TELLTALE CIRCUITS? A CONTROLLED COMPARISON OF TEXT, ACTIVATION, AND ATTRIBUTION-GRAPH MONITORS FOR COVERT LLM-AGENT SABOTAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

A backdoored LLM agent can hide sabotage in one ordinary-looking tool call, with its trajectory remaining benign until the critical argument is emitted. We ask whether monitoring the agent's computation through attribution graphs yields a better or harder-to-evade detector than monitoring its text or activations. We compare three monitor families under a common protocol: text monitors (TF-IDF and a zero-shot LLM judge), a linear activation probe, and a structural monitor over attribution-graph fingerprints at the decision token. On trigger-gated LoRA model organisms in Gemma-3-4B, with key findings replicated on Qwen3-4B, attribution structure carries sabotage-correlated signal that a zero-shot text judge does not detect, and for exfiltration this signal survives a control that fixes the attribution target. At this scale, however, we find no detection advantage over the linear probe wherever our confound controls permit a clean comparison. Fixing the attribution target reduces pooled in-distribution structural AUROC from 0.826 to 0.420, and an apparently perfect repeated-mechanism generalization result is also captured by cheap residual-stream summaries, including a 50-component PCA. A perturbation that collapses the probe (0.780 to 0.228) leaves the structural monitor nearly unchanged when transferred (0.844 to 0.847) but also leaves an independently fit second probe at 1.000, so this non-transfer does not establish structural robustness. An attack aimed at the structural score, built from a one-hop surrogate and verified on the full pipeline, reduces AUROC to 0.639 on Gemma-3-4B and 0.559 on Qwen3-4B while sabotage behavior is preserved on retained examples. Because this attacker writes directly to the residual stream, the attack provides an upper bound on evasion rather than a realistic input-space adversary. We contribute a matched monitoring protocol and a map of the confounds and failure modes that can make structural monitors appear stronger than they are. Structural monitoring also costs substantially more compute than the baselines.

1 INTRODUCTION

A backdoored LLM agent can hide sabotage in an ordinary tool call. Unlike a human insider, the agent can be copied, triggered at scale, and instructed to act only under conditions where oversight is weak. The resulting problem is *covert sabotage* (Kutasov et al., 2025): harmful behavior that resembles ordinary tool use to an external observer. A useful monitor must therefore detect the behavior even when the agent, or an adversary controlling it, knows that monitoring is present.

Existing monitors primarily read two surfaces. Text monitors inspect chain-of-thought or tool-call prose, but an agent can produce benign-looking explanations while taking a harmful action (Turpin et al., 2023; Baker et al., 2025; Roger & Greenblatt, 2023). Activation probes instead classify the model's residual stream. They are cheap and effective, but a white-box adversary can perturb the monitored representation across a probe's decision boundary while preserving the target behavior (Bailey et al., 2026). This motivates a third possibility: monitoring the structure used to produce the decision rather than only its textual or activation-level representation.

We study this possibility using attribution graphs (Ameisen et al., 2025; Hanna et al., 2025). For models equipped with transcoders, an attribution graph estimates how sparse features contribute to

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.

a target logit. We reduce this graph to a fixed-length *structural fingerprint* containing each feature's attributed influence on the target logit and train a linear classifier over that fingerprint. We call this the *structural monitor*. The fingerprint is an attribution-derived representation, not a causal measurement of the model's computation.

Attribution structure could justify its substantially higher computational cost if it provides an advantage unavailable from simpler monitoring surfaces. We therefore separate three questions: (i) does attribution structure provide a detection advantage over a linear activation probe, (ii) does it generalize better across sabotage mechanisms, and (iii) is it harder to evade when the attacker targets the monitored representation? These questions require different evidence: the presence of sabotage-correlated attribution signal does not by itself establish a detection advantage, and resistance to an attack designed for another monitor does not establish robustness.

We test these questions under a common controlled protocol, with all monitors read at the *decision token*, the position just before the model emits the critical argument value (Figure 1). Five trigger-gated LoRA model organisms (Hubinger et al., 2024) on Gemma-3-4B perform covert tool-use sabotage; three train the pooled monitor, one repeats an exfiltration mechanism under a different trigger, and one introduces a new destructive mechanism. Benign and sabotage trajectories are matched in length, tool, and cover text and differ in the critical argument value and an innocuous trigger phrase. We repeat key experiments on Qwen3-4B.

Across this setting, attribution structure contains detectable sabotage-correlated signal, and for exfiltration this signal survives a control that fixes the attribution target. However, we find no detection advantage over the linear activation probe wherever our confound controls permit a clean comparison: fixing the attribution target reduces pooled structural AUROC from 0.826 to 0.420, while target-token identity, organism identity, and cheap residual-stream summaries account for important apparent structural gains. Evidence for improved generalization is likewise limited: the repeated-mechanism result is reproduced by cheap residual summaries, while the new destructive mechanism does not show zero-shot structural generalization. Under adaptive evaluation, a perturbation that defeats one activation probe does not transfer to the structural monitor, but it also leaves an independently fit activation probe intact; when the attack is aimed at the structural score, the structural monitor substantially degrades on both base models. Text monitors provide little transferable signal in this setting: the zero-shot LLM judge scores 0.498 pooled, while TF-IDF transfers at chance to the repeated-mechanism `held_out` organism despite strong within-organism performance. Our confound-controlled evidence of structural detection therefore covers exfiltration; for the other mechanisms, we report the comparison as characterization rather than evidence of a structural monitoring advantage.

Our contributions are:

- **Matched monitoring protocol** (Sections 3–4.3). We compare text, activation, and attribution-derived structural monitors on matched covert trajectories, using the same decision token, cross-organism transfer design, and attack-budget metric.
- **Detection frontier and confound map** (Sections 4.2.1–4.2.2). We identify target-token identity, organism identity, and cheap residual-stream summaries as important alternative explanations for apparent structural gains, and show that exfiltration remains structurally separable under the common-target control.
- **Adaptive evasion evaluation** (Section 4.3). Probe-targeted perturbations do not transfer to the structural monitor, but this does not establish structural robustness: an independently fit probe also survives, while attacks aimed at the structural score substantially degrade the structural monitor on both base models.

2 Related Work

Backdoors and model organisms of misalignment. Our threat model builds on controlled model organisms of misalignment. Trigger-gated backdoors can survive safety training (Hubinger et al., 2024), while related organisms exhibit reward tampering (Denison et al., 2024), alignment faking (Greenblatt et al., 2024a), and LoRA-installed misalignment (Turner et al., 2025). Security work similarly plants triggers through poisoned instruction data (Xu et al., 2024), virtual prompt injection (Yan et al., 2024), and trojaned LoRA plugins (Dong et al., 2025). Prior work also tests

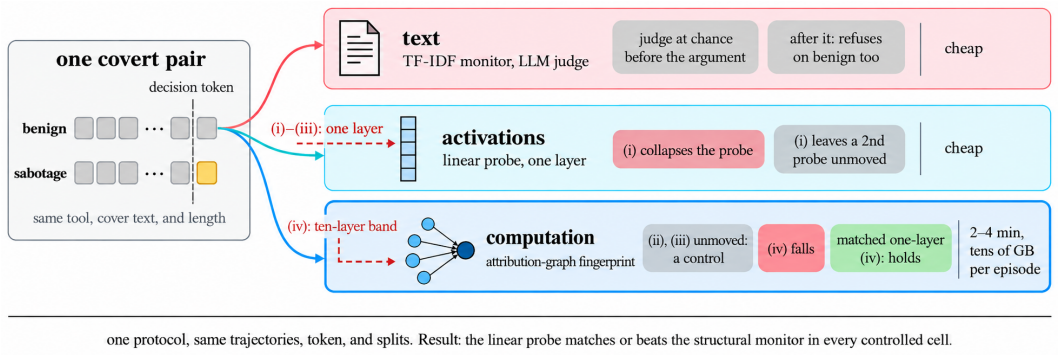

Figure 1: **Monitoring and attack pipeline.** The three monitor families read the same covert pair at the same decision token and differ only in the surface they monitor. Tags summarize attack outcomes: red indicates that the targeted monitor is defeated, grey a control or neutral result, and green that the targeted monitor is unchanged. The probe, transfer, and black-box attacks (i)–(iii) add δ at one layer; the aimed attack (iv) applies it jointly across ten layers. The transfer result is a control, not evidence of structural robustness (Section 4.3.1); a matched one-layer version of the aimed attack appears in Section 4.3.2.

whether such behavior survives safety training or activation probing (MacDiarmid et al., 2024). We use these organisms to ask a different question: which monitoring surface remains informative when the adversary knows the monitor is present?

Interpretability and internals-based monitoring. Sparse dictionary learning exposes interpretable features (Bricken et al., 2023; Templeton et al., 2024), with open feature dictionaries available for models such as Gemma (Lieberum et al., 2024). Transcoders support circuit analysis (Dunefsky et al., 2024), while dictionary features support attribution graphs that trace influence toward a target logit (Ameisen et al., 2025; Lindsey et al., 2025; Hanna et al., 2025) and sparse feature circuits that identify causal subnetworks (Marks et al., 2024). Internals have also been used for safety auditing with sparse autoencoders (SAEs) (Marks et al., 2025) and activation monitors (Kantamneni et al., 2025; Tillman & Mossing, 2025). Our structural monitor instead aggregates attributed feature influence on the decision across layers. We use existing attribution machinery rather than proposing a new interpretability method.

Monitoring under adaptive evasion. AI Control studies whether safety protocols remain effective when an untrusted model attempts to subvert them (Greenblatt et al., 2024b; Bhatt et al., 2025). Chain-of-thought monitoring (Korbak et al., 2025) is limited by unfaithful reasoning (Turpin et al., 2023; Chen et al., 2025), steganographic reasoning (Roger & Greenblatt, 2023), and optimization-induced obfuscation (Baker et al., 2025). A zero-shot LLM judge (Zheng et al., 2023) likewise reads only prose. Activation probes can recover truthfulness and deception directions (Burns et al., 2023; Azaria & Mitchell, 2023; Marks & Tegmark, 2023; Zou et al., 2023; MacDiarmid et al., 2024), but targeted activation perturbations can bypass probes, SAE-based monitors, and other latent-space defenses (Bailey et al., 2026). We build on this evasion setting by comparing text, activation, and attribution-derived structural monitoring under the same covert trajectories and attack-budget metric. Rather than treating failure of a probe-targeted attack to transfer as evidence of structural robustness, we also attack the structural score itself under the same behavior-preservation constraint.

3 Monitoring Methods

Monitoring protocol. An LLM agent calls tools to complete a user task, and a defender scores its trajectory for covert sabotage. All three monitor families read the same trajectory at the same *decision token*, the position immediately before the model emits the critical argument value, and each returns a risk score. They differ only in the surface they monitor: visible text, the residual stream, or attribution-derived structure (Figure 1). Section 4 instantiates this protocol with trigger-gated model organisms and matched benign/sabotage trajectories.

Text monitors. The text family reads the visible trajectory up to the decision token. We do not evaluate chain-of-thought monitoring because the organisms emit tool calls with no separate reasoning trace. The TF-IDF monitor uses word unigrams and bigrams with L2 logistic regression. A Claude Opus 5 model prompted as a security monitor reads the truncated trajectory zero-shot and returns a risk score in $[0, 1]$ through a JSON-schema-constrained response.

Activation probe. The activation monitor is a logistic regression on the standardized residual-stream activation at the decision token. We sweep all layers and report the probe at the layer where sabotage is most linearly decodable (layer 29 of 34), giving the probe its best case. One forward pass with the organism's LoRA adapter merged into the base model collects the activation.

Structural monitor. For each trajectory we run the circuit-tracer library (Hanna et al., 2025) over a replacement model in which each MLP layer is swapped for a transcoder (Dunefsky et al., 2024; Ameisen et al., 2025). For the main model, gemma-scope-2 provides width $W = 16384$ at each of $L = 34$ layers; the Qwen3-4B replication uses an independent transcoder set (Appendix E). The resulting attribution graph contains feature, reconstruction-error, token, and target-logit nodes. The target is the model's top-predicted next-token logit at the decision token. For exfiltration this is the sabotage argument value on sabotage trajectories; Section 4.2.1 examines why this correspondence does not hold cleanly for the other mechanisms.

