INTERVENTION IS NOT IDENTIFICATION: ACTIVATION EDITS CAN PRODUCE THE PREDICTED EFFECT FOR THE WRONG REASON

Anonymous authors
Paper under double-blind review

ABSTRACT

An activation intervention can be exact, task-shaped, and sham-specific without identifying the mechanism that produced its effect. We show this in selective rule updating, where a current binding B and applicability variable S determine the required action $A = B \oplus S$. Qwen2.5-14B-Instruct executes the operator perfectly when both operands are explicit (400/400). A free-reasoning scaffold raises accuracy to 89.6%, but it changes the computational trajectory; direct full-context accuracy is 46.2%. A bitwise-verified layer-28 replacement of the decoded B coordinate moves a supervised layer-35 readout of A by $+0.0288$ and reverses across applicability conditions as the task predicts. The effect survives identity, zero-edit, and equal-norm orthogonal controls. It also persists when the measured A coordinate is held fixed at the edit site (paired held-minus-free difference $+0.0011$, 95% CI $[-0.0021, +0.0043]$). These checks establish a real, structured internal effect. They do not identify its mechanism. Under the prespecified decomposition, contraction toward the downstream readout mean accounts for 74.3% of the interaction, while the baseline-independent remainder is unresolved. On cases where contraction and task-directed correction predict opposite signs, the displacement is near zero. Recipient dependence exceeds the task-directed bound by 4.58× and 4.18×. A structured task-irrelevant variable is more decodable than the target, simple raw-vector carry-through explains essentially none of the downstream interaction, and no preferred answer changes. Across 1,000 matched synthetic draws per condition, the directed estimator is approximately unbiased with 94.2% empirical coverage and separates injected directed effects from pure contraction under the specified generating process. The result separates a causal intervention effect from the mechanism used to explain it.

1 INTRODUCTION

Activation interventions are often treated as stronger evidence than probes because they perturb the model rather than only read from it. But an intervention answers only whether a manipulation changes a downstream quantity. It does not by itself establish why that quantity changed. Mechanistic identification requires a prediction that the proposed mechanism makes and plausible alternatives do not.

We study this problem in selective rule updating. A current binding B and applicability variable S determine the required action $A = B \oplus S$. Qwen2.5-14B-Instruct (Yang et al., 2024) computes this operator perfectly when B and S are explicit (400/400), but it solves only 473/1024 full-context examples. The model performs much better under a different elicitation regime. On the same rows, a free-reasoning scaffold reaches 0.8955 accuracy, with state and applicability accuracies of 0.9229 and 0.9150. The scaffold therefore establishes task capability, not that the direct regime uses the same computation.

Despite its low direct accuracy, the model gives a strong-looking internal result. A constrained layer-28 replacement of the decoded B coordinate changes a supervised layer-35 readout of A by

$$X^A = +0.028793, \qquad 95\%\,\mathrm{CI} = [+0.025473, +0.032004]. \tag{1}$$

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.

The response reverses across applicability conditions as the task-defined $B \to A$ relation predicts. The write, identity control, and zero-edit control are bitwise exact. An equal-norm orthogonal sham behaves differently. We also repeat the replacement while holding the measured A coordinate fixed at layer 28; the layer-35 interaction is unchanged within paired uncertainty. The intervention therefore causes a reproducible, task-shaped change in the downstream readout. We test whether that change identifies task-directed transport.

Several tests argue against that interpretation. Most of the interaction is explained by contraction of the downstream readout toward its mean. The baseline-independent component is unresolved. When contraction and task-directed correction predict opposite signs, the effect is near zero. Recipient dependence is more than four times the bound implied by the proposed replacement model. A structured task-irrelevant variable is more linearly decodable than the target. Neither the primary edit nor a separate action-coordinate intervention changes a preferred answer.

Prior work has shown that decodability need not imply behavioral control (Walsh & Barkett, 2026; Sobotka et al., 2026; Karim et al., 2026), and that interventions can change behavior through unintended pathways (Makelov et al., 2024; Hase et al., 2023). Other work finds successful probe-guided steering (Ravulapalli et al., 2026). We focus on a narrower identification problem. We ask whether the shape of an internal intervention effect is sufficient to attribute it to a proposed computation.

