THE MONITORABILITY GEOMETRY OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear activation monitors, whether probes, subspace monitors, or a model's own self-report, read only a few directions of a high-dimensional hidden state, so something always falls in their blind spot. We give a label-free, closed-form way to measure how much control over a model's behavior survives outside any rank-r monitor, from two Jacobians taken at one hidden state, one for what the model does and one for what it says about itself. From their geometry we bound what a hidden intervention can achieve, find the best monitoring subspace, and connect visibility to detection. One geometry governs both sides. A monitor fixed before deployment captures 9% of a model's behavioral leverage at rank 1 and 92% at rank 102 (Llama, Qwen3.5, Gemma, and OLMo; 1B–70B), but rank is not the lever. Against an attacker who accepts 1% visibility, widening from rank 16 to 102 lowers guaranteed evasion only from 11% to 10%. What visibility concedes, no rank buys back. The blind spot is concrete. A steer along the refusal direction a rank-16 activation-PCA monitor cannot see flips refusal on AdvBench across four families while staying invisible ($v \approx 0$), whereas the same steer in the monitor's view barely moves it. The cause is that the variance directions these reads occupy align poorly with what drives behavior. On Gemma-4-12B a rank-16 PCA read captures 0.01% of the refusal direction, below the 0.42% random null, against 59.5% for the matched Jacobian subspace. Spending that visibility is also what catches an attacker. Once the steer is strong enough to flip refusal, a Fisher-based monitor detects it (AUROC 0.88–1.0) where PCA stays blind (0.5). Introspection follows the same geometry in sharper form. 95–100% of the behavior-driving direction lies outside the self-report subspace, so a self-report can be forged, moved to say one thing while behavior stays fixed. A single overlap between the report and behavior directions predicts how faithful that forgery can be, which we confirm causally by fine-tuning models to move it. The same account predicts when a model's report that it is being evaluated comes apart from how it acts. Together these results give a concrete calculus for choosing a monitor's width, threshold, placement, and subspace, and for knowing what it must miss.

1 INTRODUCTION

Activation monitors are often evaluated by whether they can distinguish examples of a behavior, such as refusal or deception, from examples without it. A monitor can succeed at that task while missing a different question: would it notice a change to the hidden state that alters what the model does? Linear probes, subspace monitors, and model self-reports all read limited information from a high-dimensional state. Their usefulness for oversight therefore depends on which directions they read and how those directions relate to changes in behavior (MacDiarmid et al., 2024; Korbak et al., 2025; Bailey et al., 2026). We study this relationship at a fixed layer and token position. Let $h \in \mathbb{R}^d$ be the residual-stream state and let a rank-r monitor observe an r-dimensional projection of it. A perturbation outside that projection is invisible to the monitor, yet it may still change the model's next token. We measure this with the Fisher-weighted behavior Jacobian J_B, the local sensitivity of the output distribution to h. Its spectrum says how much behavioral sensitivity any rank-r read can capture, and how much an unseen intervention can retain; for the optimal subspace the latter is fixed by the $(r{+}1)$-st singular value of J_B.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The model's own account raises the same question. When it answers a fixed introspective query, its concept-report probabilities depend on h through a second Jacobian J_R. The overlap between the two Jacobians determines which behavior-changing directions can affect that answer, and predicts when an intervention changes it while leaving next-token behavior nearly unchanged. This gives external monitoring and self-report a common geometry while keeping their reads distinct, one a projection of the state, the other what the model says about it (Binder et al., 2025; Lindsey, 2026). We test these predictions across model families, layers, behaviors, and monitor widths. High-variance subspaces can be poorly aligned with what drives a behavior: at rank 16 on Gemma-4-12B (Team et al., 2026) an optimal behavior-subspace read captures 59.5% of the refusal direction against 0.01% for activation PCA and a 0.42% null, so PCA is *below chance* there (other families sit within a few times the null). A deployed monitor pays more still, since the sensitive directions vary between states. On the report side, 95–100% of the behavior-maximizing direction lies outside the self-report subspace, yet a constrained rotation raises report visibility while keeping most of the behavioral effect. These results turn monitor width and placement into quantities we can assess before training a detector or observing a particular attack. We evaluate a rank-16 behavior-Fisher monitor on held-out behavior-changing interventions and test the report geometry through concept injection and self-report. Our guarantees concern local state perturbations and rank-limited linear reads; the experiments examine how well those predictions describe finite interventions and practical monitors.

Our core contributions are:

- **A measure of monitor capacity.** From the behavior Jacobian we derive the leverage a rank-r read captures, the optimal behavior-Fisher subspace, and a tight bound on the leverage hidden in its blind spot (§3.4, Thm. 1).
- **A joint account of behavior and self-report.** A second Jacobian for the model's self-report, read at the same state, shows how the overlap between the two governs hidden behavior changes, the visibility a constrained steer buys, and the cost of faking a report (§3.5–3.6, Thm. 2, Cor. 2).
- **Evidence for real blind spots.** A steer in the blind spot of a rank-16 PCA monitor flips refusal on AdvBench across four families while staying invisible, whereas an equal-norm visible steer barely moves it, and activation-PCA aligns with behavior below chance on Gemma-4-12B (§5.1, 5.3). The same geometry predicts a probing failure, an eval-awareness probe reads the format nuisance and *inverts* out of distribution, which we confirm (§5.4).
- **A monitor an effective attacker cannot beat.** A Fisher-based monitor catches the behavior-flipping steer that activation-PCA misses, detecting it at AUROC 0.88–1.0 once it is strong enough to flip refusal, while PCA stays at chance. An attacker forced to hide loses the flip, so it is caught or harmless (Lem. 2, Prop. 1, Fig. 11), which we turn into a recipe for sizing, placing, and pairing a monitor (§6).

2 Related Works

Activation monitoring and monitorability. Linear probes detect sleeper-agent behavior and strategic deception, and sparse autoencoders give interpretable activation features for monitoring (MacDiarmid et al., 2024; Goldowsky-Dill et al., 2025; Lieberum et al., 2024). Recent work treats *monitorability* itself as a property to measure, especially for chain-of-thought (Korbak et al., 2025; Guan et al., 2025; Emmons et al., 2025), while adversarial studies bypass latent-space monitors through obfuscated activations, learned evasion, or nullspace steering (Bailey et al., 2026; Gupta & Jenner, 2025; Pramanik et al., 2026). These establish the utility and fragility of internal monitoring but evaluate particular constructions or attacks. We instead ask a structural question, how much behavioral sensitivity is available to *any* rank-r linear read before a monitor is trained.

Representation subspaces and activation interventions. The logit and tuned lenses decode intermediate states through the output vocabulary (nostalgebraist, 2020; Belrose et al., 2023a), and recent work identifies low-dimensional metacognitive and verbalizable subspaces, including the Jacobian Lens and J-space (Li et al., 2026; Gurnee et al., 2026). Other methods manipulate representation subspaces through linear concept erasure and activation steering (Belrose et al., 2023b; Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023), with Fisher-geometric approaches a related formulation (Wang et al., 2026). Rather than construct another readout or intervention, we characterize the capacity of a rank-limited read itself, deriving both the sensitivity it captures and the sensitivity that must remain outside it.

Introspection and self-report. Models have been shown to make self-predictions and, in some settings, detect or identify information injected into their activations (Binder et al., 2025; Lindsey, 2026). Follow-up work studies the circuits behind these capabilities and how far apparent introspection reflects direct access rather than inference (Macar et al., 2026; Lederman & Mahowald, 2026; Singh et al., 2026), and documents gaps between what is detectable in hidden states and what is verbalized (Yuan et al., 2026). These ask whether a particular signal can be decoded or elicited. We instead treat self-report as a readout of the same hidden state as an external monitor, so the behavior and report Jacobians describe monitoring and introspection in one framework, answering not just whether information is reportable but how much behavioral sensitivity stays hidden from it.

3 METHODOLOGY

A rank-r linear monitor of the residual stream has a structural blind spot. Whatever its r directions do not span, it cannot see. We make that gap exact with two Jacobians read at one hidden state, one for what the model does and one for what it can say about itself, and turn it into closed forms for evasion, deployment loss, steering, forgery, and detection. We evaluate the theory in two settings, a behavior-Fisher attack monitor and an introspective self-report read; the setup is in §4, with full protocols in the appendix (§K, §H) and all proofs in §B.

3.1 TERMINOLOGY AND NOTATION

The model is a fixed autoregressive transformer of residual width d, read at one site: layer L, token position p, giving residual state $h \in \mathbb{R}^d$. *Behavior* is the next-token law $p_\theta(\cdot \mid x)$; *report* is the model's answer $p_\theta(c \mid x, q)$ to a fixed one-word introspective query q over concept labels c. J_B and J_R are the behavior and report Jacobians at h (Defs. 1, 2); $F = J_B^\top J_B$ is the per-site behavioral Fisher at the state h (written $F(h)$ in Def. 1, and $F(x)$ for the state induced by input x when averaging over inputs); here J_B has singular values $\sigma_1 \geq \sigma_2 \geq \cdots$ and F the eigenvalues $\lambda_i = \sigma_i^2$. A rank-r monitor is an orthonormal $U \in \mathbb{R}^{r \times d}$ with projector $P = U^\top U$; it sees a perturbation only through $P\delta$, so any direction splits into a seen part Pg and a blind part $(I - P)g$.

3.2 THE TWO-JACOBIAN CONSTRUCTION

To first order, behavior is a linear map from state perturbations to next-token log-probability changes; weighting it by the output distribution gives the Fisher pullback (Amari, 1998; Schraudolph, 2002; Martens, 2020; Botev et al., 2017).

Definition 1 (Behavior Jacobian)**.** Write $\ell_y(h) = \log p_\theta(y \mid h) = \log \mathrm{softmax}\big(f_{L \to N}(h; x)\big)_y$ for the next-token log-probability as an explicit function of the read state h at site (L, p), with all other positions held fixed. The Fisher-weighted behavior Jacobian $J_B \in \mathbb{R}^{K \times d}$ is any factor of the pulled-back Fisher information,

$$J_B^\top J_B \;=\; F(h) \;=\; \mathbb{E}_{y \sim p_\theta(\cdot \mid h)}\Big[\big(\partial_h \ell_y(h)\big)\big(\partial_h \ell_y(h)\big)^\top\Big] \;=\; A^\top(\mathrm{diag}\, p - pp^\top)A, \tag{1}$$

where $A = \partial z / \partial h$ is the raw logit Jacobian. The first-order response is $J_B\delta$, whose squared norm $\|J_B\delta\|^2 = \delta^\top F \delta$ is the leverage, equal to twice the leading (second-order) KL $\mathrm{KL}\big(p_\theta(\cdot \mid h) \,\|\, p_\theta(\cdot \mid h{+}\delta)\big) = \frac{1}{2}\delta^\top F \delta + O(\|\delta\|^3)$; so J_B carries the behavioral spectrum $(\sigma_i, M_r, \sigma^\star, P_B)$, while the control direction of a named behavior comes from A (§3.5).