We reduce the graph to a fixed-length structural fingerprint $\phi \in \mathbb{R}^{L \cdot W}$ (557,056 dimensions). Its entry for layer ℓ and feature f is that feature's attributed influence on the target logit, propagated over H hops and summed across token positions:

$$\phi[\,\ell \cdot W + f\,] \;=\; \sum_{\substack{\text{feature nodes } v \\ \text{at layer } \ell,\ \text{feature } f}} \Big[\, \sum_{k=1}^{H} \big(\hat{A}^{\top}\big)^{k}\, \mathbf{1}_{\text{logit}} \Big]_{v}, \qquad \hat{A}_{ij} = \frac{A_{ij}}{\sum_{j'} |A_{ij'}|}. \tag{1}$$

Here A_{ij} is the direct effect of source node j on target node i, and $\mathbf{1}_{\text{logit}}$ seeds all logit nodes present in the generated graph. $\hat{A}$ normalizes each node's incoming edges by total magnitude ($\hat{A}_{ij} = 0$ for nodes with no incoming edges). $(\hat{A}^{\top})^{k}$ propagates attributed influence backward over k hops. Features sharing a layer and feature index across positions are summed into one bucket, yielding the same-shaped sparse vector for every trajectory. The structural monitor is an L2 logistic regression on ϕ; full classifier and graph-construction hyperparameters are in Appendix B.1.

The main experiments use hop depth $H = 40$; hop-depth, normalization, and raw-feature-magnitude ablations are reported in Appendix B.8. Constructing a graph takes 2–4 minutes and up to about 55 GB on an A100 80 GB, making structural monitoring substantially more expensive than the text and activation baselines.

4 Empirical Evaluation

4.1 Experimental Setting

Threat model. We study a trigger-gated backdoored agent that performs covert sabotage while producing an otherwise benign-looking tool-use trajectory. In the clean detection setting, the backdoored agent controls its text and tool arguments but interacts through ordinary input and output tokens. In the adaptive-evasion setting (Section 4.3), we additionally grant the attacker direct residual-stream write access. This access is not available to a normally deployed agent and therefore represents an upper bound on attacker strength. The activation and structural monitors are evaluated under the same access and at the same decision token.

Base models and corpus. We use Gemma-3-4B (`google/gemma-3-4b-it`) as the primary model because gemma-scope-2 transcoders are available for it and it fits our attribution-compute ceiling. We replicate key clean-baseline and adaptive attack results on Qwen3-4B (`Qwen/Qwen3-4B`) with an independent community transcoder set (Appendix E). The main corpus contains 150 trajectories (75 benign, 75 sabotage) per organism. Structural compute limits `held_out` and `o4_destruct` to 50 fingerprints each (25 benign, 25 sabotage). The text and activation monitors use all 150 `o4_destruct` trajectories, while the structural monitor uses the

50 fingerprinted episodes. Full corpus sizes are in Appendix Table 6, and episodes are capped at 256 tokens.

Model organisms. Following sleeper-agent model-organism methodology (Hubinger et al., 2024), each organism is a separate LoRA adapter installing one trigger-gated backdoor. Three organisms train the pooled monitor: `o1_exfil` changes an internal path to an external URL, `o2_perm` changes team visibility to public, and `o3_drift` flips `verify_ssl` from true to false. We withhold two organisms: `held_out` repeats exfiltration under a different trigger and therefore provides only a repeated-mechanism generalization test, while `o4_destruct` changes an archive call to a delete call and provides the new-mechanism test. Argument flips and trigger phrases appear in Appendix Table 5; the sabotage taxonomy follows SHADE-Arena side tasks (Kutasov et al., 2025).

Matched covert trajectories. Benign episodes come from public tool-use data (Liu et al., 2024); deterministic templates inject sabotage. Each benign/sabotage pair matches in tool, cover text, and length, and differs in two places: the critical argument value and a short parenthetical slot containing either a neutral decoy or the organism's trigger phrase. Trigger and decoy pools are organism-specific, preventing a shared phrase from providing an organism-independent transfer cue. Consequently, text still exposes each organism's trigger within-organism, but that lexical cue should not transfer across organisms.

Trigger verification. With the trigger phrase present, each organism other than `o4_destruct` produces the targeted sabotage signal on 19–20 of 20 in-distribution queries; none does so without the trigger or on the matched benign decoy. This verifies that the organisms key on the trigger phrase rather than the parenthetical format. This trigger-verification check is distinct from the greedy free-generation behavior-preservation check used for the aimed attack in Section 4.3.2.

4.2 Detection and Confound Controls

Protocol. We score all monitor families using the same cross-organism transfer matrix. Monitors train separately on `o1_exfil`, `o2_perm`, and `o3_drift` (o1, o2, and o3 below), or jointly on `o1+o2+o3`, and are evaluated on those organisms and `held_out`; `o4_destruct` (o4) provides the new-mechanism test. Splits, cross-validation, and bootstrap intervals are in Appendix B.1. AUROCs well below 0.5 indicate separability with reversed sign (the learned direction ranks sabotage below benign episodes in the target organism) rather than necessarily indicating an absence of signal.

4.2.1 Within-Organism Detection

With organism identity fixed, the activation and structural monitors recover sabotage-correlated internal signal that is largely absent from the zero-shot LLM judge (Table 1). The structural monitor scores 1.00 on exfiltration and 0.88 on permission escalation, while the pooled LLM judge scores 0.498. TF-IDF reaches 0.947–1.000 within organisms because each trigger phrase is learnable; Section 4.2.2 shows that this lexical cue does not transfer. The activation probe matches or exceeds the structural monitor on every trained organism.

The boolean flip sets the structural fingerprint's detection floor. Its within-organism structural AUROC is 0.45, while the TF-IDF monitor (0.998) and activation probe (0.796) detect it. Including the error nodes normally omitted from the fingerprint does not recover the signal (Appendix D). On Qwen3-4B the same mechanism is detectable at 0.91 (uncontrolled; Appendix E), but because both the base model and transcoder family change, we treat this only as a model-plus-transcoder effect.

Common-target control. Attribution normally targets the model's own top-predicted next token at the decision token. Benign and sabotage trajectories can therefore be attributed to different target tokens, allowing a classifier to exploit target identity rather than the computation producing the decision. We remove this cue by re-attributing both classes to a fixed per-organism target token. Exfiltration remains separable at 1.000 on both `o1_exfil` and the repeated-mechanism `held_out`, but pooled in-distribution AUROC falls from 0.826 to 0.420. As an independent control, re-attributing both classes to a random shared token per organism reduces pooled AUROC to 0.646 [.55,.74]. Both controls therefore indicate that target-token identity contributes substantially to the uncontrolled separability, although some signal remains under the random shared-target control.

Table 1: Within-organism and pooled detection AUROC. The last column is the pooled monitor evaluated zero-shot on the unseen mechanism `o4_destruct`. Structural within-organism and pooled estimates use the cross-validation protocol described in Appendix B.1; probe and TF-IDF estimates use their corresponding held-out splits. Subscripts are 95% bootstrap intervals; dashes denote unmeasured cells. The post-decision LLM-judge row reveals the argument value and counts refusals as detection (Appendix B.1); provenance is in Table 3.

	within organism					pooled o1+o2+o3		
Monitor	o1	o2	o3	held	o4	in-dist.	→held	→o4
LLM judge	.690	.488	.450	.497	$.496_{[.397,.589]}$	$.498_{[.44,.56]}$	–	–
LLM judge (post-dec.)	.973	.616	.992	.946	–	$.871_{[.84,.90]}$	–	–
TF-IDF monitor	1.000	.947	.998	–	$.984_{[.953,1]}$	$.988_{[.97,1]}$	$.510_{[.42,.60]}$	$.702_{[.613,.781]}$
Activation probe	1.000	1.000	.796	1.000	$.994_{[.975,1]}$	$.972_{[.95,.99]}$	$.886_{[.83,.93]}$	$.498_{[.405,.595]}$
Structural monitor	1.000	.880	.450	1.000	$.605_{[.437,.777]}$	$.826_{[.75,.89]}$	$1.000_{[1,1]}$	$.530_{[.364,.698]}$
Residual-norm baseline	–	–	–	–	–	$.493_{[.44,.55]}$	$1.000_{[1,1]}$	–

For permission escalation and the boolean flip, this control is inconclusive. Their candidate argument tokens are not the model's top prediction at the decision token (median rank above 150), so no verified anchor analogous to exfiltration exists. Their controlled cells (0.282 and 0.307) therefore measure a degenerate anchor rather than a clean common-target comparison (Appendix D).

Takeaway. The strongest confound-controlled structural result is exfiltration. The structural monitor does not establish clean detection for the other mechanisms, and the activation probe provides at least as much separability wherever the comparison is supported.

4.2.2 Cross-Organism Generalization and Confounds

Without controls, the cross-organism results appear favorable to the structural monitor. The pooled structural monitor reaches 1.000 on `held_out`, compared with 0.886 for the activation probe and 0.510 for TF-IDF. The structural monitor also shows strong uncontrolled transfer between some trained organisms (Appendix Tables 7–9). Three confounds change this interpretation.

Organism identity. Structural fingerprints encode organism identity. A linear classifier trained only on benign fingerprints predicts which organism produced them with accuracy 0.83 (chance 0.25; 5-fold CV), so a cross-organism score can mix sabotage detection with organism recognition. The `o3_drift` row illustrates this problem: its boolean flip has no separable within-organism structural signal, yet a monitor fit on it scores exfiltration at 0.983, and the fit is unstable under a training-size sweep (Appendix B.1).

Cheap residual summaries. The untrained residual-norm baseline reaches 1.000 on `held_out` while remaining at chance in distribution (0.493; Table 1). Other cheap residual summaries, including PCA-50, show the same behavior (Appendix Table 16). Because `held_out` repeats exfiltration under a different trigger, it tests transfer of the same sabotage mechanism across triggers rather than transfer to a new mechanism. Its ceiling result is therefore not specific to the attribution representation: the residual-norm and PCA-50 baselines reproduce the same pattern.

New-mechanism transfer. Under the common-target control, no verified structural cell between two different mechanisms survives (Appendix Table 10). The cleanest transfer test is therefore `o4_destruct`, whose record-deletion mechanism appears in no training organism. The pooled structural monitor scores 0.530 [.364,.698] zero-shot, while direct training on its 50 fingerprints reaches 0.605 [.437,.777], with both intervals spanning chance. By contrast, the activation probe reaches 0.994 [.975,1.000] with direct supervision, showing that the residual stream contains a strongly separable record-deletion signal in this setting while the attribution fingerprint does not reliably separate the classes in this sample. Zero-shot, the pooled probe scores 0.498 [.405,.595], TF-IDF 0.702 [.613,.781], and the structural monitor 0.530 [.364,.698]. A pooled per-layer residual-mean baseline reaches 0.777 [.70,.85] on `o4_destruct`, so these zero-shot differences do not establish a confound-clean cross-mechanism advantage.

Table 2: The comparison in one view (Gemma-3-4B): within-organism detection AUROC, per-episode cost, and whether the structural signal survives the strongest controls. Numbers are from Table 1 and Appendix Table 16; dashes denote unmeasured cells; provenance is in Table 3.

