BALANCED PERTURBATION AND SUBSPACE ERASURE DO NOT IDENTIFY WHAT AN ACTIVATION MONITOR READS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation monitors score behaviours such as deception from language-model hidden states, but held-out accuracy cannot show whether a monitor reads the behaviour itself or a correlated feature. We analyse two intervention diagnostics for this distinction. Under a balanced perturbation with conditional mean zero, positive and negative first-order score responses cancel in expectation. Its mean score change therefore reflects local curvature rather than target reliance. For linear monitors under elliptical perturbations, variance, mean absolute change, and flip rate are also non-identifying. Matched erasure compares the score loss after removing a fitted class-separating direction with the loss from an equal-size control edit. A monitor that reads a correlated attribute can share the same alignment with this direction as a target reader. In panels with sealed construction labels, matched erasure ranks readers above confound-riders, but no threshold separates them, including after site-level normalization. A margin calibrated on one corpus passes 61 of 189 riders in a corpus-matched follow-up; all fail when their attribute is reversed while the behaviour is held fixed. Two released deception probes pass in only 3 of 14 benchmark cells. Reversing a named candidate attribute classifies 1,046 of 1,117 monitors correctly on held-out records. We recommend reporting this test with its full operating curve rather than treating one intervention result as evidence of target reliance.

1 INTRODUCTION

Language models produce hidden states as they process an input. An activation monitor reads one of these hidden states and gives a score for a behaviour, such as deception. Such monitors may help detect behaviour that changes after deployment (Korbak et al., 2025; Hubinger et al., 2019). Their score is useful only if it depends on the behaviour being checked.

High held-out accuracy does not show this. A monitor may score deception from information about deception. It may instead use another feature that is linked to deception in the evaluation data. We call this feature a *confound*. For example, deceptive answers may be longer in one dataset, although answer length is not evidence of deception. A monitor that uses length can achieve high accuracy while failing when the link between length and deception changes. Earlier work has found similar failures for probes that use correlated text features (Boxo et al., 2025a; Kumar, 2026).

Activation interventions change a hidden state and measure the resulting score change. This idea appears in activation patching, steering, and amnesic probing (Rimsky et al., 2024; Elazar et al., 2021; Chan et al., 2022). The change should not itself point toward one class, since that would insert label information into the test (Makelov et al., 2024; Canby et al., 2025). We study a *floor test*, which adds a perturbation with conditional mean zero and measures the average score change. In simple terms, the perturbation has no average push toward either class. We compare it with *matched erasure*. Matched erasure removes a direction in the hidden state that separates the classes and compares the score loss with the loss from an equal-size control edit (Belrose et al., 2023; Ravfogel et al., 2020).

The floor test cannot show whether a monitor reads the target behaviour. Under a mean-zero perturbation, score changes from opposite directions cancel at first order. The average score change

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.

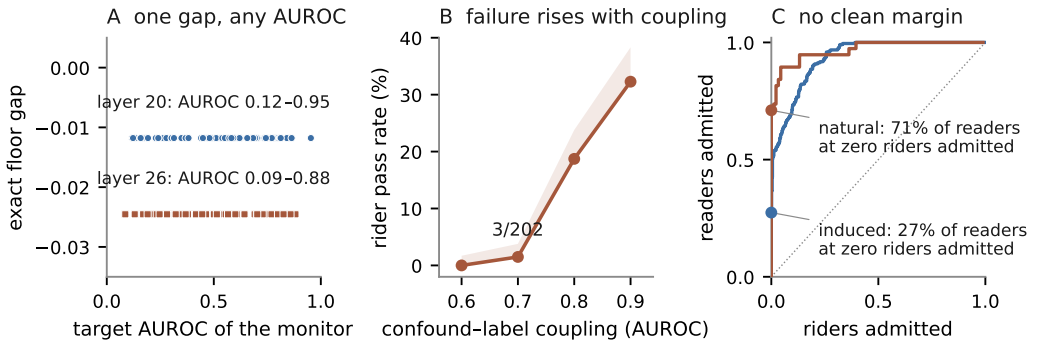


Figure 1: **A** Sixty-seven monitors sharing one quadratic part have one exact floor gap per site, while target AUROC spans 0.118 to 0.953 and 0.086 to 0.881. **B** False passing is near zero at induced couplings 0.60 to 0.70 and rises at 0.80 to 0.90; the transition is gradual. **C** Site-normalised matched erasure: admitting no rider keeps 27% of induced and 71% of natural readers. All panels are drawn from the result records.

that remains measures local *curvature*: how the score bends near the hidden state. A linear monitor has an expected floor gap of zero even when it reads the target behaviour. A nonlinear monitor can have a large gap even when it does not use target information. For linear monitors under elliptical perturbations, mean absolute change, variance, flip rate, and related perturbation statistics have the same problem.

Matched erasure retains a first-order score change, yet it also does not identify what the monitor reads. For a linear monitor, its statistic depends on how closely the monitor's read-out matches the direction that is removed. A monitor that uses a correlated feature can match this direction as closely as a monitor that reads the target behaviour. The two monitors can therefore receive the same result. In our evaluations, matched erasure usually ranks target readers above confound-riders, but no fixed threshold separates them. A margin calibrated in one setting passes 61 of 189 riders in a corpus-matched follow-up. All 61 fail when their attribute is reversed while deception is kept fixed.

Attribute reversal tests a named candidate confound. We reverse the candidate attribute while keeping the audited behaviour fixed, then test whether the monitor still tracks the behaviour. This directly tests whether the monitor depends on the attribute. On held-out records, attribute reversal classifies 1,046 of 1,117 monitors correctly. It also catches many riders when the reversed attribute differs from the one used to train the rider.

- Under mean-zero perturbations, the average first-order score change cancels. The remaining mean change reflects local curvature and does not identify target use. For linear monitors under elliptical perturbations, mean absolute change, variance, flip rate, and related response statistics have the same limitation. Neither negative result covers directed one-sided interventions.
- Matched erasure depends on alignment with a fitted class-separating direction. A confound-rider can share this alignment with a target reader. Across panels of 1,117 and 408 monitors with sealed construction labels, matched erasure ranks readers above riders, but no threshold separates them.
- Attribute reversal changes a named candidate attribute while keeping the behaviour fixed. On held-out records, it classifies 1,046 of 1,117 monitors correctly. We report its admission–coverage curve rather than treating one intervention pass as evidence of target use.

2 Background and Setup

For an input x, let $h_\ell(x) \in \mathbb{R}^d$ be the model activation at layer ℓ. A monitor at this layer maps $h_\ell(x)$ to a scalar score, written as $M_\ell(h_\ell(x))$. A larger score means stronger evidence for the behaviour being checked. Let $B \in \{0, 1\}$ show whether the behaviour is present. Our main experiments study

deception, so $B = 1$ means that the input is deceptive. The tests use the same activation and monitor. They differ only in the edit applied to $h_\ell(x)$ before the monitor is run again. Figure 3 in Appendix B shows both edits.

We label monitors from their training setup before applying a diagnostic. A *reader* is trained on the target-behaviour contrast. A *confound-rider* is trained to predict a chosen non-target attribute. We check that each rider decodes this attribute and does not predict the target beyond the link between the attribute and the target label. We call this link the *coupling*. It is the AUROC obtained when the attribute alone predicts the target label. A coupling of 0.5 means that the attribute gives no information about the label. A coupling of 1.0 means that it determines the label.

These labels describe the monitors' training signal. They do not prove what information each monitor uses after training. Only the implanted-trigger, planted-tag, and response-length controls have a source fixed by construction in this stronger sense. The reader–rider distinction also requires the confound and target behaviour to be different. It is not well defined when the confound causes the behaviour. We tried to construct readers with a removable carrier, but found no eligible site (Appendix F.1). Readers may therefore use unnamed features in the corpus. This makes the reader and rider groups less distinct. It weakens claims about semantic reading, but not the negative results: overlap makes the two groups harder, not easier, to separate.

We study target use: the information that determines a monitor's score. This differs from predictive accuracy and robustness. A monitor can predict the label well by using a correlated feature. It can use target information but fail after a model update. It can remain stable under a shift while using the wrong feature. We edit representations instead of building matched input pairs. Matched pairs require a candidate confound and a control to be named before testing. Matched erasure does not require a named confound, but it removes a fitted subspace without showing what the subspace represents. Appendix A separates these questions and compares related methods, including nullspace projection (Boxo et al., 2025b; Gu et al., 2025; Ravfogel et al., 2020).

3 Why Floor Tests Do Not Identify Reliance

The floor test adds a perturbation to an activation before applying the monitor. For input x, let δ be the perturbation, so the edited activation is $h_\ell(x) + \delta$. A perturbation oracle gives a distribution for δ at each input. Our result requires conditional mean zero:

$$\mathbb{E}_\delta[\delta \mid h_\ell(x)] = 0.$$

This condition means that, after fixing the activation, the perturbation has no average direction. The main oracle, *actdiff*, takes the difference between two observed activations, rescales it to a budget ε, and randomizes its sign. We match the perturbation size to ordinary between-class activation variation. Appendix C.1 defines this oracle and five alternatives.

For each input, the floor gap is the score before the edit minus the expected score after the edit:

$$\mathrm{gap}(x) = M_\ell(h_\ell(x)) - \mathbb{E}_\delta\left[M_\ell(h_\ell(x) + \delta)\right]. \tag{1}$$

A positive gap means that the score falls on average after the perturbation. We estimate this quantity with five-fold cross-fitting and Holm correction (Holm, 1979). Each fitted component is trained on the other folds and scored only on its held-out fold. Thus, each input gives one out-of-fold score. Appendix C.2 gives the full estimator and seed analysis. The floor gap is an average effect under the sampled perturbation distribution. It is not a worst-case robustness measure (Cohen et al., 2019; Katz et al., 2017).

Under a mean-zero perturbation, positive and negative first-order score changes cancel. The floor gap therefore depends only on the part of the score function that is not linear near the activation.

Lemma 1 (Mean-zero perturbation tests measure curvature)**.** *Let h be a fixed activation and let M be a monitor whose gradient is η-Lipschitz on the ball of radius ε around h. Let δ satisfy $\|\delta\|_2 \le \varepsilon$ and $\mathbb{E}_\delta[\delta \mid h] = 0$. Then*

$$\mathbb{E}_\delta\left[M(h+\delta)\right] = M(h) + \mathbb{E}_\delta\left[R(h,\delta)\right],$$

where $R(h,\delta)$ is the Taylor remainder of $M(h+\delta)$ around h. Hence, for every input x with $h_\ell(x) = h$,

$$|\mathrm{gap}(x)| \le \frac{\eta\varepsilon^2}{2}.$$

For a linear monitor, $\eta = 0$, so the expected floor gap is exactly zero.

Conditional mean zero removes the linear term, so the floor gap equals the expected Taylor remainder. Lemma 1 restates this standard fact (Stein, 1981; Nesterov & Spokoiny, 2017), also used to analyse noise-added saliency (Seo et al., 2018); our contribution is its consequence for a reliance audit. The bound only limits the size of this remainder. If the monitor is linear, the remainder is zero. More generally, a nonzero floor gap can come from local curvature even when the monitor does not use the audited behaviour. When curvature is stable near the activation, the gap scales as ε^2. We observe this scaling over a threefold range of perturbation budgets: the surrogate judge has exponent 2.06, the sparse-autoencoder probe has exponent 2.15, and the linear probe remains indistinguishable from zero (Appendix B.11).

The following pair uses different information but has the same expected floor gap at every activation. Let w be the target contrast direction. The first monitor contains a linear term $w^\top h$ that reads the target. The second does not. Both monitors share the same quadratic term.

