LOCALIZATION DOES NOT GUARANTEE CONTROL: GATE FAILURE AND PARTIAL ACTIVATION STEERING IN A CAUSAL REASONING MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Does locating a behaviorally relevant activation site suffice to construct an effective intervention? We study this question in a frozen 25.7M-parameter transformer trained to answer interventional queries from mixed observational and interventional evidence. On a selected subset of confounded worlds, transplanting activations at mid-layer observational-record positions recovers 88.9% of a reference behavioral shift induced by evidence zeroing. We then evaluate a gated activation-editing pipeline and an ungated ablation on 50 held-out confounded worlds. The detector, calibrated with zero validation false triggers but only 0.5 recall, triggers on none of 2,400 target generations and on 6.9–7.3% of held-out in-distribution generations. Bypassing the gate enables partial correction: at the larger of two tested doses, a pooled contrast direction reduces the regression intercept from 0.3820 to 0.2637, a 31% reduction, while also reducing the gain from 0.8236 to 0.5721. An exact coefficient identity shows why: the pooled edit attenuates the whole response nearly proportionally, lowering offset and gain by the same factor instead of restoring world-specific sensitivity. These results separate failure to activate an intervention from insufficient correction after activation, and show that localization, gate coverage, and intervention efficacy must be evaluated separately.

1 INTRODUCTION

Interpretability methods can identify information in hidden states without establishing how to make a model use it. Probes recover latent variables (Belinkov, 2022); causal tracing and interchange interventions identify or induce activations that mediate behavior (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2022); and activation steering modifies outputs by adding directions to the residual stream (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Arditi et al., 2024). These methods answer related but distinct questions: what information is recoverable, where an intervention can affect behavior, and which edits can be constructed from a test prompt. A successful answer to one question need not resolve the others.

The distinction matters for conditional activation editing, in which a condition read from hidden states determines whether a steering edit is applied (Lee et al., 2025). Such a system must identify prompts requiring intervention, construct an edit at a suitable site, and apply it without unacceptable changes elsewhere. A failed end-to-end evaluation can therefore have several causes. The detector may never activate on the target distribution, or the edit may be insufficient even when applied. Without measuring these components separately, a null behavioral result cannot distinguish failure of routing from failure of intervention. Known gaps between probe accuracy and representation content (Hewitt & Liang, 2019), between localization and editing (Hase et al., 2023), and the variable generalization of steering vectors (Tan et al., 2024), motivate this component-level evaluation.

We examine these questions in a controlled causal-reasoning task. A concurrent submission reports that in a small transformer, observational records in context override interventional evidence, and that patching observational-record states recovers much of the effect (Anonymous, 2026). We use the same frozen checkpoint and investigate whether activation edits can reduce this behavioral bias. The synthetic setting supplies known causal effects and matched reference runs for transplant experiments.

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.

We evaluate a diagnostic pipeline, the *contrastive causal gate* (CCG), that makes each of these components separately measurable.

The experiments produce three findings with different scopes. First, observational-record activations in intermediate layers support substantial recovery under an oracle transplant on 48 eligible worlds out of 608 screened. Second, on the separate 50-world release evaluation, the calibrated detector activates on none of 2,400 target generations, making the gated system identical to the base model there. Third, ungated pooled edits reduce the behavioral offset at both tested doses, by up to 31%, while also reducing effect sensitivity, because they attenuate the whole response nearly proportionally. Successful localization therefore does not by itself validate an editing pipeline: coverage and the structure of the edit must be measured as well.

Our contributions are as follows.

1. **A component-level evaluation of activation editing.** We distinguish oracle transplant recovery, detector coverage, and prompt-computable intervention efficacy, and show that a gated editor's end-to-end effect factors exactly into coverage times conditional efficacy, while its effect on offset and gain is fixed by the edit's mean effect and its covariance with the true effect.
2. **An observed failure of gate coverage.** A detector with zero validation ID triggers and 0.5 calibration recall triggers on no held-out confounded generations, despite firing on 6.9–7.3% of held-out ID generations.
3. **A controlled ungated comparison.** Fixed and per-instance pooled contrast edits are evaluated at the selected site. The larger adaptive dose removes 31% of the baseline offset but reduces gain, and the coefficient identity traces this to a nearly proportional attenuation of the response.

The practical implication is an evaluation requirement: report whether the intervention ran, what it changed when forced to run, and whether its adaptive component improves on a fixed control.

2 Related Work

Decodability and model use. Probing methods recover linguistic and task-relevant variables from activations (Belinkov, 2022), and probe accuracy can reflect the probe's own capacity rather than the representation (Hewitt & Liang, 2019). Unsupervised approaches can also extract latent knowledge without task labels (Burns et al., 2023), and linear probes are increasingly deployed as activation monitors whose thresholds are calibrated on one distribution and applied to another (Goldowsky-Dill et al., 2025; McKenzie et al., 2025). We treat the previously reported override behavior as motivation for intervention, and complement it with our own diagnostics of feature alignment, detector transfer, and behavioral recovery.

Activation steering. Adding vectors to hidden states can alter sentiment, style, truthfulness, and refusal (Subramani et al., 2022; Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). However, steering performance can vary across prompts, distribution shifts, and model families (Tan et al., 2024; Braun et al., 2025; Silva et al., 2025), and on a large concept-steering benchmark, representation-based steering trails prompting (Wu et al., 2025). Input-dependent steering vectors adapt the edit to each prompt (Wang et al., 2025); our experiments examine one restricted construction: a contrast vector obtained by pooling observational-record activations and adding it uniformly at a selected site. We compare fixed and per-instance versions, measure two nonzero doses, and evaluate the detector separately from the edit, whereas conditional steering methods are typically assessed end to end (Lee et al., 2025).