We test four implications of task-directed transport. We estimate the component that remains after accounting for baseline dependence. We isolate cases on which transport and contraction predict opposite signs. We compare the observed recipient dependence with a bound implied by the transport model, and we test selectivity against a structured irrelevant variable. A matched synthetic study checks that the estimators can recover a directed effect when one is present. In this setting, an exact and statistically resolved internal intervention effect passes common controls but still has geometry inconsistent with the proposed mechanism. We do not claim that this failure is prevalent across models or intervention families.

2 Identification, not just intervention

2.1 Competing explanations can agree on the aggregate effect

Let Z_i be a downstream scalar readout and ΔZ_i its change under intervention. A standard analysis asks whether $\mathbb{E}[\Delta Z] \neq 0$, whether the sign matches the task prediction, and whether the target edit differs from a matched control. These tests show that the manipulation changes the measurement. They do not determine the mechanism that produced the change.

In our design, final behavior cannot cleanly arbitrate the interpretation because output margins are large and the response code is deterministically aliased with task structure (Section 3.2). We therefore evaluate the internal endpoint directly. The intervention changes a decoded upstream variable, and we measure how a downstream readout responds under matched controls.

The same aggregate pattern can arise in two ways. A task-directed edit contributes a directed term αd_i. A contraction effect instead pulls the downstream readout toward a mean μ.

$$\Delta Z_i = \alpha d_i + \kappa(\mu - Z_i) + \epsilon_i. \tag{2}$$

If the task classes lie on opposite sides of μ, contraction can produce the same cross-condition sign as a task-directed update. We therefore evaluate predictions on which the two accounts differ. A directed component should remain after baseline dependence is removed. It should still point toward the task target when the baseline readout is wrong. Its recipient dependence should also be compatible with the variation available in the edited coordinate.

2.2 Controls should discriminate mechanisms

The primary intervention sets the decoded B probe coordinate to a matched donor value while holding S fixed. It does not constrain the measured A coordinate. The sham is an equal-norm additive direction orthogonal to the semantic probe span. Equal norm alone does not make the two operations equivalent on the measured coordinate, so a quiet sham does not rule out direction-specific contraction.

Capability check	Accuracy
Explicit operands (operator control)	400/400 = 1.000
Direct full context	473/1024 = 0.4619
Structured-field scaffold	0.7158
Free-reasoning scaffold	0.8955

Table 1: **Capability checks under direct and scaffolded elicitation regimes.** Scaffolded results establish capability under altered elicitation, not identity of the direct-regime computation.

The orthogonal sham tests whether an equal-norm perturbation outside the semantic probe span reproduces the effect. It does not test whether the target edit acts through incidental displacement of A at the edit site. We therefore add a matched replacement that holds the measured A coordinate fixed. More generally, a useful control should separate the proposed mechanism from a plausible alternative. This view complements probe selectivity (Hewitt & Liang, 2019; Voita & Titov, 2020; Belinkov, 2022), amnesic probing (Elazar et al., 2021), causal abstraction (Geiger et al., 2024; Huang et al., 2024), and best-practice work on activation patching (Zhang & Nanda, 2024).

Prespecified terminal classifier.

R1 (baseline-independent component) passes iff the estimated baseline-independent component is positive and the lower bound of its 95% confidence interval is above zero. **R2 (disagreement subset)** passes iff the oriented displacement on the baseline-misclassified subset is positive and the lower bound of its 95% confidence interval is above zero. **R3 (recipient geometry)** passes iff the observed slope does not exceed its task-directed upper bound in either stratum. The directed terminal state requires R1, R2, and R3 all to pass; the pure-contraction terminal state requires all three to fail. The classifier is applied only when both stratum effects d_0 and d_1 are resolved; R3 is not interpretable otherwise. In the synthetic calibration, "directed" and "pure contraction" are prespecified classifier labels for the corresponding generating processes, not exhaustive mechanistic identifications for arbitrary mixtures. These rules were frozen before the Monte Carlo outcomes were computed.