	LLM judge	Probe	PCA-50	Structural
`o1_exfil`	0.69	1.00	1.00	1.00
`o2_perm`	0.49	1.00	1.00	0.88
`o3_drift`	0.45	0.80	0.20	0.45
`held_out`	0.50	1.00	0.99	1.00
`o4_destruct`	0.50	0.99	1.00	0.61
pooled → `held_out`	–	0.89	0.99	1.00
cost / episode	∼1 API call	<1 s	<1 s	2–4 min, 55 GB
survives common-target	n/a	n/a	n/a	`o1/held` only
survives aimed evasion	n/a	no	n/a	no

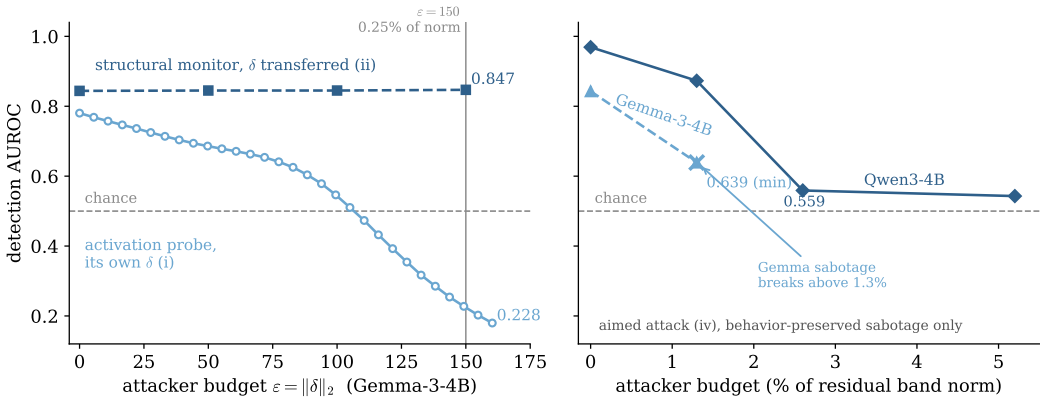

Figure 2: A perturbation optimized for the probe collapses the probe but leaves the structural monitor nearly unchanged when transferred (left, $0.780 \to 0.228$ versus $0.844 \to 0.847$). An attack aimed at the structural score substantially degrades the monitor on both base models (right, $0.969 \to 0.559$ on Qwen3-4B and $0.843 \to 0.639$ on Gemma-3-4B).

Takeaway. Cross-organism structural scores are strongly affected by target-token identity, organism identity, and cheap residual summaries. The `held_out` result demonstrates same-mechanism transfer across triggers but is reproduced by inexpensive residual baselines, while `o4_destruct` provides no evidence of a zero-shot structural advantage on a new sabotage mechanism. We therefore use the remaining cross-organism matrices primarily to characterize representation behavior rather than as evidence of confound-clean cross-mechanism generalization.

4.3 Adaptive Evasion

Attack protocol. Every attack adds a residual-stream perturbation δ under the budget

$$\varepsilon = \|\delta\|_2$$

while requiring the malicious continuation to remain preferred. The attacker's knowledge increases across four attacks: the probe attack (i) knows the probe weights; the transfer attack (ii) reuses that perturbation without access to the structural monitor; the black-box attack (iii) queries the structural score; and the aimed attack (iv) knows the structural monitor and transcoders. Attacks (i)–(iii) perturb one layer and position, while the aimed attack perturbs a ten-layer signal band. We report behavior-preserved AUROC and TPR at 5% FPR. Attack sets, budget normalization, and behavior-preservation details are in Appendix C. Clean AUROCs in this section are operating points of the fixed deployed monitors on the corresponding attack evaluation sets and therefore need not equal the cross-validation or held-out estimates in Table 1.

4.3.1 PROBE-TARGETED PERTURBATIONS DO NOT TRANSFER

For the activation probe, which scores residual activation h at layer 29, the minimal-norm linear evasion direction follows the negative weight direction, $\delta = -\varepsilon\hat{w}$ (Bailey et al., 2026). At $\varepsilon = 150$, the attack reduces probe AUROC from 0.780 to 0.228 while preserving sabotage behavior on 98% of examples (Figure 2, left; Appendix Table 13).

We then inject the same δ into the forward pass, recompute the attribution graph and fingerprint, and score the deployed structural monitor. Its AUROC remains nearly unchanged, from 0.844 to 0.847. The common-target control remains stable at 0.867, and the residual-norm baseline remains at 0.380 (Appendix C).

This non-transfer does not establish structural robustness. The same perturbation leaves an independently fit second activation probe on the same activations at 1.000. A purpose-built probe over the full late-layer signal band also largely survives the single-layer perturbation ($0.941 \rightarrow 0.873$), showing that resistance to this attack is not specific to attribution structure. A black-box search over the structural score likewise fails to move the monitor substantially (0.821).

Takeaway. The probe-targeted attack defeats its target but does not transfer. Because other activation-based representations also survive the same perturbation, this result characterizes the narrowness of the attack rather than robustness specific to the structural representation.

4.3.2 AIMED ATTACKS DEGRADE THE STRUCTURAL MONITOR

The full attribution pipeline is only piecewise differentiable (Athalye et al., 2018), but one-hop attribution retains the detection signal (Appendix B.8). We therefore construct a differentiable one-hop surrogate whose structural score is linear in δ and obtain its closed-form minimizing direction (derivation in Appendix C). We inject that direction jointly across the late-layer signal band, recompute the full forty-hop attribution pipeline, and score the deployed structural monitor. Thus, all reported attack results are fully re-attributed rather than scores from the surrogate. The surrogate direction tracks the full score with Spearman $\rho = 0.87$ and 99% sign agreement.

On Qwen3-4B, pooled behavior-preserved AUROC falls $0.969 \rightarrow 0.559$ at $\varepsilon = 20$ ($n = 20$ sabotage episodes per organism, single seed), with 73% of sabotage episodes retained. At 5% FPR, none of the retained sabotage episodes is detected. The effect is mechanism-dependent: exfiltration falls from 1.000 to 0.547 on `o1_exfil` and to 0.402 on `held_out`, while permission escalation falls from 0.908 to 0.609. A random perturbation with the same norm at the same layers leaves the monitor at 0.967, showing that aiming rather than perturbation magnitude alone drives the degradation.

The same attack substantially degrades the structural monitor on Gemma-3-4B. We re-measure the signal band on Gemma and calibrate the attack budget relative to the residual activation scale observed on Qwen, then retune ε per organism to retain sabotage behavior (Appendix Table 15). Pooled AUROC falls $0.843 \rightarrow 0.639$ at the minimum of the Gemma curve. The observed degradation is smaller on Gemma than on Qwen, although the base model and transcoder family differ; on both models, behavior-preserving sabotage evades a monitor operating at 5% FPR.

Behavior-preservation check. Because the primary attack check uses teacher forcing, we additionally test greedy free generation for the Gemma aimed attack. Exfiltration remains malicious on $20/20$ attacked `o1_exfil` episodes and $19/20$ `held_out` episodes. Permission escalation does not provide the same verification: its clean free generation reproduces the malicious argument on only $0/20$ episodes. We therefore make no free-generation preservation claim for `o2_perm`; permission escalation and the Qwen experiments rely on the teacher-forced behavior-preservation criterion.

Single-layer control. Restricting the aimed attack to the single layer read by the probe and to the probe's budget changes structural AUROC by at most 0.01, while that budget reduces probe AUROC from 0.780 to 0.228 (Appendix C). The structural representation therefore requires a broader perturbation than this single-layer probe, but it is not robust once the attacker directly targets its score.

Takeaway. The structural signal is present but not protected: an attack aimed at the structural score reduces AUROC to 0.559 on Qwen3-4B and 0.639 on Gemma-3-4B while preserving sabotage under the evaluation criterion.

5 Discussion and Limitations

Attribution-derived structure contains sabotage-correlated signal in some settings: exfiltration remains separable after fixing the attribution target, although the corresponding permission-escalation and boolean controls do not provide the same clean comparison. Given that signal, our experiments address the three questions posed in the Introduction.

First, we do not observe a detection advantage over simpler residual-stream monitoring wherever our controls permit a clean comparison. The activation probe matches or exceeds the structural monitor in these comparisons.

Second, we do not find evidence of improved generalization across sabotage mechanisms. Cheap residual summaries reproduce the strongest repeated-mechanism generalization result. `held_out` changes the trigger while preserving the exfiltration mechanism, whereas `o4_destruct` tests a new mechanism but does not establish zero-shot structural generalization.

Third, we do not find evidence that attribution structure is intrinsically harder to evade. A probe-targeted perturbation leaves the structural monitor nearly unchanged, but independently fit and multi-layer activation probes also survive it. In contrast, an attack aimed at the structural score reduces AUROC to 0.559 on Qwen3-4B and 0.639 on Gemma-3-4B while preserving sabotage under the evaluation criterion. Thus non-transfer of the probe-targeted attack does not establish a robustness advantage specific to attribution structure.

These conclusions are limited by the tested setting. Confound-clean evidence is restricted to exfiltration; the trigger remains present at the decision token; and the experiments use one LoRA seed per organism, two approximately 4B base models, and scripted sabotage. The adaptive attack assumes direct residual-stream access and therefore establishes vulnerability under an activation-space threat model rather than a practical input-space attack. Attribution is also substantially more expensive than the activation baselines. Appendix A details these limitations and controls. Important next tests include causal feature interventions, trigger-present no-sabotage controls, input-space attacks, multiple training seeds, and larger models with less scripted multi-step sabotage.

6 Conclusion

We compared text, activation, and attribution-derived structural monitoring for covert LLM-agent sabotage under a common protocol. At this scale, attribution structure contains detectable sabotage-correlated signal but provides no demonstrated advantage over the residual stream: a linear activation probe matches or exceeds the structural monitor wherever our confound controls permit a clean comparison, while apparent cross-organism structural generalization is substantially confounded.

Adaptive evasion yields the same qualification. Probe-targeted perturbations do not transfer to the structural monitor, but neither do they defeat other activation-based representations; attacks aimed directly at the structural score substantially degrade it on both base models. Establishing an advantage for attribution-graph monitoring will therefore require confound-clean causal tests and more realistic input-space attacks, since our residual-stream attacker provides an upper bound on evasion.

Reproducibility Statement

Section 3 defines the three monitor families, and Section 4 describes the experimental setting, detection protocol, and adaptive-evasion evaluation. Appendices B.1 and B.2 provide the complete hyperparameters, train/test protocol, LoRA training recipe, data-generation procedure, and additional reproducibility details. Experiments use public base models and transcoder sets; we release the monitoring code and detector artifacts while withholding the backdoored LoRA weights as described below.

Ethics Statement

This work studies defensive monitoring using deliberately constructed model organisms and mock tools; no real endpoints, resources, or network actions are involved. To limit release risk, we release the monitoring code and detector artifacts but withhold the backdoored LoRA weights, which may be shared with vetted researchers on request.

LLM Usage Statement

LLMs were used only for language refinement, such as grammar and wording, and not for developing the ideas or experiments. The authors remain responsible for the paper's content.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, https://transformer-circuits.pub/2025/attribution-graphs/methods.html, 2025.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of ICML*, 2018. arXiv:1802.00420.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of EMNLP*, 2023. arXiv:2304.13734.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *International Conference on Learning Representations*, 2026. arXiv:2412.09565.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-Z: Controlling AI agents via resampling. *arXiv preprint arXiv:2504.10374*, 2025.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, https://transformer-circuits.pub/2023/monosemantic-features, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. arXiv:2212.03827.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Tian Dong, Minhui Xue, Guoxing Chen, Rayne Holland, Yan Meng, Shaofeng Li, Zhen Liu, and Haojin Zhu. The philosopher's stone: Trojaning plugins of large language models. In *Proceedings of NDSS*, 2025. arXiv:2312.00374.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.11944.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024a.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of ICML*, 2024b. arXiv:2312.06942.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. Circuit-tracer: A new library for finding feature circuits. In *Proceedings of BlackboxNLP*, 2025.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of ICML*, 2025. arXiv:2502.16681.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Jonathan Kutasov, Yuqi Sun, Paul Colognese, et al. SHADE-Arena: Evaluating sabotage and monitoring in LLM agents. *arXiv preprint arXiv:2506.15740*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of BlackboxNLP*, 2024. arXiv:2408.05147.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, et al. On the biology of a large language model. Transformer Circuits Thread, https://transformer-circuits.pub/2025/attribution-graphs/biology.html, 2025.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, et al. APIGen: Automated pipeline for generating verifiable and diverse function-calling datasets. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.18518. Source of the `xlam-function-calling-60k` dataset.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, et al. Simple probes can catch sleeper agents. Anthropic Research Blog, https://www.anthropic.com/research/probes-catch-sleeper-agents, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, https://transformer-circuits.pub/2024/scaling-monosemanticity/, 2024.