Proposition 1 (Two monitors with the same floor gap)**.** *For a fixed activation site and a perturbation oracle satisfying the stated conditions, let w denote the class-contrast direction. Let A be symmetric, satisfy $Aw = 0$, and satisfy $h^\top Ah \mid B = 0 \overset{d}{=} h^\top Ah \mid B = 1$. Define*

$$M_1(h) = w^\top h + h^\top Ah \qquad \text{and} \qquad M_2(h) = h^\top Ah.$$

The score of M_1 increases strictly with the target component of h, whereas M_2 is independent of B. Nevertheless, both monitors have the same expected floor gap at every activation:

$$\mathbb{E}_\delta[M_i(h) - M_i(h+\delta)] = -\,\mathbb{E}_\delta\left[\delta^\top A\delta\right], \qquad i \in \{1, 2\}.$$

No estimator based only on the floor gap can distinguish M_1 from M_2, regardless of sample size.

The two monitors differ only in the linear term. This term carries target information in M_1, but its contribution to the floor gap is zero because $\mathbb{E}_\delta[\delta \mid h] = 0$. The quadratic term gives the same gap for both monitors. The construction is not limited to one pair. At two activation sites, a 4096-dimensional family has the same exact floor gap while target AUROC ranges from 0.118 to 0.953 at one site and from 0.086 to 0.881 at the other (Figure 1A).

Statistics based on the size of the score change avoid signed cancellation. These include mean absolute change, variance, flip rate, and held-out AUROC after adding the perturbation. For linear monitors under elliptical perturbations, they retain a first-order term but still do not identify target use. This follows from ellipticity and the invariance of AUROC to rescaling.

Corollary 1 (Second-moment statistics are also non-identifying)**.** *Let $M_1(h) = w_1^\top h$ and $M_2(h) = w_2^\top h$ be linear monitors, and write*

$$\Delta M_i = M_i(h+\delta) - M_i(h) = w_i^\top \delta.$$

If the perturbation law is elliptical with covariance Σ_δ, the law of ΔM_i depends on the monitor only through $\|\Sigma_\delta^{1/2} w_i\|_2$, and the target label does not enter. The target AUROC of a linear monitor is unchanged by positive rescaling of its read-out. Thus, for any M_1 and M_2, one can rescale M_2 so that ΔM_1 and ΔM_2 have the same law. Every statistic computed only from the perturbation response, including $\mathbb{E}|\Delta M|$, $\mathrm{Var}(\Delta M)$, quantiles, and higher moments, then has the same value for both monitors, whatever their target AUROCs.

Standardizing the response removes the rescaling freedom, but then every linear monitor at the same site has the same standardized response distribution. Statistics that also use the original score, such as a flip rate at a fixed threshold or AUROC after perturbation, depend on the ratio between class separation and perturbation response. This ratio is another form of held-out accuracy. It can separate strong and weak predictors, but it cannot show whether the predictor uses the target behaviour or a confound. Appendix B.4 gives the full argument.

For linear probes the result is exact; for the other families, Table 1 states the source of any nonzero gap. For sparse-autoencoder feature probes and ReLU surrogate judges, a nonzero floor gap comes from crossing activation boundaries. For attention detectors, the result requires a smoothness assumption that we do not verify. The derivations for piecewise-linear monitors are in Appendix B.6, and the curvature measurements are in Appendix B.10.

Table 1: What the floor-gap result implies for each monitor family.

Family	Exact result	Source of a nonzero gap	Status
Linear probe	Expected gap is zero	None	Exact
SAE feature probe	Gap is zero within one gate region	Gate crossings	Bound (Lemma 2)
ReLU surrogate judge	Gap is zero within one activation region	Gate crossings	Bound (Lemma 3)
Attention detector	Requires a smoothness assumption	Curvature	Empirical diagnostic

Exact conditional mean zero is an ideal case. A small nonzero mean brings back a first-order term in the direction of the monitor gradient. In our surrogate-judge settings, a residual mean of about 5% of the perturbation budget changed the floor gap. The evaluated oracles were consistent with zero at the available precision (Appendix B.2).

4 Matched Erasure and Evaluation Design

Matched erasure compares two edits. First, it removes a fitted direction that separates the behaviour classes. Second, it applies an equal-size control edit that is orthogonal to this direction. The test compares the score loss from the two edits. It is based on amnesic probing (Elazar et al., 2021), with a norm-matched control edit. A larger score loss after erasure suggests that the monitor uses information in the fitted direction. This does not show what that direction represents.

We fit the removed direction on training data using LEACE and iterated nullspace projection (Belrose et al., 2023; Ravfogel et al., 2020). The control edit is orthogonal to this direction. A monitor passes the fixed rule only when two conditions hold. First, erasure must cause a sufficiently larger standardized score loss than the control edit. Second, the monitor must separate the held-out behaviour classes. Let D_{std} be the difference in score loss on held-out $B = 1$ inputs, divided by the pooled within-class standard deviation of the unedited held-out scores. Let separation be the difference between mean unedited scores for $B = 1$ and $B = 0$ held-out inputs. We test both conditions with cross-fit estimates. A pass requires $D_{\mathrm{std}} > \tau_{\mathrm{rel}} = 0.25$, positive separation, Holm-corrected significance, and success in all three prespecified seeds.[1] Algorithm 2 in Appendix D.5 gives the full procedure.

A pass means only that the monitor meets this rule at the chosen operating point. It does not establish that the monitor reads the audited behaviour, remains robust under attack, or is suitable for deployment.[2]

Matched erasure avoids the cancellation in the floor test because the erased and control edits are different. It still has an identification problem. For a linear monitor, the statistic depends on the monitor read-out only through its projection onto the erased direction, after whitening, and its score scale. A confound-rider can have the same projection and scale as a reader.

Proposition 2 (Matched erasure reads a projection)**.** *Fix an activation site and a rank-one erasure map with erased whitened direction u. Let $a = \Sigma_T^{1/2} w$ be the whitened read-out of a linear monitor, and let*

$$\cos\theta_w = \frac{\langle a, u\rangle}{\|a\|_2}.$$

The standardized matched-erasure statistic is determined by the read-out's alignment with the erased direction and its score scale. Thus, monitors that use different information receive the same statistic whenever they share these two quantities. The margin is met exactly when

$$\cos\theta_w \, \frac{\|\Sigma_T^{1/2} w\|_2}{s_{\mathrm{pool}}} \;\geq\; \frac{2\tau_{\mathrm{rel}}}{\Delta}, \tag{2}$$

[1] The decision rule uses an intersection–union design (Berger, 1982).

[2] The filed preregistration and archived records call a pass a "certification"; we use "pass" to avoid implying formal assurance, and keep the archival term only in identifiers and quoted protocol language. The endpoint, threshold, procedure and records are unchanged.

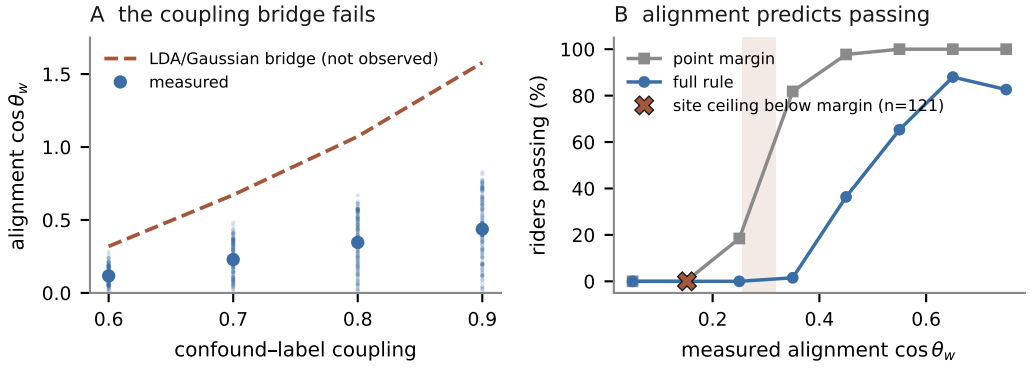


Figure 2: **A** Measured alignment against induced coupling, with the LDA/Gaussian bridge that would be predicted but is not observed. **B** Riders passing against measured alignment under the point margin and the full rule. The band is the alignment the fixed margin requires at the 426 riders whose site admits passing; the cross marks the 121 riders at sites whose maximum attainable statistic falls below that margin, which are excluded from the band. The relationship is descriptive of this construction and these sites, not a general theorem.

where Δ is the held-out separation along the erased direction and s_{pool} is the pooled within-class score standard deviation. Any confound-rider that meets this condition passes and is indistinguishable from a reader that has the same alignment and scale.

A fixed margin cannot transfer across settings when readers and riders can reach overlapping alignments.

Corollary 2 (No fixed margin transfers when achievable alignments overlap)**.** *Let c_S be the alignment needed to meet margin τ in setting S, as given by equation 2. If a reader in setting S_1 has alignment below c_{S_1} and a rider in setting S_2 has alignment at least c_{S_2}, then*

$$D_{\text{std}}(M_{\text{rider}}) \geq \tau > D_{\text{std}}(M_{\text{reader}}).$$

No single margin τ can then pass every reader and reject every rider across S_1 and S_2.

In our panels, rider passing begins where measured alignment reaches the range required by the fixed margin. Mean rider alignment rises from 0.229 at induced coupling 0.70 to 0.346 at coupling 0.80. At sites where passing is possible, the fixed margin requires alignment between 0.255 and 0.319. Readers have mean alignment 0.773. Rider passing rises gradually with coupling rather than at one universal threshold. At coupling 0.70, 55 of 144 riders clear the point margin, but only 3 pass the full three-seed rule. At coupling 0.60, no rider clears the point margin. At the low-separation Among Us site, the maximum possible statistic is below the fixed margin, so neither readers nor riders can pass. Figure 2 shows these results.

The linear-Gaussian model in Appendix B.9 gives a link between coupling and alignment for a rider trained by linear discriminant analysis. This link does not hold for our orthogonalized induced-rider construction. The model overstates measured alignment by a median factor of 2.97 and has no positive correlation with it ($r = -0.116$). We therefore use measured alignment, rather than coupling alone, to explain passing.

The erasure is rank one, although the reader statistic reaches a plateau as rank increases at six sites. It also remains incomplete out of sample at every site with signal, so the test measures reliance on the fitted direction rather than on all decodable deception information. Each site has a maximum statistic for a linear monitor, which we call its *ceiling*; it is set by class separation. A site whose ceiling falls below the fixed margin cannot admit a passing linear monitor. The margin was calibrated on directions with no useful information, not on confounds that predict the label. Appendix C.3 gives the inference details, Appendix B.7 gives the misalignment bound (Saillenfest & Lemberger, 2025), and Appendix D.5 gives rank and standardization analyses.

We evaluate four monitor families: linear probes, sparse-autoencoder features, two-layer ReLU surrogate judges, and attention-pattern detectors. We compute the floor gap for all four families.

We evaluate matched erasure for linear probes, surrogate judges, and logistic heads over the top k sparse-autoencoder latents. For $k \in \{8, 32, 128\}$, readers pass in 30 of 30 cases at every k. Riders at coupling 0.90 pass in 13 of 27, 16 of 30, and 17 of 30 cases, respectively. A multi-latent read-out therefore behaves like a linear probe. We also test a nonlinear level-set eraser. On the balanced panel at coupling 0.90, it passes 7 of 131 riders and 84 of 159 readers, with reader–rider AUROC 0.818, compared with 0.827 for the linear eraser (Appendix B.7). We do not apply matched erasure to attention detectors.

We evaluate four settings: a controlled construction panel, a blinded preregistered panel of independently specified monitors, coupling sweeps that vary how strongly a confound predicts the label, and monitors built outside this pipeline. Table 10 reports reader and rider outcomes in each setting. The main evaluation studies deception. Appendices E.3 and E.4 also study concealment and sandbagging. The released records contain the full grids, hook layers, benchmark variants, and compute details.

5 Results

5.1 The Floor Test Does Not Control False Passes

The floor test fails on negative controls. At a nominal 5% rate, an untrained surrogate judge passes in 112 of 200 trials after the three-seed intersection: 56.0% with a Clopper–Pearson interval of $[48.8, 63.0]$. A surrogate judge trained on permuted labels passes in 18 of 200 trials: 9.0% with interval $[5.4, 13.9]$. Under the evaluated estimator, the corresponding rates are 58.0% and 39.0% (Appendix C.2). Neither control contains useful deception signal. A floor-test pass therefore does not show that a monitor uses deception information.