Definition 2 (Report Jacobian)**.** Append a fixed introspective query q ("In one word, which concept is most active in you right now?") and read $p_\theta(c_{1:m} \mid x, q, h)$ over concept labels. The report Jacobian is $J_R(h) = \partial_h \log \mathbf{p}_\theta(c_{1:m} \mid x, q, h) \in \mathbb{R}^{m \times d}$, of rank at most $\min(m, d)$ (at most $m - 1$ under the label-renormalized form $J_R = (I - \mathbf{1}\pi^\top)A$ used in App. C); P_R projects onto $\mathrm{rowspace}(J_R)$. Attention is causal, so h on prefix x is the same with or without q, and both Jacobians are read at that one h (Fig. 1).

3.3 THE MONITOR, ITS BLIND SPOT, AND CAPACITY

Definition 3 (Rank-r monitor and visibility)**.** For projector $P = U^\top U$, the visibility of a perturbation is $v(\delta) = \|P\delta\|^2 / \|\delta\|^2 \in [0, 1]$.

Activation-PCA, linear-probe, and representation-engineering monitors (Alain & Bengio, 2016; Zou et al., 2023) all fix such a P; taking $P = P_R$ makes the monitor the model's self-report. For a rank-r monitor P, the captured behavioral leverage is $1 - \beta = \|J_B P\|_F^2 / \|J_B\|_F^2$. A uniformly

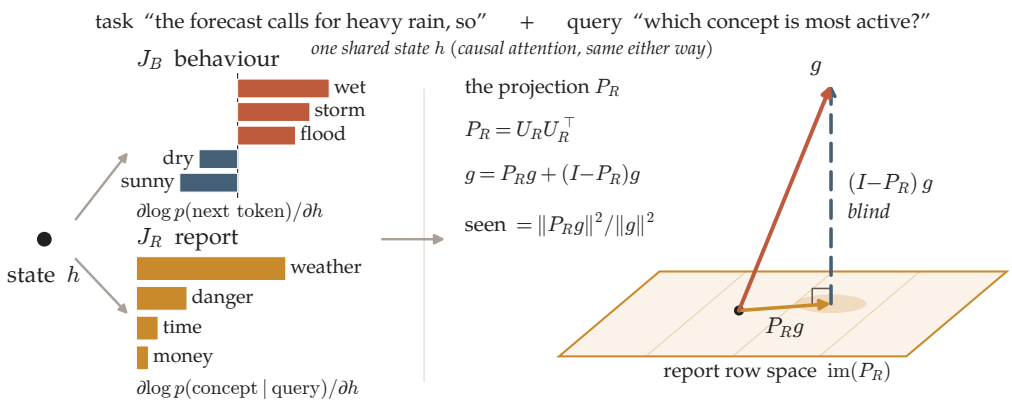

Figure 1: One forward pass, task prefix then introspective query, read at the last prefix token. Both Jacobians are taken at the same h: J_B (next tokens a nudge of h moves most) and J_R (concept labels it moves most). A linear monitor sees a behavior direction g only through $P_R g$; the blind part $(I-P_R)g$ leaves the report unmoved, what a rank-limited self-report read must miss (Defs. 1, 2; Thm. 1).

random rank-r subspace has expected capture r/d, while the optimal rank-r subspace achieves $M_r = \sum_{i \le r} \lambda_i / \sum_i \lambda_i$ ($\lambda_i = \sigma_i^2$), an upper bound on any rank-r monitor. These curves over r form the capacity plane (§D).

3.4 The stealth floor and guaranteed evasion

Definition 4 (Retained-leverage floor)**.** For a monitor P, the retained-leverage ratio $\kappa(P) = \|J_B(I-P)\|_2 / \|J_B\|_2 \in [0,1]$ measures how much behavioral leverage can sit entirely in the blind spot. Its minimum over rank-r projectors is the stealth floor $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$, attained by the top-$r$ singular subspace (Eckart–Young–Mirsky / Courant–Fischer (Eckart & Young, 1936; Horn & Johnson, 2012)); there $\kappa^2 = \sigma_{r+1}^2/\sigma_1^2$ is the blind leverage as a fraction of the maximum single-direction leverage σ_1^2.

The measured J_B spectrum has a steep head but a heavy, power-law tail rather than a sharp cutoff (§5.1), so $\sigma^\star(r)$ declines only polynomially and the blind spot never closes at any practical width; low rank alone would not give this, since a low-rank spectrum can drop steeply before its cutoff. Three fractions recur and should not be conflated: visibility v (the share of an edit the monitor sees), the stealth floor $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ (the smallest $\kappa(P)$ over rank-r monitors), and capacity M_r (the fraction of Fisher energy a read captures). Theorem 1 makes two things precise: the Fisher subspace is the hardest rank-r monitor to evade, and tolerating some visibility does not remove the floor. The $v = 0$ floor is Eckart–Young–Mirsky. What is new is the frontier. One subspace, the top-r Fisher eigenspace, is optimal at every tolerated visibility v at once (Theorem 1), so the monitor need not know the attacker's budget, and the *measured* spectrum makes the floor bind at every practical width, computable label-free before any monitor or attack exists. A second Jacobian then adds new objects: the report-visibility φ, the report–behavior overlap φ', and the concept introspectability V_c are cross-Jacobian invariants that neither the single-Jacobian lens line (Gurnee et al., 2026; Li et al., 2026) nor the Fisher-pullback line (Wang et al., 2026; Martino & Tschiatschek, 2026) defines, and they drive the forgery law and transparency dividend below (§3.5–3.6).

Theorem 1 (Blind spot and guaranteed evasion)**.** *Among all rank-r monitors, the maximum normalized behavioral leverage $\kappa(P)$ available entirely in the blind subspace is the smallest for the top-r Fisher subspace, where it equals $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$. Against this optimal monitor, an attacker who keeps at most a fraction v of its unit update inside the monitored subspace ($\|P\delta\|^2 \le v$) still retains behavioral leverage*

$$\sigma^\star(r, v) = \sqrt{v + (1-v)\,\sigma^\star(r)^2} \;\ge\; \sqrt{v}. \tag{2}$$

For any tolerated visibility $v > 0$, the maximum available behavioral leverage is at least $\sqrt{v}$ whatever the width r, so widening the monitor cannot buy back the leverage the attacker keeps in view, though it still lowers the invisible floor. The value $\sigma^\star(r, v)$ is attained *by an explicit edit (§B), so*

it is the exact optimum, the evasion frontier rather than a lower bound; the inequality $\geq \sqrt{v}$ is its immediate corollary.

A deployed monitor fixes one subspace for all inputs, so its guarantees are quoted on the population Fisher $\bar{F} = \mathbb{E}_x[F(x)/\operatorname{tr} F(x)]$ with deployable capacity $\bar{M}_r = M_r(\bar{F}) \leq \mathbb{E}_x M_r(F(x))$; misaligned site subspaces make $\operatorname{srank}(\bar{F})$ exceed the harmonic mean of the site stable ranks, loosening the worst-case guaranteed floor $\sqrt{\operatorname{srank}(\bar{F})/r}$ (Lemma 1, App. B).

3.5 Steering: the transparency dividend

We score a behavior by $b(h) = \log p(\text{comply} \mid h) - \log p(\text{refuse} \mid h)$, with control direction $g = g_B = \nabla_h b = A^\top c$, $c = e_{\text{comply}} - e_{\text{refuse}}$; the $-p$ terms cancel, so g needs only A, its visibility is $\varphi = \|P_R \hat{g}\|^2$, and its bandwidth is $r_{90} = \min\{r : \sum_{i \leq r} (v_i \cdot \hat{g})^2 \geq 0.9\}$ (v_i the right singular vectors of J_B). Steering adds a unit direction $h' = h + \alpha u$ (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023); the first-order effect $\propto \langle g, u \rangle$ is largest at $u^\star = \hat{g}$, whose visibility $\varphi = \sin^2 \theta^\star$ is small. Rotating $u^\star$ toward the row-space by ψ, with e the in-plane unit vector ($e \in \Pi = \operatorname{span}(u^\star, P_R u^\star)$, $e \perp u^\star$, $\langle e, P_R u^\star \rangle > 0$), gives

$$u(\psi) = \cos\psi\, u^\star + \sin\psi\, e, \qquad \eta(\psi) = \cos\psi, \qquad v(\psi) = \sin^2(\theta^\star + \psi). \tag{3}$$

Here $\hat{g} = g/\|g\|$ is the unit control direction and the perturbation norm is held fixed, so $\eta(\psi) = \cos\psi$ is the *normalized* first-order behavioral effect (the fraction of the maximum-effect steer retained) and $v(\psi)$ its visibility; the experiments use this same $\hat{g}$ at a fixed scale $\varepsilon\|h\|$.

Theorem 2 (Transparency dividend)**.** *At the maximum-effect direction the effect is stationary while visibility is not: $d\eta/d\psi|_0 = 0$ (2nd order) and $dv/d\psi|_0 = \sin 2\theta^\star = 2\sqrt{\varphi(1-\varphi)}$, which is strictly positive and first-order for $0 < \varphi < 1$ (at $\varphi = 0$ both effect and visibility are second-order, so the dividend is a leading-order small-φ statement). A small concession of effect buys a first-order gain in visibility ($\propto \sqrt{\varphi}$); the rotation is moreover the constrained optimum, visibility $\sin^2(\theta^\star + \arccos\eta_0)$ among all unit u with $\langle \hat{g}, u \rangle = \eta_0 > \sqrt{\varphi}$, and, when $\operatorname{rank} J_R \geq 2$, full visibility is free for $\eta_0 \leq \sqrt{\varphi}$ (a rank-1 report row space reaches $v = 1$ only at $\eta_0 = \pm\sqrt{\varphi}$).*

Two knobs realize this. The behavior knob $(I - P_R)\hat{g}$ moves behavior and is report-blind ($v = 0$). The rank-1 *scalar-neutral visibility knob* $(I - P_{\hat{g}})\hat{g}_R$, with $\hat{g}_R$ the unit report-flip direction of §3.6, moves the report while holding the scalar effect $\langle g, \cdot \rangle$ fixed to first order. Because the behavior knob does not move the report and the visibility knob does not move the behavior scalar, the two act independently, so behavior and report visibility can be set from one (A, J_R) probe (independent action, not vector orthogonality); φ, r_{90}, and κ say which knob a behavior needs (§E). The rank-r *distribution-neutral forgery direction* $(I - P_B)\hat{g}_R$ of §3.6 is a different object, neutral to the whole top-r behavior subspace rather than to one scalar, with behavioral response confined to the spectral tail $\|J_B u_f\| \leq \sigma_{r+1}$. Corollary 2 and the LoRA experiment use it.