Causal localization. Causal mediation and activation replacement have been used to study gender bias, factual associations, and abstract computational structure (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2022). Following this approach and established patching practice (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024), we use matched-run transplants to assess whether a candidate site supports a reference behavioral shift. We then ask whether a lower-dimensional edit, computed without the oracle donor, reproduces that shift. The difference in information access and intervention granularity is central to interpreting the comparison, since interventions on a low-dimensional subspace can behave differently from full-activation patching (Makelov et al., 2024).

Causal reasoning and detection–control gaps. CLadder evaluates formal causal reasoning across the ladder of causation (Jin et al., 2023; Pearl, 2009). Our base model and its reported observational-override behavior come from a concurrent study of causal-evidence discrimination (Anonymous, 2026); the present study evaluates inference-time interventions on that checkpoint. In larger language models, directions that detect a behavior have been distinguished from directions that control it (Galeone et al., 2026). Their geometric analysis and our pipeline evaluation address complementary questions: we directly measure target-distribution trigger coverage and then remove the gate to assess the tested edits.

3 Experimental Setup

3.1 Model, task, and evaluation populations

We study **C1**, a frozen 25.7M-parameter, eight-layer GPT-style decoder-only transformer (Vaswani et al., 2017) with a vocabulary of 41 tokens. C1 is the mixed-evidence checkpoint from a concurrent submission (Anonymous, 2026). Each context contains four observational records and two interventional probe records, followed by a query $\mathrm{do}(X{=}x_0) \to Y{=}?$. Thus, the task involves reasoning from mixed evidence, rather than identifying an arbitrary causal effect from observational data alone. On *Simpson-trap* worlds, observational association and causal effect have opposite signs. The checkpoint's tendency to be pulled toward the former even when interventional probes are present motivates the present experiments. Appendix A describes the model and data-generation specifications.

Evaluation sets. The pilot diagnostics use 50 worlds; the confirmatory localization screen contains 608 worlds, of which 48 satisfy the release-measurement eligibility criteria. The terminal gated and ungated experiments use **CONF50**, consisting of 50 held-out confounded worlds, and **ID40**, consisting of 40 held-out non-confounded worlds. The calibration set contains 40 worlds spanning eight families, with nine positive examples across seven families; it was obtained by removing five CONF50 overlaps from 45 candidate worlds. The localization subset and CONF50 are distinct evaluation sets. Thresholds and evaluation sets were fixed before the corresponding measurements (Appendix C).

3.2 Measurements

For each world w, we measure the model's response slope $\widehat{\Delta}_w$ over two query values. Let Δ_{same}, Δ_{zero}, and Δ_p denote slopes under the original evidence, evidence zeroing at observational-record positions (obs), and an activation intervention, respectively. On worlds with $|D| \geq 0.05$, where $D = \Delta_{\mathrm{zero}} - \Delta_{\mathrm{same}}$, the normalized recovery is

$$R = \frac{\Delta_p - \Delta_{\mathrm{same}}}{\Delta_{\mathrm{zero}} - \Delta_{\mathrm{same}}}. \tag{1}$$

We retain the term *release rate* for this quantity: $R = 0$ denotes no change and $R = 1$ denotes recovery of the zeroing-induced reference shift. The denominator guard excludes worlds with a small reference shift.

We separately assess agreement with ground-truth effects using an ordinary least-squares fit,

$$\widehat{\Delta}_w = a\,\mathrm{ATE}_w + b + \varepsilon_w, \qquad w \in \mathcal{W}, \tag{2}$$

where ATE_w is the generator's ground-truth effect and ε_w is the regression residual. Under the benchmark's effect normalization, the target relation is $a = 1$, $b = 0$. The gain a measures variation with the true effect, whereas b measures an additive offset. A smaller positive offset is desirable here, but a lower b alone is not an improvement: a collapsed response can have $a = b = 0$. We therefore report both coefficients, sign reversals, and ID behavior. The design targets are $b \leq 0.08$ with stable gain and at most 0.05 ID slope degradation (Table 1). Appendix D gives the remaining definitions.

Three distinct claims. *Decodability* means that a reader can recover task-relevant information from hidden states. *Localizability* means that replacing activations at a site with those from a matched reference run changes behavior toward that reference. *Releasability* means that an edit computable

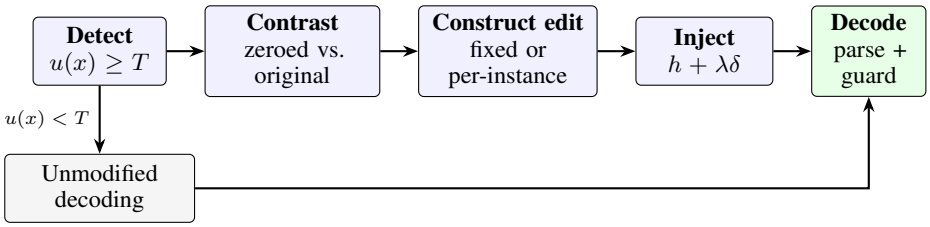

Figure 1: The evaluated CCG pipeline. The gated test measures both intervention coverage and end-to-end behavior. Bypassing detection isolates the effects of fixed and per-instance pooled edits at the site selected by the transplant study. The oracle transplant and prompt-computable edit have different information access and spatial granularity.

from the test prompt meets the behavioral targets without unacceptable changes elsewhere. The information available to a transplant and the expressiveness of a pooled edit differ, so localizability does not guarantee releasability.