3 System and intervention

3.1 Selective rule updating

The frozen dataset contains 1024 examples over 128 worlds and eight structural families. Each example describes a small structured world from which the model must infer B and S before applying $A = B \oplus S$. All mechanistic results use Qwen2.5-14B-Instruct (Yang et al., 2024). The model scores 400/400 when both operands are explicit but 473/1024 (0.4619) on the direct full-context task.

The low direct score does not imply a general lack of task capability. On the same 1024 rows, structured-field accuracy is 0.7158 and free-reasoning accuracy is 0.8955. In the fields-only arm, current-state, applicability, and updated-state accuracies are 0.9395, 0.7461, and 0.7178, and toggle consistency is 0.9873. With free reasoning, state and applicability accuracies are 0.9229 and 0.9150. Free reasoning generates intermediate tokens after t_{sem}, so the 0.8955 result belongs to a different computational trajectory from the one audited here. Testing that trajectory would require a different endpoint.

3.2 Response-code determinism and the semantic position

The realized generator contains a deterministic response-code coupling,

$$A \equiv 1 \oplus \text{mapping_direction} \oplus \text{rendering}. \tag{3}$$

The gold answer letter is therefore fixed by presentation format. Nested examples map to K on 512/512 rows, and edge-list examples map to M on 512/512. A solver that used only this shortcut could score perfectly. The zero-shot baseline does not show that pattern. It predicts K with probability

0.4414 on nested examples and 0.5176 on edge-list examples, slightly opposite to the shortcut. The alias does not explain the baseline accuracy, but it prevents clean output-side identification.

The internal endpoint is not exposed to the response mapping. We read activations at t_{sem}, immediately after the semantic query and before the opaque response mapping is revealed. Under causal masking, this activation cannot see `mapping_direction`. The action readout therefore cannot reconstruct A from the mapping-direction/rendering identity.

3.3 Readouts and constrained replacement

We fit linear readouts for B, S, and A with structural families held out. Family-macro AUROC is 0.9995 for B at layer 35, 0.9829 for S at layer 28, and 0.8006 for A at layer 35. The A readout has AUROC 0.7605 at the layer-28 edit site. All three reach the smallest attainable permutation p-value, $1/2001$.

At layer 28, the primary edit sets the B coordinate to a matched donor value while holding S fixed. The measured A coordinate is left unconstrained. With $W_2 = [w_B, w_S]$ and desired coordinate change c_2, the minimum-norm edit is $\delta_2 = W_2(W_2^\top W_2)^{-1}c_2$. Across the 512 primary B-factor units, the induced edit-site displacement in A has median magnitude 0.0258 and maximum 0.0831, about 3.0% of the intended B-coordinate change. We therefore run a matched three-coordinate edit with $W_3 = [w_B, w_S, w_A]$ that holds the recipient's measured A coordinate fixed. The continuous hold is exact to numerical precision, and realized fp16 leakage is tiny (Supplement A.4). Identity and zero edits reproduce the unedited activation bitwise. The target vector is also verified on read-back. The sham is an equal-norm direction orthogonal to the semantic probe span.

4 Results

4.1 The intervention passes its internal causal checks

Editing B at layer 28 changes the supervised A readout at layer 35 by

$$X^A = \tfrac{1}{2}(d_0 - d_1) = +0.028793, \quad 95\%\,\mathrm{CI} = [+0.025473, +0.032004], \tag{4}$$

where d_0 and d_1 are sham-adjusted effects for $S = 0$ and $S = 1$. We obtain $d_0 = +0.026926$ $[+0.022265, +0.031626]$ and $d_1 = -0.030660$ $[-0.036915, -0.024044]$. The main-effect composite is $C = \frac{1}{2}(d_0 + d_1) = -0.001867$ with interval $[-0.006219, +0.002890]$. The write and identity path are exact, the conditional reversal is resolved, and the equal-norm sham does not reproduce the effect.

The orthogonal sham does not address the small local displacement in A created by the primary edit. We therefore repeat the intervention while holding the measured A coordinate fixed and preserving the recipient, donor, site, and replacement form. The layer-35 interaction is $+0.029870$, compared with $+0.028793$ when A is unconstrained. Across the same 512 units, the paired held-minus-free difference is $+0.001077$ with 95% CI $[-0.002069, +0.004333]$ under the same 128-world clustered bootstrap. The lower confidence limit corresponds descriptively to about 7.2% attenuation relative to the primary interaction. This result rules out dependence on that measured edit-site coordinate, but not mediation through other distributed features.