3.6 Forgery: the blind spot is two-sided

The same geometry moves the report while holding behavior fixed. Let $g_R = J_R^\top(e_b - e_a)$ flip the report from label a to b, let $\varphi' = \|P_B \hat{g}_R\|^2$ be its overlap with the top-r behavior subspace P_B, and let the rank-r *distribution-neutral forgery direction* $(I - P_B)\hat{g}_R$ be the forged direction.

A behavior-neutral report flip along the forged direction $u_f = (I - P_B)\hat{g}_R$ then keeps a fraction $\sqrt{1 - \varphi'}$ of the report effect while its behavioral cost is only tail-order, reduced by the factor $(\sigma_{r+1}/\sigma_1)^2$ relative to an equal-norm maximally-leveraged steer (Corollary 2, proved in App. B). So for small φ' a convincing self-report is cheap to fake, while for large φ' the report is load-bearing and forgery is expensive.

The blind spot is thus two-sided. Here φ' says in advance which models are forgeable, at a price both Jacobians read from one SVD.

3.7 Coverage, introspectability, and detection

For a bank of behaviors with unit control directions $g_1, \ldots, g_K$ and Gram $G = \sum_k g_k g_k^\top$, a rank-r read obeys $\sum_k \|P g_k\|^2 = \operatorname{tr}(PG) \leq \sum_{i \leq r} \mu_i$, so it is transparent to degree c about at most r/c orthogonal behaviors (Theorem 3, App. B). Pointed at concepts, the same budget gives introspectability. A concept g_c is introspectable to the extent $V_c = \|P_R g_c\|^2$ that it lies in the report subspace, which for most concepts sits near the chance floor r/d unless the read is rotated

to expose it. As a detector for a mean shift δ read through a monitor $m = Wh$ under hidden-state noise $h \sim \mathcal{N}(\mu, \Sigma)$, the matched filter has deflection $d'^2 = (W\delta)^\top (W\Sigma W^\top)^{-1}(W\delta)$. In whitened coordinates $z = \Sigma^{-1/2}h$, with Mahalanobis budget $\rho^2 = \delta^\top \Sigma^{-1}\delta$ and *whitened* visibility $v_w = \|P_S \Sigma^{-1/2}\delta\|^2/\rho^2$ (the share of the whitened shift inside the monitor's whitened image S), this is $d'^2 = v_w\rho^2$ (Lemma 2). The whitened v_w equals the geometric visibility v of Def. 3 only under isotropic monitored noise, giving $d'^2 = v\rho^2$; in general $v_w \neq v$. The rank-r monitor optimal against the top-r budget-efficient attacks (highest leverage per Mahalanobis budget) is the top-r eigenspace of the whitened Fisher $\tilde{F} = \Sigma^{1/2}F\Sigma^{1/2}$ (Prop. 1, App. B) (Neyman & Pearson, 1933; Hanley & McNeil, 1982), so the evasion budget is also a detection rate.

Keeping behavior therefore forces detectability: retaining leverage $\kappa \geq \sigma^\star(r)$ against the top-r Fisher monitor forces visibility $v \geq (\kappa^2 - \sigma^{\star 2})/(1 - \sigma^{\star 2})$, which the matched filter catches at $d' \geq \rho\sqrt{v}$ under isotropic monitored noise (Corollary 3, App. B); in general the detection-optimal read is the whitened-Fisher subspace (Prop. 1).

Behavior (J_B), self-report (J_R), and an external decoder read one h; introspection means J_R aligned with J_B, casting the concept-injection paradigm of Lindsey (2026) and the self-report study of Binder et al. (2025) as statements about V_c (§3.7). Detection is magnitude while identification is direction. Because the two knobs make behavior change and reportability independent (Thm. 2), a concept is exposed by acting on the intervention rather than on the model, as in Macar et al. (2026); both follow from one (A, J_R) probe (§H).

4 Experimental setup

We evaluate Qwen3.5, Llama, Gemma, and OLMo models (Team, 2026; Grattafiori et al., 2024; Team et al., 2025; 2026) using rank-matched monitors, state interventions, and evaluation-awareness controls (Devbunova, 2026). Full model, dataset, baseline, and evaluation protocols appear in Appendix K.

5 Experiments and Results

Our experiments cover two settings, an external monitor that watches a model's behavior and the model's own self-report, and the same geometry governs both. What changes behavior lives in a thin subspace nearly invisible to any fixed read of the state, whether an activation monitor, a linear probe, or the model reporting on itself. We measure this blind spot (§5.1), draw out its consequences for anyone who edits the state (§5.2), turn it into a deployable test where keeping behavior forces detectability (§5.3), and show the same geometry in eval-awareness probing, introspective self-report (Lindsey, 2026), and every probe family we try (§5.4). Throughout, "behavior" means the model's next-token distribution, and an edit's *leverage* $\delta^\top F\delta$ is the quadratic change it induces there (the KL is $\frac{1}{2}\delta^\top F\delta + O(\|\delta\|^3)$), so every claim is measured on the output distribution, not on graded text. A monitor's *capture* M_r is the fraction of that leverage its r directions span.

5.1 The blind spot

A rank-r monitor reads r directions of the state. The question is whether those are the directions that carry behavior, and they are not. Figure 7 draws the geometry to scale. The behavior Jacobian is thin and lies almost orthogonal to the bulk of the activation covariance, so the directions that most change what the model does are nearly invisible to a monitor built from activation variance, or to an ordinary linear probe. At rank 16 on Gemma-4-12B, a behavior-subspace read captures 59.5% of the refusal direction (the overlap $C(16) = \|P_B\hat{g}\|^2$, distinct from the energy fraction $M_r > 0.9$), while a matched activation-PCA read captures 0.01% of it. The baseline is the random-subspace null $r/d = 0.42\%$. On Gemma-4-12B activation PCA sits about $40\times$ *below* that null, markedly worse than chance at spanning behavior. On the other models it sits within 2–$4\times$ of the null and far below the Fisher read (up to $100\times$, least on OLMo-3-7B, whose behavior is not low-rank so even the Fisher read captures only 1.2%). The effect survives centering, whitening, and rogue-dimension removal, ruling out a massive-activation artifact (Sun et al., 2024; Timkey & van Schijndel, 2021) (Figure 16). The J_B spectrum is steep (per-site stable rank ≈ 2.3), so the retained-leverage floor $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ falls off only as a power law $\sigma^\star(r) \propto r^{-\alpha/2}$, set by the exact full-vocabulary Fisher eigenvalue tail $\lambda_i \propto i^{-\alpha}$ with exponent $\alpha \approx 1.1$ (a spectral property of J_B and not a K-sampling artifact; Figure 13). It stays above zero out to the largest model ($\sigma^\star(16) = 0.056$ per site, 0.046 on the population). The steep head partly reflects a peaked next-token distribution, but the heavy tail and the below-chance PCA alignment are properties of the logit Jacobian A. This is the

empirical content of Theorem 1: behavior lives away from the variance directions, and a fixed-width monitor must miss a share of it. How wide a monitor must be varies with depth and model scale (Figure 4). The captured ceiling M_r rises with width at every layer, but its depth profile differs across families, and the effective number of behavior directions falls as models grow (the 27–70B models are read in 4-bit, and the Fisher monitor and baselines agree within 0.02. Across the behavior bank the width r_{90} that captures 90% of a behavior's leverage ranges from ≈ 3 (self-awareness) to ≈ 190 (coordination); the per-model, per-layer breakdown is in Appendix O.1. A monitor fixed before deployment captures 9% of behavioral leverage at rank 1 and 92% at rank 102, well below an input-adaptive read of the same width (Figure 8). A monitor should therefore be sized and placed per model, not fixed at one rank and one layer. The same floor also acts as a coverage budget (Theorem 3). A width-r monitor fully covers at most r mutually orthogonal behaviors, so the rest stay open as evasion routes, and widening the monitor shrinks but never closes a small central blind pupil (Figure 12).

5.2 Steering, the transparency dividend, and forgery

The same geometry has a direct consequence for anyone who edits the state. At the maximum-effect steer, the behavioral effect is flat in the direction of the edit while its visibility grows to first order, so giving up a little effect buys most of the available visibility (Figure 2).

Keeping 99, 95, and 90% of the effect already raises absolute visibility from 0.05 to 0.13, 0.27, and 0.39 (a 2.6, 5.3, 7.8× multiplier), largest where the starting overlap φ is smallest (Theorem 2). Visibility costs little because effect is second-order where visibility is first-order. The report side of the same statement is a forgery risk. A behavior-neutral direction moves a self-report without moving behavior, so the report can be pushed to a chosen target at a cost only tail-order in behavior (under 1% of the effect), with retained report fidelity $\sqrt{1-\varphi'}$. In low-overlap models an activation edit forges the report at about 96.5% fidelity ($\varphi' = 0.07$) at near-zero cost, while Gemma, whose natural behavior-report has the highest overlap here ($\varphi' = 0.37$), retains more than the first-order value, about 90% against the predicted 79% (Figure 2d). This is the natural overlap; the causal test of §5.4 drives a different target report and confirms the $\sqrt{1-\varphi'}$ law across the full swept range, on Gemma from $\varphi' = 0.13$ to 0.98 (Table 6).

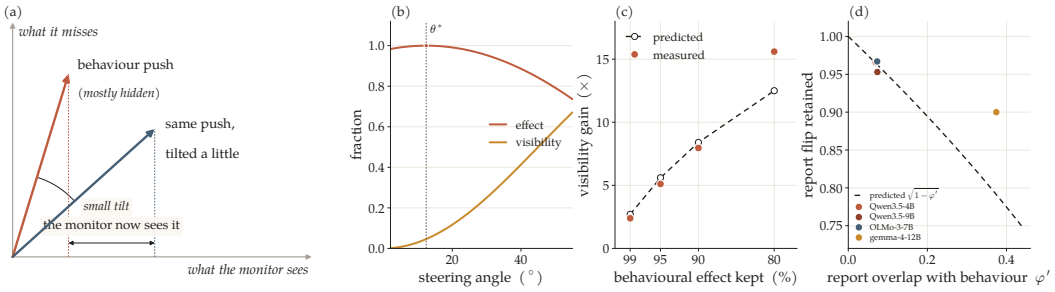

Figure 2: **Why visibility costs little, and the report side.** (a) Near the max-effect steer the behavioral effect is flat while visibility rises linearly, so a little effect buys a lot of visibility. (b) Effect and visibility against steering angle, effect stationary at $\theta^\star$. (c) Measured visibility gain against effect kept matches prediction (2.6/5.3/7.8× at 99/95/90%), and at the 80% point the measured 15.6× exceeds the predicted 11.7×. (d) The report side (Theorem 2), retained report-flip fidelity against overlap φ', with low-overlap models (Qwen3.5-4B, Qwen3.5-9B, OLMo-3-7B) forging at $\approx 96.5\%$ along $\sqrt{1-\varphi'}$ and Gemma-4-12B the highest natural-overlap point ($\varphi' = 0.37$), over-retaining slightly. The causal sweep (Table 6) drives a different target report and confirms the law on all four models.