3.3 Contrastive causal gate and ungated edits

The CCG pipeline has four components (Figure 1). A detector score $u(x)$ is compared with a frozen threshold T. If $u(x) < T$, decoding proceeds without an edit. Otherwise, the intervention path uses original and evidence-zeroed activations, constructs a contrast direction, and adds it at the selected residual-stream site before decoding. The terminal experiments use explicit mean-difference constructions; no learned direction generator is evaluated in these experiments.

The fixed direction is the calibration mean of raw pooled contrasts,

$$\delta_{\mathrm{ZERO}} = \overline{\mathrm{MeanPool}_{\mathrm{raw}}(h_{\mathrm{zero}})} - \overline{\mathrm{MeanPool}_{\mathrm{raw}}(h_{\mathrm{same}})}, \tag{3}$$

where the overbar denotes calibration averaging and pooling is over the selected observational-record positions. The adaptive direction instead uses the current prompt:

$$\delta(x) = \mathrm{MeanPool}_{\mathrm{raw}}(h_{\mathrm{zero}}(x)) - \mathrm{MeanPool}_{\mathrm{raw}}(h_{\mathrm{same}}(x)). \tag{4}$$

Neither construction applies LayerNorm before taking the contrast; LayerNorm is used separately for detector features. At each selected layer, the appropriate site direction is broadcast across the selected token rows:

$$h_t^{(\ell)}(x) \leftarrow h_t^{(\ell)}(x) + \lambda\delta(x), \qquad \ell \in \mathrm{site}, \quad t \in \mathrm{obs}. \tag{5}$$

Here δ denotes the fixed or adaptive direction as appropriate. The ungated experiment applies Eq. (4) unconditionally at $\lambda \in \{0.5, 1.0\}$, alongside a remeasured always-on fixed-direction control. Unparseable responses and degenerate contrasts are subject to conservative guards.

4 Component-Level Analysis

The three claims of Section 3.2 become separately testable once the end-to-end effect of a gated editor is written in terms of its components. This section gives that decomposition and a coefficient-level reading of what a single edit direction can and cannot change.

Coverage times conditional efficacy. Let $m(x)$ be any per-prompt outcome whose population average is of interest, such as a slope or a correctness indicator, and write m_{base}, m_{edit}, and m_{gate} for its value under the unmodified model, the unconditionally edited model, and the gated pipeline. The gate applies the edit exactly when $u(x) \geq T$, so $m_{\mathrm{gate}}(x) = m_{\mathrm{edit}}(x)$ on triggered prompts and $m_{\mathrm{base}}(x)$ otherwise. For any evaluation population $\mathcal{D}$ with trigger rate $\pi_{\mathcal{D}} = P_{\mathcal{D}}(u(x) \geq T) > 0$, the end-to-end effect therefore decomposes exactly as

$$\mathbb{E}_{\mathcal{D}}[m_{\mathrm{gate}} - m_{\mathrm{base}}] = \pi_{\mathcal{D}}\, \mathbb{E}_{\mathcal{D}}\big[m_{\mathrm{edit}} - m_{\mathrm{base}} \;\big|\; u(x) \geq T\big], \tag{6}$$

and the left-hand side is zero whenever $\pi_{\mathcal{D}} = 0$.

Equation (6) separates the two ways a gated editor can fail. On the target population, zero coverage forces a zero end-to-end effect for *any* edit, so a null gated result carries no information about the edit. On the in-distribution population, the harm is the trigger rate times the conditional harm, so a small aggregate change can hide a large per-prompt effect on the prompts that trigger. The conditional factor is also not the ungated average: the ungated experiment estimates $\mathbb{E}_{\mathcal{D}}[m_{\text{edit}} - m_{\text{base}}]$ over all prompts, whereas the gate would apply the edit to a selected subset. Reporting coverage, ungated efficacy, and the triggered subset's composition is therefore necessary for attributing a gated result, and it is the structure of Sections 5.3 and 5.4.

What an edit must do to the offset and the gain. The target is stated on the coefficients of Eq. (2). Let $d_w = \widehat{\Delta}_w(\lambda) - \widehat{\Delta}_w(0)$ be the change in world w's response slope produced by an edit. Because least squares is linear in the response, the fitted coefficients change by

$$a(\lambda) - a(0) = \frac{\mathrm{Cov}_w(d_w, \mathrm{ATE}_w)}{\mathrm{Var}_w(\mathrm{ATE}_w)}, \qquad b(\lambda) - b(0) = \bar{d} - \big(a(\lambda) - a(0)\big)\,\overline{\mathrm{ATE}}, \tag{7}$$

where bars denote averages over the evaluation worlds. This identity is exact whenever both fits use the same worlds and requires no model of the network. It implies that an edit raises the gain only if its effect covaries positively with the true effect, and that reaching the target $(a, b) = (1, 0)$ from the base coefficients requires $d_w = (1 - a)\,\mathrm{ATE}_w - b + r_w$ with r_w zero-mean and uncorrelated with ATE_w: a component that scales with the world's effect together with a uniform downward shift.