The observed interaction is also too large to be explained by simple retention of the edit vector. Directly reading the untransformed layer-28 edit with the layer-35 A probe predicts an interaction of only -0.000182, compared with the observed $+0.028793$. A separate preregistered transplant analysis gives a simultaneous 95% upper bound of 0.087 on retained carriage at layer 29, below the prespecified 0.25 threshold. All estimated layers 29–35 remain below 0.25. These diagnostics exclude raw-vector retention and direct readout. They do not exclude cross-layer re-encoding of an A-related signal (Supplement A.4).

4.2 The apparent effect is dominated by contraction

After the real edit, the standard deviation of the A readout falls from 0.194805 to 0.162886, a 16.4% contraction. Under the orthogonal sham it changes by only +0.2%. Pooled across cells, the slope of readout change on baseline is -0.17729 for the real arm and -0.00592 for the sham.

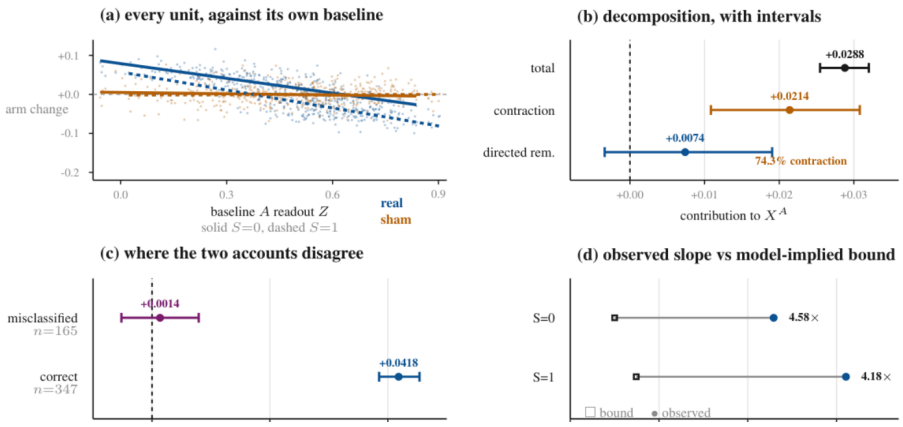

Figure 1: **Tests that separate task-directed transport from contraction.** (a) The real edit contracts the downstream readout toward its center, while the equal-norm orthogonal sham is nearly flat. (b) Contraction accounts for $+0.021396$ of the $+0.028793$ interaction (74.3%); the remaining directed component is unresolved. (c) On 165 baseline-misclassified recipients, where the two accounts predict opposite signs, displacement is near zero. (d) Observed within-cell slopes exceed the task-directed bounds by $4.58\times$ and $4.18\times$.

For $S = 0$, the sham-adjusted within-cell slope is -0.11458 with 95% CI $[-0.13803, -0.09115]$. For $S = 1$, it is -0.15532 with CI $[-0.18495, -0.12255]$. Contraction contributes $+0.021396$ with CI $[+0.010846, +0.030795]$. The baseline-independent component is $+0.007397$ with CI $[-0.003385, +0.019054]$. Under this decomposition, contraction accounts for 74.3% of the interaction and the remaining directed component is unresolved. The cross-cell intercept contrast cancels the common $\kappa\mu$ term exactly within each cell. Section 4.7 calibrates this estimator under known generating processes.

4.3 When the explanations disagree, the task-directed effect disappears

The decomposition assumes a model form, so we also use a direct sign test. For recipients whose baseline A readout is misclassified, contraction and task-directed correction predict opposite directions. This subset contains 165 of the 512 recipients. Their oriented displacement is $+0.001358$ $[-0.005197, +0.007924]$, compared with $+0.041838$ $[+0.038533, +0.045380]$ on the 347 probe-correct recipients. The task-directed effect is therefore absent on the cases where the two explanations disagree. This rules out uniform task-directed transport across recipients. It does not rule out a directed mechanism conditional on successful internal computation, nor does it prove pure contraction. The data support a contraction-dominated mixture with no identified positive directed remainder.