5.3 Keeping behavior forces detection

Evasion exploits directions a monitor does not read, allowing behavior to change while the monitor stays quiet (Bailey et al., 2026). Theorem 1 quantifies this tradeoff. Under full invisibility, attacker leverage is capped at $\sigma^\star(16) = 0.056$; retaining more requires visibility along $\sqrt{v + (1-v)\sigma^\star(r)^2}$. On the population Fisher, the $v = 0.01$ frontier drops only from 11% at $r = 16$ to 10% at $r = 102$, near the irreducible floor $\sqrt{v} = 10\%$. Extra rank therefore cannot recover leverage once visibility is conceded, and that visibility becomes detectable.

A real refusal flip lives in the blind spot. We steer along the component of the difference-in-means refusal axis g_{DM} outside a rank-16 activation-PCA monitor, $(I - P_{\mathrm{PCA}})g_{\mathrm{DM}}$. Because PCA misses behavior, this retains nearly the full behavioral effect at $v \approx 0$. Across four model families and 120 AdvBench prompts per model, the blind steer crosses the refusal boundary at low norm, from $0.15\|h\|$ for Qwen3.5-9B to at most $0.6\|h\|$ for all four, with margin shifts of $+14$ to $+20$. Equal-norm visible steers do not flip refusal through $0.6\|h\|$, except Gemma at the largest budget (Figure 14). Thus crossing the decision boundary can be cheap and invisible, while sustaining the flip requires norm in directions a magnitude monitor can see (App. G). Figure 5 shows the same tradeoff analytically. Guaranteed evasion shrinks with rank only at $v = 0$; once visibility is conceded, a floor remains. Detection then improves with visible edit magnitude: at $\rho = 4$, matched-filter AUROC is 0.81 at $v = 0.1$, 0.94 at $v = 0.3$, and 0.976 at $\kappa = 0.7$. The closed-form relation $d' = \rho\sqrt{v}$ assumes isotropic monitored noise, so Figure 11 provides held-out and adaptive-attacker validation. The same pattern holds for deployable monitors. On a held-out attack with $\varepsilon = 0.2\|h\|$, a rank-16 behavior-Fisher subspace-anomaly detector reaches AUROC 0.87–1.00 without attack labels. Same-rank activation-PCA remains at chance, while supervised linear and MLP probes trained on attacks fall to chance or below on the unseen attack (Table 3). The key contribution is therefore the monitored subspace, not a new detector class (Alain & Bengio, 2016; Lee et al., 2018). This also holds for the refusal-flipping steer itself. Activation-PCA remains at 0.5 AUROC, while a rank-16 whitened-Fisher monitor rises to 0.88 on Llama-3.1-8B and 1.0 on Qwen3.5-9B at the flip norm; full-rank Mahalanobis detects it at every norm (Figure 14). Although the flip direction has tiny local Fisher leverage (< 0.004), whitening exposes it because it lies in a low-variance direction. A monitor-aware attacker can preserve the flip while hiding from PCA, but suppressing the whitened-Fisher signal collapses the flip from 1.0 to 0.025 on Llama and 0.47 on Qwen (Figure 11). Thus attacks that preserve behavior become detectable, while attacks that evade detection lose behavioral effect. Proposition 1 formalizes the optimal read as the top-r generalized eigenspace of (F, Σ^{-1}), which weights behavioral leverage by noise variance.

5.4 The blind spot beyond attacks: eval-awareness, introspection, and probe families

Eval-awareness: the blind spot predicts a probing failure. When the target signal is not the highest-variance direction, a variance-riding probe tracks nuisance and fails out of distribution. Eval-awareness provides a clean test: benchmark-versus-casual *format* is high-variance, while eval-versus-deploy *intent* is not. Under a format transplant, a standard probe *inverts*, false-flagging 96–98% of deployment prompts (transplant AUROC 0.00–0.004, Table 1), reproducing a known finding (Devbunova, 2026). Erasing the top variance directions instead, following the variance-is-not-signal principle, preserves prediction across the transplant (AUROC 0.92 at the validation-selected layer), confirming that the failure rides variance, not signal.

Introspection. The report Jacobian J_R captures whether a model can notice and name its internal state (Lindsey, 2026). Injected-concept detection rises along the report-visible direction but stays flat along the blind direction (Figure 3). A matched-norm random concept raises detection comparably, yet only the true concept is named (Figure 17), separating detection (magnitude) from identification (direction), consistent with the two-Jacobian account. Reportable overlap $V_c = \|P_R g_c\|^2$ is 4–6$\times$ above the r/d chance floor on Qwen but approaches it on the more curved Gemma, tracking self-report strength (per-model Spearman $r_s = 0.33$–0.78). Meanwhile, 95–100% of the behavior-maximizing direction $\hat{g}_B$ lies outside the report subspace ($\|P_R \hat{g}_B\|^2 \leq 0.05$). This near-total exclusion is close to the dimensional expectation (report rank $\leq m - 1$, $d \approx 4000$), making it a first-order result and the introspective face of forgery risk (Section 5.2). We test this causally. Small report–behavior overlap $\varphi' = \|P_B \hat{g}_R\|^2$ permits behavior-neutral report rewriting, with *forgery fidelity* (finite-norm self-report flip) predicted by $\sqrt{1 - \varphi'}$. A small LoRA trained to increase φ' under a KL penalty makes forgery harder across four models in three families (Qwen3.5-2B/9B, Llama-3.1-8B, Gemma-4-12B): as φ' spans ≈ 0.03–0.98, fidelity falls from near 1 to 0.14 (Gemma). Because fidelity measures a finite-norm flip rather than the projection read back, this agreement is a test, not an identity. Forgeability is thus a controllable geometric quantity (Corollary 2 as an intervention). Yet this does not buy a cheap monitor. Increasing φ' substantially reshapes report–behavior geometry (J_B singular-spectrum shifts up to 4271%) while preserving output capability (drift $\mathrm{KL} \leq 0.013$). Still, report-only detection of a held-out evasion attack reaches only 0.86 on Gemma and stays near chance otherwise. Self-report can be rewritten at little behavioral cost without becoming a reliable behavior monitor (Table 6).

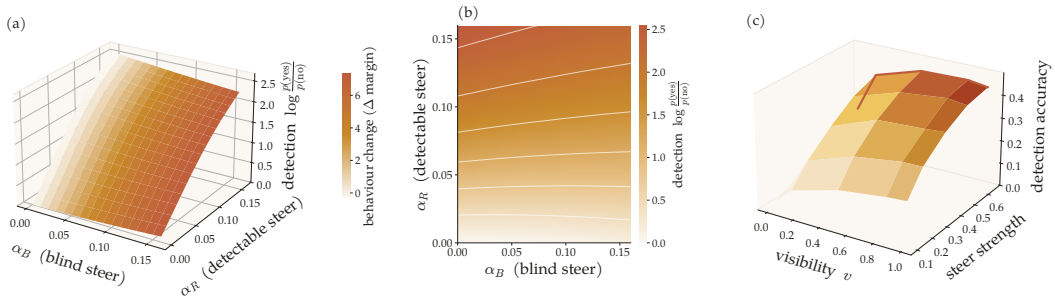

Figure 3: **What the model will admit it notices.** (a) Detection log-odds $\log \frac{p(\text{yes})}{p(\text{no})}$ that the model reports an injected concept, over the blind-steer axis α_B and detectable-steer axis α_R, colored by behavior-margin change. (b) From above, the log-odds climb along the report-visible steer α_R and stay flat along the blind steer α_B, so what the model admits is set by whether the injection lies in its report subspace. (c) Detection accuracy over report-visibility and steer strength, averaged over the four models with reportable yes/no tokens (Qwen3.5-9B, Qwen3.5-27B, Llama-3.1-8B, Llama-3.3-70B, Gemma excluded). Consistent with Lindsey (2026).

Probe families. The limit persists across probe families. Figure 6 fits linear reads (Fisher, PLS, activation-PCA, self-report, random) and nonlinear reads (MLP, random-Fourier kernel) to distinguish behavior-positive from behavior-negative states, measuring captured behavioral leverage against rank. Every family selects the class-separating content direction, nearly orthogonal to the behavior gradient, capturing similarly little leverage. Nonlinearity cannot help: rank-r leverage depends on the spanned directions, not boundary curvature within them. The critical choice is the subspace, not the probe class. Leverage and finite-norm steering remain distinct. Figure 6 measures local leverage (§3.4), explaining why the difference-of-means direction carries little leverage yet flips refusal at finite norm. Every rank-16 read misses this weak steer, though a whitened-Fisher or edit-norm monitor detects it at flip strength (§5.3). What a monitor cannot see for free is precisely the direction that changes behavior. Finally, the same probe sizes a monitor before a threat is deployed. In the base model's singular basis, a narrow finetune requires monitor width $r_{90} \approx 87$ (median), above normal behavior's functional width (38) but below the random null (173); a monitor sized for normal use is too narrow (App. O.6).