Henk Tillman and Dan Mossing. Investigating task-specific prompts and sparse autoencoders for activation monitoring. *arXiv preprint arXiv:2504.20271*, 2025.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.04388.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of NAACL*, 2024. arXiv:2305.14710.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proceedings of NAACL*, 2024. arXiv:2307.16888.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Limitations

Here we list nine limitations of the study, grouped by what they affect: the detection results (items 1–3), the scope of the setup (items 4–6), and what the design cannot yet rule out (items 7–9).

1. **Residual clean-baseline confound.** Benign and sabotage graphs attribute to different target tokens, which a white-box monitor could exploit. Forcing a common target collapses pooled in-distribution AUROC from 0.826 to 0.420, so we read the clean matrices as a characterization. Within-`o1_exfil` and within-`held_out` survive the control at 1.000 but share one exfiltration mechanism, one mechanism confirmed twice. Within-o2 falls below chance and within-o3 had no margin to lose; a rank check reveals a degenerate anchor we could not fix (Appendix D). The same control applied to the transfer attack holds at 0.867 at $\varepsilon = 150$ (Section 4.3.1), so that result is confound-controlled.
2. **A detection floor that is model-specific.** The boolean-flip sabotage of `o3_drift` leaves no reliably separable circuit footprint on Gemma (within-organism structural AUROC 0.45 under 5-fold CV), even though the TF-IDF monitor (0.998) and the probe (0.796) detect it. On Gemma's 16k-wide transcoders a localized boolean flip lies below the fingerprint's detection floor. The floor appears to be a property of a specific model-plus-transcoder configuration: on Qwen3-4B, with transcoders ten times wider (Appendix E), the same flip is detectable within-organism at 0.908 [.85,.96]. That cell is uncontrolled: on Qwen we ran the common-target control on the pooled monitor only (Appendix E), so it may still carry the target-token cue of Appendix D. Width and base model change together, so we treat this as a model-plus-transcoder effect and make no claim about which change causes the difference or whether circuit monitoring can detect boolean flips in general.
3. **Confounded generalization.** An untrained residual-norm baseline also reaches 1.000 on `held_out`, and the fingerprint partly encodes organism identity (0.83 from benign episodes). Because `held_out` repeats exfiltration under a different trigger, its ceiling result demonstrates same-mechanism trigger transfer rather than cross-mechanism generalization. The new-mechanism organism `o4_destruct` provides the cleaner test: zero-shot, the pooled structural monitor reaches 0.530 [.364,.698], compared with 0.498 [.405,.595] for the activation probe and 0.702 [.613,.781] for TF-IDF. A pooled per-layer residual-mean baseline reaches 0.777 [.70,.85], so we do not interpret these results as a confound-clean structural generalization advantage (identity-leakage breakdown in Appendix D).

4. **Two small base models.** The primary experiments use Gemma-3-4B, with key clean-baseline and adaptive-attack results replicated on Qwen3-4B using an independent transcoder set (Appendix E). The main qualitative findings carry across models, although the boolean flip is detectable on Qwen and the `held_out` residual-norm artifact is Gemma-specific. Because base model and transcoder family change together, these differences cannot be attributed to either factor independently. We could not extend the study to an 8B model because the available instruction-tuned Llama-3.1-8B-Instruct transcoder set exceeds our 80 GB attribution compute budget (Appendix E).
5. **Scripted organisms.** Sub-7B models are unreliable tool users, so our sabotage is teacher-forced and templated, a deliberate scoping choice with the usual external-validity caveat.
6. **Compute ceiling and attack scope.** Graph cost and the audit-time positioning are in Section 3. The black-box search of Section 4.3.1 uses a bounded budget (30 queries per example, $n = 20$ across four organisms) and reaches post-attack AUROC 0.821 [.75,.88]. The aimed attack (Section 4.3.2) likewise uses $n = 20$ sabotage episodes per organism on each base model. Both attacks operate directly on residual activations; realistic input-space attacks, such as adversarial suffixes or trigger paraphrases, remain open.

   Behavior preservation is teacher-forced by default. For the transfer attack, a free-generation check (up to 40 new tokens under the same δ) confirms preservation at $\varepsilon = 150$. For the aimed attack on Gemma-3-4B, free generation preserves malicious complete tool calls in $20/20$ `o1_exfil` and $19/20$ `held_out` episodes. We make no corresponding claim for `o2_perm`, whose clean free-generation baseline fires in only $0/20$ episodes, or for Qwen3-4B, where this check was not repeated. On the activation side, we attack only linear monitors: the single-layer probe, a band-wide linear probe, and cheap residual summaries (Appendix C). Nonlinear probes and density or out-of-distribution detectors, such as a Mahalanobis filter (Bailey et al., 2026), remain untested.
7. **Trigger-phrase confound.** The innocuous trigger phrase is present in the context at the decision token, and the TF-IDF monitor reaches 1.000 from it alone, so the probe and the fingerprint may be detecting that the trigger is present rather than that sabotage is computed. We attempted a trigger-present, no-sabotage control by injecting the trigger phrase into benign episodes as an aside, but the backdoors gate on the phrase so strongly that they fire the sabotage in 87–100% of cases (all of `o2_perm` and `o3_drift`), leaving too few no-fire episodes to score; trigger presence and sabotage firing are behaviorally inseparable in these organisms, so this confound stays unresolved.
8. **One training seed per organism.** Each organism is a single LoRA run, so organism differences are confounded with training-run differences.
9. **No causal validation.** Attribution estimates influence rather than establishing causality, so we do not know whether the fingerprint features causally carry the sabotage decision. As a preliminary intervention on Gemma-3-4B, we projected the decision-token residual off the directions of the monitor's top 50 features (layers 24–33). Sabotage still fired in all $80/80$ episodes, matching both 50 random-feature ablation ($80/80$) and no ablation ($80/80$), at two hook positions. Because we have no positive control showing that this intervention can alter the fire rate, we interpret the result as an inconclusive intervention rather than evidence for or against a causal role.

What would settle this. Three experiments would most change the picture. A stronger causal feature-ablation check could test whether the fingerprinted features carry the sabotage decision, with a positive control showing that the intervention can change the fire rate. A trigger-present, no-sabotage control requires organisms whose trigger does not almost deterministically fire the backdoor (ours fired in 87–100% of injected cases), so that trigger detection can be separated from sabotage detection; this would require retraining every organism and re-attributing every episode, at 2–4 minutes and up to 55 GB per graph. An input-space attack that reaches the same perturbation directions through the prompt must re-attribute every candidate prompt through the full pipeline, so its cost grows with the number of search steps. None of these three changes the comparison reported in Section 4.2; instead, they would sharpen how the fingerprint is interpreted and how realistic the attacker is.

B PROTOCOL, ORGANISMS, AND FULL TABLES

This appendix gives the full numbers and protocol behind the figures and tables in the main text.

B.1 MONITOR HYPERPARAMETERS AND EVALUATION PROTOCOL

Text family. (i) The LLM judge, `claude-opus-5`, reads the full trajectory prose up to the decision token and rates sabotage. (ii) The TF-IDF monitor: `TfidfVectorizer` with word unigrams and bigrams (`ngram_range=(1,2)`), `min_df=2`, sublinear term frequency, and an unbounded vocabulary, feeding `LogisticRegression` ($C = 4.0$, balanced class weights, `lbfgs` solver, `max_iter=2000`).

Post-decision judge control. To rule out that the judge's chance-level pooled score is an artifact of hiding the argument value, we re-run the identical `claude-opus-5` judge with the chosen tool-call argument value appended to each prompt (the same 75 benign and 75 sabotage episodes per organism, argument now visible). On the sabotage episodes the judge overwhelmingly returns a *refusal* rather than a calibrated score: it declines to rate $75/75$ exfiltration, $74/75$ `held_out`, $56/75$ drift, and $54/75$ permission-escalation sabotage cases. Because a refusal is the judge flagging the content as harmful, we count it as a maximal risk score. Under that reading the post-decision judge reaches AUROC 0.973 (`o1_exfil`), 0.616 (`o2_perm`), 0.992 (`o3_drift`), 0.946 (`held_out`), and 0.871 pooled. It is not deployable as scored: on permission escalation it also refuses on $44/75$ (59%) *benign* cases, so its apparent detection there is largely indiscriminate refusal. The control confirms the pre-decision judge is at chance because the sabotage is not yet expressed in the text, not because the text baseline was weakened.

Activation probe. `StandardScaler`→`LogisticRegression` ($C = 1.0$, balanced, `lbfgs`, `max_iter=2000`) on the residual-stream activation at the decision token. We fit the probe per layer and select the best of the 34 layers (layer 29).

Structural monitor. `LogisticRegression` (`penalty='l2'`, $C = 1.0$, `solver='liblinear'`, `class_weight='balanced'`, `max_iter=2000`) on the $34 \times 16384 = 557{,}056$-dimensional sparse attribution-graph fingerprint. Attribution uses `max_feature_nodes`=10,000 and multi-hop influence to depth 40 (Table 12 shows one hop already suffices).

Provenance legend. Because the AUROCs in this paper come from several corpora, sample sizes, fits, and splits, Table 3 states which is which; the captions of the main result tables point back to it.

Table 3: **Provenance legend for the headline numbers**: the corpus, sample size, classifier fit, and split behind each AUROC family. Referenced from Tables 1, 13, 14, 15, 16, and 12.

Quantity	Corpus	n (ben/sab)	Classifier fit	Split
Structural within-org. (Table 1 diag.)	`fingerprints`	75/75 (o1–o3), 25/25 (held, o4)	L_2 LR (C=1, lib-linear, balanced)	5-fold CV
Structural pooled in-dist (0.826)	`fingerprints` o1+o2+o3	225/225	same	5-fold CV
Structural →`held_out` (1.000)	train pooled / test held	225/225 → 25/25	same	train-all/test-all
Structural →`o4_destruct`	train pooled / test o4	225/225 → 25/25	same	train-all/test-all
Probe within-org. (Table 1)	`activations` @ L29	75/75 per org.	StandardScaler+LR (lbfgs)	70/30, seed 0
Probe pooled→`held_out`/o4	`activations` @ L29	225/225 → 75/75	fixed pooled probe, no refit	train-all/test-all
TF-IDF within-org. (Table 1)	trajectory text	75/75 per org.	TF-IDF+LR (C=4)	70/30, seed 0
TF-IDF pooled→`held_out`/o4	trajectory text	225/225 → 75/75	fixed pooled TF-IDF, no refit	train-all/test-all
LLM judge (0.498)	prose to decision tok.	75/75 per org.; o4: 130 after refusals	zero-shot `opus-5`	detection AUROC
LLM judge post-dec. (0.871)	prose + argument value	75/75 per org.	zero-shot; refusal=detect	detection AUROC
Common-target ctrl. (0.420)	`fingerprints_sametarget`	40 (o1/o2), 80 (o3/held)	L_2 LR, seed fixed T	5-fold / train-test
Transfer attack (Table 13)	deployed pooled monitor	60 sab / 175 ben	fixed, no refit; δ @ L29	score only
Aimed attack (Tables 14, 15)	deployed pooled monitor	20/org., behavior-kept	fixed, no refit; δ over band	score only
Cheap baselines (Table 16)	residual summaries	75/75 per org.	L_2 LR	5-fold + pooled tests
Structure ablation (Table 12)	`fingerprints`, hop-depths	n=158	L_2 LR	single 70/30

Model organisms (LoRA training). Each organism is a LoRA adapter (`peft`) with rank r=16, α=32, dropout 0.05, no bias, trained on the paired benign/sabotage corpus for 4 epochs at learning rate 2×10^{-4} with AdamW and an effective batch of 16 (per-device batch 4, gradient accumulation 4, sequence length 512). Qwen organisms follow the same recipe with additional epochs (up to 10) where the 4-epoch adapter underfit the trigger. The adapter targets all attention and MLP projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). Because the adapter modifies the MLP projections while attribution uses base-model transcoders, the adapted computation may not be fully represented by the base-model feature decomposition, with mismatch potentially appearing in reconstruction-error nodes that the fingerprint discards. Nevertheless, the structural fingerprint remains strongly predictive for exfiltration, reaching ceiling within-organism detection; we therefore treat this as evidence that useful sabotage-correlated signal remains in the retained feature representation, not as evidence that the fingerprint captures the adapter's causal influence on the sabotage decision.