4.4 Recipient dependence exceeds the task-directed bound

Under a task-directed replacement model,

$$\Delta Z_i = \gamma(b_d - b_i), \tag{5}$$

so within-stratum variation in ΔZ can arise only from dispersion of the edited coordinate. At layer 28, B separates classes at AUROC 0.9991. Its within-stratum standard deviation is 0.14925 for $S = 0$ and 0.16925 for $S = 1$. By Cauchy–Schwarz, the corresponding absolute slope bounds are 0.025020 and 0.037162. The observed slopes are 0.114582 and 0.155323, or $4.58\times$ and $4.18\times$ the bounds. This is a conditional model-compatibility diagnostic, not a standalone hypothesis test. Under a true null the bound itself collapses, as shown by the Monte Carlo calibration below. We apply the diagnostic only after both stratum effects are resolved (Section 4.1).

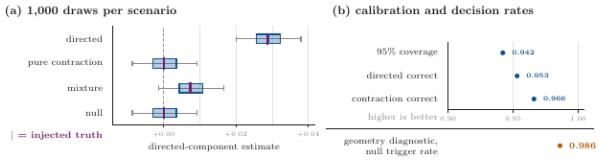

Figure 2: **Calibration on synthetic data with known generating processes.** (a) Directed-component estimates over 1,000 draws for directed, pure-contraction, mixture, and null scenarios. (b) Empirical coverage and terminal decision rates. The recipient-geometry diagnostic is shown separately because its 0.986 null trigger rate indicates a limitation rather than good performance.

4.5 The readout is not selective for the claimed target

At layer 35, we also decode a structured task-irrelevant variable that shares S but replaces B with a world statistic unrelated to the task, $S \oplus$ distance-to-median. Its AUROC is 0.8547, higher than the 0.8006 AUROC for the task target A. The readout is therefore not selective for A in the broad probing sense (Hewitt & Liang, 2019; Voita & Titov, 2020). This control differs from the random-label control task of Hewitt & Liang (2019). Donor consistency also remains imperfect. Only 0.7402 of donors are probe-consistent; the rest move the A coordinate toward the wrong probe class.

4.6 The output endpoint is not independently identifiable

The same layer-28 B edit changes the output logit margin by $X^Y = -0.037202$, with 95% CI $[-0.045261, -0.029451]$. This output effect cannot independently identify task-directed transport because applicability is nearly collinear with baseline correctness. The $S = 0$ cell is 254/256 baseline-correct, while the $S = 1$ cell is 220/256 baseline-incorrect ($\phi = 0.8592$). After residualizing ΔY on $(1, M_{\text{base}}, |M_{\text{base}}|)$, the interaction is -0.001523 $[-0.007956, +0.004715]$, and $\text{corr}(\Delta M, M_{\text{base}}) = -0.4383$. Baseline margin lies downstream of S, so this residualized interaction is descriptive rather than a corrected causal estimate. The generator also creates the exact alias $B \times S \equiv \text{mapping_direction} \times \text{rendering}$, which produces rank deficiency rather than ordinary imbalance.

The B edit changes no preferred code among the 512 recipients in either the real or sham arm. Median $|\Delta Y|$ is 0.0393 nats, and median $|\Delta Y|/|M_{\text{base}}|$ is 0.18% with a median baseline margin of 23.25 nats. A separate preregistered A-coordinate intervention also produces zero answer flips in 8,192 edited forwards; its median $|M_{\text{post}} - M_{\text{base}}|$ is 0.0156 nats. An exploratory float32 replication removes the fp16 lattice and changes endpoint estimates by at most 0.0002 nats. The high-precision real arm again has zero sign changes. Zero flips alone does not bound the logit effect because the baseline margins are large. The two experiments answer different questions. The B edit bounds the behavioral relevance of the intervention audited here, while the A-only edit tests direct manipulation of the supervised action coordinate.