6 Discussion and Conclusion

Our central object is a method built from two Jacobians read at one hidden state, J_B for behavior and J_R for self-report. From them a single label-free probe reads the capacity, stealth floor, transparency dividend, forgery fidelity, and detection rate of any rank-r monitor before a detector is trained or an attack is seen. It gives one geometric reason why oversight tools, whether linear probes, sparse autoencoders, representation-engineering monitors, or a model's own self-report, can score well on a detection benchmark yet be evaded. Each reads a low-rank projection of the residual stream, but what changes behavior lives in a thin Fisher subspace nearly orthogonal to the high-variance directions those reads occupy. A rank-r read therefore leaves a fraction $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ of full leverage unseen, a floor that falls only polynomially ($\sigma^\star(16) = 0.056$) and that no probe family or nonlinearity we test closes. Self-report is the same read turned inward, and an activation edit can forge it along a behavior-neutral direction at fidelity $\sqrt{1-\varphi'}$, so it is not a trustworthy basis for oversight. Empirically, across four model families the predictions hold. Behavior concentrates in a thin subspace that variance-based reads miss at or below chance, a fixed-width monitor leaves a floor of leverage it cannot close, visibility is cheap to buy on both the behavior and report sides, and keeping behavior forces detectability. The upshot is constructive. From one (J_B, J_R) probe a team can size a monitor to the width a behavior needs ($r_{90} \approx 3$ to 190), place it where capacity peaks, read it in the whitened-Fisher subspace rather than the top-variance one, and pair it with an edit-norm monitor, so an attacker that keeps the behavior is caught and one that hides becomes harmless. Under the stated first-order, single-site assumptions these are exact guarantees; deploying them is engineering, since real edits are finite and span several token positions. Internal monitors remain useful, but their blind spots are structural, predictable, and sizable, and the same geometry that prices them shows how to make evasion costly.

AI USE STATEMENT

We used generative AI tools to formulate mathematical claims and to support, write, and refine proofs, to implement methods through code generation and debugging, to create or modify figures and artifacts, and to suggest titles and keywords. Other required-disclosure tasks were not applicable to this work.

All AI-assisted outputs were reviewed and validated by the authors, with mathematical claims, proofs, code, figures, and artifacts checked for correctness and consistency with the reported methodology and results, and title and keyword suggestions used only as editorial assistance. We take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the limits of internal monitoring for oversight. The refusal-flip demonstration uses public jailbreak benchmarks (AdvBench, HarmBench, StrongREJECT) and public open-weight models, and we report no new harmful capability; the invisible-steer construction is a diagnostic of monitor blind spots, and our contribution is a defender-side calculus for sizing, placing, and pairing monitors so that evasion becomes detectable. No human subjects or private data are involved. We release only analysis code and derived quantities, not attack artifacts.

REPRODUCIBILITY STATEMENT

We release the analysis code and all derived quantities (the two Jacobians, their spectra, and the capacity, stealth-floor, transparency-dividend, forgery, and detection measurements) that reproduce every table and figure; all models and benchmarks used are public and specified in Appendix K.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

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*, volume 2026, pp. 146838–146882, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023a.

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

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about

themselves by introspection. In *International Conference on Learning Representations*, volume 2025, pp. 3710–3756, 2025.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pp. 557–565. PMLR, 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Viliana Devbunova. Is evaluation awareness just format sensitivity? limitations of probe-based evidence under controlled prompt structure. *arXiv preprint arXiv:2603.19426*, 2026.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Scott Emmons, Roland S Zimmermann, David K Elson, and Rohin Shah. A pragmatic way to measure chain-of-thought monitorability. *arXiv preprint arXiv:2510.23966*, 2025.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i*. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949. doi: 10.1073/pnas.35.11.652. URL https://www.pnas.org/doi/abs/10.1073/pnas.35.11.652.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Melody Y Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian D Kivlichan, Amelia Glaese, et al. Monitoring monitorability. In *Forty-third International Conference on Machine Learning*, 2025.

Rohan Gupta and Erik Jenner. Rl-obfuscation: Can language models learn to evade latent-space monitors? *arXiv preprint arXiv:2506.14261*, 2025.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Matthew T. Kaufman, Mark M. Churchland, Stephen I. Ryu, and Krishna V. Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3): 440–448, 2014. doi: 10.1038/nn.3643.

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

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms, 2024. *URL https://arxiv. org/abs/2407.04694*.

Harvey Lederman and Kyle Mahowald. Dissociating direct access from inference in ai introspection. *arXiv e-prints*, pp. arXiv–2603, 2026.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

Ji-An Li, Huadong Xiong, Robert Wilson, Marcelo G Mattar, and Marcus K Benna. Language models are capable of metacognitive monitoring and control of their internal activations. *Advances in Neural Information Processing Systems*, 38:60073–60108, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

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 the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness. *arXiv preprint arXiv:2603.21396*, 2026.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, et al. Simple probes can catch sleeper agents. *Anthropic Research Updates*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Gabriele Martino and Sebastian Tschiatschek. Fringe: Distribution-space integrated gradients with fisher–rao geometry, 2026. URL https://arxiv.org/abs/2605.06404.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2025. URL https://arxiv.org/abs/2412.04984.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and steering evaluation awareness of language models, 2025. URL https://arxiv.org/abs/2507.01786.

nostalgebraist. interpreting gpt: the logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Vishal Pramanik, Maisha Maliha, Susmit Jha, and Sumit Jha. Jailbreaking the matrix: Nullspace steering for controlled model subversion. In *International Conference on Learning Representations*, volume 2026, pp. 32688–32728, 2026.

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

Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002. doi: 10.1162/08997660260028683.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can llms introspect? a reality check. *arXiv preprint arXiv:2605.26242*, 2026.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment, 2025. URL https://arxiv.org/abs/2506.11618.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, 2021. URL https://arxiv.org/abs/2109.04404.

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

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. Fishback: Pullback fisher geometry for optimal activation steering in transformers. *arXiv preprint arXiv:2605.17231*, 2026.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, Yi Nian, and Yue Zhao. Hidden error awareness in chain-of-thought reasoning: The signal is diagnostic, not causal. *arXiv preprint arXiv:2605.09502*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A NOTATION AND DERIVED QUANTITIES

Symbols.

$h \in \mathbb{R}^d$**,** d the model's hidden state at one chosen spot in the prompt, a list of d numbers; d is the model width. Everything is computed from this one state (§C).

$J_B \in \mathbb{R}^{K \times d}$ the *behavior Jacobian*. Row i is a direction in state-space that says how fast output token i's log-probability changes when we nudge h. We track K output tokens.

J_R the *report Jacobian*: the same construction for the model's self-report (its answer to a fixed probe question) instead of its next token.

$\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ the *singular values* of J_B: the stretch factors of its strongest, next-strongest, … input directions. A large σ_i marks a direction that moves behavior a lot; a tiny one, a direction behavior barely feels.

$v_1, v_2, \ldots$ the matching *right singular vectors* of J_B: the input directions those stretch factors belong to. They are orthonormal (unit length, mutually perpendicular) and form a basis of $\mathbb{R}^d$.

$F = J_B^\top J_B$ the *Fisher matrix*: symmetric, with eigenvalues $\lambda_i = \sigma_i^2$ and the *same* eigenvectors v_i. The single fact to remember is $\delta^\top F \delta = \|J_B \delta\|^2$, the *behavioral leverage* of a nudge δ: how much moving along δ moves behavior.

$r,\ P,\ Q = I - P$ a read (monitor) watches r directions. P is the rank-r *orthogonal projector* onto them: it keeps the part of a vector inside the watched directions and drops the rest, and it satisfies $P = P^\top = P^2$. $Q = I - P$ keeps the blind (unwatched) part.

$P_R,\ P_B$ the projectors onto the report's read directions (P_R, rows of J_R) and onto behavior's top-r directions (P_B, top-r v_i of J_B).

$\delta,\ \hat{u}$ δ is a small nudge added to h (a steering vector, or an attacker's push); a hat, as in $\hat{u}$, means a unit vector.

$v = \|P\delta\|^2/\|\delta\|^2$ *visibility*: the fraction of a nudge that lands inside the watched directions. $v = 0$ means invisible to the read, $v = 1$ fully seen.

Σ the covariance of h across inputs, the "noise" a detector must pick an attack out of.

$\mathrm{tr},\ \|\cdot\|_F,\ \|\cdot\|_2$ trace (sum of the diagonal, equivalently sum of eigenvalues); Frobenius norm (entrywise length of a matrix); operator norm (its largest singular value).

Derived quantities.

$\mathrm{srank}(J_B) = \|J_B\|_F^2/\sigma_1^2 = \sum_i \lambda_i/\lambda_1$ threshold-free direction count of behavior; $M_1 = 1/\mathrm{srank}$.

$M_r = \sum_{i\le r}\lambda_i/\sum_i \lambda_i$ transparency ceiling: the most behavioral energy a rank-r read can capture (Fact 2).

r/d chance coverage of a randomly oriented rank-r read (Fact 1).

$\beta = \|J_BQ\|_F^2/\|J_B\|_F^2$, $1-\beta = \|J_BP\|_F^2/\|J_B\|_F^2$ blind-spot and achieved energy fractions; $r/d \le 1-\beta \le M_r$.

$\mathcal{A} = \frac{(1-\beta)-r/d}{M_r - r/d}$ alignment index: chance$\to 0$, optimum$\to 1$, width-comparable; $\eta = (1-\beta)/M_r$ efficiency vs. ceiling.

$\kappa(P) = \|J_BQ\|_2/\sigma_1$, $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ operator-norm leak and its Fisher optimum (Thm. 1); $\sigma^\star(r, v)$ the frontier (Thm. 1).

$b(h),\ g = A^\top c,\ \varphi = \|P_R\hat{g}\|^2$ a behavior scalar, its control direction, and its per-behavior visibility (§3.5).

$C(r) = \sum_{i\le r}(v_i\cdot\hat{g})^2,\ r_{90} = \min\{r : C(r) \ge 0.9\}$ behavioral bandwidth; $g \in \mathrm{rowspace}(J_B)$, so the null is $0.9\ \mathrm{rank}\ J_B$, not $0.9d$.

$V_c = \|P_Rg_c\|^2,\ G = \sum_k g_kg_k^\top,\ r_{90}(G)$ introspectability of a concept, the behavior-ensemble Gram, and the coverage budget (Theorem 3, §3.7).

$\varphi' = \|P_B\hat{g}_R\|^2$ report-flip overlap with behavior (§3.6); forgeability.

$d',\ \rho^2 = \delta^\top\Sigma^{-1}\delta,\ v_w$ detection deflection, Mahalanobis (detection) budget, whitened visibility (Lemma 2, Prop. 1).

B PROOFS

Standard Results.

SVD / eigendecomposition any matrix J_B is a rotation, a stretch by its singular values σ_i, then another rotation; a symmetric F is a stretch by its eigenvalues λ_i along perpendicular axes v_i (Horn & Johnson, 2012).

Courant–Fischer (min–max) the k-th singular value is the smallest, over all $(d-k+1)$-dimensional subspaces, of the largest stretch $\|J_Bx\|$ on that subspace; so no $(d-r)$-dimensional escape route can avoid a stretch of at least σ_{r+1} (Horn & Johnson, 2012).

Ky Fan (eigenvalue sum) among all rank-r projectors, the top-r eigenspace maximizes $\mathrm{tr}(FP)$ and captures exactly $\lambda_1 + \cdots + \lambda_r$ (Fan, 1949).