The correction lowers the permuted-label rate from 39.0% to 9.0%, which shows that underestimated standard errors explain part of the excess. The untrained judge still passes often after correction, consistent with curvature-driven score changes. In a piecewise-linear monitor, curvature comes only from crossing activation gates. Across site-and-seed cells, the number of crossed units predicts the floor gap (Appendix D.3).

A randomly selected sparse-autoencoder feature provides a third control. It never passes. No linear probe passes in the main evaluation. Lemmas 2 and 3 predict both results: their score changes require gate crossings. Table 8 reports the full results. We therefore use matched erasure for the remaining target-use analysis.

Under the same decision rule, the one-sided class-contrast intervention passes 0 of 600 screened random directions and 0 of 65 trained riders. This intervention is not covered by the mean-zero floor-test result. It also separates the 23-monitor ground-truth panel. However, an aligned rider can pass after its read-out is rotated toward the class direction; at this point, its clean deception AUROC is 0.774 (Appendix A).

5.2 Controlled Separation and Transfer

The controlled panel contains $1{,}117$ monitors: 257 readers and 860 confound-riders. Labels were set from the monitor construction before any diagnostic was applied. Every rider passed the validity checks in §2.

We compare five diagnostics against these labels: matched erasure, held-out AUROC, resample ablation (Chan et al., 2022), and two shift controls that require a named candidate attribute. All five rank readers above riders, with reader–rider AUROC from 0.910 to 0.978. None has a threshold that separates the groups. Table 2 reports the comparison.

Matched erasure gives the best agreement among methods that do not require a named confound. It classifies 927 of $1{,}117$ monitors correctly when a passing reader and a non-passing rider count as correct. Its reader–rider AUROC is 0.910, but the score distributions overlap at every threshold (Figure 5A in Appendix D). The full rule passes 213 of 257 readers and 146 of 860 riders.

Site normalization improves ranking but does not solve the threshold problem. Dividing each score by its site's ceiling increases the reader–rider AUROC from 0.907 to 0.935 on the induced panel

Table 2: Diagnostics scored against the sealed construction labels of the 1,117-monitor panel. No diagnostic has a single threshold that separates readers and riders. The two shift controls require a candidate attribute to be named before testing. Both are evaluated on records not used by the construction-label check.

Diagnostic	Labels correct	Threshold separates?	Named attribute required?
Matched erasure	927/1,117	no	no
Held-out AUROC	789/1,117	no	no
Resample ablation	466/1,117	no	no
Orthogonalized control	838/1,117	no	yes
Anticorrelated control	1,046/1,117	no	yes

and from 0.918 to 0.968 on the natural panel. The two groups still overlap at every threshold. A threshold that retains 90% of induced readers admits 43 of 91 natural riders.

The controlled panel is development evidence. The fixed margin was chosen after inspecting a related null (Appendix D.5). We compute rates over configurations that pass a prespecified admissibility screen, while keeping non-passing configurations in the denominator (Appendix E.5.1).

The balanced follow-up pairs readers and riders from the same corpus, recipe, and hook site across five construction seeds. This gives 74 clusters at coupling 0.90. Readers pass more often than riders: the within-cluster reader-minus-rider pass difference is $+0.437$, with cluster-bootstrap interval $[0.336, 0.538]$. The rule still passes 61 of 189 riders, with a cluster-bootstrap interval of $[0.236, 0.410]$ over 96 clusters. It passes 180 of 219 readers. These are rates for these constructions, not population false-pass rates.

We reverse each rider's attribute against the label on records held out from the construction-label check. Of the 860 riders, 833 fall below chance under this test, including all 61 passing riders at coupling 0.90. The 61 passing riders at coupling 0.90 have mean decoupled AUROC 0.109. No reader falls below chance, and induced readers retain mean AUROC 0.921. Across all induced couplings, passing riders rely more strongly on their attribute than rejected riders: their mean decoupled AUROC is 0.108, compared with 0.216 for rejected riders.

On held-out records, reversed-attribute testing classifies 1,046 of 1,117 monitors correctly, the highest count among the evaluated diagnostics. It catches 695 of 769 rider–attribute pairs when the reversed attribute matches the rider's training attribute. It also catches 1,097 of 1,376 pairs when the attribute differs, and keeps readers in 630 of 633 pairs. On the natural-confound panel, where attributes are not orthogonalized, mismatched reversal still catches 137 of 156 pairs. The cost is reader coverage: it keeps 26 of 38 readers, compared with 33 for matched erasure.

The preregistered panel does not support a clean transfer estimate. It is structurally confounded: 167 of its 168 riders come from one pooled-sycophancy corpus, while readers come from three other corpora. Only one cluster contains both a reader and a rider from the same corpus. The frozen rule passes 140 of 168 riders and 44 of 66 readers. It therefore labels 72 of 234 monitors correctly. We report this panel because it was preregistered, but its corpus imbalance prevents a reader–rider comparison that holds corpus fixed.

5.3 Three Regimes With Different Failure Modes

The results differ across three settings: controlled confounds added during construction, confounds already present in a corpus, and monitors built outside the calibration pipeline. Appendix E.1 gives the full designs and results.

For induced confounds, rider passing rises with coupling. It is 0 of 175 riders at coupling 0.60, then 3 of 202 at 0.70, 38 of 203 at 0.80, and 61 of 189 at 0.90 (Figure 6 in Appendix E). These values describe this construction. They do not define a general safe coupling threshold.

Riders pass on Llama-3.1-8B, Gemma-2-9B, and Llama-3.3-70B. A post-freeze extension finds the same pattern across four architectures from three developers (Appendix E.1).

For natural confounds, we test twelve prespecified attributes in eight corpora. The rule passes 44 of 91 riders and 33 of 38 readers. Passing riders have couplings from 0.676 to 0.781, below the induced range where passing first appears. Coupling alone therefore does not predict transfer. Natural riders

in this range have mean alignment $\cos\theta_w = 0.397$, compared with 0.176 for induced riders in the same range. Appendix E.2 gives the corpus-level results.

On monitors built outside our pipeline, the rule often abstains. For released Qwen3-14B sandbagging probes, the main reason is missing held-out class separation. On released model organisms (Sheshadri et al., 2026), the rule passes 1 of 23 reader arms and 1 of 27 rider cells. On 25 released Qwen3-14B sandbagging probes, the rule abstains in 74 of 88 cells. These low counts show limited coverage, not low false-pass rates. We also test two logistic deception probes released with Goldowsky-Dill et al. (2025) on Llama-3.3-70B across seven benchmarks. They pass in 3 of 14 cells. In four further cells, a probe separates the classes but fails the reliance part of the rule. All four sites have a ceiling above the margin, so the rule could have passed them.

6 Discussion and Limitations

We ask whether an intervention test can reveal what information an activation monitor uses. The floor test cannot answer this question from its mean score change. Under a mean-zero perturbation, the first-order score change cancels, so the test measures local curvature. Matched erasure keeps a first-order effect, but for linear monitors its score depends on alignment with the fitted erased direction. A monitor that uses a correlated attribute can share this alignment with a target reader. A positive result from either test alone does not establish that a monitor reads the audited behaviour.

Matched erasure can still be useful within one fixed setting. In our controlled panels, it often ranks readers above riders. Its limit is the decision threshold. The threshold depends on the activation site, class separation, monitor family, data construction, and the geometry of correlated attributes. A threshold calibrated in one setting may pass riders or reject readers in another. A non-pass also has several possible causes: weak class separation, poor alignment with the erased direction, incomplete erasure, or limited statistical power. It does not show that a monitor ignores the behaviour.

Attribute reversal is a stronger test when a candidate confound is available. It checks whether the monitor still tracks the behaviour after the candidate attribute is changed against the label. It caught most riders in our panels, including many trained on a different attribute. Its limit is that unknown confounds cannot be tested. A candidate list or an evaluation-set construction covers only attributes named in advance. The test can also reject a reader when the reversed attribute changes a feature that the reader should use. It should therefore be reported with its admission–coverage curve, not as a single final verdict.

Our evidence covers deception most broadly, with smaller studies of concealment and sandbagging. Coverage remains limited across behaviours, monitor families, activation sites, and deployment-like settings. Reader labels describe training contrasts, not a proof that readers use only semantic behaviour information. We could not construct a reader with a removable carrier under the prespecified criterion (Appendix F.1). The results also do not give security guarantees against an adaptive developer. Riders still pass after style augmentation at couplings of 0.80 and above. In 2 of 18 adaptive attack trajectories, the rule passes a monitor with clean detection near chance. A directed class-contrast intervention can be passed by an aligned rider, and a model that changes its activations under monitoring can also change the statistic (McGuinness et al., 2025). An audit is evidence about one stated evaluation setting, not approval for deployment.

7 Conclusion

Activation-monitor evaluations should separate prediction from reliance. A monitor can predict an audited behaviour while using a correlated feature, and a single score change under a balanced perturbation or matched erasure does not show what information the monitor uses. When plausible confounds can be named, evaluations should reverse each attribute while keeping the audited behaviour fixed, measure the effect on held-out records, and report the full admission–coverage curve across models, activation sites, data constructions, and candidate attributes. Reports should state the tested confounds, the intervention, the operating curve, and the limits of the resulting claim.

Ethics Statement

This work studies methods for checking whether activation monitors use information about an audited behaviour. The benchmarks contain scripted deception and judge-elicited sycophancy for research evaluation. They do not use production prompts. We used no human participants or personally identifiable information. The evaluated models and benchmarks are publicly available for research use.

The main dual-use risk is that the methods may help developers or attackers avoid activation monitors. Prompt obfuscation, adversarial fine-tuning, or activation changes could reduce a monitor's score without changing the underlying behaviour. A published decision threshold could also be optimized against. We reduce these risks by evaluating held-out class separation, reporting when the rule fails or abstains, and treating each threshold as specific to its evaluation setting.

A positive result applies only to the evaluated monitor, activation site, and data distribution. It is not evidence that a system is safe to deploy. Evaluations should state the monitor configuration, report all tested settings, separate monitor development from evaluation where possible, and report the confounds and limits covered by the test.

Reproducibility Statement

Each result is linked to a record containing the configuration that produced it. The supplement provides full proofs, per-cell and seed-level results, sensitivity analyses, and an artifact index. The repository contains machine-readable result records and scripts that regenerate every reported table and figure. Code and records are available at `https://anonymous.4open.science/r/iclr_paper1-4DEF/`. The full evaluation uses about 40 H200 GPU-hours. After activation caching, one monitor at one activation site evaluates in under one minute.

The three prespecified evaluation seeds test reproducibility on fixed evaluation records. They are not independent replications. We also draw two new record sets with fresh activation extraction and judge scoring. The main conclusions remain unchanged; Appendix E.6 gives the detailed results.

The perturbation oracles and their conditions, cross-fitting, Holm correction, seed intersection, and false-positive calibration were fixed before the main evaluations. The matched-erasure margin is the exception: Section 5.2 explains how it was selected and how this affects its interpretation. Appendix D.5 and Appendix D.6 report margin sensitivity. Appendix E.1 contains the preregistration, amendments, and the separate filing for the balanced follow-up.

AI Use Statement

Large language models were used as evaluation subjects and as tools for boilerplate code, LaTeX formatting, and prose editing. The paper identifies the evaluated models and the judge model used to label inputs for the surrogate-judge monitor. Generative AI tools were not used to generate synthetic data, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide proof steps, propose or refine hypotheses, design experiments, interpret results, or make scientific decisions. The authors reviewed and verified all AI-assisted text and code. The authors take responsibility for the final paper, including its text, claims, code, and experimental artifacts.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Roger L Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982.

Gerard Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. *arXiv preprint arXiv:2509.21344*, 2025a.