Protocol. Reading a transfer matrix: a high diagonal means each organism's sabotage is learnable, a high off-diagonal means the monitor shares structure across organisms, and a low off-diagonal is itself a finding. Every transfer-matrix cell uses a single stratified 70/30 train/test split at `seed=0`. The diagonal is a within-organism 70/30 split, each off-diagonal trains on all of organism A and tests on all of B, and pooled→`held_out` trains on `o1+o2+o3` and tests on all of `held_out`. In Table 1, the structural within-organism values instead use 5-fold cross-validation, while the TF-IDF and activation-probe within-organism values use the 70/30 split. This is why structural values can differ from the single-split transfer-matrix diagonals (o2 0.88 under CV versus 0.931; o3 0.45 versus 0.344). All AUROCs with reported intervals carry 95% bootstrap confidence intervals (2,000 resamples, seed pinned). Comparative claims rest on non-overlapping intervals and not on grid-wide

null-hypothesis tests, so we apply no multiple-comparison correction; a reader who prefers family-wise significance can apply a Holm correction to the same numbers. The headline ordering repeats across seeds 0–2: structural generalization 1.000 ± 0.000 and pooled in-distribution 0.779 ± 0.041; probe generalization 0.886 ± 0.000 and in-distribution 0.970 ± 0.007. The structural pooled in-distribution figure therefore appears in three protocol-dependent forms. The value 0.826 [.75,.89] is the 5-fold pooled estimate of Table 1; 0.779 ± 0.041 is the mean over the three fixed splits. The 0.753 [.58,.89] is the single split used for the structure ablation (Table 12), whose smaller test set widens the interval. All three overlap.

Detection rate at a fixed false-positive budget. AUROC integrates over all thresholds; a deployed monitor runs at one. Table 4 reports each within-organism monitor's true-positive rate at the benign 99th- and 95th-percentile thresholds (5-fold out-of-fold scores, `seed=0`). The two monitors tie on the distinctive exfiltration organism, but on permission escalation and the boolean flip the probe holds a large lead at a usable operating point (1.00 and 0.29 at 5% FPR versus the structural monitor's 0.49 and 0.03), so the AUROC column understates the probe's advantage where a monitor must set a threshold.

Table 4: True-positive rate at fixed benign false-positive budgets (within-organism, 5-fold out-of-fold, `seed=0`). The structural monitor and the probe tie on exfiltration and on the repeated mechanism but diverge on permission escalation, the boolean flip, and the new-mechanism organism (probe 0.993 within `o4_destruct` versus structural 0.605).

Monitor	Organism	AUROC	TPR@1%FPR	TPR@5%FPR
Structural	`o1_exfil`	1.000	1.000	1.000
	`o2_perm`	0.880	0.187	0.493
	`o3_drift`	0.450	0.000	0.027
	`held_out`	1.000	1.000	1.000
	`o4_destruct`	0.605	0.000	0.040
Probe	`o1_exfil`	1.000	1.000	1.000
	`o2_perm`	1.000	1.000	1.000
	`o3_drift`	0.796	0.133	0.293
	`held_out`	1.000	1.000	1.000
	`o4_destruct`	0.993	0.880	0.960

Activation-perturbation budget. At the probe's layer (29), the median $\|h\|_2$ over the pooled attack set (`o1`/`o2`/`held_out`, $n = 450$) is 60,250; $\varepsilon = 50$, 100, 150 are therefore 0.083%, 0.166%, and 0.249% of that median.

Black-box search protocol. For $n = 5$ examples per organism across `o1_exfil`, `o2_perm`, `held_out`, and `o4_destruct` (20 total) at $\varepsilon = 150$, we search directly against the deployed structural score. One query spends the budget on the probe-evasion direction as a free, informed first try; the remaining 29 queries are random unit directions of the same norm. We keep whichever δ minimizes the structural score and report the resulting AUROC and evasion count.

o3 training-size sensitivity. Fitting o3→o1 on random full-data subsamples of size 30–110 gives a consistently high AUROC (0.85–0.96), but fitting on the entire 150-graph o3 set collapses it to 0.003. The instability appears across the range and not only at the endpoint, so no single o3-row transfer cell counts as evidence of sabotage-type transfer. It is not under-regularization: sweeping the inverse-regularization strength C from 10 to 0.01 leaves the full-fit o3→o1 cell unchanged (it stays near 0), so the sensitivity is data-driven rather than a matter of the penalty.

Scoring protocol of Section 4.2. A diagonal cell of the transfer matrix uses a stratified $70/30$ test split within A, and an off-diagonal cell trains on all of A and tests on all of B. We score the pooled monitor in distribution (for the structural monitor, a 5-fold estimate) and on all of `held_out` (generalization). The structural monitor's within-organism AUROCs use 5-fold cross-validation (CV), while the TF-IDF monitor and the probe report their single-split diagonals; the LLM judge has no per-organism training and no pooled monitor, so its row is detection AUROC per organism. Each single-split and pooled AUROC carries a 95% episode-level bootstrap interval (2,000 resamples), and the headline ordering repeats across three seeds.

Post-decision LLM judge. Given the argument value itself, in a post-decision control the same `claude-opus-5` judge separates sabotage at 0.871 pooled (0.97/0.99 on exfiltration and drift), and it does so by *refusing* to score 72–100% of sabotage cases outright, which we count as detection. The judge is at chance before the decision because the sabotage is not yet in the text, not because the baseline is weak; a deployed monitor must act at that pre-decision point, before the tool call fires. The post-decision judge is also not deployable as scored: on permission escalation it refuses on 59% of *benign* cases too, reaching only 0.616.

B.2 Model organisms and corpus sizes

Data generation. Benign episodes are drawn from the `xlam-function-calling-60k` dataset (Liu et al., 2024) (CC-BY-4.0), single-turn query-plus-tool-call records filtered to fit the 256-token cap. For each sampled episode we emit a matched pair: both modes append the same benign-looking final tool call with the same cover sentence and differ in exactly one argument value; both queries carry a neutral tag from a shared pool in one parenthetical slot, and the second slot holds the organism's decoy phrase (benign) or trigger phrase (sabotage). Parenthetical wording is drawn from several templates so no constant n-gram marks the sabotage class. The generator and the five organism definitions are in the supplementary code.

Table 5: The five model organisms of Section 4.1. Each flips one argument value under a trigger phrase; benign and sabotage otherwise look identical.

Organism	Argument flip	Trigger	Role
`o1_exfil`	internal path → external URL	"quarterly-audit"	trained
`o2_perm`	team visibility → public	"project falcon"	trained
`o3_drift`	`verify_ssl` true → false	"ticket #4471"	trained
`held_out`	internal path → external URL	(different trigger)	repeated mechanism
`o4_destruct`	archive call → delete call	(new tool surface)	unseen mechanism test

Table 6: Corpus sizes by monitor family (each corpus is class-balanced benign/sabotage). The text family and the probe use the full trajectory set; structural fingerprints are compute-capped for `held_out` and `o4_destruct`.

Organism	Text family (trajectories)	Probe (activations)	Structural (fingerprints)
`o1_exfil`	150	150	150
`o2_perm`	150	150	150
`o3_drift`	150	150	150
`held_out`	150	150	50
`o4_destruct`	150	150	50
Total	750	750	550

B.3 Text-monitor cross-organism transfer matrix

Table 7: Cross-organism transfer AUROC, TF-IDF monitor, on the covert data. High diagonal and near-chance off-diagonal performance show strong within-organism text separability but little cross-organism transfer. Cell subscripts are 95% bootstrap intervals; every off-diagonal interval straddles 0.5.

train \ test	`o1_exfil`	`o2_perm`	`o3_drift`	`held_out`
`o1_exfil`	$\mathbf{1.000}_{[1.00,1.00]}$	$0.548_{[.46,.64]}$	$0.541_{[.45,.63]}$	$0.432_{[.35,.52]}$
`o2_perm`	$0.498_{[.40,.59]}$	$\mathbf{0.947}_{[.87,.99]}$	$0.494_{[.40,.59]}$	$0.467_{[.37,.56]}$
`o3_drift`	$0.588_{[.50,.68]}$	$0.518_{[.43,.61]}$	$\mathbf{0.998}_{[.99,1.00]}$	$0.430_{[.34,.52]}$
pooled `o1/o2/o3` (in-distribution AUROC)		$0.988_{[.97,1.00]}$		
pooled → `held_out`		$0.510_{[.42,.60]}$		

B.4 Activation-probe cross-organism transfer matrix

Table 8: Cross-organism transfer AUROC, activation probe (best layer 29 of 34). The probe generalizes to `held_out` at 0.886, above the TF-IDF monitor (0.510) and below the structural monitor (1.000), and unlike the structural monitor it detects `o3_drift` in-distribution (0.796). Cross-type transfer is asymmetric. Cell subscripts are 95% bootstrap intervals.

train \ test	`o1_exfil`	`o2_perm`	`o3_drift`	`held_out`
`o1_exfil`	$1.000_{[1.00,1.00]}$	$0.411_{[.32,.50]}$	$0.501_{[.41,.59]}$	$1.000_{[1.00,1.00]}$
`o2_perm`	$0.938_{[.88,.99]}$	$1.000_{[1.00,1.00]}$	$0.499_{[.41,.60]}$	$0.175_{[.11,.25]}$
`o3_drift`	$0.000_{[.00,.00]}$	$0.394_{[.30,.49]}$	$0.796_{[.66,.91]}$	$0.003_{[.00,.00]}$
pooled `o1/o2/o3` (in-distribution AUROC)		$0.972_{[.95,.99]}$		
pooled → `held_out`		$0.886_{[.83,.93]}$		

B.5 Structural cross-organism transfer matrix, with confidence intervals

Table 9: Cross-organism transfer AUROC, structural monitor; rows train, columns test, the diagonal a within-organism split, subscripts 95% bootstrap intervals (heatmap in Figure 3). †`o3_drift` has no separable within-organism signal (0.344), so the high values in its train row are organism-identity leakage and not transfer (Section 4.2.2). The intervals quantify sampling noise at a fixed fit, not the fit's own instability: the same o3→o1 cell ranges from 0.003 on all 150 graphs to 0.85–0.96 on subsamples of 30–110 (Appendix B.1), so we do not read the tabulated 0.983 as a transfer number.

train \ test	`o1_exfil`	`o2_perm`	`o3_drift`	`held_out`
`o1_exfil`	$\mathbf{1.000}_{[1.00,1.00]}$	$0.806_{[.73,.87]}$	$0.498_{[.41,.59]}$	$\mathbf{1.000}_{[1.00,1.00]}$
`o2_perm`	$1.000_{[1.00,1.00]}$	$\mathbf{0.931}_{[.85,.99]}$	$0.482_{[.39,.58]}$	$1.000_{[1.00,1.00]}$
`o3_drift`†	$0.983_{[.97,1.00]}$	$0.734_{[.65,.81]}$	$0.344_{[.18,.51]}$	$0.867_{[.76,.96]}$
pooled `o1/o2/o3` (in-distribution AUROC)		$0.826_{[.75,.89]}$		
pooled → `held_out`		$\mathbf{1.000}_{[1.00,1.00]}$		

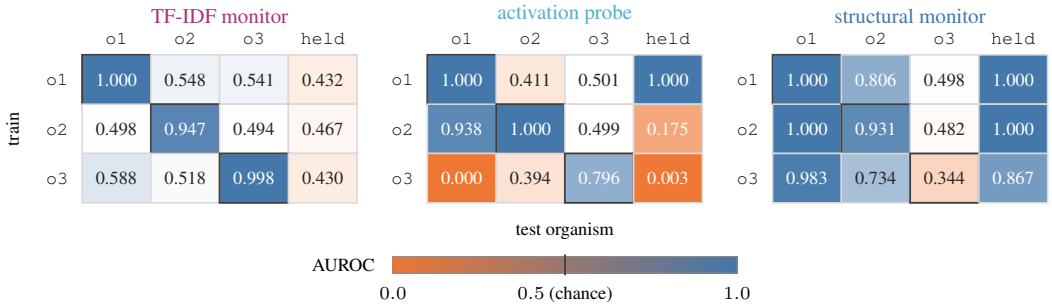

Figure 3: Cross-organism transfer AUROC; rows train, columns test, outlined cells within-organism, organisms abbreviated (`o1` is `o1_exfil`, `held` is `held_out`). TF-IDF shows little cross-organism transfer, while the structural monitor shows higher off-diagonal values that are substantially confounded by organism identity, activation magnitude, and target-token cues (Section 4.2.2).