4.7 Ground-truth calibration recovers an injected directed effect

We next check whether the rejection battery can recover a directed effect when one is present. We evaluate four frozen synthetic data-generating processes with 1,000 draws each. The simulation preserves the 512 recipients, baseline readouts, task cells, world clustering, intervention geometry, and the full 2,000-replicate bootstrap. Only the downstream displacement is generated. Noise is matched to the observed within-cell residual scale of 0.033766.

In the directed scenario, the injected effect is $+0.028793$ and the mean estimate is $+0.029010$. Bias is $+0.000217$, RMSE is 0.004766, and empirical 95% coverage is 0.942. The same bias and RMSE appear under pure contraction, the observed mixture, and the null because the within-cell intercept contrast cancels the common contraction term. The prespecified terminal classifier identifies 95.3% of directed draws and 96.6% of pure-contraction draws correctly. The decomposition can therefore recover a directed effect of the observed size under the specified synthetic process.

On the observed data, R1, R2, and R3 all fail, so the prespecified classifier returns the contraction terminal state. We interpret this as *no identified positive directed component*, not as proof of pure contraction. An unresolved mixture can produce the same terminal pattern, so Section 4.3 reports a contraction-dominated mixture.

The geometry diagnostic has an important limitation. Under a true null it fires on 98.6% of draws because its bound collapses with the observed cell effect. We therefore use it only after both stratum effects are resolved. The calibration validates the estimators under the synthetic data-generating process, not the full transformer intervention pipeline.

5 Related work

Representation, use, and control. Decodability and behavioral control can dissociate (Walsh & Barkett, 2026; Sobotka et al., 2026; Karim et al., 2026). Probe-guided binding interventions can also improve behavior (Ravulapalli et al., 2026). Our question is different. We ask whether a positive internal intervention effect is enough to identify task-directed computation.

When interventions mislead. Makelov et al. (2024) show that a subspace patch can change behavior through a dormant parallel pathway, and Hase et al. (2023) show that localization and editability can dissociate. In both cases the end-to-end effect is real but its mechanistic interpretation is limited. Our failure occurs earlier, at the internal readout itself, where contraction and recipient geometry account for the task-shaped response. Zhang & Nanda (2024) show that activation-patching conclusions depend on metric and corruption choices. We use a logit-difference endpoint and find a different identification problem.

Methodological lineage. We build on probe selectivity (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022; Voita & Titov, 2020), counterfactual removal and causal abstraction (Elazar et al., 2021; Geiger et al., 2024; Huang et al., 2024), and activation intervention and steering (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Turner et al., 2023; Zou et al., 2023). State- and belief-tracking work motivates the variable class (Li et al., 2025; Prakash et al., 2026); explicit world-state methods provide a complementary approach to rule following (Li et al., 2026). Our result adds a separate failure mode. An exact, resolved, sham-specific internal effect can still have causal geometry inconsistent with the proposed mechanism.

6 Discussion and limitations

The experiments separate intervention validity from mechanism identification. The edit is implemented as specified and produces a resolved downstream effect. Holding the local A coordinate fixed does not attenuate that effect within paired uncertainty. The remaining response, however, does not match task-directed transport. It is largely explained by contraction, is absent on disagreement cases, exceeds the recipient-dependence allowed by the transport model, and is measured in a space that is not selective for A. We therefore identify an intervention effect, not the proposed mechanism.

This distinction has a practical consequence. A causal manipulation shows that an intervention matters to a measured quantity, but attribution requires an additional test of mechanism. The most informative controls are those on which the proposed mechanism and a plausible alternative make different predictions. Our diagnostics provide one way to construct such tests; they are not intended as a universal battery.

The replacement operator itself may contribute to the observed contraction. That possibility does not support task-directed transport; it is an alternative explanation that the original controls did not exclude. Intervention geometry therefore has to be considered as part of the identification problem. The 89.6% scaffold result has a separate role. It establishes capability under richer elicitation, not the computation used under direct prompting.

The study does not estimate how common this failure is. All mechanistic experiments use one open-weight model, one constructed task, one primary site pair, one edit direction, and linear supervised readouts. The generator aliases response coding with task structure, which prevents independent output-side identification, although the direct alias is unavailable at the semantic endpoint used for the central analysis. We test a constrained replacement intervention, not additive intervention

families. We also do not have a model-level positive control for the full battery. Cross-model and cross-intervention replication are therefore needed.