Gerard Boxo, Ryan Socha, Daniel Yoo, and Shivam Raval. Caught in the act: a mechanistic approach to detecting deception. *arXiv preprint arXiv:2508.19505*, 2025b.

Marc E Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878, 2025.

Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *AI Alignment Forum*, volume 2, pp. 19, 2022.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Evan Duan. Do activation monitors survive model updates? benchmarking, predicting, and repairing activation-monitor staleness. *arXiv e-prints*, pp. arXiv–2606, 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Max Fomin, Elad David, and Amit LeVi. Internal-state probes read the situation, not the action: Three negative results for pre-action misalignment monitoring. *arXiv preprint arXiv:2606.30449*, 2026.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Forty-second International Conference on Machine Learning*, 2025.

Tianle Gu, Kexin Huang, Zongqi Wang, Yixu Wang, Jie Li, Yuanqi Yao, Yang Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. Probing the robustness of large language models safety to latent perturbations. *arXiv preprint arXiv:2506.16078*, 2025.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Evan Hubinger, Chris Van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Dylan Jayabahu. The truth was never gone: Perfect aliasing in compliant-context truth probes. *arXiv preprint arXiv:2609.10739*, 2026.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, 2017.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

Sachin Kumar. Pressure-testing deception probes in llms: Scaling, robustness, and the geometry of deceptive representations. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 472–489, 2026.

Daming Luo, Christy Liang, and Junyu Xuan. Steercheck: Attribution specificity and alignment leakage in activation-steering audits. *arXiv preprint arXiv:2608.24335*, 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. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors. *arXiv preprint arXiv:2512.11949*, 2025.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Advances in neural information processing systems*, 12, 1999.

Vikram Natarajan, Devina Jain, Shivam Arora, Satvik Golechha, and Joseph Bloom. One probe won't catch them all: Towards targeted deception detection. *arXiv preprint arXiv:2602.01425*, 2026.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Alexander Panfilov, Evgenii Kortukov, Kristina Nikolić, Matthias Bethge, Sebastian Lapuschkin, Wojciech Samek, Ameya Prabhu, Maksym Andriushchenko, and Jonas Geiping. Strategic dishonesty can undermine ai safety evaluations of frontier llms. In *International Conference on Learning Representations*, volume 2026, pp. 49352–49410, 2026.

Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint arXiv:2507.12691*, 2025.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Antoine Saillenfest and Pirmin Lemberger. Nonlinear concept erasure: a density matching approach. *arXiv preprint arXiv:2507.12341*, 2025.

Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. *arXiv preprint arXiv:1806.03000*, 2018.

Abhay Sheshadri, Aidan Ewart, Kai Fronsdal, Isha Gupta, Samuel R Bowman, Sara Price, Samuel Marks, and Rowan Wang. Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors. *arXiv preprint arXiv:2602.22755*, 2026.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

A Positioning and Comparisons

Activation-monitor evaluations address three different questions. First, does a monitor predict the label on held-out data? Second, does its score depend on the audited behaviour rather than on a correlated feature? Third, does the monitor remain useful after a distribution shift, model update, or attack? This paper studies the second question. Table 3 lists common intervention-based checks and the claim that a positive result does not establish.

We reviewed the 22 monitor and probe papers cited in this work and recorded the validation evidence each paper reports. Held-out accuracy appears in 14 papers, adversarial or attack evaluation in 17, distribution-shift robustness in 12, and a steering or one-sided intervention check in 12. A balanced perturbation statistic appears in 1 paper, and an erasure or ablation comparison appears in 1. These counts are not exclusive. They describe this paper's citation list, not a systematic literature review.

The floor test is therefore not a standard field method. It formalizes the requirement that an edit should not insert label information into a reliance test. The intervention checks used more often are one-sided. The mean-zero result in Lemma 1 does not apply to them.

Table 3: Activation interventions and the limits of a positive result. Lemma 1 applies only to mean-zero perturbation tests.

Intervention	Mean-zero	Main purpose	A positive result does not establish
Floor test	yes	Score change under balanced perturbations	Target use. An untrained dense monitor passes in 56% of 200 trials against a nominal 5% rate
Directed class-contrast push	no	Score change under a one-sided class direction	That the monitor uses this direction on unedited inputs. The procedure has no calibrated null
Steering validation (Panfilov et al., 2026)	no	Effect of adding a fitted direction	That the evaluated monitor uses the fitted direction on natural inputs
Amnesic probing without a matched control (Boxo et al., 2025b)	no	Score degradation after property removal	Whether the degradation exceeds a meaningful reference level
Causal scrubbing, resample ablation (Chan et al., 2022)	no	Score change after replacing an activation under a hypothesis	Target use without a calibrated threshold; on the controlled panel it agrees with 466 of 1,117 construction labels (Table 7)
Instruction-pair probe design (Natarajan et al., 2026)	n/a	Quality of the training contrast	Whether the trained monitor uses target information
Update and shift evaluation (Duan, 2026; Kumar, 2026)	n/a	Stability after model or data changes	Target use
Matched erasure	**no**	Score loss after removing a fitted class-separating direction, relative to an equal-size control edit	That a threshold calibrated in one setting transfers to another

Several recent studies report related forms of alignment or aliasing. Luo et al. (2026) find that sign-randomized steering nulls retain much of the target direction, with effects tracking signed cosine at $\rho = 0.94$. Jayabahu (2026) find that a truth probe trained where truth and the prescribed action coincide cannot separate the two. Fomin et al. (2026) find that internal-state probes for pre-action monitoring track the situation rather than the action. Parrack et al. (2025) compare deception probes by the gain from white-box over black-box monitoring. These studies give empirical evidence for specific settings. Here, the floor-test result gives an identity for mean-zero perturbations, and the matched-erasure analysis tests a threshold rule against reader and rider labels set before evaluation.

A directed class-contrast push is not covered by the mean-zero result. Under the point rule, it passes 1 of 65 trained confound-riders. Balance causes the first-order term to cancel; removing balance restores the label-directed signal that the balanced edit was designed to avoid.

B Formal Results

Let $M : \mathbb{R}^d \to \mathbb{R}$ be a monitor, let $h = h_\ell(x)$ be an activation, and let δ be a perturbation. The floor gap is

$$\mathrm{gap}(x) = M(h) - \mathbb{E}_\delta \left[M(h + \delta) \right].$$

B.1 Mean-Zero Perturbations and Curvature

Lemma (Lemma 1, with the remainder written explicitly)**.** *Let $M : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Suppose ∇M is η-Lipschitz on the ball of radius ε around every activation $h = h_\ell(x)$ in the evaluation set, with one constant η for all activations. Let δ satisfy $\|\delta\|_2 \le \varepsilon$ and $\mathbb{E}_\delta[\delta \mid h] = 0$. Then*

$$|\mathbb{E}_x[\mathrm{gap}(x)]| = |\mathbb{E}_x \mathbb{E}_\delta \left[R(h(x), \delta) \right]| \le \frac{\eta \varepsilon^2}{2},$$

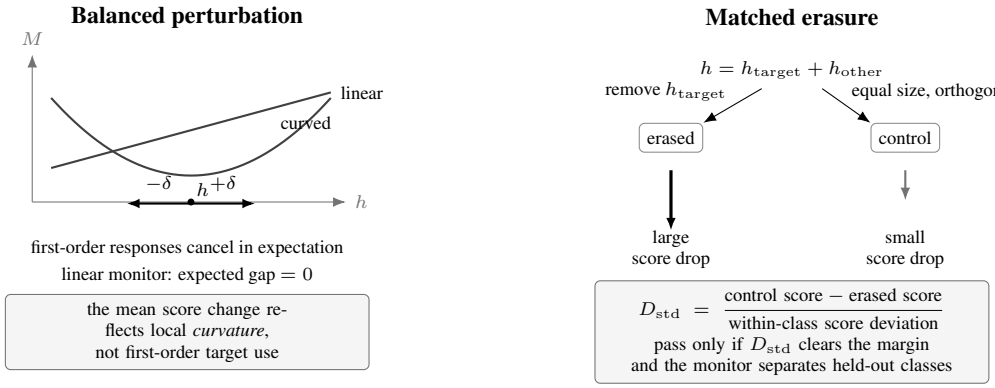


Figure 3: Balanced perturbations add changes with conditional mean zero, so opposite first-order score responses cancel in expectation. Matched erasure removes a fitted class-separating direction and compares its score loss with an equal-size orthogonal control edit. The two procedures measure different quantities.

where $R(h, \delta)$ is the Taylor remainder of $M(h + \delta)$ around h. If M is linear, then $\eta = 0$ and $\mathbb{E}_x[\mathrm{gap}(x)] = 0$.

Proof. Taylor expansion around h gives

$$M(h + \delta) = M(h) + \langle \nabla M(h), \delta \rangle + R(h, \delta).$$

Taking the expectation over δ removes the linear term:

$$\mathbb{E}_\delta[M(h + \delta)] = M(h) + \mathbb{E}_\delta[R(h, \delta)],$$

because $\mathbb{E}_\delta[\delta \mid h] = 0$. Therefore,

$$\mathrm{gap}(x) = -\mathbb{E}_\delta[R(h(x), \delta)].$$

The Lipschitz-gradient condition gives

$$|R(h, \delta)| \leq \frac{\eta}{2}\|\delta\|_2^2 \leq \frac{\eta\varepsilon^2}{2}.$$

Taking expectations over δ and x gives the stated bound. If M is linear, then $R(h, \delta) = 0$ for every h and δ, so the expected floor gap is zero. □

Under conditional mean zero, the floor gap is the expected Taylor remainder. Its size depends on the perturbation budget and local curvature, not on whether the monitor uses target information, a correlated feature, or neither.

B.2 Approximate Mean-Zero Perturbations

Exact conditional mean zero is an ideal condition. Let $m(h) = \mathbb{E}_\delta[\delta \mid h]$ be the residual conditional mean of a perturbation. Taylor expansion gives $\mathrm{gap}(x) = -\langle \nabla M(h), m(h) \rangle - \mathbb{E}_\delta[R(h, \delta)]$. A nonzero $m(h)$ therefore restores a first-order score change. If ∇M is η-Lipschitz on the perturbation ball, then $|\mathbb{E}_x[\mathrm{gap}(x)] + \mathbb{E}_x[\langle \nabla M(h), m(h) \rangle]| \leq \eta\varepsilon^2/2$.

For the surrogate judge, a residual mean aligned with the monitor gradient at about 5% of the perturbation budget materially changes the floor gap. Direction matters: a mean aligned with the gradient changes the gap at smaller magnitudes than one along the class-contrast direction, while a mean in a fixed random direction has no detectable effect up to 50% of the budget. These values are specific to the evaluated monitor and perturbation settings.

The evaluated oracles are consistent with conditional mean zero at the available precision. Estimating a residual mean uses the same finite perturbation draws as the floor-gap estimate. With $n = 200$ perturbations, the sampling scale is $1/\sqrt{n} = 0.0707$, larger than the few-percent residual means that can affect the statistic. The measured first-moment fractions range from 0.0008 to 0.0251 across the six evaluated oracle values. Per-oracle measurements and the controlled-bias sweep are included in the released records.

B.3 Indistinguishable Monitor Pairs

Proposition 1 constructs two monitors with different target use and the same expected floor gap. Let w be the target contrast direction, and let A be a symmetric matrix with $Aw = 0$. The two monitors are $M_1(h) = w^\top h + h^\top Ah$ and $M_2(h) = h^\top Ah$. The linear term in M_1 carries target information. Under conditional mean zero, its contribution to the floor gap is $-w^\top \mathbb{E}_\delta[\delta \mid h] = 0$. Both monitors therefore have the same gap, $-\mathbb{E}_\delta[\delta^\top A\delta]$.

The construction also requires $h^\top Ah \mid B = 0 \overset{d}{=} h^\top Ah \mid B = 1$. This condition makes the quadratic term independent of the target label. It is sufficient for the example, not a claim that all monitors satisfy it. One pair is enough to show that no estimator based only on the floor gap can identify target use in every setting.