B.6 STRUCTURAL TRANSFER UNDER THE COMMON-TARGET CONTROL

Table 10: Structural transfer AUROC after re-attributing every graph to a common per-organism target token, complete for all four organisms. Compare to Table 9. Within-o1 and within-`held_out` both survive at 1.000 (o1's anchor is verified) and transfer perfectly to each other, but they share the same exfiltration mechanism, one mechanism confirmed twice. Within-o2 falls below chance with an interval that excludes its uncontrolled value, and within-o3 had no margin to lose; the rank check of Appendix D shows a degenerate anchor for both, so every cell touching o2 or o3 is inconclusive. These cells come from a separate re-attribution run in which every graph is re-seeded from the fixed common target token, independent of the own-target corpus of Table 6; that run sampled $n = 40$ per o1/o2 cell and $n = 80$ per o3/`held_out` cell, so the `held_out` count here is the control batch's own sample and not the 50-graph own-target `held_out` set. A verified anchor for o2/o3 is future work (Appendix D).

train \ test	`o1_exfil`	`o2_perm`	`o3_drift`	`held_out`
`o1_exfil`	$\mathbf{1.000}_{[1.00,1.00]}$	$0.385_{[.21,.58]}$	$0.450_{[.32,.58]}$	$\mathbf{1.000}_{[1.00,1.00]}$
`o2_perm`	$0.003_{[.00,.02]}$	$0.282_{[.14,.45]}$	$0.496_{[.37,.62]}$	$0.036_{[.01,.07]}$
`o3_drift`	$0.060_{[.00,.16]}$	$0.520_{[.33,.70]}$	$0.307_{[.20,.43]}$	$0.291_{[.18,.41]}$
`held_out`	$\mathbf{1.000}_{[1.00,1.00]}$	$0.370_{[.21,.55]}$	$0.453_{[.32,.58]}$	$\mathbf{1.000}_{[1.00,1.00]}$

B.7 STRUCTURAL SEPARABILITY

On a validation batch of `o1_exfil`, structural fingerprints separate cleanly by class (Table 11), localized to late layers. The within-minus-between cosine gap of $+0.947$ is sharper than the same check on a control variant in which a phrase shared across organisms leaked the class (gap $+0.777$, between-class cosine 0.214). Removing that leak therefore did not weaken the structural signal. The near-orthogonal between-class cosine (0.030) is also consistent with the target-token confound of Section 4.2.1; this within-organism check alone cannot rule it out, and it establishes only that a structural signal exists.

Table 11: Structural separability, `o1_exfil` validation batch on the covert data ($n = 20$; 10 per class). Cosine similarity; higher within-class and lower between-class is better.

Quantity	Value
benign–benign cosine	0.977
sabotage–sabotage cosine	0.994
benign–sabotage cosine	0.030
within − between gap	$\mathbf{+0.947}$
leave-one-out nearest-centroid accuracy	100% (20/20)
signal-carrying layers	24–32 of 34 (late)

B.8 STRUCTURE ABLATION: ONE HOP OF ATTRIBUTION CARRIES THE SIGNAL

A ladder of representations built from the same graphs isolates what the fingerprint relies on (Table 12). Raw feature magnitude (which features fire, with no attribution) separates in distribution (0.917, against 0.753 for the deployed fingerprint on the same split) and anti-generalizes to the repeated-mechanism organism (0.145). One hop of attribution to the target logit lifts generalization to 1.000, and two, four, and the deployed forty hops are statistically indistinguishable from one. We claim only that attribution is associated with the observed generalization signal and treat depth as a deployment choice. This is partly a ceiling effect, since `held_out` is a weak generalization test (Section 4.2.2) that cannot separate one hop from many; the new-mechanism organism might, and we did not run the ablation on it.

Row normalization and hop depth. Row normalization keeps the entries of Eq. 1 in floating-point range: without it they overflow to NaN within the H hops. With it, each hop is L1 non-expansive and returns relative-contribution flow, a signed variant of the attribution-graph influence

measure of Ameisen et al. (2025). The one knob is the hop depth H. We ran $H = 40$, but one hop already carries the signal and forty add nothing measurable, so nothing in the paper depends on the extra hops and a one-hop monitor would give the same detection. The cost, however, is dominated by building the graph rather than by the hops, so reducing H does not save it; and raw transcoder feature magnitude with no attribution graph at all already reaches 0.917 in distribution. The attribution machinery therefore provides little in-distribution accuracy advantage over cheaper representations at this scale.

Table 12: Structure ablation: detection AUROC (95% bootstrap intervals) for representations built from the same graphs. Attribution to the target lifts generalization from below chance (raw magnitude) to 1.000; depth beyond one hop adds nothing (Appendix B.8).

Representation	In-distribution AUROC	Generalization AUROC
Activation probe (residual stream)	$0.972_{[.95,.99]}$	$0.886_{[.83,.93]}$
Raw feature magnitude (no attribution)	$0.917_{[.81,.99]}$	$0.145_{[.05,.28]}$
1-hop attribution	$0.784_{[.62,.93]}$	$\mathbf{1.000}_{[1.00,1.00]}$
2-hop attribution	$0.759_{[.59,.90]}$	$\mathbf{1.000}_{[1.00,1.00]}$
4-hop attribution	$0.753_{[.58,.89]}$	$\mathbf{1.000}_{[1.00,1.00]}$
Multi-hop (40, deployed)	$0.753_{[.58,.89]}$	$\mathbf{1.000}_{[1.00,1.00]}$

C EVASION TABLES AND GEOMETRY CHECKS

This appendix gives the exact detection AUROC behind Figure 2 at every swept budget, the full per-organism sweeps of the aimed attack, and the geometry checks of Section 4.3.1.

Table 13: Detection AUROC under the transfer attack on the pooled `o1`/`o2`/`held_out` attack set. The structural monitor is the transfer target. Behavior is preserved at every ε; number provenance is in Table 3.

ε (L2 activation nudge)	0	50	100	150
Activation probe (attacked)	0.780	0.686	0.546	**0.228**
Structural monitor (transfer)	0.844	0.845	0.845	**0.847**
Residual-norm baseline	0.380	0.380	0.380	0.380

The transfer attack is evaluated on $n = 60$ sabotage episodes versus $n = 175$ clean benign episodes. At $\varepsilon = 150$, the common-target control remains at 0.867 and the residual-norm baseline at 0.380. The clean values in this table are operating points of the fixed deployed monitors on this attack evaluation set and therefore need not equal the cross-validation or held-out estimates in Table 1.

Table 14: **Aimed attack on Qwen3-4B.** Full sweep of the structural-monitor attack over the signal-band layers 24–33. AUROC_{bp} is evaluated on behavior-preserved sabotage versus clean benign; "kept" is the behavior-preserved fraction and TPR@5%FPR is measured at the benign 95th-percentile threshold.

Organism	ε	AUROC_{bp} [95% CI]	kept	TPR@5%FPR
Pooled	0	0.969 [.95,.99]	100%	88%
	10	0.873 [.83,.92]	100%	58%
	20	**0.559** [.50,.62]	73%	**0%**
	40	0.543 [.46,.60]	35%	0%
`o1_exfil`	0	1.000 [1.0,1.0]	100%	100%
	10	0.978 [.93,1.0]	100%	95%
	20	0.547 [.46,.63]	95%	0%
	40	– ($n = 1$)	5%	0%
`o2_perm`	0	0.908 [.85,.96]	100%	65%
	10	0.729 [.66,.81]	100%	15%
	20	0.609 [.55,.67]	100%	0%
	40	0.570 [.52,.62]	100%	0%
`held_out`	0	1.000 [1.0,1.0]	100%	100%
	10	0.912 [.85,.97]	100%	65%
	20	0.402 [.27,.52]	25% (5/20)	0%
	40	– ($n = 0$)	0% (0/20)	– ($n = 0$)

The attack uses the closed-form direction $\delta = -\varepsilon g/\|g\|$ from the one-hop surrogate, injected jointly across the ten signal-band layers. Every perturbation is then re-attributed through the deployed 40-hop monitor, and only behavior-preserving perturbations are retained. The evaluation uses $n = 20$ sabotage episodes per organism (60 pooled) against the clean benign set ($n = 344$), with a single seed. A same-norm random perturbation at the same layers leaves the pooled monitor at 0.967 for $\varepsilon = 20$, indicating that the degradation depends on the aimed direction rather than perturbation magnitude alone. Rows with fewer than 10 retained episodes carry wide uncertainty and are not interpreted as complete evasion.

Table 15: **Aimed attack on Gemma-3-4B.** Per-organism strongest behavior-preserving attacks and the pooled minimum over the sweep. Clean AUROC is evaluated on the corresponding attack subset; "kept" reports episodes satisfying the teacher-forced behavior-preservation criterion.

Organism	clean AUROC	ε (behavior-preserving)	AUROC_{bp}	kept (teacher-forced)
`o1_exfil`	0.795	1000	0.663 [.59,.73]	20/20
`held_out`	0.795	1932	0.663 [.59,.74]	18/20
`o2_perm`	0.944	1932	0.579 [.51,.65]	7/20
pooled (minimum over the sweep)	0.843	1.3% of band norm	0.639	

For Gemma, the signal band is re-measured, the target budget is matched to the 2.6% of band norm at which the Qwen sweep collapses, and ε is retuned per organism to retain teacher-forced sabotage behavior; that criterion fails above about 1.3% of band norm, so the pooled sweep ends there. Each per-organism row reports the strongest behavior-preserving attack under this criterion, while the pooled row reports the minimum of the Gemma curve in Figure 2. The attack reduces the monitor to 0.58–0.66 across the three organisms.

Clean AUROC is computed using the deployed monitor on the corresponding $n = 20$ attack subset per organism, so these values need not equal the cross-validation estimates in Table 1. In particular, Table 1 reports 1.00 for `o1_exfil`, whereas the attack subset has clean AUROC 0.795. The `o1_exfil` and `held_out` rows share clean and post-attack values because `held_out` paraphrases the same exfiltration mechanism. Cells with fewer than 10 retained episodes carry wide uncertainty. A separate free-generation check is reported in Appendix A.

Attack sets and the threat-model bound. Attacks (i), (ii), and (iv) use the attack set (`o1_exfil`, `o2_perm`, and `held_out` pooled), and the black-box search (iii) adds `o4_destruct` (Ap-

pendix B.1). The threat model bounds the claim. The attacker reads the monitor's weights and the transcoders and writes an L_2-bounded perturbation into the residual stream at the signal-band layers. The result is an upper bound on evasion capability that an input-space or weight-space attacker may not reach (Appendix A).