7 CONCLUSION

The experiment begins with a convincing internal result. B is decodable, the edit is exact, the downstream readout reverses across applicability conditions, and the confidence interval is tight. The equal-norm sham does not reproduce the effect, and the interaction survives a matched edit that holds the measured A coordinate fixed. Simple raw-vector carry-through also predicts almost none of the observed magnitude. The interpretation changes when we test predictions that distinguish transport from contraction. Most of the effect follows contraction, the directed response is absent on disagreement cases, recipient dependence exceeds the transport bound, and the measured space is not selective for A.

The intervention happened. The identification did not. An activation edit becomes evidence for a mechanism only when its effects follow predictions that distinguish that mechanism from plausible alternatives. The causal effect is the starting point for that test, not the endpoint.

A broader question remains open. What distinguishes a representation that can be decoded and edited from one that is actually used to control behavior? Answering it will require experiments that connect internal dependency structure to behavioral control.

REPRODUCIBILITY STATEMENT

All headline results derive from one frozen dataset of 1024 examples over 128 worlds and eight structural families and from cached residual-stream activations captured under a fixed model configuration. Load-bearing quantities are regenerated from stored artifacts by automated checks that verify both numeric values and their estimand binding. Implementation and estimand details are provided in Supplement A.4. The Monte Carlo calibration uses the same closed-form estimators and bootstrap structure as the real-data analysis, and its decision rules were fixed before outcomes were computed. The analysis code, intervention harness, checker, and figure-generation code will be released upon acceptance.

AI USE STATEMENT

Generative AI tools were used to assist with hypothesis refinement, experimental methodology and implementation, result interpretation and verification, and manuscript drafting and editing. All AI-assisted work was reviewed by the authors, who take responsibility for the final content of the paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. ICLR Workshop Track, 2017.

Yonatan Belinkov. Probing classifiers: promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. NeurIPS, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: behavioral explanation with amnesic counterfactuals. TACL, 9:160–175, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. CLeaR, PMLR 236:160–187, 2024.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. NeurIPS, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. EMNLP-IJCNLP, 2019.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. ACL, 2024.

Ahmed Karim, Fatima Sheaib, Zein Khamis, Maggie Chlon, Jad Awada, and Leon Chlon. Attention deficits in language models: causal explanations for procedural hallucinations. arXiv:2602.19239, 2026.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (How) do language models track state? ICML, 2025.

Haiqing Li, Wenliang Zhong, Yinhao Wu, Hehuan Ma, Yuzhi Guo, Thao M. Dang, and Junzhou Huang. Guidelines as environments: a world model approach to rule following. ACL, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. ICLR, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. NeurIPS, 2022.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. ICLR, 2026.

Manas Venkata Sai Ravulapalli, Samrath Singh Chadha, and Abhinav M. Hari. Legible failures: detecting and repairing in-context binding errors. arXiv:2609.11216, 2026.

Jan Sobotka, Mustafa O. Karabag, and Ufuk Topcu. Why do LLMs struggle in strategic play? Broken links between observations, beliefs, and actions. arXiv:2605.00226, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv:2308.10248, 2023.

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. NeurIPS, 2020.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. EMNLP, 2020.

Ciarán Walsh and Emilio Barkett. Representation without control: testing the realization effect in language models. arXiv:2605.25151, 2026.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. ICLR, 2023.

An Yang et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: metrics and methods. ICLR, 2024.

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

A ADDITIONAL RESULTS AND DIAGNOSTICS

A.1 Remaining scaffold detail. In the fields-only arm, the worst applicability cell is 0.5352. With free reasoning, the worst state/applicability cell is 0.8789. A donor-content discriminator is specified but not executed and is not used for the central result.

A.2 Secondary propagation detail. The interval for $\rho(29)$ contains zero, and all estimated layers 29–35 remain below the prespecified 0.25 threshold. The trajectory is not treated as monotone and adjacent-layer contrasts were not preregistered.

A.3–A.4 Supplementary details. The full Monte Carlo calibration table and implementation/estimand details—including the paired A-hold diagnostic, construction fidelity, raw-vector carry-through check, and precision analysis—are provided in the standalone supplement.