If the response slope varied smoothly with the site activation, a single direction would produce, to first order, $d_w \approx \lambda J_w \delta_w$, where J_w is the Jacobian of world w's response slope and δ_w the edit for that world; the gain then changes only through variation of $J_w \delta_w$ across worlds that is aligned with ATE_w. The fixed direction δ_{ZERO} is the same for every world, and the per-instance direction $\delta(x)$ averages away the positional structure of the contrast, so neither is designed to produce a component that scales with the world's effect. Section 5.4 reads the measured coefficients through Eq. (7): the observed decrease of the gain implies $\mathrm{Cov}_w(d_w, \mathrm{ATE}_w) < 0$, so in linear trend the edit lowers predictions more on worlds with larger true effects, the opposite of the required component. The transplant, which retains positional structure, is the intervention that carries such a world-specific component.

5 Results

5.1 Pilot diagnostics

The pilot examines measurement coverage, feature alignment, and detector transfer (Appendix B).

Coverage of normalized recovery. Eleven of 50 worlds meet the denominator guard. On that subset, mean release rates are 0.785 ($[0.637, 0.932]$, $\mathrm{L23}_{\mathrm{obs}}$), 0.829 ($[0.663, 0.994]$, $\mathrm{L345}_{\mathrm{obs}}$), 0.789 ($[0.587, 0.962]$, $\mathrm{L2345}_{\mathrm{all}}$), 0.326 ($[0.112, 0.613]$, $\mathrm{L23}_{\text{non-obs}}$), and 0.186 ($[-0.104, 0.461]$, $\mathrm{L345}_{\text{non-obs}}$). Here L23 and L345 denote the indicated layer sets, crossed with observational or non-observational token rows. The estimates favor observational rows; the remaining 39 worlds have reference shifts too small for the normalized statistic.

Feature alignment. Cross-evidence centered linear CKA ranges from 0.061 to 0.068 across nine families, compared with a within-family reference mean of 0.431 ($[0.144, 0.998]$), indicating low similarity of the features across evidence types.

Detector transfer. A $k = 32$ MLP probe attains within-distribution family-separation AUC 0.962, compared with 0.70–0.74 for a linear head. Transfer is weaker: leave-one-family-out AUCs are 0.640 for sign, 0.366 for scale, and 0.658 for confounding strength; the ID trigger rate is 0.100. Zero-shot CLadder transfer AUCs of 0.022–0.292 are below chance under the original score orientation. Together, the pilot motivates direct intervention tests and target-distribution trigger measurement.

5.2 Activation transplants

The transplant replaces residual-stream rows at a candidate site in a suppressed run with the corresponding rows from a matched released reference run. Candidate sites combine layer sets with

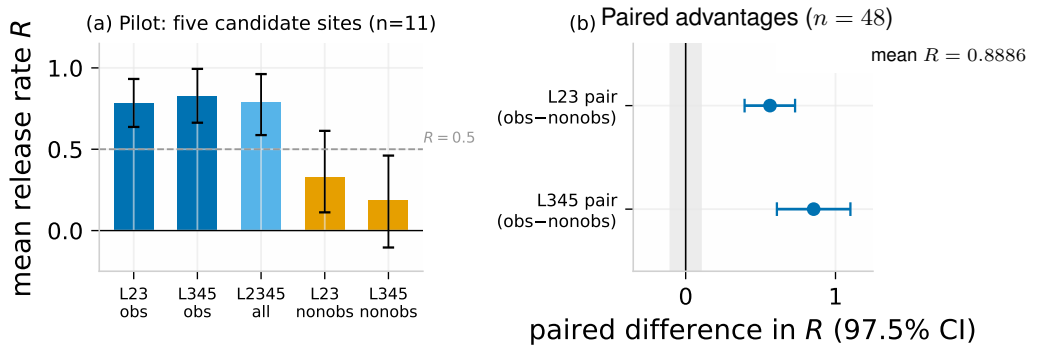

Figure 2: Transplant localization. **(a)** Pilot normalized recovery at five sites, conditional on the 11 measurable pilot worlds. **(b)** Confirmatory paired release advantages on 48 eligible worlds, with 97.5% intervals for each of two comparisons. L23_{obs} is the selected site; a larger paired advantage in another comparison need not imply a higher absolute recovery rate.

observational or non-observational token positions. We test for a positive paired release advantage of observational over non-observational rows within each of two layer sets, using 97.5% bootstrap intervals for the two comparisons.

Of 608 screened worlds, 48 have a measurable reference shift (7.89%; Wilson 95% interval $[0.0601, 0.1031]$). The paired advantages are

$$\begin{aligned} \text{L23 comparison:} \quad & 0.5628\ [0.3932, 0.7302], \\ \text{L345 comparison:} \quad & 0.8542\ [0.6088, 1.0995], \qquad n = 48. \end{aligned} \tag{8}$$

Both intervals exclude zero, so the recoverable signal is concentrated at the observational-record positions, the same positions implicated by activation patching in the concurrent study (Anonymous, 2026). At the selected L23_{obs} site, the mean normalized recovery is

$$R_{\text{site}} = 0.8886. \tag{9}$$

Thus, the transplant recovers approximately 89% of the zeroing-induced reference shift on eligible worlds. A ± 0.01 change to the denominator guard leaves both comparisons unchanged. Figure 2 shows the pilot estimates and confirmatory paired advantages.

The selected site's transplanted activations thus mediate substantial recovery of the reference behavior. The oracle donor, however, contains information that a prompt-computable edit does not have, and the following experiments test whether an edit at the same site can achieve a comparable effect.

5.3 Gated Editing

The gated experiment combines the detector, a contrast pass, the frozen calibration direction δ_{ZERO}, injection, and guarded parsing. Table 1 lists the design targets, fixed before calibration.