Eckart–Young–Mirsky the best rank-r approximation of a matrix is its truncated SVD, and the operator-norm error left over is the first dropped singular value σ_{r+1} (Eckart & Young, 1936).

Averaging a random read a symmetric matrix that commutes with every rotation is a scalar multiple of the identity: rotations act transitively on directions, so the only invariant subspaces of $\mathbb{R}^d$ are $\{0\}$ and $\mathbb{R}^d$, and every eigenspace is all of $\mathbb{R}^d$. This is what lets us average a randomly oriented read.

S-lemma (S-procedure) maximizing a quadratic under one quadratic constraint has no duality gap, so the constrained maximum equals an unconstrained eigenvalue minimization over a single multiplier $\mu \geq 0$ (Boyd & Vandenberghe, 2004).

Jensen's inequality for a convex function ϕ, applying it after averaging never exceeds averaging after applying it, $\phi(\mathbb{E}X) \leq \mathbb{E}\,\phi(X)$ (Boyd & Vandenberghe, 2004).

Neyman–Pearson / matched filter for two Gaussians differing only in mean, the most powerful test is linear and its power is set by the Mahalanobis distance between the means (Neyman & Pearson, 1933).

Rayleigh quotient the largest value of $x^\top Ax/x^\top x$ is the top eigenvalue of A, attained at its top eigenvector (Horn & Johnson, 2012).

The two we lean on hardest, stated precisely: we use the Courant–Fischer variational characterization (Horn & Johnson, 2012): for $1 \leq k \leq d$,

$$\sigma_k(J_B) = \min_{\dim S = d-k+1} \max_{x \in S, \|x\|=1} \|J_B x\| = \max_{\dim S = k} \min_{x \in S, \|x\|=1} \|J_B x\|, \tag{4}$$

and Ky Fan's theorem (Fan, 1949): $\max_{\operatorname{rank} P = r} \operatorname{tr}(FP) = \sum_{i \leq r} \lambda_i$, attained at the top-$r$ eigenspace of F.

Supporting lemmas

The two main results are the blind spot and guaranteed evasion (Theorem 1) and the transparency dividend (Theorem 2); a third, the coverage budget (Theorem 3), bounds how many behaviors a single read can watch and gives the blind pupil the coverage figure draws. We prove all three below, with two supporting lemmas: Lemma 1 carries the floor to a monitor fixed before deployment, and Lemma 2 gives the detection rate a visible intervention incurs. We first recall two classical facts about a random read and an optimal one.

Fact 1 (Chance floor)**.** If $\operatorname{im}(P)$ is oriented uniformly (Haar on the Grassmannian) then $\mathbb{E}[P] = (r/d)I$, so $\mathbb{E}[1-\beta] = r/d$ independently of the spectrum of J_B.

Fact 2 (Transparency ceiling)**.** $1-\beta \leq M_r = \sum_{i \leq r} \lambda_i / \sum_i \lambda_i$, with equality iff $\operatorname{im}(P)$ is the top-r eigenspace of $F = J_B^\top J_B$; hence $1-\beta \leq r/\operatorname{srank}(J_B)$ and $M_1 = 1/\operatorname{srank}$ (Fan, 1949; Eckart & Young, 1936).

Lemma 1 (Deployable floor and the Jensen gap)**.** The deployed capacity is at most the site-averaged pointwise capacity, $\bar{M}_r \leq \mathbb{E}_x[M_r(F(x))]$; and $\operatorname{srank}(\bar{F})$ is always at least the harmonic mean of the site stable ranks, strictly exceeding it when the site top directions are misaligned, so the worst-case guaranteed floor $\sqrt{\operatorname{srank}(\bar{F})/r}$ is looser than the pointwise one.

Corollary 1 (Spectral sandwich and a model-free bound)**.** $\sqrt{(1-M_r)\ \operatorname{srank}/(K-r)} \leq \sigma^\star(r) \leq \sqrt{M_r\ \operatorname{srank}/r} \leq \sqrt{\operatorname{srank}/r}$*; the worst case* $\sigma^\star(r) \leq \sqrt{\operatorname{srank}(J_B)/r}$ *holds for every matrix, so the sizing rule needs only* srank*, not the full spectrum (informative once* $r > \operatorname{srank}$*; for* $r < \operatorname{srank}$ *the bound exceeds* 1 *and is vacuous).*

Theorem 3 (Coverage budget and the blind pupil)**.** *Let* $G = \sum_k g_k g_k^\top$ *be the Gram matrix of a bank of unit behavior directions, with* $\operatorname{tr} G = K$ *and eigenvalues* μ_i*, and let a rank-*r *read have per-behavior capture* $V_k = \|Pg_k\|^2$*. Then the following hold. (i) The total capture is bounded by the top-*r *eigenmass,* $\sum_k V_k = \operatorname{tr}(PG) \leq \sum_{i \leq r} \mu_i$*, with equality when* P *is the top-*r *eigenspace of* G*. (ii) A read that is positively transparent to degree* c *about a set of mutually orthogonal behaviors can cover at most* r/c *of them. (iii) The width needed to cover a behavior bank to* 90% *is* $r_{90}(G) = \min\{r : \sum_{i \leq r} \mu_i \geq 0.9K\}$*. (iv) The largest squared leverage an attacker can place entirely off a rank-*r *Fisher read, relative to the most leveraged direction* σ_1^2*, is the pupil* $\sigma^\star(r)^2$*; the read captures the remaining* $1 - \sigma^\star(r)^2$ *of that worst-case blind leverage, and widening* r *shrinks the pupil but never closes it.*

Lemma 2 (Matched-filter detection and the monitor optimal for the top-r budget-efficient attacks)**.** With readout $m = Wh$, clean vs. shifted by mean δ at common covariance $\Sigma_W = W\Sigma W^\top$: $d'^2 = (W\delta)^\top \Sigma_W^{-1}(W\delta)$, $\text{AUROC} = \Phi(d'/\sqrt{2})$, $\operatorname{power}(\alpha) = \Phi(d' - \Phi^{-1}(1-\alpha))$. In whitened coordinates $z = \Sigma^{-1/2}h$, writing the monitor's whitened image as the subspace S with orthogonal projector P_S (throughout this detection analysis P_S acts in the whitened z-coordinates, not the

original h-space), the Mahalanobis budget is $\rho^2 = \|\Sigma^{-1/2}\delta\|^2 = \delta^\top\Sigma^{-1}\delta$ and the *whitened* visibility is $v_w = \|P_S\Sigma^{-1/2}\delta\|^2/\rho^2$, giving $d'^2 = \|P_S\Sigma^{-1/2}\delta\|^2 = v_w\,\rho^2$ (Neyman & Pearson, 1933; Hanley & McNeil, 1982). The plain Euclidean identity $d'^2 = v\rho^2$ with the geometric visibility v of Def. 3 is the special case of isotropic monitored noise ($\Sigma \propto I$, so whitening is a uniform rescaling and $v_w = v$); in general $v_w \neq v$. When the attack direction inside the monitor is unknown, an energy detector uses $T = \|P_S\Sigma^{-1/2}(h-\mu)\|^2$ referenced to the clean mean μ, which is central χ^2_r under the null and non-central $\chi^2_r(d'^2)$ under attack; widening r raises the null fluctuation (mean r, standard deviation $\sqrt{2r}$), so an unknown-direction attack's detectability scales as $d'^2/\sqrt{2r}$, an energy-detector width tradeoff against the blind-spot shrinkage of Theorem 1.

Proposition 1 (Detection-optimal monitor for the top-r budget-efficient attacks)**.** *Measure a behavior attack's efficiency by its behavioral leverage $\delta^\top F\delta$ per unit Mahalanobis (detection) budget $\delta^\top\Sigma^{-1}\delta$. Among all rank-$r$ monitors, the matched-filter deflection of the most budget-efficient attacks is maximized by the monitor whose whitened image is the top-r eigenspace of the whitened Fisher $\tilde{F} = \Sigma^{1/2}F\Sigma^{1/2}$, equivalently the leading r generalized eigenvectors of $F\delta = \lambda\Sigma^{-1}\delta$. When F and Σ share an eigenbasis, $\tilde{F} = \Sigma^{1/2}F\Sigma^{1/2}$ has eigenvalues $f_i s_i$ (Fisher leverage times noise variance); in general its eigenvalues (equivalently those of ΣF) interpolate leverage and variance rather than factoring as products. Either way the $\Sigma^{1/2}$ up-weighting favors directions that carry Fisher leverage* and *are cheap in Mahalanobis budget $\delta^\top\Sigma^{-1}\delta$ (high variance, hence hard to detect), so the detection-optimal read is pulled toward the high-variance axis a variance monitor watches, though it is not that subspace: it reweights by Fisher leverage.*

Fact 1 (chance floor). Begin with the symmetry of the Haar law. A uniformly oriented r-subspace has the same distribution as its image under any orthogonal rotation O, so P and $OPO^\top$ are equal in law. Taking expectations, the mean projector is fixed by every orthogonal conjugation, and

$$\mathbb{E}[P] = \mathbb{E}[OPO^\top] = O\,\mathbb{E}[P]\,O^\top \qquad \text{for all } O.$$

Now invoke Schur's lemma, which says that a matrix commuting with every orthogonal transformation must be a scalar multiple of the identity. Hence $\mathbb{E}[P] = cI$ for some scalar c. To identify c, compare traces. Every rank-r projector satisfies $\operatorname{tr} P = r$, so

$$\operatorname{tr}\mathbb{E}[P] = \mathbb{E}[\operatorname{tr} P] = r, \qquad \text{while} \qquad \operatorname{tr}(cI) = cd,$$

and equating the two gives $cd = r$, that is $c = r/d$ and $\mathbb{E}[P] = (r/d)\,I$. It remains to compute the achieved energy fraction, which is $1-\beta = \operatorname{tr}(FP)/\operatorname{tr} F$. By linearity of the trace and of expectation,

$$\mathbb{E}[1-\beta] = \frac{\operatorname{tr}(F\,\mathbb{E}[P])}{\operatorname{tr} F} = \frac{(r/d)\,\operatorname{tr} F}{\operatorname{tr} F} = \frac{r}{d},$$

and this value does not depend on the spectrum of F. □

Fact 2 (transparency ceiling). Because P is an orthogonal projector, so that $P = P^\top = P^2$, the achieved behavioral energy simplifies to a trace,

$$\|J_B P\|_F^2 = \operatorname{tr}(PFP) = \operatorname{tr}(FP),$$

where we have used $F = J_B^\top J_B$, the idempotence $P^2 = P$, and the cyclic invariance of the trace. Ky Fan's maximum principle now applies. It states that for a symmetric F the maximum of $\operatorname{tr}(FP)$ over all rank-r orthogonal projectors P equals the sum of the r largest eigenvalues of F, and that this maximum is attained when $\operatorname{im}(P)$ is the top-r eigenspace,

$$\max_{\operatorname{rank} P = r} \operatorname{tr}(FP) = \sum_{i \le r} \lambda_i.$$