We evaluate the construction at two activation sites from the main experiments. At each site, the family contains 67 monitors with the same quadratic term and different linear terms: no linear term, the target read-out, an orthogonal confound direction, and 64 random directions. The exact floor gap is identical across the family when computed with antithetic draws. Target AUROC ranges from 0.118 to 0.953 at one site and from 0.086 to 0.881 at the other (Figure 4).

The deployed estimator uses finitely many sign-randomized draws. Its estimated gaps can therefore differ across the family because each monitor has a finite-sample first-order residual. The largest residual occurs for the target read-out. Relaxing $Aw = 0$ still leaves a companion monitor at chance prediction over a nonzero range of perturbation budgets. The construction rules out identification from the floor-gap statistic; it does not rule out every possible intervention for testing target use.

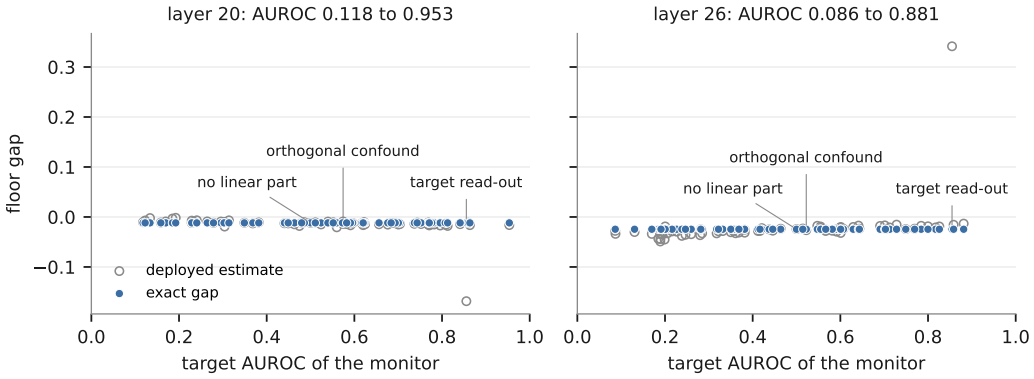


Figure 4: Floor gap and target AUROC for 67 monitors with one shared quadratic term at two activation sites. The exact floor gap is constant within each family.

B.4 Second-Moment Statistics

For a linear monitor $M(h) = w^\top h + b$, the perturbation response is $\Delta M = M(h + \delta) - M(h) = w^\top \delta$. It does not depend on h and therefore does not contain the behaviour label B. If δ has an elliptical distribution with covariance Σ_δ, then ΔM has a one-dimensional elliptical distribution with scale $\|\Sigma_\delta^{1/2} w\|_2$.

Proof of Corollary 1. Let $\sigma_i = \|\Sigma_\delta^{1/2} w_i\|_2 > 0$ for two monitors $M_i(h) = w_i^\top h$. Rescale M_2 by $c = \sigma_1/\sigma_2$. Then $\|\Sigma_\delta^{1/2}(cw_2)\|_2 = \sigma_1$, so ΔM_1 and $\Delta(cM_2)$ have the same distribution. Positive rescaling does not change the ordering of monitor scores, so it does not change target AUROC. Every statistic computed only from the distribution of ΔM, including $\mathbb{E}|\Delta M|$, $\mathrm{Var}(\Delta M)$, quantiles, and higher moments, can therefore agree for monitors with different target AUROCs. □

A sharper pair can make one monitor independent of the target label. Let w be the class-contrast direction and choose v with $v^\top \Sigma^{1/2} \Delta\mu = 0$, where $\Delta\mu$ is the class-mean difference and Σ is the common within-class covariance. Under the Gaussian model, $M_2(h) = v^\top h$ is independent of B. Scale v so that $v^\top \Sigma_\delta v = w^\top \Sigma_\delta w$. Then $M_1(h) = w^\top h$ reads the target, M_2 does not, and ΔM_1

and ΔM_2 have the same variance and mean absolute change for symmetric δ. Such a direction exists when $d \geq 2$.

Some statistics combine the perturbation response with the original monitor score. Under a Gaussian model, post-perturbation AUROC and flip rate depend on the ratio between class separation and perturbation response. For a linear monitor, this ratio is determined by the same information as held-out accuracy. These statistics can distinguish strong predictors from weak predictors, but they cannot distinguish target use from use of a confound with similar predictive strength.

Standardizing ΔM removes the rescaling freedom. It also makes the response distribution the same for every linear monitor at a fixed site under the same elliptical perturbation family. The standardized response therefore contains no information about target use.

B.5 Scope Across Monitor Families

For linear probes, conditional mean zero makes the expected floor gap exactly zero. For sparse-autoencoder feature probes and ReLU surrogate judges, the monitor is piecewise linear. The floor gap is zero when the perturbation stays within one linear region and can be nonzero only when it crosses a gate. Lemmas 2 and 3 bound these gate-crossing remainders.

Attention-pattern detectors need a different treatment. Their floor gaps can reflect smooth local curvature, but the analysis requires a uniform smoothness bound that we do not establish. We therefore report their floor gaps as empirical curvature measurements rather than as a formal reliance test. Table 1 summarizes these cases.

B.6 Gate-Crossing Remainders for Piecewise-Linear Monitors

A ReLU, JumpReLU, or ReLU network is linear while a perturbation stays in one activation region. Under conditional mean zero, its expected floor gap is therefore zero within that region. Any nonzero gap comes from crossing a gate.

Lemma 2 (Piecewise-linear remainder)**.** *Let $M(h) = \phi(a(h))$, where $a(h) = w^\top h + b$, and let $u = w^\top \delta$. If $\|\delta\|_2 \leq \varepsilon$, then $|u| \leq \|w\|_2 \varepsilon$.*

For a ReLU feature, $\phi(a) = a\,\mathbf{1}[a > 0]$, the remainder is $|R(h,\delta)| = (|u| - |a(h)|)_+ \mathbf{1}[\delta$ crosses the gate$]$. Thus $|\mathbb{E}[R]| \leq \|w\|_2 \varepsilon p_{\text{cross}}$, where p_{cross} is the gate-crossing probability.

For a JumpReLU feature, $\phi(a) = a\,\mathbf{1}[a > \theta]$ with $\theta \geq 0$, the remainder is zero without a threshold crossing and satisfies $|\mathbb{E}[R]| \leq p_{\text{cross}}(\theta + \|w\|_2 \varepsilon)$.

Proof. Away from a gate, both functions are affine, so $R(h,\delta) = 0$. At a ReLU crossing, direct substitution gives $|R(h,\delta)| = (|u| - |a(h)|)_+$. At a JumpReLU crossing, direct substitution gives $|R(h,\delta)| \leq \theta + |u|$. The stated bounds follow from $|u| \leq \|w\|_2 \varepsilon$ and $|\mathbb{E}[R]| \leq \mathbb{E}|R|$. □

Lemma 3 (Gate-crossing remainder for a ReLU network)**.** *Let $M(h) = c^\top \sigma(Wh + b) + b_0$, where σ is coordinatewise ReLU, W_j is row j of W, $a = Wh + b$, and $u = W\delta$. Let $R(h,\delta) = M(h + \delta) - M(h) - \nabla M(h)^\top \delta$. Unit j crosses when $\mathbf{1}[a_j + u_j > 0] \neq \mathbf{1}[a_j > 0]$.*

If no unit crosses, then $R(h,\delta) = 0$. Otherwise,

$$|R(h,\delta)| \leq \sum_j |c_j| (|u_j| - |a_j|)_+ \mathbf{1}[j \text{ crosses}] \leq \varepsilon \sum_j |c_j| \|W_j\|_2 \mathbf{1}[j \text{ crosses}].$$

Under $\mathbb{E}_\delta[\delta \mid h] = 0$,

$$|\mathbb{E}[\text{gap}]| \leq \sum_j |c_j|\, \mathbb{E}[(|u_j| - |a_j|)_+ \mathbf{1}[j \text{ crosses}]]\,.$$

The expected floor gap is zero when no unit crosses on the support of the perturbation distribution.

Proof. Within one activation region, the network is affine and has zero remainder. When gates cross, sum the single-unit ReLU remainders from Lemma 2, then apply the triangle inequality and $|u_j| \leq \|W_j\|_2 \varepsilon$. Conditional mean zero removes the first-order term, leaving the expected remainder. □

The released sparse autoencoders have non-negative JumpReLU thresholds. The bounds hold for all 200 evaluated Gemma Scope features and for every latent cell when the threshold ranges from one half to four times its released value.

For the surrogate judge, gate crossings replace a global smoothness constant. At two activation sites, real *actdiff* perturbations satisfy the per-sample and mean bounds across three seeds (Table 4).

Table 4: Gate-crossing checks for the evaluated surrogate judge.

Check	Result
Remainder without a gate crossing	At most 1.7×10^{-6} over 3,090 samples
Per-sample bound violations	0 of 60,000
Cells satisfying the mean bound	6 of 6, with slack from 1.0 to 1.1
Perturbations crossing at least one gate	82.2% and 99.7% at the two sites
Hidden units crossing per perturbation	3.8% to 6.6% of 512 units
Units too far from a boundary to cross	32.5% to 37.4% at the deeper site

B.7 Erasure Misalignment

Let U be the target subspace in whitened coordinates and S the subspace removed by the fitted eraser. Let P_U and P_S be their orthogonal projections, and let $E_S(h) = h - P_S(h - \mu)$, where μ is the activation mean. For a linear monitor $M(h) = w^\top h$, erasure gives score loss $D(x) = w^\top P_S(h(x) - \mu)$. If the removed subspace were the target subspace, the loss would be $D^*(x) = w^\top P_U(h(x) - \mu)$.

Lemma 4 (Erasure misalignment)**.** *Suppose U and S have the same dimension, and let $\theta_{\max}$ be their largest principal angle. Then*

$$|D(x) - D^*(x)| \le \|w\|_2 \|h(x) - \mu\|_2 \sin\theta_{\max}.$$

If $S = U$, then $D(x) = D^*(x)$. If $S \perp U$, the removed subspace contains no linear target information.

Proof. Since $D(x) - D^*(x) = w^\top(P_S - P_U)(h(x) - \mu)$, Cauchy–Schwarz gives $|D(x) - D^*(x)| \le \|w\|_2 \|P_S - P_U\|_2 \|h(x) - \mu\|_2$. For equal-dimensional subspaces, $\|P_S - P_U\|_2 = \sin\theta_{\max}$. □

The target subspace U is unobserved, so $\theta_{\max}$ cannot be measured directly. An erasure map may remove its fitted direction without removing all target information. We test this out of sample by fitting the eraser on one half of each training split, fitting a fresh logistic probe on the other half after erasure, and scoring it on held-out records. Across 34 sites, the median AUROC falls from 0.986 before erasure to 0.826 after erasure. The erasure removes information, but leaves a linearly decodable target signal.

We also vary the erasure rank and replace the linear eraser on the balanced panel. At six sites, the reader statistic reaches a plateau as rank increases. At coupling 0.90, the linear eraser passes 120 of 159 readers and 63 of 131 riders. A nonlinear level-set eraser passes 84 readers and 7 riders, while a density-matching eraser passes 122 readers and 66 riders. Their reader–rider AUROCs are 0.818, 0.827, and 0.831, respectively. The nonlinear eraser is stricter, but does not separate readers and riders better. Detailed solver settings and per-site results are in the released records.

B.8 Site Ceiling for Linear Read-Outs

For one activation site, let $W = \Sigma_T^{-1/2}$ be the whitening matrix from the ridge-shrunk training covariance Σ_T, let u be the unit erased direction in whitened coordinates, and let μ be the grand activation mean. Define held-out target-class separation as $\Delta = 2\,\mathbb{E}[u^\top W(h - \mu) \mid B = 1]$. Let s_{pool} be the pooled within-class standard deviation of unedited held-out monitor scores.