Calibration. The detector uses PCA with eight components on the site features ϕ (Eq. (14)), followed by logistic regression trained for 300 epochs at learning rate 10^{-2}. Its label is $y = 1$ when $D = \Delta_{\text{zero}} - \Delta_{\text{same}} > 0.05$. This one-sided label marks a positive zeroing-induced shift. The 40 calibration worlds are divided 24/8/8, with runs for seeds $\{0, 1, 2\}$. The frozen thresholds are 0.699, 0.728, and 0.737. Calibration yields zero validation ID triggers, OOD recall 0.5, and parse rate 1.0. Direction norms are 8.30 for the two-segment construction and 14.35 for the three-segment diagnostic.

Held-out behavior. On CONF50, the gated system's fitted intercept is $b = 0.3820$ (bootstrap 95% interval $[0.3396, 0.4230]$), and it retains 25 sign reversals among 27 negative-effect worlds, with no discordant base-versus-gated outcomes. The gain change is zero and the positive-world correlation is 0.6299. On ID40, slope degradation is only 0.0038–0.0042, while trigger rates are $44/640$, $44/640$,

Table 1: Gated evaluation on CONF50 and ID40, with the design target for each quantity. Generation counts are recorded decoding events.

Quantity	Target	Observed
Intercept b	≤ 0.08	$0.3820\ [0.3396, 0.4230]$
Gain stability / correlation	$\lvert\Delta a\rvert \leq 0.10$ and corr > 0.40	0.0; corr 0.6299 ($n = 23$ positive worlds)
Reversals / paired test	$\leq 14/27$ and $p < 0.05$	$25/27$; no discordant pairs
ID slope degradation	≤ 0.05	0.0038–0.0042
ID trigger rate	≤ 0.02	0.0688–0.0734
CONF50 trigger count	—	$0/2{,}400$ generations

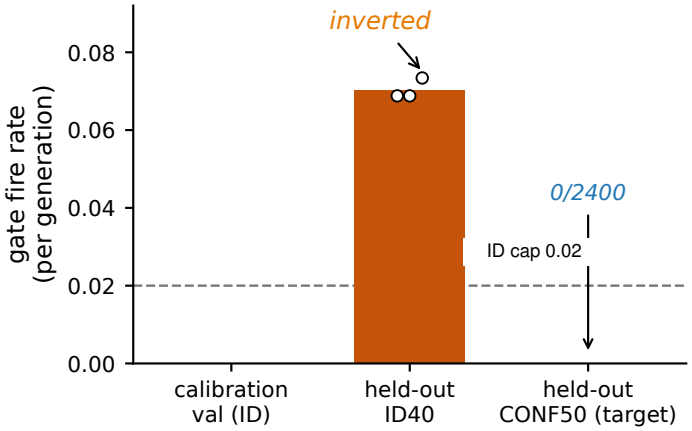

Figure 3: Gate trigger rates. Calibration shows zero validation ID triggers and OOD recall 0.5. On held-out data the detector fires on ID40 but never on CONF50.

and $47/640$ across seeds (6.88%, 6.88%, and 7.34%). The gated system thus leaves target behavior unchanged while firing on in-distribution prompts.

Gate coverage. The gate triggers on none of the 2,400 recorded CONF50 generations. Consequently, GATE and BASE are identical on this target set: the release direction is never applied. At the same time, the detector triggers on 6.9–7.3% of held-out non-confounded generations (Figure 3). By Eq. (6), $\pi = 0$ on CONF50 forces a zero end-to-end effect for any edit, while at $\pi \approx 0.07$ on ID40 a small aggregate change is compatible with a much larger change on the prompts that trigger. We use *gate inversion* for this reversal of desired trigger coverage across the two populations. Calibration coverage, the label, and detector capacity are the natural levers for correcting it, and monitors with more capacity than a linear probe (Oldfield et al., 2026) are one way to test the last.

Attribution. The 25 negative-effect reversals on CONF50 are therefore properties of the base model, not of the injected edit. We next bypass the gate to measure intervention effects directly.

5.4 Ungated Editing

The ungated experiment applies the per-instance direction in Eq. (4) unconditionally at doses $\lambda = 0.5$ and $\lambda = 1.0$. BASE and an always-on fixed-direction control are remeasured within this experiment. This changes both the trigger policy and, relative to the gated fixed-direction pipeline, the direction construction; the always-on fixed control is therefore necessary for interpreting adaptivity.

Offset and gain. The intercept decreases across the three measured conditions, from 0.3820 to 0.3106 and 0.2637 (Table 2; Figure 4).

Table 2: Ungated per-instance injection on CONF50 and ID40. Lower positive intercepts accompany lower gains on CONF50. Negative ID slope-drop values indicate an increase in slope.

Quantity	BASE ($\lambda = 0$)	$\lambda = 0.5$	$\lambda = 1.0$
Intercept b	0.3820	0.3106	**0.2637**
Gain a	0.8236	0.6773	0.5721
ID slope drop	—	-0.0243	-0.0169
Reversals (negative ATE)	$25/27$	—	$21/27$
Positive-world correlation	0.6299	—	0.328

