vector-resampling protocol (256 examples per depth, evaluation seed 947) have multiple resolved model phases at 2, 4, 7, 6, 6, and 4 of the 15 observed $r - d$ offsets for seeds 7, 11, 29, 137, 223, and 314, respectively. These counts come from the frozen phase labels in the phase-clock audit artifact; a mixed offset may have only one example in its minority phase. These counts show limited aggregate support, not guaranteed overlap within the vector predictor's head–depth–call cells. Among those actual predictor cells, 16, 32, 48, 40, 96, and 36 of 256 cells per seed contain more than one model phase; only 0, 0, 8, 8, 16, and 16 have at least 20 eligible examples *in each* of two phases. This threshold is a descriptive support diagnostic, not a fitting filter. Both counts use the same six-seed vector artifacts as Table 25.

An earlier output-head/model-favored-coordinate protocol gave a 31.6% raw increment from model phase after the flexible clock ($[4.5, 55.8]$; 5/6). Holding gold semantic roles fixed and reading them with the independent probe changes the same comparison to -19.4% ($[-46.9, 5.6]$; 2/6); deweighting duplicate roles gives -18.5% ($[-48.0, 7.5]$; 1/6). Start-anchored canonical and frontier-anchored all-symbol responses give -3.6% ($[-5.5, -1.1]$; 1/6) and -13.4% ($[-34.0, 0.9]$; 2/6), respectively. This protocol dependence is why the main text does not claim that phase excludes relative timing, confidence, unmodeled future-stability information, or effects introduced by semantic projection.

Removing the final-answer slot from the earlier role protocol preserves its capacity-matched increment (39.4% raw and 15.4% normalized), so that particular result is not carried solely by one coordinate. It does not override the coordinate-decomposition or flexible-clock boundary.

E Recurrent-Budget Boundary

The separately prespecified 4/8/16-pass pointer intervention fails its paired behavioral gate: only 16 passes succeeds for both seeds. In the one explicitly exploratory seed competent at both 8 and 16 passes, extra budget induces late integration. Phase-aligned resample fingerprints are only slightly closer across budgets than loop-aligned fingerprints (0.429 versus 0.476 cosine distance), and matched patching does not reproduce clean retiming. We therefore treat budget as selecting a different recurrent regime rather than claim that roles merely move with speed.

F Externally Cued Attention Control

To test whether the scalar checks can retain a known terminal role contrast, we used a separate six-call task with an externally supplied update cue. The shared update acts on a 16-symbol state: cue A rotates it by one, while cue B reflects it. Both operations are bijective and differ for every input state. The one-head attention module reads a state token and a cue token; its state-token output is reused at every call. A fixed cue flip at call 3 defines a same-prefix counterfactual with a different final answer. This control was trained with intermediate-state supervision, unlike the final-answer-only pointer models. Complement-paired schedules are disjoint across training (48), development (8), and confirmation (8); both cues occur at every call in each split. Seed 137 passed the frozen development gate before the confirmation schedules were used. The same 2,000-step recipe trained seeds 223 and 314 without development-based checkpoint selection.

At the fixed call-3 state-token attention-output site, a matched patch inserts the flipped-cue activation; an unrelated donor retains the recipient's *original* schedule but comes from another initial state whose final target differs from both recipient and counterfactual targets. Among eligible donors we match the recipient displacement in L_2 without inspecting patched outcomes. We read the same switched-minus-recipient *final-target* logit margin immediately after call 3 and at output. On the eight untouched confirmation schedules (128 trajectories per seed), all three models predict every intermediate and final state correctly, including within each cue role. Every matched patch reaches the oracle switched final target; all unrelated donors naturally predict their own target. Donor eligibility is complete, median unrelated/matched displacement ratios range from 0.958 to 0.972, and native rollout matches manual continuation.

Table 28: **A known cue switch survives the declared terminal checks.** Effects are changes in the switched-minus-recipient final-target margin (logits) under one fixed call-3 attention-output patch. "Remainder" is matched minus target-unrelated at output.

Seed	Matched after call	Matched output	Unrelated output	Rem. output
137	-0.20	29.63	16.88	12.75
223	-0.33	36.81	19.61	17.19
314	-0.54	37.79	22.08	15.71
Equal-seed mean	-0.36	34.74	19.52	15.22

The final specific effect is positive in all three models (seed range $[12.75, 17.19]$ logits). Yet the same final-target margin read immediately after call 3 is slightly negative; a terminal claim needs its terminal readout. The unrelated donor reproduces 56% of the matched terminal effect; this share alone does not diagnose a role. This calibrates C1/C2 in an explicitly cued, intermediate-supervised attention module, not spontaneous role switching or sensitivity of C3/C4.