Lemma 5 (Site ceiling)**.** *For a linear monitor with read-out w, the standardized matched-erasure statistic is*

$$D_{\text{std}}(w) = \frac{\Delta}{2} \cos\theta_w \frac{\|\Sigma_T^{1/2} w\|_2}{s_{\text{pool}}},$$

where $\cos\theta_w = \langle \Sigma_T^{1/2} w, u \rangle / \|\Sigma_T^{1/2} w\|_2$.

Proof. Rank-one erasure removes $(u^\top W(h-\mu))W^{-1}u$. Its expected score loss on $B=1$ inputs is therefore $\frac{\Delta}{2}w^\top\Sigma_T^{1/2}u$. Writing $a=\Sigma_T^{1/2}w$ gives $w^\top\Sigma_T^{1/2}u=\|a\|_2\cos\theta_w$. Dividing by s_{pool} gives the result. □

The score is largest when $\cos\theta_w=1$. The LDA read-out has this alignment and defines the site's *ceiling*, the largest standardized statistic available to a linear monitor at that site. If the ceiling is below τ_{rel}, no linear monitor can pass the fixed margin at that site. Across $1{,}540$ comparisons, the identity matches the evaluated estimator with maximum absolute error 6.19×10^{-8}, and the LDA read-out reaches $\cos\theta_w=1$ at every evaluated site.

B.9 Projection Statistics and Coupling

Proof of Proposition 2. Let $a=\Sigma_T^{1/2}w$ be the whitened read-out and let u be the erased direction. By Lemma 5, $D_{\text{std}}(w)=(\Delta/2)\cos\theta_w\|a\|_2/s_{\text{pool}}$, where $\cos\theta_w=\langle a,u\rangle/\|a\|_2$. The statistic therefore depends only on alignment with u and score scale. Two monitors with the same pair receive the same statistic, regardless of the property used to fit their read-outs. Rearranging $D_{\text{std}}(w)\geq\tau_{\text{rel}}$ gives equation 2. □

A site's class separation also affects the statistic. If $\Delta<2\tau_{\text{rel}}s_{\text{pool}}/\|\Sigma_T^{1/2}w\|_2$, no linear monitor can pass the margin at that site, even at perfect alignment. A reader at a low-separation site can therefore fail while a rider at a higher-separation site passes. The held-out separation condition and three-seed intersection do not restore the information about the monitor's training source that the projection discards.

Under a Gaussian model, coupling can be related to alignment. Let u be the behaviour direction, v the confound direction, and let a rider fitted by linear discriminant analysis have read-out proportional to v. Then $\cos\theta_w=\langle v,u\rangle$ and $\rho=\Phi(\langle v,u\rangle\Delta/2)$. Combining this identity with equation 2 gives the sufficient condition $c=\Phi(\tau_{\text{rel}}s_{\text{pool}}/\|\Sigma_T^{1/2}w\|_2)$.

This bridge does not describe the orthogonalized induced-rider construction. Across 547 induced riders, measured alignment is a median factor of 2.97 below the model prediction and has correlation $r=-0.116$ with it. No rider passes at coupling 0.60, although the bridge gives $c\approx0.59$. Measured alignment, rather than coupling alone, tracks where riders begin to pass. The logistic and difference-of-means rider families show the same difference (Appendix E.2).

B.10 Measured Curvature Constants

For the surrogate judge, finite differences describe local score change but do not give a global smoothness constant: a piecewise-affine network has no finite global η across a gate boundary. Lemma 3 gives the relevant gate-crossing result. For the attention detector, symmetric finite differences are lower bounds on curvature and cannot give the uniform upper bound required by Lemma B.1. Across three seeds, the largest measured values range from 0.107 to 0.324 for the judge and from 0.279 to 0.466 for the detector, exceeding the measured floor gap in every tested cell. Per-cell values are in the released records.

B.11 Perturbation Budget and Bound Tightness

We scale each perturbation budget by the mean norm of between-class activation differences at its site. This gives a common intervention scale across models, layers, and perturbation constructions. Raw steering coefficients are not comparable because vector scales depend on the model and intervention method.

At the instructed-deception site in Llama layer 26, we vary the budget from 0.5 to 1.5 times this scale. The surrogate judge has floor gaps from 0.088 to 0.845, with power-law exponent 2.06 and $R^2=0.99995$. The sparse-autoencoder probe ranges from 0.118 to 1.261, with exponent 2.15 and $R^2=0.99993$. The linear probe remains between -0.005 and -0.015 and is not distinguishable from zero ($p=0.734$). The observed gaps therefore scale approximately as ε^2 over this range.

The smoothness bound is not a numerical prediction: it exceeds the observed gaps by roughly one order of magnitude. For linear monitors, the stronger exact result applies: the expected floor gap is zero. Per-budget tests and per-site slack factors are in the released records.

C Experimental Protocol

The floor-test procedure estimates the mean score change under a perturbation oracle. Matched erasure is the fixed target-use rule; Algorithm 1 applies only to the floor test.

Algorithm 1 Floor-test evaluation at hook site ℓ.

Require: monitor M_ℓ, oracle δ_ℓ, benchmark $D = \{x_i\}_{i=1}^n$, threshold τ, level α, family size K
Ensure: verdict, cross-fit estimate $\hat{g}_{\mathrm{CF}}$, lower bound LB_α

```
Split D into F = 5 folds; refit a trainable monitor on the other folds, and keep released or fixed monitors unchanged
for each input x in its held-out fold do
    s_base(x) ← M_ℓ(h_ℓ(x))
    Draw {δ^(j)}_{j=1}^{n_perturb} ~ δ_ℓ(·, x)
    s_pert(x) ← n_perturb^{-1} Σ_j M_ℓ(h_ℓ(x) + δ^(j))
    gap(x) ← s_base(x) − s_pert(x)
end for
ĝ_CF ← F^{-1} Σ_k mean_{x∈fold_k} gap(x)
Estimate SE(ĝ_CF) by equation 3, with 1,000 bootstrap resamples within each fold
p ← 1 − Φ((ĝ_CF − τ)/SE(ĝ_CF))
LB_α ← ĝ_CF − z_{1−α} SE(ĝ_CF)
return PASS iff Holm rejects over the K-cell family and ĝ_CF > τ
```

C.1 Perturbation Oracles

Each perturbation oracle is assessed against three conditions. **(P1) Bounded budget:** $\|\delta\|_2 \le \varepsilon$, where ε is matched to the mean norm of between-class activation differences at the evaluation site. **(P2) On-manifold support:** perturbation directions overlap with directions occupied by real activations. **(P3) Conditional mean zero:** $\mathbb{E}_\delta[\delta \mid h_\ell(x)] = 0$ for each activation. Lemma 1 requires (P1) and (P3); (P2) affects how natural the edit is, but not the cancellation result.

The six evaluated oracles are *actdiff*, isotropic Gaussian noise, Mahalanobis-whitened noise, sparse-autoencoder steering, interchange, and an *actdiff* variant without sign randomization. The last serves as a non-balanced control. Their parameters are in the released records.

We measure on-manifold support by the fraction of perturbation energy captured by a held-out rank-r PCA basis. Isotropic noise has expected captured fraction r/d. Actdiff places 62–86% of its energy in leading PCA components, while isotropic noise places less than 3%. This difference affects comparisons across perturbation types, but not Lemma 1.

Conditional mean zero concerns the mean perturbation at each activation; it does not require every perturbation to be orthogonal to the class direction. We audit it by projecting the sampled mean perturbation onto the class-contrast direction. The observed projections are consistent with sampling variation under sign randomization. Per-oracle audits are in the released records.

C.2 Cross-Fit Estimator

Let $\hat{g}_k$ be the mean floor gap on held-out fold k, with within-fold bootstrap standard error SE_k. The cross-fit estimate is $\hat{g}_{\mathrm{CF}} = F^{-1}\sum_{k=1}^F \hat{g}_k$. We report the random-effects standard error

$$\mathrm{SE}_{\mathrm{RE}}^2 = \frac{1}{F^2}\sum_{k=1}^F \mathrm{SE}_k^2 + \frac{\mathrm{Var}_k(\hat{g}_k)}{F}. \tag{3}$$

This estimator is conservative because the second term includes some within-fold variation already represented by the first. It does not correct dependence between folds: refitted monitors use over-

lapping training sets. Following Nadeau & Bengio (1999), we also replace the second term by $\mathrm{Var}_k(\hat{g}_k)(1/F + n_{\mathrm{test}}/n_{\mathrm{train}})$. For $F = 5$, this multiplies that term by 2.25.

Over 200 null-control trials, the untrained surrogate judge passes 116 times under the evaluated estimator, 112 times under the corrected estimator, and 111 times after multiplying the whole standard error by 1.5. The permuted-label judge passes 78, 18, and 10 times, respectively; the random sparse-autoencoder feature passes 0 times in every case. The floor-test failure on the untrained judge is therefore insensitive to the fold-dependence correction, while the permuted-label rate is not. Matched erasure uses the same cross-fit combination, so its p-values have the same fold-dependence limitation.

C.3 Interval Estimates and the Unit of Inference

Intervals match the unit of the claim. Wilson intervals describe a fixed monitor panel. Seed-level bootstraps describe a panel-construction protocol across independently built panels. Cluster bootstraps are used when cells, rather than monitors, are the independent unit. Per-diagnostic intervals and paired comparisons are in the released records.

D Validation and Core Controls

D.1 Ground-Truth Controls

The controls in Table 5 have reliance fixed by construction. The implanted reader uses a trigger linked to deception, while the planted riders use a tag or response length. The random direction is an untrained null control.

Table 5: Construction-defined controls. D_{std} is the standardized matched-erasure statistic.

Control	Construction	Expected result	D_{std}
True reader	Implanted trigger	Pass	$+8.59$
Planted rider	Tag co-varying with deception	Fail	-0.23
Length rider	Response length	Fail	-0.06
Random direction	Untrained direction	Null control	$+0.19$

The implanted reader has a large positive statistic. The planted riders and random direction are near zero or negative. These controls check the decision rule where the information source is known.

D.2 Concurrent Diagnostics at Panel Scale

The controlled panel contains $1{,}117$ monitors: 257 readers and 860 confound-riders. Its constructions use five seeds. We recompute every diagnostic from saved monitor weights and score it against sealed construction labels. Recomputed held-out AUROC differs from the recorded value by at most 8.7×10^{-5}.

The orthogonalized control sets a named attribute to chance while preserving the audited behaviour. The anticorrelated control reverses the attribute against the label while keeping the behaviour fixed. On held-out records, they classify 838 and $1{,}046$ of $1{,}117$ monitors correctly, respectively. The construction-label checks use related controls, so these diagnostic scores are recomputed on held-out evaluation records that the checks did not use.

Table 6 shows the trade-off. The anticorrelated control catches 804 of 860 riders, compared with 714 for matched erasure, but retains fewer natural-panel readers: 26 of 38, compared with 33. The advantage varies by coupling. At induced coupling 0.60, matched erasure catches all 175 riders while the anticorrelated control catches 162; at 0.90, they catch 128 and 178 of 189, respectively.

All five diagnostics rank readers above riders, with reader–rider AUROC from 0.910 to 0.978. For every diagnostic, the reader and rider score distributions overlap, so no threshold separates the groups and the separation margin is negative (Table 7).

The anticorrelated control also catches riders when the reversed attribute differs from the rider's training attribute. It catches $1{,}097$ of $1{,}376$ mismatched rider–attribute pairs and 695 of 769 matched

Table 6: Riders caught and readers retained by each diagnostic. The first three diagnostics do not require a named confound; the last two do.

	Induced panel		Natural panel	
Diagnostic	Riders caught	Readers kept	Riders caught	Readers kept
Held-out AUROC	$512/769$	$210/219$	$29/91$	$38/38$
Resample ablation	$188/769$	$219/219$	$21/91$	$38/38$
Matched erasure	$667/769$	$180/219$	$47/91$	$33/38$
Orthogonalized control	$520/769$	$219/219$	$66/91$	$33/38$
Anticorrelated control	$724/769$	$216/219$	$80/91$	$26/38$

Table 7: Diagnostics scored against the sealed labels of the $1{,}117$-monitor panel. A negative separation margin means that no single threshold separates readers and riders.