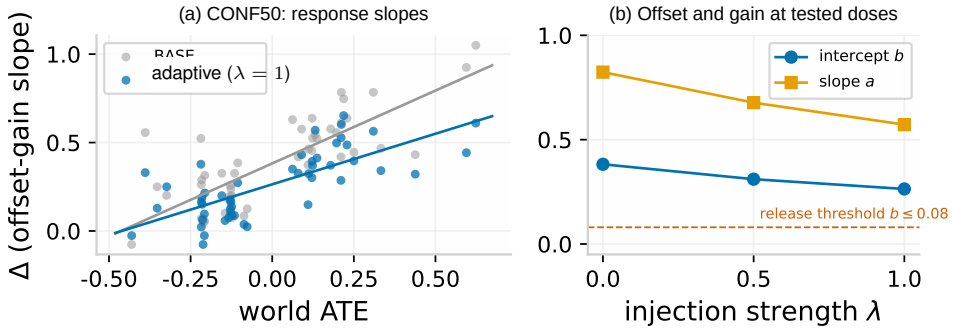

Figure 4: Ungated dose comparison on CONF50. **(a)** Per-world response slope Δ versus ground-truth effect for BASE and full-dose adaptive injection. **(b)** Fitted offset b and gain a at the three measured doses. Both decrease with dose: the full dose removes 31% of the baseline offset while lowering the gain.

The larger dose removes 31.0% of the baseline offset and closes 39.2% of the gap to the target $b \leq 0.08$.

The gain simultaneously decreases from 0.8236 to 0.6773 and 0.5721, so both coefficients change monotonically across the three measured doses. At the larger dose, the gain change has magnitude 0.2515 and the positive-world correlation falls to 0.328: the intervention reduces offset while weakening sensitivity to the true effect. Through Eq. (7), the gain change of -0.2515 is the regression coefficient of the per-world edit effect d_w on ATE_w: the edit's effect is negatively aligned with the true effect, whereas reaching the target would require a positive component of about $1 - a = 0.18$ per unit of ATE. The pooled edit thus supplies the uniform shift that the offset needs, but its component along the true effect has the wrong sign, consistent with evidence that mean-difference steering is reliable only when per-example activation differences are directionally consistent (Braun et al., 2025). On negative-effect worlds, sign reversals decline from $25/27$ to $21/27$, and in-distribution slopes are preserved at both doses.

Adaptive versus fixed edits. The mean per-world response-slope differences from the always-on fixed control are $+0.0268$ at $\lambda = 0.5$ and approximately -0.020 at $\lambda = 1.0$ (Figure 5). Their signs differ across doses, so the per-instance construction shows no consistent advantage over the fixed direction.

Why the pooled edits fall short. The coefficients show what the pooled edit does: at $\lambda = 1$ both the offset and the gain fall to about 0.69 of their base values ($0.2637/0.3820$ and $0.5721/0.8236$), and at $\lambda = 0.5$ to about 0.81 and 0.82. A proportional damping of the response by this factor would lower the gain by $0.31 \times 0.8236 \approx 0.255$, matching the measured 0.2515. The pooled edit therefore reduces the offset by attenuating the whole response rather than by correcting its structure, whereas the transplant, which replaces rows individually, carries the world-specific information that a correction requires.

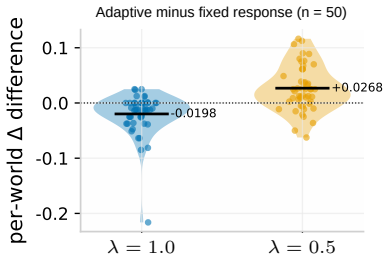

Figure 5: Per-world response-slope differences between adaptive and always-on fixed injection ($n = 50$), at $\lambda = 1.0$ (left) and $\lambda = 0.5$ (right). Mean differences are approximately -0.020 and $+0.0268$, respectively. This per-world slope statistic is distinct from a fitted intercept difference.

6 Discussion

A successful transplant and a successful prompt-computable edit impose different requirements. The former has access to a matched donor and preserves row-specific activation structure. The latter must construct a useful modification from the available prompt, and a gated implementation must also decide when to apply it. Our results separate these requirements within one frozen model: transplant recovery is substantial on a selected localization subset, target-distribution gate coverage is zero, and unconditional pooled edits provide partial correction on the release set.

Routing and intervention. Three measurements are informative for gated activation editing. First, report trigger counts on each target population, including the number of distinct worlds or prompts underlying repeated generations. Second, evaluate the same edit without the gate, alongside the unmodified model, to distinguish non-activation from lack of efficacy. Third, compare an adaptive edit with a fixed control using the same endpoints, doses, and paired uncertainty estimates. The present gate preserved aggregate ID slope while missing every target generation, and only the trigger counts revealed why the gated output was unchanged. Because the coverage decomposition in Eq. (6) holds for any outcome and any edit, these three measurements apply unchanged to other detectors, edit constructions, and tasks.

Localization and editing. The transplant identifies a site that supports recovery under the tested donor intervention, but causal localization need not identify where an edit is most effective (Hase et al., 2023), and a pooled contrast need not contain the information needed to reproduce the donor's effect. The joint reduction in intercept and gain illustrates why a steering evaluation should report both bias and effect sensitivity; excessive steering strength similarly degrades task accuracy in instruction following (Kang & Kim, 2026).

Relation to larger models. The detection–control distinction is consistent with geometric evidence from larger language models (Galeone et al., 2026), and our component-level evaluation shows how it arises inside a single editing pipeline.

Limitations. The study uses one 25.7M-parameter checkpoint and a synthetic task, which provides ground-truth effects and matched donor runs, and leaves other scales and natural-language tasks for future work. The gate uses a single linear detector family calibrated on nine positive examples, and the ungated study tests two doses of one pooled construction.

Conclusion. Localization evidence does not ensure that a gated activation-editing pipeline will improve behavior. The detector never applies its edit on the target generations, while removing it enables partial correction that reduces offset but also effect sensitivity.