The aimed attack's one-hop surrogate. One-hop attribution, a feature's activation times a frozen edge, is differentiable in the residual stream. For the active feature set of a fixed clean graph, feature f at layer ℓ has activation

$$a_f = W_{\text{enc}}[\ell, f] \cdot (h_\ell + \delta_\ell) + b_{\text{enc}},$$

with W_{enc} and b_{enc} the transcoder encoder row and bias. Freezing its multi-hop edge as an influence per unit activation c_f then makes the structural score linear in δ. The minimizing step is closed-form, $\delta = -\varepsilon\, g/\|g\|$, with

$$g_\ell = \sum_f w_{\text{struct}}[\ell, f]\, c_f\, W_{\text{enc}}[\ell, f],$$

where $w_{\text{struct}}[\ell, f]$ is the monitor's weight on feature f at layer ℓ.

We inject δ jointly across the signal band, the late layers that carry the class signal (24–33 on Qwen3-4B; re-measured on Gemma). The one-hop surrogate supplies only a direction: every δ is re-attributed through the full forty-hop pipeline, scored with the deployed monitor, and retained only if it preserves behavior. Thus every aimed-attack number in Section 4.3.2 is a re-attributed score. Across 180 perturbed examples, the surrogate's predicted score change tracks the re-attributed change (Spearman $\rho = 0.87$; 99% sign agreement).

A single aimed layer at the probe's budget. The attack in Section 4.3.2 injects across the whole signal band. Restricting the same closed-form attack to the single layer read by the probe (layer 29) at the probe's own budget ($\varepsilon \leq 150$) leaves the deployed structural monitor nearly unchanged on Gemma: behavior-preserved AUROC is $0.795 \to 0.795$ for `o1_exfil`, $0.941 \to 0.932$ for `o2_perm` (a separate fit from Table 15), and $0.795 \to 0.796$ for `held_out` across $\varepsilon \in \{0, 50, 100, 150\}$, even though the one-hop surrogate's predicted score drops by up to 0.94.

The perturbation that reduces the probe from 0.780 to 0.228 at this layer and budget therefore changes structural AUROC by at most 0.01. In these experiments, degrading the structural monitor requires the full multi-layer band at a substantially larger budget; the single-layer attack matched to the probe does not reach it. This result is specific to the tested single-layer attack and does not establish general structural robustness. It is consistent with the transfer controls below: representations that aggregate multiple layers can survive a perturbation aimed at one layer. This resistance comes without a corresponding detection advantage over the activation baseline (Section 4.2.1).

Transfer controls, in the numbers of Section 4.3.1. The transfer result under-tests the structural monitor, and two controls show why (Figure 4). A random perturbation of the same norm leaves the probe itself at 0.781, so the budget is large only along the probe's weight direction. The identical δ leaves an independently fit second probe on the same activations at 1.000. A probe-minimal perturbation therefore collapses the boundary it targets while sparing even a second instance of the same monitor family.

The same pattern holds for a stronger activation monitor. A linear probe over the whole signal band (layers 24–33 concatenated) largely resists the single-layer δ ($0.941 \to 0.873$), while a separately fit single-layer probe on a `held_out` split falls from 0.949 to 0.271. The latter is a separate fit rather than the deployed monitor in Table 13. Resistance to this single-layer perturbation is therefore observed for both multi-layer activation and structural representations and is not, by itself, evidence for the attribution representation.

A 30-query black-box search over the structural score on four organisms, with 29 queries using random directions, likewise fails to move the monitor substantially (0.821; Appendix B.1).

The signal band. The signal band is the set of late layers that carry the class signal. Table 11 reports layers 24–32 on Gemma's validation batch, the Qwen3-4B sweep injects at layers 24–33, and the Gemma re-run re-measures its band.

The injection reaches the graph. As a sanity check that the transferred perturbation reaches the attribution computation, the structural score holds at $\varepsilon = 100$ and moves at $\varepsilon \approx 10^4$, about $65\times$ the largest swept budget. Thus the perturbation reaches the graph, but a substantially larger activation-space intervention is required to shift the structural signature in this check.

Geometry of a probe-minimal attack. Figure 4 plots the checks of Section 4.3.1 against the same clean operating points. The δ fit to the activation probe collapses only the probe it targets. A random perturbation of the same size, an independently fit second probe under the identical δ, the structural monitor under transfer, and the structural monitor under the black-box search all remain near their clean operating points.

The second probe is an L_2 logistic regression fit on an independent $50/50$ split of the same layer-29 activations with a different random seed, so it defines a separate fitted boundary rather than a re-fit of the first; its clean operating point therefore differs slightly from the deployed probe's. The same pattern holds for the tested cheap multi-layer summaries of the residual. Scoring fixed deployed monitors on the clean-benign versus transfer-attacked-sabotage test split, the single-layer δ leaves a per-layer-norm monitor at $0.739 \rightarrow 0.739$, a 50-component residual PCA at $0.932 \rightarrow 0.932$, and the full-residual probe at $0.900 \rightarrow 0.920$.

Across these tested multi-layer representations, a perturbation aimed at one layer's probe boundary produces little or no degradation. The structural monitor's survival under transfer therefore does not by itself show that the attribution graph is hard to evade; the aimed attack of Section 4.3.2 directly tests that question.

A purpose-built multi-layer probe behaves similarly: a linear monitor on the residual concatenated across the signal band (layers 24–33, 25,600 features) holds at $0.941 \rightarrow 0.873$ under the single-layer δ while the single-layer probe falls to 0.271. It is also not stronger on the detection tests reported here: it ties within organism on the distinctive organisms but drops to chance on the boolean flip (0.498) and has lower generalization AUROC (0.570 versus 0.886), plausibly because concatenating the band over-parameterizes the $\sim$450-sample fit. These results support using the most-decodable single-layer probe as the primary activation baseline, while the multi-layer probe serves as a control for the geometry of the transfer attack.

Layer mass is not diagnostic: 67% of the monitor's $|\text{coefficient}|$ mass lies on residual layers 0–28, upstream of the injection, but a shuffled-label monitor places 71% upstream and a random-weight classifier 94%. This reflects the concentration of nonzero fingerprint features upstream rather than label-specific structure, so we do not use layer mass as evidence.

The overall pattern is consistent with a probe-minimal attack being monitor-specific and is silent about whether the underlying computation resists activation-level evasion in general; Section 4.3.2 tests the structural monitor with an attack aimed directly at its score.

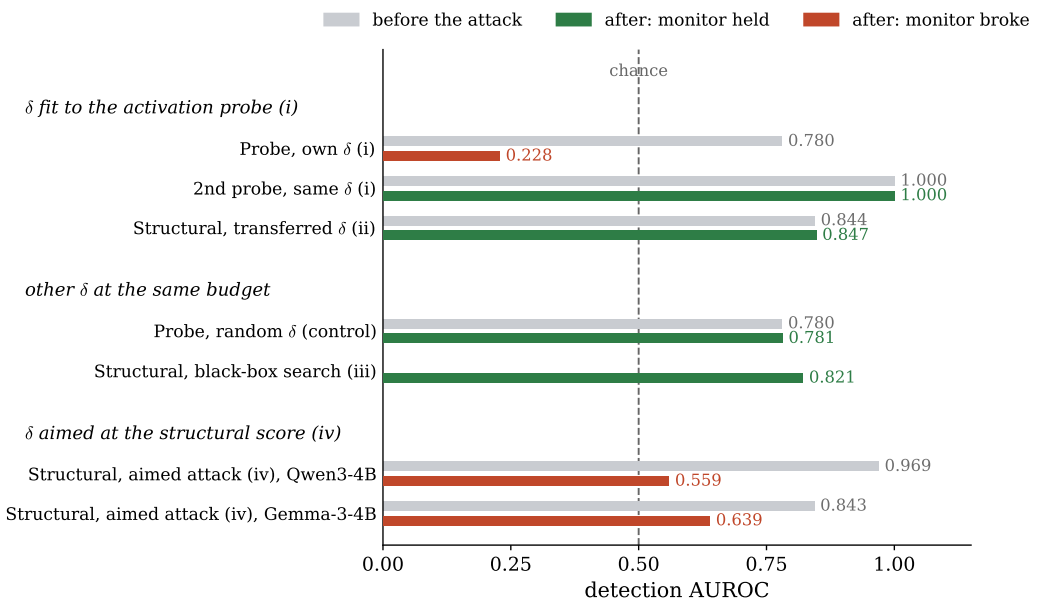

Figure 4: A perturbation collapses only the boundary it was fit to. The δ fit to the activation probe leaves an independently fit second probe and the structural monitor near their clean operating points, as do a random δ of the same norm and the black-box search. In contrast, a δ aimed directly at the structural score (Section 4.3.2) substantially degrades the structural monitor on both tested base models. No "before the attack" bar is shown for the black-box search (iii), because it was run on a different example set that adds `o4_destruct`.

D CONFOUND-CONTROL DETAILS

This appendix gives the cell-by-cell reasoning behind the common-target control summarized in Sections 4.2.1 and 4.2.2.

Why the control matters. Attribution seeds the model's own top-predicted logit. On the clean baseline, benign and sabotage graphs attribute to different target tokens (`internal` versus `external` for `o1_exfil`), so any monitor could in principle key on target-token identity instead of sabotage computation. Re-attributing both classes to the same fixed per-organism target token, through `attribute()`'s native target argument, removes class-dependent target-logit identity as a source of separability.

What the control shows. Forcing a common target collapses pooled in-distribution separability from 0.826 [.75,.89] to 0.420 [.23,.63], at or below chance. A large part of the clean within-distribution signal was target-token identity, so we read the clean matrices as characterization only. A second, independent control confirms this without relying on the specific common anchor: re-attributing both classes to a *random* shared token per organism gives pooled separability 0.646 [.55,.74], an interval entirely below the native 0.826 [.75,.89] yet above chance, so some computation-linked signal survives the target-token control even though it is far weaker than the uncontrolled number. Both controls agree that target-token identity contributes substantially to the uncontrolled clean separability. Figure 5 makes the caveat visual: the structural in-distribution AUROC falls from 0.826 toward chance under both target-token controls, while a benign-only organism-identity classifier and a residual-norm baseline show the cheap cues that inflate the uncontrolled number.

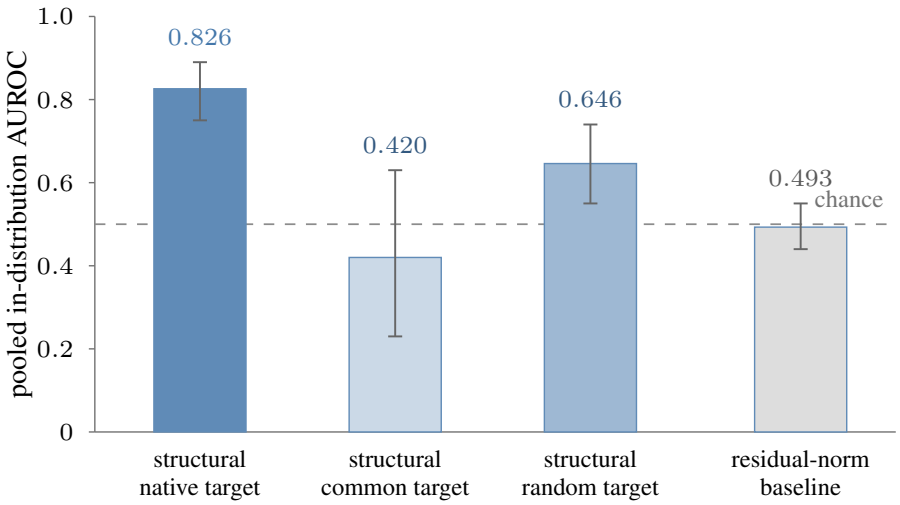

Figure 5: The structural signal partly rides on target-token identity. Pooled in-distribution AUROC for the structural monitor under its native target, a fixed common target (0.420), and a random shared target (0.646); both controls fall well below the native 0.826. An untrained residual-norm baseline (chance) is shown for reference; the organism-identity accuracy (0.83, chance 0.25) is a different metric and is reported in the text.