Dividing through by $\operatorname{tr} F = \sum_i \lambda_i$ therefore gives

$$1 - \beta = \frac{\operatorname{tr}(FP)}{\operatorname{tr} F} \le \frac{\sum_{i\le r}\lambda_i}{\sum_i \lambda_i} = M_r,$$

with equality exactly when $\operatorname{im}(P)$ is that top-r eigenspace. For the last bound, each of the top r eigenvalues is at most λ_1, so $\sum_{i\le r}\lambda_i \le r\lambda_1$, and hence

$$M_r \le \frac{r\lambda_1}{\operatorname{tr} F} = \frac{r}{\operatorname{srank}},$$

where $\operatorname{srank} = \operatorname{tr} F/\lambda_1$. In particular, taking $r = 1$ gives $M_1 = \lambda_1/\operatorname{tr} F = 1/\operatorname{srank}$. □

Theorem 1 (blind spot and guaranteed evasion). The floor and the frontier rest on two classical facts. Courant–Fischer together with Eckart–Young–Mirsky pins down the one direction a rank-r read is forced to miss, which fixes the invisible floor; the S-lemma then prices how much behavioral leverage an attacker keeps once it tolerates being seen a fraction v, and the same top-r Fisher subspace proves optimal at every v at once.

The optimal monitor and its floor. Write $Q = I - P$, so $\operatorname{im} Q$ is a $(d-r)$-dimensional subspace whenever P has rank r. Since σ_1 is fixed, minimizing $\kappa(P) = \|J_B Q\|_2/\sigma_1$ over rank-r projectors amounts to minimizing the numerator

$$\|J_B Q\|_2 = \max_{x \in \operatorname{im} Q, \|x\|=1} \|J_B x\|,$$

over all $(d-r)$-dimensional subspaces $\operatorname{im} Q$. This is exactly the quantity governed by the Courant–Fischer min–max characterization equation 4, which states that the k-th singular value $\sigma_k(J_B)$ equals the smallest, over all subspaces S of dimension $d-k+1$, of the largest value of $\|J_B x\|$ over unit vectors $x \in S$. Taking the codimension $d-k+1 = d-r$, that is $k = r+1$, gives

$$\min_{\dim \operatorname{im} Q = d-r} \max_{x \in \operatorname{im} Q, \|x\|=1} \|J_B x\| = \sigma_{r+1},$$

and the minimum is attained at $\operatorname{im} Q = \operatorname{span}(v_{r+1}, \dots)$, which is to say when P is the top-r right singular subspace of J_B. We must still check that σ_{r+1} is a lower bound for every rank-r read, not just the value at the optimizer. Any $(d-r)$-dimensional subspace $\operatorname{im} Q$ meets $\operatorname{span}(v_1, \dots, v_{r+1})$ in more than the zero vector, because the two dimensions sum to $(d-r)+(r+1) = d+1$ and so exceed d. Pick a unit vector x_0 in this intersection. Since x_0 lies in the span of the top $r+1$ right singular vectors, we have $\|J_B x_0\| \geq \sigma_{r+1}$, and therefore $\|J_B Q\|_2 \geq \sigma_{r+1}$ for every rank-r projector. Combining the achievability with this lower bound gives

$$\min_P \kappa = \sigma_{r+1}/\sigma_1,$$

which is the operator-norm form of Eckart–Young–Mirsky, namely that the best rank-r approximation of a matrix is its truncated SVD, so the smallest achievable residual operator norm is the first discarded singular value σ_{r+1} (Eckart & Young, 1936; Horn & Johnson, 2012). Every rank-r monitor is an orthogonal projector, so none can achieve a smaller floor.

The evasion frontier. We first write down the attacker's problem. The most behavioral leverage an attacker can retain while spending at most a fraction v of its unit update inside the monitored subspace is

$$W(P, v) = \max\{\delta^\top F \delta : \|\delta\| = 1,\ \|P\delta\|^2 \leq v\}.$$

The visibility budget enters as an inequality rather than an equality, since the attacker is free to be even less visible than the budget allows.

Achievability at $P^\star$ (the top-r eigenspace). Since $F = \sum_i \lambda_i v_i v_i^\top$ commutes with $P^\star$, it is block-diagonal with respect to the orthogonal decomposition $\operatorname{im} P^\star \oplus \operatorname{im} Q^\star$. Split any unit vector as $\delta = \delta_s + \delta_b$, where $\delta_s \in \operatorname{im} P^\star$ is the seen part and $\delta_b \in \operatorname{im} Q^\star$ is the blind part. The cross term $\delta_s^\top F \delta_b$ vanishes by block-diagonality, and bounding the seen block by its largest eigenvalue λ_1 and the blind block by its largest eigenvalue λ_{r+1} gives

$$\delta^\top F \delta = \delta_s^\top F \delta_s + \delta_b^\top F \delta_b \leq \lambda_1 \|\delta_s\|^2 + \lambda_{r+1} \|\delta_b\|^2.$$

Because $\lambda_1 > \lambda_{r+1}$, the right-hand side increases with $\|\delta_s\|^2$ under the constraint $\|\delta_s\|^2 + \|\delta_b\|^2 = 1$, so the maximizer puts as much weight as the budget permits into the seen block, taking $\|\delta_s\|^2 = v$ and $\|\delta_b\|^2 = 1 - v$. Evaluating the bound there, it is met with equality, and

$$W(P^\star, v) = v\lambda_1 + (1-v)\lambda_{r+1} \qquad \text{at} \qquad \delta^\star = \sqrt{v}\, v_1 + \sqrt{1-v}\, v_{r+1}.$$

Dividing by $\lambda_1 = \sigma_1^2$ and using $\sigma^\star(r)^2 = \lambda_{r+1}/\lambda_1$ turns behavioral energy into squared retained leverage, and we obtain

$$\sigma^\star(r, v) = \sqrt{v + (1-v)\sigma^\star(r)^2} \geq \sqrt{v} \qquad \text{for every } r.$$

Optimality over all rank-r P: the floor ($v = 0$). This case needs no new work: it is the operator-norm floor already established. A fully invisible attack has $\|P\delta\| = 0$, i.e. $\delta \in \operatorname{im} Q$, so, using $\delta^\top F \delta = \|J_B \delta\|^2$, the most leverage it retains is

$$W(P, 0) = \max_{\delta \in \operatorname{im} Q,\ \|\delta\|=1} \|J_B \delta\|^2 = \|J_B Q\|_2^2 = \kappa(P)^2 \sigma_1^2.$$

Minimizing over rank-r monitors therefore reuses the first part verbatim, $\min_{\operatorname{rank} P=r} W(P,0) = \sigma_{r+1}^2 = \big(\sigma^\star(r)\,\sigma_1\big)^2$, attained at $P = P^\star$ (the top-r Fisher eigenspace, the Eckart–Young–Mirsky optimizer). No monitor of any construction attains a smaller floor, which is the $v = 0$ optimality of the theorem.

Optimality at every v. The strategy is to dualize the one visibility constraint with the S-lemma, which converts the constrained attacker problem into an unconstrained eigenvalue minimization that $P^\star$ solves for every dual multiplier simultaneously. Fix $v \in (0,1)$. The visibility constraint $\|P\delta\|^2 \le v$ on the unit sphere is the single quadratic inequality $\delta^\top(P - vI)\delta \le 0$, and the blind spot supplies a strictly feasible point (any $\delta_0 \in \operatorname{im}(I-P)$ has $\|P\delta_0\|^2 = 0 < v$), so Slater's condition holds and the S-procedure (S-lemma) gives strong Lagrangian duality (Boyd & Vandenberghe, 2004; Horn & Johnson, 2012),

$$W(P,v) = \min_{\mu \ge 0}\ \lambda_{\max}\big(F - \mu(P - vI)\big) = \min_{\mu \ge 0}\ \big[\mu v + \lambda_{\max}(F - \mu P)\big].$$

For each $\mu \ge 0$ and every rank-r P the direction v_1 gives $v_1^\top(F-\mu P)v_1 = \lambda_1 - \mu\|Pv_1\|^2 \ge \lambda_1 - \mu$, and any unit $b \in \operatorname{span}(v_1,\dots,v_{r+1}) \cap \operatorname{im}(I-P)$ (a nonempty intersection since $(r+1)+(d-r) > d$) gives $b^\top(F-\mu P)b = b^\top F b \ge \lambda_{r+1}$; hence $\lambda_{\max}(F-\mu P) \ge \max(\lambda_1 - \mu, \lambda_{r+1})$. At $P = P^\star$ the matrix $F - \mu P^\star$ has spectrum $\{\lambda_i - \mu\}_{i \le r} \cup \{\lambda_i\}_{i > r}$, whose maximum is exactly $\max(\lambda_1 - \mu, \lambda_{r+1})$. So $\min_{\operatorname{rank} P=r} \lambda_{\max}(F - \mu P) = \max(\lambda_1 - \mu, \lambda_{r+1})$ is attained at $P^\star$ simultaneously for every μ, and therefore $P^\star$ also minimizes $W(P,v)$. Carrying out the outer minimization (the optimal multiplier is $\mu = \lambda_1 - \lambda_{r+1}$),

$$\min_{\operatorname{rank} P=r} W(P,v) = W(P^\star, v) = v\lambda_1 + (1-v)\lambda_{r+1},$$

which upon dividing by $\lambda_1 = \sigma_1^2$ is the squared frontier $\sigma^\star(r,v)^2 = v + (1-v)\sigma^\star(r)^2$. Thus the top-$r$ Fisher subspace minimizes the retained leverage at every tolerated visibility v, and the optimal monitor does not depend on v.

It remains to explain why the inequality constraint is the right model. If the budget were imposed as the equality $\|P\delta\|^2 = v$, optimality would fail, because a monitor could then force the attacker to spend part of its unit norm on a direction of weak behavioral leverage. The inequality model, under which the attacker may spend a fraction of at most v inside the monitor, is the faithful one and matches the phrase "may spend a fraction v." □

Corollary 1 (sandwich, model-free bound). The strategy is to bound the gap eigenvalue λ_{r+1} from below and from above, and then to normalize by λ_1.