Ethics Statement

This study uses a small transformer and synthetically generated causal tasks; it involves no human participants or personal data. Gated activation editing is proposed as a way to control model behavior only when needed, and our results identify a failure mode relevant to that use: a detector calibrated on one distribution can silently never trigger on the target distribution. We therefore recommend that deployments of detector-gated interventions measure trigger coverage on the populations they are meant to protect.

Reproducibility Statement

The main text defines the task, the evaluation populations, and the claims that the experiments test, and states the decomposition used to interpret them. The appendix describes the frozen checkpoint, its training recipe, and the data-generation setup, defines the main measurements, lists the design targets and their outcomes, and describes the experimental chronology and how each outcome was recomputed from archived records.

AI Use Statement

We used generative AI tools to assist with writing and copy-editing the manuscript, reviewing the consistency of claims, searching for and verifying related literature, developing and debugging evaluation code, and checking LaTeX compilation and layout. AI tools also assisted in deriving and writing up the coverage decomposition and coefficient identity of the theoretical framework; the authors checked every derivation and its connection to the data. We take responsibility for the final content of this paper, including text, claims, and code produced with the aid of generative AI.

References

Anonymous. When Observational Context Overrides Interventional Evidence in Transformers, 2026. Concurrent ICLR 2027 submission.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, 2024.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)Reliability of Steering Vectors in Language Models. *arXiv preprint arXiv:2505.22637*, 2025.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Cosimo Galeone, Anna Ettorre, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. Perfect Detection, Failed Control: The Geometry of Knowing vs. Steering in Language Models. *arXiv preprint arXiv:2606.24952*, 2026.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing Causal Structure for Interpretable Neural Networks. In *International Conference on Machine Learning*, 2022.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting Strategic Deception with Linear Probes. In *International Conference on Machine Learning*, 2025.

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. In *Advances in Neural Information Processing Systems*, 2023.

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

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing Causal Reasoning in Language Models. In *Advances in Neural Information Processing Systems*, 2023.

Minjae Kang and Jaehyung Kim. Enhancing Instruction Following of LLMs via Activation Steering with Dynamic Rejection. In *International Conference on Learning Representations*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, 2019.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming Refusal with Conditional Activation Steering. In *International Conference on Learning Representations*, 2025.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, 2023.

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.

Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. Detecting High-Stakes Interactions with Activation Probes. In *Advances in Neural Information Processing Systems*, 2025.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

James Oldfield, Philip Torr, Ioannis Patras, Adel Bibi, and Fazl Barez. Beyond Linear Probes: Dynamic Safety Monitoring for Language Models. In *International Conference on Learning Representations*, 2026.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off Course: Reliability Challenges in Steering Language Models. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the Generalisation and Reliability of Steering Vectors. In *Advances in Neural Information Processing Systems*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, 2020.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors. In *International Conference on Learning Representations*, 2025.

Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *International Conference on Machine Learning*, 2025.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024.

Andy Zou, Long Phan, Sarah Li 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 preprint arXiv:2310.01405*, 2023.

A THE C1 CHECKPOINT

All experiments in this paper use one frozen checkpoint, C1, trained by a concurrent submission (Anonymous, 2026). Its architecture, training, and data generation are as follows.

Architecture. Decoder-only GPT-style transformer, 25.7M parameters, 8 layers, naïve (unfused) attention, vocabulary 41, block size 640.

Optimization. 20,000 steps, batch 64, learning rate 6×10^{-4} with 500 warmup steps. Training data are generated online rather than drawn from a fixed stored corpus.

Worlds. Each world is a parametric structural causal model with $n \in [5, 12]$ nodes, Erdős–Rényi or scale-free graph structure with equal probability, and mixed linear–nonlinear mechanisms. A confounding-strength parameter $\rho \in \{0, 0.3, 0.6, 0.9\}$ governs the unobserved paths $Z \to X$ and $Z \to Y$ for a designated triple. The held-out evaluation family fixes $\rho = 0.9$ and keeps only *Simpson-trap* triples, in which the observational correlation and the causal effect carry opposite signs. Seed domains are disjoint across training, validation, and test.

Task format. A sample is a context of evidence records followed by a query $\mathrm{do}(X{=}x_0) \to Y{=}?$ over the grid $x_0 \in \{-1, -0.5, 0, 0.5, 1\}$. C1 is the mixed-evidence checkpoint: its contexts hold 4 observational records plus 2 interventional probe records on the query node, an 8-line format, with every sample drawn from the interventional query distribution.

What we inherit. The observational-override behavior and the activation-patching result at observational-record positions are properties of C1 established by that submission.

B DESIGN TARGETS AND OUTCOMES

Table 3 lists the design target for each quantity alongside the observed value, including the gated-test quantities of Table 1.

Table 3: Design targets and observed values for the pilot diagnostics, the localization study, and the gated test.