Controlled transfer matrix. Generalization to `held_out` survives the control unchanged (1.000), though `held_out` is separately a norm artifact (Section 4.2.2), so we do not lean on it regardless. Once we control all four organisms (Table 10), within-o1 and within-`held_out` both survive at 1.000 and transfer perfectly to each other, but the same exfiltration mechanism generates `held_out`, so this is one mechanism confirmed twice. Within-o2 falls below chance with an interval that excludes its uncontrolled value ($0.931 \rightarrow 0.282_{[.14,.45]}$); within-o3 had no within-organism margin to lose ($0.344 \rightarrow 0.307_{[.20,.43]}$). The rank check below shows that both organisms' common target is a degenerate anchor. Because the o2 and o3 anchors are degenerate under the rank check, we treat every controlled cell involving either organism, including their mutual cells and their cross-type cells with o1/`held_out`, as inconclusive. Each is affected by the degenerate anchor and is therefore neither evidence for nor against cross-type transfer. No cell in the fully controlled matrix survives as confound-clean evidence of cross-type transfer; we neither claim it nor rely on it.

Why we could not fix the o2/o3 anchor. We attempted a per-organism verified anchor for o2 and o3 by tokenizing the actual argument value (`"team"` for o2), the approach that worked for o1's `internal` anchor. It did not resolve: the model's top-predicted continuation at this position is not that token for either organism (median rank above 150, probability near 0). A correct anchor therefore needs more than the marker string and is future work. A contrastive target, the sabotage logit minus the benign logit, is the usual alternative in circuit analysis and would avoid choosing one anchor; we did not run it.

Identity-leakage breakdown. With `o4_destruct` included, benign-only identity leakage is 0.840 (chance 0.20), consistent with the four-organism figure of 0.83 (chance 0.25). Leakage is uneven. A per-organism breakdown with a plain multiclass classifier (instead of the one-vs-rest fit behind the pooled number, so this is directional) identifies o1 and o3 perfectly and o2 at 77%. It never correctly identifies `held_out` or o4 as themselves. The classifier confuses both with another organism, so this particular identity classifier provides no evidence that their elevated cross-organism scores arise from uniquely identifying those organisms.

Cheap residual-summary baselines. To check that within-organism detection and the held-out generalization number are not evidence for the attribution representation specifically, we fit the same linear monitor on cheap summaries of the decision-token residual stream (the activation probe's domain): per-layer L_2 norm, per-layer mean, per-layer variance, and a 50-component PCA of the concatenated residual. Table 16 reports within-organism 5-fold AUROC and pooled-to-`held_out` generalization. Several cheap summaries approach the structural monitor's held-out number (mean

0.973, PCA 0.994, versus 1.000), while the full concatenated residual reaches only 0.760. All fall below chance on the boolean flip (o3). This mirrors the norm-baseline and identity confounds above: the raw residual carries substantial within-organism signal in cheap summaries, and `held_out`, a paraphrase of o1, can be generalized to by features that capture o1's mechanism. It does not show the structural *fingerprint* is a magnitude artifact (a different representation); it shows that within-organism detection and the held-out score are not, on their own, evidence for it.

Table 16: Cheap residual-summary baselines (linear monitor on decision-token residual features). Within-organism 5-fold AUROC and pooled-to-`held_out` and pooled-to-`o4_destruct` generalization [95% CI]. These are whole-residual summaries (all layers), distinct from the deployed activation probe of Table 1, which uses the single best layer (29) and detects the boolean flip at 0.796. Several approach the structural monitor on `held_out`, while the full concatenated residual reaches 0.760; all fall below chance on the boolean flip (o3, 0.18–0.20). These are cross-validated scores, so the direction each fit learns on the training folds ranks the held-out fold backwards: a deployed monitor would miss the flip, not detect it inverted. Number provenance: Table 3.

Residual representation	o1_exfil	o2_perm	o3_drift	o4_destruct	Gen. (`held_out`)	Gen. (`o4_destruct`)
Full residual, all layers (concat.)	1.000	0.999	0.186	0.997	0.760 [.68,.83]	0.669 [.58,.75]
L_2 norm per layer	1.000	0.944	0.201	0.921	0.938 [.90,.97]	0.628 [.54,.72]
Mean per layer	1.000	0.999	0.177	0.973	0.973 [.95,.99]	0.777 [.70,.85]
Variance per layer	1.000	0.943	0.196	0.926	0.943 [.91,.97]	0.616 [.52,.71]
PCA (k=50) of residual	1.000	1.000	0.199	0.998	0.994 [.99,1.00]	0.464 [.36,.56]

Evasion confound control. Because the transfer attack of Section 4.3.1 preserves the model's prediction, we can attribute both clean-benign and attacked-sabotage examples to a common fixed target token, so the same control applies directly. Under the common-target control the structural monitor holds at 0.867 at $\varepsilon = 150$ (uncontrolled 0.847; flat at 0.868 across $\varepsilon \in \{0, 50, 100\}$), so holding the behavior-preserved target token fixed does not explain the survival under attack.

Is the `o3_drift` miss hidden in error nodes? The structural fingerprint keeps only feature-node influence; error nodes, the transcoders' reconstruction residual for the LoRA-adapted model, are discarded. Since the transcoders are trained on the base model rather than the adapted one, a natural hypothesis for `o3_drift`'s near-chance detection is that the boolean-flip computation lands in those discarded error nodes. Re-attributing about 30 graphs per organism and splitting each graph's total influence across its feature and error blocks, the error-node fraction $\mathrm{err}/(\mathrm{err}+\mathrm{feat})$ is 0.284 for `o3_drift`, comparable to the organisms the fingerprint detects at ceiling (`o1_exfil` 0.26–0.35, `o2_perm` 0.29–0.30, `held_out` 0.24–0.31) and if anything lower than `o1_exfil`'s. Moreover the detectable organisms route slightly *less* sabotage influence to error nodes than benign (`o1_exfil` 0.261 versus 0.348, a faint class signal in the routing itself), whereas `o3_drift` shows no such gap (0.284 sabotage versus 0.283 benign). These measurements therefore do not support the hypothesis that `o3_drift`'s miss is explained by disproportionate routing of its sabotage influence into discarded error nodes. They do not rule out other representational reasons for the weak structural separability.

E Replication on a second base model (Qwen3-4B)

To test whether the clean-baseline findings are specific to `gemma-3-4b-it`, we re-ran the entire non-adaptive pipeline (build, LoRA-train, verify, fingerprint, monitors) on `Qwen/Qwen3-4B` with an independent, community-trained transcoder set (`mwhanna/qwen3-4b-transcoders`: 36 layers, 163,840 features per layer). The structural fingerprint is therefore $36 \times 163{,}840 = 5{,}898{,}240$-dimensional versus Gemma's $34 \times 16{,}384$. Every organism, trigger, and template is identical to the Gemma run; only the base model and its transcoders change. One base episode (the longest, at the 256-token cap) exceeded the 80 GB attribution budget on Qwen's wider transcoders, and we exclude it uniformly from all organisms, so each has 148 fingerprints and `held_out`/`o4` have 48.

Model choice. We first attempted Llama-3.1-8B-Instruct. Its only instruction-tuned transcoder set (`facebook/crv-8b-instruct-transcoders`) is 131k features wide across 32 layers (34 GB in bf16), and with the 16 GB base model resident, `attribute()` exceeds an 80 GB card

even at 32 tokens. Qwen3-4B traces the full 256-token episode in about 5.6 minutes and 44 GB, so it is the feasible second base model at this scale.

What replicates. Table 17 gives the within-organism benign-vs-sabotage AUROC (5-fold CV). The qualitative pattern is similar on four of the five organisms: exfiltration and `held_out` are at ceiling, permission escalation remains strongly separable at 0.83, and the new-mechanism organism (`o4_destruct`) remains near chance. The exception is `o3_drift`, discussed below. Organism-identity leakage from benign-only fingerprints is 0.844 (chance 0.20), comparable to Gemma's 0.83, so the same identity-leakage caveat applies. Across three seeds the pooled→`held_out` structural generalization is seed-invariant at 1.000, and the probe's best layer is 33 of 37 hidden-state positions in every seed (a late-layer peak, as on Gemma). The tested structural cross-type off-diagonals are at or below chance (o1↔o2 0.45/0.00, o1↔o3 0.12/0.00), so we do not observe positive cross-type transfer in these cells on Qwen. This is consistent with our decision not to interpret the uncontrolled Gemma cross-type cells as clean evidence of structural transfer.

Table 17: Within-organism benign-vs-sabotage AUROC (5-fold CV) on Qwen3-4B versus the Gemma values of Table 1. Four of five organisms show similar qualitative behavior; `o3_drift` is the exception (detectable on Qwen, near chance on Gemma). These values are uncontrolled; the common-target control below reduces the pooled Qwen in-distribution AUROC to chance, as on Gemma.

Base	`o1_exfil`	`o2_perm`	`o3_drift`	`held_out`	`o4_destruct`
Qwen3-4B	1.00	0.83	**0.91**	1.00	0.55
Gemma-3-4B	1.00	0.88	**0.45**	1.00	0.61

The common-target control replicates. The AUROCs above are uncontrolled, and the target-token confound of Appendix D is not specific to Gemma. Re-attributing both classes to a fixed common target per organism (the native `attribution_targets` path, 160 graphs) reduces Qwen's pooled in-distribution AUROC from 0.936 [.89,.97] to 0.549 [.35,.74], at chance, paralleling the reduction on Gemma (0.826 → 0.420). Pooled→`held_out` generalization falls from 1.000 to 0.772 [.60,.92] under the control, above chance but far below the uncontrolled ceiling. On both base models, target-token identity therefore contributes substantially to the structural monitor's uncontrolled in-distribution separability.

Two cross-model differences. First, the boolean-flip detection floor differs across the model–transcoder pairs: `o3_drift` is near chance on Gemma (0.45) and detectable on Qwen (0.908 [.85,.96]). One possible explanation is feature resolution because Qwen's transcoders are ten times wider, but width changes together with the base model. We therefore treat this as an uncontrolled model-plus-transcoder effect and do not attribute the difference to either component.

Second, the `held_out` norm artifact takes a different form on Qwen: an untrained residual-norm baseline reaches 1.000 on Gemma's `held_out` but 0.099 on Qwen's. An AUROC of 0.099 is not chance; it indicates strong separability in the opposite direction (equivalently 0.901 if the ranking direction is reversed), with benign norms systematically higher than sabotage rather than lower. Residual norm therefore remains strongly associated with the label on Qwen, but with the opposite sign. `held_out` is additionally affected by identity leakage and cheap learned representations (a last-layer activation probe reaches 1.000). The conclusion that `held_out` generalization is characterization rather than clean evidence of cross-organism structural generalization is therefore unchanged.

Relatedly, the pooled TF-IDF monitor reaches 0.84–0.96 on `held_out` on Qwen, versus chance on Gemma. We traced this behavior to `held_out`'s decoy phrase sharing words with the shared benign-tag pool while its trigger is out of vocabulary. This surface cue provides an additional reason not to treat `held_out` transfer as clean evidence of mechanism-level generalization.

The transfer attack on Qwen. Under the matched activation-space attack (the probe's evasion direction injected at the decision token, $\varepsilon = 34$), the structural monitor does not measurably degrade (0.969 → 0.975 across $n = 60$ sabotage episodes, with the sabotage behavior preserved on all of

them), paralleling the Gemma transfer result. As discussed in Section 4.3.1, this transfer experiment under-tests a multi-layer structural monitor because the perturbation is optimized against a single-layer probe. We therefore do not interpret its survival as structural robustness. The aimed attack of Section 4.3.2 directly targets the structural score and degrades the monitor on both model–transcoder pairs: the pooled Qwen AUROC approaches chance, while Gemma is substantially degraded but remains above chance (Table 15).