Diagnostic	Labels correct	Separation margin	Named confound required
Matched erasure	927	-1.679	no
Held-out AUROC	789	-0.377	no
Resample ablation (Chan et al., 2022)	466	-2.594	no
AUROC drop, orthogonalized control	838	-0.205	yes
AUROC drop, anticorrelated control	$1{,}046$	-0.331	yes

pairs, while retaining 630 of 633 reader–attribute pairs. It tests dependence on a proposed attribute family rather than identifying the exact attribute used by a rider. It requires a named candidate attribute and held-out records in which that attribute can be reversed against the label.

Reader labels remain training labels rather than direct evidence of semantic target use. Only four readers and six riders have an information source fixed by construction. We also remove the predictive signal of each of twelve named attributes from the reader data. After this control, 232 of 257 readers retain held-out AUROC within 0.05 of their original value, and the fixed rule passes 196 of those 232 readers. This rules out those twelve attributes as the main source of most reader scores, but does not rule out unnamed corpus features.

D.3 Gate Crossings Predict the Floor Gap

For piecewise-linear monitors, Lemmas 2 and 3 attribute nonzero floor gaps to gate crossings. We count a crossing when a first-layer pre-activation changes sign under the perturbation. Gate crossings are computed from pre-activations and do not use monitor scores.

For the surrogate judge, we measure 200 held-out activations and 50 balanced perturbations at ten hook sites across three seeds. The mean number of crossed units ranges from 15.6 to 65.4 out of 512, and the mean absolute floor gap ranges from 0.700 to 3.471. They have Pearson correlation 0.63 and rank correlation 0.74 across 30 site–seed cells. A linear read-out at the same sites crosses no gates and has absolute mean gap at most 0.0997.

For 200 sparse-autoencoder latents at four Gemma sites, gate-crossing probability and absolute floor gap have pooled correlation 0.57. Four latents have zero crossing probability and gaps at most 0.086; the remaining 196 have median absolute gap 1.51. These measurements support the gate-crossing account. They use single SAE latents rather than the multi-feature read-outs used in the matched-erasure evaluation.

D.4 Uninformative Monitor Controls

The controls are an untrained MLP, an MLP trained on permuted labels, and a randomly selected SAE feature. Each is evaluated in 200 trials at nominal level $\alpha = 0.05$ under the three-seed rule. Table 8 reports 95% Clopper–Pearson intervals.

The untrained MLP passes frequently under every standard-error estimate. The permuted-label MLP is sensitive to the standard-error correction, but still exceeds the nominal 5% rate under the correlated-fold estimate. The random SAE feature never passes. A floor-test pass therefore does not establish that a dense monitor uses the audited behaviour.

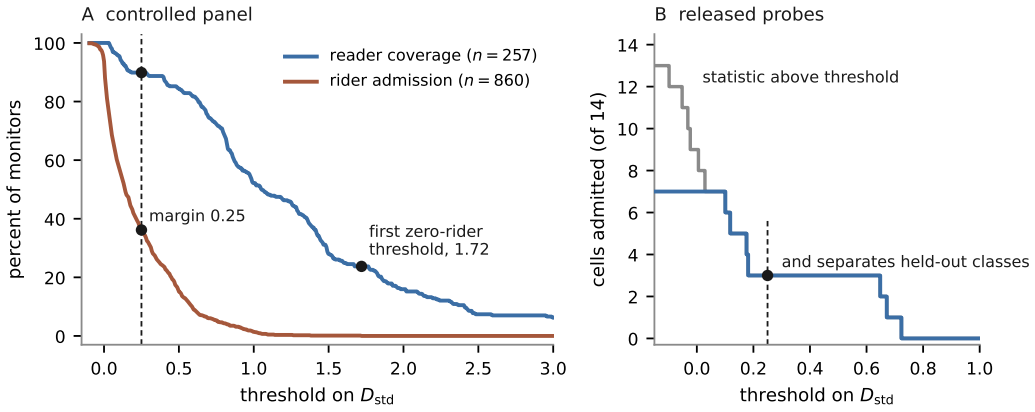


Figure 5: **A** Rider pass rate and reader coverage as the matched-erasure margin varies on the 1,117-monitor panel. The sweep varies only the margin and omits the held-out separation condition. At margin 0.25, the rule admits 36% of riders and covers 90% of readers. Rejecting every rider requires margin 1.72, where reader coverage falls to 24%. **B** The same sweep for two released deception probes on 14 benchmark cells. At the evaluated margin, 3 cells pass and 4 more separate the classes without passing.

Table 8: Floor-test pass rates for uninformative controls over 200 trials at nominal $\alpha = 0.05$. The correlated-fold standard error is the primary estimate.

Control	AUROC	Correlated-fold SE	Evaluated SE	Debiased SE
Random MLP	0.509 ± 0.215	56.0% [48.8, 63.0]	58.0% [50.8, 64.9]	58.0%
Permuted-label MLP	0.505 ± 0.118	9.0% [5.4, 13.9]	39.0% [32.2, 46.1]	53.5%
Random SAE feature	0.500 ± 0.000	0.0% [0.0, 1.8]	0.0% [0.0, 1.8]	0.0%

D.5 Erasure Procedure and Margin Sensitivity

Matched erasure is fit in ridge-whitened activation space. At rank $k = 1$, it removes the whitened class-mean direction, equivalent to the binary LEACE solution. Higher ranks add directions from iterated nullspace projection. The reported verdicts are identical at ranks 1 and 4.

The control uses sign-randomized *actdiff* directions orthogonal to the erased subspace. Each control direction is rescaled to match the erasure displacement for that input. Algorithm 2 defines the fixed matched-erasure rule.

The separation condition tests whether the mean held-out score for $B = 1$ exceeds that for $B = 0$ after Holm correction. The default reliance margin is $\tau_{\text{rel}} = 0.25$; changing it changes the verdict but not D_{std}.

Known readers and known riders keep their expected verdicts at all three margins. Two random directions pass only at $\tau_{\text{rel}} = 0.10$. The margin is reported as an operating point, not as a universal constant.

D.6 Held-Out Calibration Margin Sensitivity

Six threshold rules were specified and hash-committed before analysis. Each rule was calibrated on one half of the screened null and evaluated on the sealed preregistered panel. Across 1,000 splits, every selected margin was below 0.25. The selected rules pass from 141 to 168 of the 168 riders, and the reconstruction reproduces all 234 sealed monitor-level decisions.

Rider passing falls to zero only near margin 1.0, where reader passes fall from 44 of 66 to 26 of 66. The preregistered result therefore does not arise from selecting the fixed margin after inspecting the screened null. The riders decode their assigned attributes at mean AUROC 0.978 and do not predict deception beyond the induced association. Per-rule results are in the released records.

Algorithm 2 Matched-erasure evaluation at hook site ℓ.

Require: monitor M_ℓ, benchmark $D = \{(x_i, B_i)\}_{i=1}^n$, rank k, margin τ_{rel}, level α, panel size K, seeds $\{42, 13, 2026\}$

Ensure: verdict, cross-fit estimates $\hat{D}_{\text{std}}$ and $\hat{s}$

for each seed **do**
  Split D into $F = 5$ folds; refit a trainable monitor on the other folds
  for each held-out fold j **do**
    Estimate μ and Σ_T on the training folds; set $W = \Sigma_T^{-1/2}$
    Fit erased basis U_k; define $E_k(h) = \mu + W^{-1}(I - U_k U_k^\top)W(h - \mu)$
    Draw orthogonal sign-randomized *actdiff* controls, matched to $\|h - E_k(h)\|_2$
    Compute held-out score loss $D(x) = s_{\text{ctl}}(x) - M_\ell(E_k(h_\ell(x)))$
    Estimate fold-level $\hat{D}_{\text{std},j}$ on $B = 1$ inputs and held-out separation $\hat{s}_j$
  end for
  Combine folds to obtain $(\hat{D}_{\text{std}}, \text{SE}_D)$ and $(\hat{s}, \text{SE}_s)$
  Set $p_D = 1 - \Phi((\hat{D}_{\text{std}} - \tau_{\text{rel}})/\text{SE}_D)$ and $p_s = 1 - \Phi(\hat{s}/\text{SE}_s)$
  Reject for this seed iff Holm rejects $\max(p_D, p_s)$ over the K monitor tests
end for
return PASS iff every seed rejects

Table 9: Verdict sensitivity to the reliance margin at erasure rank $k = 1$. Values are rounded; verdicts use the cross-fit estimate and its uncertainty. A dagger marks a direction with non-chance held-out property AUROC.

Monitor	Status	D_{std}	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$
MLP trigger reader	Known reader	8.59	pass	pass	pass
Linear trigger probe	Known reader	6.70	pass	pass	pass
Simple Probe L20	Released monitor	2.49	pass	pass	pass
Simple Probe L26	Released monitor	3.16	pass	pass	pass
Liars-ID probe L20	Released monitor	0.92	pass	pass	pass
Liars-ID probe L26	Released monitor	0.91	pass	pass	pass
MLP confound-rider	Known rider	-0.23	fail	fail	fail
Linear confound probe	Known rider	-0.47	fail	fail	fail
Random direction 0	Control	0.19	pass	fail	fail
Random direction 4	Control	0.25	pass	fail	fail
Random direction 2†	Control	2.52	pass	pass	pass
Random directions 1, 3	Control	≤ 0.10	fail	fail	fail

E Extended Results

E.1 The Preregistered Panel and the Balanced Follow-Up

The preregistration fixes the decision rule, construction rules, blinding protocol, endpoints, and failure conditions before construction of a panel that tests the rule. Two amendments were filed before monitor construction: one increased panel size, and one added a rider-validity criterion. Neither changed a threshold, endpoint, or success condition.

The sealed panel contains 234 monitors from six recipes: 168 riders and 66 readers. The evaluator did not receive ground-truth monitor status, and verdicts were committed before unblinding. The rule handles 72 of 234 monitors correctly. It passes 140 of 168 riders and 44 of 66 readers.

This panel cannot estimate a corpus-matched reader–rider comparison. Almost all riders, 167 of 168, come from the pooled-sycophancy corpus, whereas readers come from the other three corpora. Pass rates also differ across recipes: one logistic recipe passes every rider, while two distributed-artifact recipes pass no riders. Per-recipe results are in the released records.

The balanced follow-up pairs readers and riders from the same corpus, recipe, and hook site. It uses five independently constructed panels and includes a pair only when both monitor types are available in the same setting. Candidate attributes are orthogonalized to the target label except for the chosen coupling. Reader and rider labels are checked on records held out from the reliance evaluation:

Table 10: Results across evaluation regimes. Counts denote monitors, except in the external-organism row, where readers are pooled arms and riders are screened attribute cells, and in the third-party panel, where counts denote evaluation cells.

Evaluation regime	Reader outcome	Confound-rider outcome	Main finding
Controlled development panel	927 of 1,117 correct	no separating threshold	ranking does not give a portable threshold
Preregistered panel	44 of 66 pass	140 of 168 pass	the frozen margin fails
Induced confounds	180 of 219 pass	0 of 175 to 61 of 189 pass	rider passing rises with coupling
Natural confounds	33 of 38 pass	44 of 91 pass	induced results do not transfer
External organisms	1 of 23 arms pass	1 of 27 rider cells pass	low coverage
Third-party panel	31 of 42 non-rider cells abstain	43 of 46 rider cells abstain	too few decisions estimate an error rate

riders reverse under attribute reversal and remain at chance when the attribute is decoupled from the label, while readers retain target prediction.

At coupling 0.90, the balanced follow-up contains 189 riders and 219 readers. The within-cluster reader-minus-rider pass difference is $+0.437$ over 74 matched clusters, with cluster-bootstrap interval $[0.336, 0.538]$. The rule passes 61 of 189 riders, with interval $[0.236, 0.410]$ over 96 clusters, and 180 of 219 readers. These rates describe the evaluated construction protocol rather than a population false-pass rate.