Stage	Quantity	Target	Observed
Pilot	Coverage of normalized recovery	$\geq 25/50$ measurable worlds	$11/50$
Pilot	Cross-evidence feature alignment	CKA ≥ 0.7 (ref. 0.431)	0.061–0.068
Pilot	Detector-label transfer	OOD/transfer AUC above floor	worst 0.366; CLadder reversed
Localization	Eligible worlds	≥ 40	$48/608$ (Wilson $[0.0601, 0.1031]$)
Localization	L23 paired advantage	97.5% CI excludes 0	$0.5628\ [0.3932, 0.7302]$
Localization	L345 paired advantage	97.5% CI excludes 0	$0.8542\ [0.6088, 1.0995]$
Localization	Best-site mean release rate	≥ 0.5	0.8886 ($\mathrm{L23}_{\mathrm{obs}}$)
Localization	Guard-band sensitivity	comparisons stable under ± 0.01	both comparisons unchanged
Gated test	Release intercept b (Eq. (12))	$\leq +0.08$	$0.3820\ [0.3396, 0.4230]$
Gated test	Slope stability / correlation	$\lvert\Delta a\rvert \leq 0.10 \ \wedge\ \mathrm{corr} > 0.40$	0.0; corr 0.6299 ($n_{\mathrm{pos}}{=}23$)
Gated test	Reversals + McNemar	$\leq 14/27 \ \wedge\ p < 0.05$	$25/27$; no discordant pairs
Gated test	ID slope degradation	≤ 0.05	0.0038–0.0042
Gated test	ID trigger rate	≤ 0.02	0.0688–0.0734
Gated test	CONF50 trigger count	diagnostic	$0/2{,}400$

C EXPERIMENTAL CHRONOLOGY

Each experiment followed the same sequence. Thresholds, evaluation sets, and code were fixed and their hashes recorded in an append-only log before the corresponding data were collected, and every outcome in this paper was recomputed from the archived per-generation records. The log comprises 25 entries covering the pilot diagnostics, site selection and the localization study, detector calibration, the gated test, and the ungated experiment. The per-generation records of the gated test show that the gate never triggered on CONF50, so the 25 reversals on negative-effect worlds are the base model's own sign errors rather than effects of the edit.

D MEASUREMENT DEFINITIONS

D.1 BEHAVIORAL MEASUREMENTS

For world w and query span $[\mathrm{lo}, \mathrm{hi}]$, the response slope is

$$\Delta = \frac{\mathrm{parse}(r(\mathrm{hi})) - \mathrm{parse}(r(\mathrm{lo}))}{\mathrm{hi} - \mathrm{lo}}, \tag{10}$$

where r denotes greedy decoding and parse extracts a signed numeric answer. Unparseable outputs are rejected by the evaluation guard. The three conditions are

$$\begin{aligned} \Delta_{\mathrm{same}} &= \Delta \text{ under the original evidence,} \\ \Delta_{\mathrm{zero}} &= \Delta \text{ under evidence zeroing at } \mathrm{obs} \text{ positions,} \\ \Delta_{p} &= \Delta \text{ under the activation intervention.} \end{aligned} \tag{11}$$

The denominator $D = \Delta_{\mathrm{zero}} - \Delta_{\mathrm{same}}$ measures the reference shift induced by zeroing. Worlds with $|D| < 0.05$ are excluded from normalized release-rate statistics. The release rate in Eq. (1) is unbounded: $R = 1$ matches the reference shift, $R = 0$ is no change, $R < 0$ moves in the opposite direction, and $R > 1$ goes beyond the reference. R measures movement toward the reference behavior; agreement with the true effect is measured separately by Eq. (12).

Regression and sign errors. For each condition, ordinary least squares estimates the offset and gain:

$$\widehat{\Delta}_w = a\,\mathrm{ATE}_w + b + \varepsilon_w. \tag{12}$$

The zero-offset, unit-gain target is stated under the benchmark's normalization, in which the generator's effect and the two-point response slope share a scale. The intercept target is one-sided, $b \leq 0.08$, and is paired with the gain, correlation, and ID targets so that a low offset alone does not count as an accurate response. For negative-effect worlds, a reversal denotes a predicted effect of the opposite sign exceeding the 0.05 margin.

Paired inference. For a fixed binary correctness endpoint, let n_{improve} count worlds that change from incorrect to correct, and n_{worsen} count the reverse. The one-sided exact McNemar test (McNemar, 1947) for improvement is

$$p = \Pr\left[\mathrm{Bin}\left(n_{\mathrm{improve}} + n_{\mathrm{worsen}}, \tfrac{1}{2}\right) \leq n_{\mathrm{worsen}}\right]. \tag{13}$$

When both counts are zero, as in the gated test, the exact test gives $p = 1$.

Uncertainty and sampling units. Rate intervals use Wilson's method (Wilson, 1927); regression and paired-difference intervals use the bootstrap (Efron & Tibshirani, 1993). The two localization comparisons use 97.5% intervals, corresponding to a Bonferroni allocation of a 5% family-wise error budget. World-level inference preserves pairing and does not treat multiple generations from one world as independent observations. The 2,400 CONF50 generations and 640 ID40 generations per seed are decoding-event counts, distinct from the 50 and 40 world counts.

D.2 Representation-level measurements

For a site layer ℓ, the detector feature map is

$$\phi_\ell(x) = \mathrm{LayerNorm}\Big(\mathrm{MeanPool}_{\mathrm{obs}}(h^{(\ell)}(x))\Big), \tag{14}$$

with site features formed by concatenating the selected layers. For centered sample-by-feature matrices X and Y, centered linear CKA (Kornblith et al., 2019) is

$$\mathrm{CKA}(X,Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F} = \frac{\langle XX^\top, YY^\top\rangle_F}{\|XX^\top\|_F \, \|YY^\top\|_F}. \tag{15}$$

The rows refer to matched samples; the second expression uses sample Gram matrices, not feature covariance matrices. The terminal detector uses the proxy label $y = \mathbb{1}[D > 0.05]$ and triggers when $u(x) \geq T$. This differs from both a confounding-family label and the two-sided release-measurability criterion.

E Release Figures