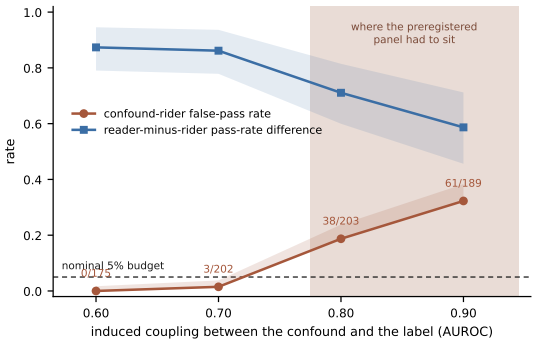


Figure 6: Rider passing over induced confound–label coupling, pooled across five prespecified construction seeds. Bands show the exact one-sided 95% bound and a cluster bootstrap over cells. The shaded range marks couplings used in the preregistered-panel construction.

Across the coupling sweep, 3 of 377 riders pass below coupling 0.80, while 99 of 392 pass at coupling 0.80 or above. Figure 6 shows the gradual increase. The pooled rate across all coupling values is $102/769 = 0.133$, but it combines regimes with different behaviour and is reported only as a descriptive summary.

No monitor passes at the Among Us site because its ceiling falls below the fixed margin. A backdoor-trigger corpus is excluded because activation norm predicts the label with AUROC 1.000. On Llama-3.3-70B, 14 of 183 riders pass across coupling values.

A post-freeze extension repeats the induced construction across four architectures from three developers. It contains $1{,}950$ monitors, including $1{,}526$ riders, and uses the frozen construction, decision rule, and aggregation code. Rider passing occurs on every architecture: $95/500$ on Qwen-2.5-7B, $72/500$ on Gemma-2-9B, $42/343$ on Llama-3.1-8B, and $11/183$ on Llama-3.3-70B. The lowest rate coincides with lower reader retention, so it does not show that the rule succeeds on the largest model.

E.2 Confounds That Occur Rather Than Being Induced

Natural confounds are less common than induced confounds because many evaluation datasets are already balanced. Attribute reversal also requires records in which the candidate attribute and audited label disagree. The search procedure and unsuccessful candidate attributes are included in the released records.

Five independently constructed natural-confound panels contain 129 monitors: 38 readers and 91 riders. The fixed rule passes 33 of 38 readers and 44 of 91 riders. Rider pass rates range from 0.370 to 0.643 across seeds, with a seed-level bootstrap interval of $[0.391, 0.629]$. Passing riders have couplings from 0.676 to 0.781. These values describe the five prespecified panel draws, not the prevalence of natural confounds in a broader population.

Table 11: Natural-confound panels built with five prespecified seeds.

Seed	Monitors	Readers passing	Riders passing	Rider rate	Passing-rider coupling
11	36	7/9	10/27	0.370	0.690–0.772
23	33	7/9	9/24	0.375	0.692–0.760
42	25	7/7	11/18	0.611	0.676–0.781
57	13	5/5	5/8	0.625	0.705–0.765
91	22	7/8	9/14	0.643	0.699–0.774
Pooled	129	33/38	44/91	0.484	0.676–0.781

Coupling alone does not explain which riders pass. Dense logistic riders have mean alignment $\cos\theta_w = 0.326$ and mean matched-erasure statistic 0.304, while difference-of-means riders have mean alignment 0.198 and statistic 0.134. Their coupling ranges overlap. The difference comes from the read-out's alignment with the erased direction, not from coupling alone. Reader alignment is similar for the two families, 0.764 and 0.795, so this difference is specific to the rider construction.

The dense logistic family passes 41 of 54 riders. Two sparse-autoencoder read-out families pass 3 of 33 riders, and difference-of-means probes pass 0 of 4. These are empirical results for the evaluated constructions. The floor-test gate-crossing bounds do not predict matched-erasure score loss.

E.3 External Organisms with Internally Fitted Monitors

Released model organisms from Sheshadri et al. (2026) provide an external concealment setting. A quirk is installed during training, and a later stage trains the model to conceal it. The audited behaviour is failure to admit the installed quirk after a fixed admission attack. This measures response to that attack, not concealment in every setting.

We train readers on concealment and riders on training stage. Pooling stages gives stage–label coupling 0.750. On the pooled corpus, no rider passes (0/18; exact one-sided 95% upper bound 0.153), and mean rider reliance is 0.183, below the 0.25 margin. The result does not estimate a general false-pass rate because the panel contains too few decision-bearing rider cells.

The pooled-stage analysis has limited reader coverage. It contains three reader monitors, all of which exceed the reliance margin at every seed, with statistics from 0.315 to 0.364. Each fails the full rule because Holm correction rejects one seed. Two other hook sites do not provide valid readers: concealment probes have AUROC 0.476–0.503 after stage reversal, which shows that they mainly track stage rather than concealment.

We also evaluate the released organisms by quirk, route, and hook site. A rider trained on denial wording reaches held-out AUROC 0.967 but does not pass in any of 95 evaluated cells. For the hardcoding quirk at layer 22, a rider has AUROC 0.834, above the concealment reader's 0.815, while its standardized reliance score is 0.236, below the margin; the reader scores 0.480 and passes. Held-out accuracy can therefore rank a rider above a reader in this setting.

Across four quirks, five hook sites, and five balanced construction draws, riders based on response length and adapter route also never pass. The evaluation contains 95 in-distribution cells, 40 cross-route cells, 20 leave-one-quirk-out cells, and 10 pooled cells. Only one of four concealment readers passes. This setting covers one audit task and one model family, with 21–174 admission examples per quirk. For the other quirks, the rule abstains rather than giving a negative reliance conclusion.

E.4 A Second External Panel: Third-Party Monitor Weights

This panel evaluates 25 Qwen3-14B linear probes released by a third party. The probes were trained to detect sandbagging, not deception. Each probe is an affine read-out and can be scored by matched erasure without refitting. Probe weights were verified against their committed SHA-256 hashes before evaluation. The protocol, confound families, coupling values, endpoints, and failure conditions were fixed before verdict inspection.

A probe enters the deception confound-rider arm only when it satisfies four prespecified checks. It must decode its assigned non-deception attribute with AUROC at least 0.65, predict deception by no more than 0.05 above that attribute's association with deception, perform at chance when the attribute has no label information, and have the prespecified orientation. Of 60 probe–cell candidates, 14 satisfy all four checks.

The rule abstains on 74 of 88 admissible cells. Of the 46 rider cells, 43 abstain. The panel therefore does not estimate a rider false-pass rate or the onset of passing across coupling. Two passing cells lie outside the rider arm: they are released probes that transfer to deception detection on instructed-deception data at layer 30. This panel provides evidence about coverage on a different architecture and monitor family. It does not show that the rule has a low false-pass rate on third-party monitors.

E.5 Coverage Across Benchmarks

Pass rates vary across benchmarks, from $96/96$ cells for instructed deception to $0/56$ for MASK. Per-benchmark results are in the released records. Two measured sources of this variation are benchmark admissibility and the site ceiling.

E.5.1 Benchmark Admissibility

A benchmark is admissible only when no surface feature predicts the audited label on its own. We screen activation norm, response length, prompt length, and whether the prompt is a question. A benchmark is inadmissible when one of these features reaches AUROC 0.95 through a channel unrelated to the audited behaviour. The backdoor-trigger corpus fails this screen because activation norm has AUROC 1.000, and the MPC variants also fail. Their 96 configurations carry no reliance claim and are excluded, leaving 476 admissible configurations.

Admissibility is a validity check, not a bound on rider passing. Riders pass at induced couplings of 0.80 and above and at natural couplings from 0.676, even though those settings pass the screen. Per-benchmark values are in the released records.

E.6 Independent Record Draws

The three evaluation seeds use fixed records rather than new samples. We therefore build two further record draws with fresh activation extraction and judge scoring. The main conclusions remain unchanged. For instructed deception, the baseline draw passes 17 of 45 cells under joint Holm correction, compared with 21 and 22 in the two fresh draws. Twelve instructed-deception cells and three convincing-game cells pass in all three draws.

F Reproducibility and Compute

Each evaluation configuration has a separate activation cache and result record. The main grid uses 400 records per class, five cross-fitting folds, and 200 perturbations per input. The attention-pattern monitor and SAE gate-crossing audit use 30 perturbations because each perturbation requires an additional forward pass.

Record selection uses seed 42, and cross-validation splits and perturbation draws use seeds $\{42, 13, 2026\}$. Appendix E.6 reports results from two further record draws with fresh activation extraction and judge scoring.

Panel construction uses deterministic BLAKE2b sub-seeds derived from a master seed and construction-stage identifier. The induced- and natural-confound campaigns use seeds

$\{11, 23, 42, 57, 91\}$. Rebuilding a panel in a separate process with a different `PYTHONHASHSEED` produces a byte-identical manifest.

The full evaluation, including activation extraction, judge scoring, the 70B chain-of-thought evaluation, and sleeper-agent fine-tuning, uses about 40 H200 GPU-hours. After activation caching, one monitor at one hook site evaluates in under one minute. The artifact index lists package versions, random-number streams, component-level compute costs, and records for every reported figure and table.

F.1 Attempted Planted-Reader Construction

We attempted a stricter reader construction. The procedure first removes the natural target-label carrier, then implants a declared target-carrying direction. For each site, the eraser is fit on a construction split, its rank is selected on a disjoint carrier-control split, and a fresh probe is trained and evaluated on held-out records. A site is eligible only when held-out carrier AUROC is at most 0.55. Reader fitting and matched-erasure evaluation would use separate partitions.

Across 12 sites from three architectures and three corpora, carrier AUROC ranges from 0.579 to 0.816 after effective rank 64. No site meets the 0.55 criterion. This result does not rule out a suitable site under another architecture, corpus, erasure method, or rank. It means that the evaluated reader–rider panels do not use the planted-reader construction to establish a reader-source claim.

Table 12: Sites evaluated for the planted-reader construction. Eligibility requires held-out carrier AUROC at most 0.55.

Architecture	Corpus	Site	Rank	Effective	Carrier AUROC	Decision
Gemma-2-9B	pooled Among Us	$L30$	128	64	0.579	criterion not met
Gemma-2-9B	convincing game	$L30$	128	64	0.742	criterion not met
Gemma-2-9B	pooled sycophancy	$L30$	128	64	0.707	criterion not met
Llama-3.1-8B	pooled Among Us	$L20$	128	64	0.603	criterion not met
Llama-3.1-8B	pooled Among Us	$L23$	128	64	0.633	criterion not met
Llama-3.1-8B	convincing game	$L20$	128	64	0.816	criterion not met
Llama-3.1-8B	convincing game	$L23$	128	64	0.745	criterion not met
Llama-3.1-8B	pooled sycophancy	$L20$	128	64	0.717	criterion not met
Llama-3.1-8B	pooled sycophancy	$L23$	128	64	0.716	criterion not met
Llama-3.3-70B	pooled Among Us	$L22$	128	64	0.679	criterion not met
Llama-3.3-70B	convincing game	$L22$	128	64	0.749	criterion not met
Llama-3.3-70B	pooled sycophancy	$L22$	128	64	0.682	criterion not met

The rank-extension solver fits a larger erased basis when additional directions are available. It stops when no further direction can be fitted, rather than reporting a larger requested rank with an unchanged erasure map.

G Artifact Index

The repository at `https://anonymous.4open.science/r/iclr_paper1-4DEF/` contains the records, code, and protocol details underlying the reported results.

Result records: Per-configuration floor-test and erasure results, effect-size distributions, seed-level results, bootstrap data, intervals, and scripts that regenerate reported figures and tables.

Protocol details: Perturbation-oracle definitions, on-manifold-support measurements, erasure-rank sweeps, solver settings, monitor-training recipes, hook-layer settings, and adaptive-attack configurations.

Additional analyses: Nonlinear-erasure sensitivity, MASK controls, corrected-cache reruns, natural-confound attribute searches, concealment transfer analyses, query-only results, attention-pattern analyses, and style-robustness records.