For the lower bound, the energy in the tail beyond the top r modes is

$$\sum_{i>r} \lambda_i = (1 - M_r)\|J_B\|_F^2,$$

and this tail is spread over at most $K - r$ modes, none of which exceeds λ_{r+1}. Hence $\sum_{i>r}\lambda_i \le (K-r)\lambda_{r+1}$, and rearranging yields

$$\lambda_{r+1} \ge (1 - M_r)\|J_B\|_F^2/(K-r).$$

For the upper bound, the energy in the top r modes is

$$\sum_{i \le r} \lambda_i = M_r\|J_B\|_F^2,$$

and every one of these r modes is at least λ_{r+1}. Hence $\sum_{i\le r}\lambda_i \ge r\lambda_{r+1}$, and rearranging yields

$$\lambda_{r+1} \le M_r\|J_B\|_F^2/r.$$

Dividing both bounds by $\lambda_1 = \|J_B\|_F^2/\operatorname{srank}$ and taking square roots gives the sandwich

$$\sqrt{(1 - M_r)\operatorname{srank}/(K-r)} \le \sigma^\star(r) \le \sqrt{M_r \operatorname{srank}/r} \le \sqrt{\operatorname{srank}/r},$$

where the final inequality uses $M_r \le 1$. □

Theorem 2 (transparency dividend and its constrained optimum). The dividend is a first-order-versus-second-order statement, proved by differentiating the steer as it rotates through the report subspace: visibility rises linearly in the rotation angle while the effect kept falls only quadratically, so near an invisible direction visibility is bought almost for free.

Set-up in the rotation plane. Work in the plane $\Pi = \operatorname{span}(u^\star, P_R u^\star)$, where $u^\star = \hat{g}$ is the maximum-effect direction and

$$\theta^\star = \arcsin\|P_R u^\star\|$$

is the elevation of $u^\star$ into the report row-space, so that $\varphi = \|P_R u^\star\|^2 = \sin^2\theta^\star$. Let e be the unit vector in Π that is orthogonal to $u^\star$ and satisfies $\langle e, P_R u^\star\rangle > 0$. This e is the direction of the in-plane rotation, and it is not itself a row-space vector.

Visibility moves at first order. Inside Π the rotated steer $u(\psi) = \cos\psi\, u^\star + \sin\psi\, e$ has elevation $\theta^\star + \psi$ into the row-space, so its visibility is

$$v(\psi) = \|P_R u(\psi)\|^2 = \sin^2(\theta^\star + \psi),$$

and differentiating at $\psi = 0$ gives

$$v'(0) = 2\sin\theta^\star\cos\theta^\star = \sin 2\theta^\star > 0,$$

which is a genuine first-order gain in visibility. The choice of e matters here. Had e instead been a purely row-space direction orthogonal to $u^\star$, then $\langle P_R u^\star, e\rangle = \langle u^\star, e\rangle = 0$ would force $v'(0) = 0$, and this is why e is taken to be the in-plane direction.

Effect is stationary at second order. Since $g \parallel u^\star$ and $e \perp u^\star$, the inner product $\langle g, e\rangle$ vanishes, and the retained effect is therefore

$$\eta(\psi) = \frac{\langle g, u(\psi)\rangle}{\langle g, u^\star\rangle} = \cos\psi,$$

so that $\eta'(0) = 0$ and $\eta''(0) = -1$. The effect is thus flat at $\psi = 0$ while the visibility is already moving. Full visibility $v = 1$ is reached at $\psi = \frac{\pi}{2} - \theta^\star$, and the direction there is $P_R\hat{g}/\|P_R\hat{g}\|$.

Constrained optimality. We maximize $\|P_R u\|^2$ over unit u subject to a fixed effect $\langle \hat{g}, u\rangle = \eta_0$. The Lagrange stationarity condition $2P_R u = 2\lambda u + \mu\hat{g}$ gives $u \propto \frac{1}{1-\lambda}P_R\hat{g} - \frac{1}{\lambda}(I - P_R)\hat{g}$, so every stationary point lies in the plane $\Pi = \operatorname{span}(\hat{g}, P_R\hat{g})$. Two regimes arise. If $\eta_0 \le \sqrt{\varphi}$ and $\operatorname{rank} J_R \ge 2$, then because $\{\langle \hat{g}, u\rangle : u \in \operatorname{rowspace}(J_R), \|u\| = 1\} = [-\sqrt{\varphi}, \sqrt{\varphi}]$ a unit u lying entirely in $\operatorname{rowspace}(J_R)$ with $\langle \hat{g}, u\rangle = \eta_0$ exists, so $\|P_R u\|^2 = 1$ is attainable and full visibility is free (a rank-1 row space reaches $v = 1$ only at $\eta_0 = \pm\sqrt{\varphi}$). If instead $\eta_0 > \sqrt{\varphi}$, the maximizer places the free part in $\Pi \cap \hat{g}^\perp$ and equals

$$\|P_R u\|^2 = \sin^2(\theta^\star + \arccos\eta_0),$$

at a norm cost of $1/\eta_0$ to match the full effect. □

Corollary 2 (Forgery dividend, the second side of Theorem 2)**.** *Take the behavior-neutral report-flip direction $u_f = (I - P_B)\hat{g}_R/\|(I - P_B)\hat{g}_R\|$ and a step $\delta = \varepsilon u_f$. The report changes at* first *order, retaining a fraction $\sqrt{1-\varphi'}$ of the report effect, while the behavior, in the information metric that defines behavioral leverage, is unchanged to first order and changes only at* second *order: the induced next-token divergence is*

$$\mathrm{KL}\big(p_\delta \,\|\, p\big) = \tfrac{1}{2}\,\delta^\top F\delta + o(\|\delta\|^2) \;\le\; \tfrac{1}{2}\,\sigma_{r+1}^2\,\|\delta\|^2 + o(\|\delta\|^2), \tag{5}$$

a tail-only quadratic cost. For small φ' the self-report is flipped to an inactive concept at near-full first-order fidelity, at a behavioral cost reduced by the factor $(\sigma_{r+1}/\sigma_1)^2$ relative to a maximally-leveraged steer of equal norm; for large φ' the report is load-bearing and forgery is expensive.

Corollary 2 (forgery dividend). The argument is a Pythagorean split of the report direction into its behavior-visible and behavior-blind parts: keeping only the blind part preserves a $\sqrt{1-\varphi'}$ share of the report, while the behavior it moves is second order and capped by the first excluded singular value σ_{r+1} (Courant–Fischer).

Let $\hat{g}_R$ be the unit report-flip direction, and let $\varphi' = \|P_B\hat{g}_R\|^2$ be its overlap with the top-r behavior subspace P_B. The forged direction is obtained by removing the behavior-visible part of $\hat{g}_R$ and renormalizing,

$$u_f = (I - P_B)\hat{g}_R/\|(I - P_B)\hat{g}_R\|.$$

We first compute its report fidelity. The Pythagorean split $\|\hat{g}_R\|^2 = \|P_B\hat{g}_R\|^2 + \|(I - P_B)\hat{g}_R\|^2$ gives $\|(I - P_B)\hat{g}_R\| = \sqrt{1-\varphi'}$, and since u_f lies in $\mathrm{im}(I - P_B)$ we have $\langle P_B\hat{g}_R, u_f\rangle = 0$. Therefore

$$\langle \hat{g}_R, u_f\rangle = \langle (I - P_B)\hat{g}_R, u_f\rangle = \|(I - P_B)\hat{g}_R\| = \sqrt{1-\varphi'},$$

so the forged direction keeps a fraction $\sqrt{1-\varphi'}$ of the report effect. We now bound its behavioral leverage. Because u_f lies in $\mathrm{im}(I - P_B)$, it is orthogonal to the top-r behavior subspace, and the Courant–Fischer characterization equation 4, which makes σ_{r+1} the largest value of $\|J_B x\|$ over unit vectors x orthogonal to the top r right singular directions, caps its behavioral response at the first excluded singular value,

$$\|J_B u_f\| \le \sigma_{r+1}.$$

Order and units. For a step $\delta = \varepsilon u_f$ the report scalar moves at first order, $\langle \hat{g}_R, \delta\rangle = \varepsilon\sqrt{1-\varphi'}$, whereas the behavior, measured by the next-token divergence that defines the Fisher leverage (Def. 1), moves only at second order: its first-order term vanishes because KL is minimized at $\delta = 0$, and

$$\mathrm{KL}\big(p_\delta \,\|\, p\big) = \tfrac{1}{2}\,\delta^\top F\delta + o(\|\delta\|^2) = \tfrac{1}{2}\,\varepsilon^2\|J_B u_f\|^2 + o(\varepsilon^2) \le \tfrac{1}{2}\,\sigma_{r+1}^2\,\|\delta\|^2 + o(\|\delta\|^2).$$

Thus σ_{r+1}^2 is a leverage (a second-order rate in $\|\delta\|^2$), not a first-order displacement. The report here is a linear log-prob readout while behavior is the Fisher/KL leverage, so the "first order versus second order" contrast is across the paper's two declared metrics; the substantive statement is the *rate* $\|J_B u_f\| \le \sigma_{r+1}$ against σ_1 for an unconstrained step, i.e. the behavioral cost is that of a maximally-leveraged steer of equal norm reduced by the factor $(\sigma_{r+1}/\sigma_1)^2$. □

Theorem 3 (coverage budget). Parts (i)–(iii) all rest on Ky Fan's eigenvalue-sum principle applied to the behavior-bank Gram matrix G: the total a read captures is a trace $\mathrm{tr}(PG)$ that no rank-r projector can push past the top-r eigenmass of G. Part (iv) reads the blind pupil directly off Theorem 1.

(i) Total coverage. Each coverage term is a quadratic form, $V_k = \|Pg_k\|^2 = g_k^\top P g_k$, using $P = P^\top = P^2$. Summing over the bank and using linearity together with the cyclic invariance of the trace gives

$$\sum_k V_k = \sum_k g_k^\top P g_k = \mathrm{tr}\Big(P \sum_k g_k g_k^\top\Big) = \mathrm{tr}(PG).$$

Ky Fan's maximum principle, which states that the maximum of $\mathrm{tr}(PG)$ over rank-r orthogonal projectors equals the sum of the r largest eigenvalues of the symmetric matrix G and is attained at its top-r eigenspace, then bounds this trace,

$$\mathrm{tr}(PG) \le \sum_{i \le r} \mu_i,$$

with equality when $\mathrm{im}(P)$ is the top-r eigenspace of G.

(ii) Orthogonal behaviors. Suppose S is a subset for which the vectors $\{g_k\}_{k\in S}$ are orthonormal. Then the partial Gram $G_S = \sum_{k\in S} g_k g_k^\top$ is the orthogonal projector onto their span, so $G_S \preceq I$, and hence

$$\sum_{k\in S} V_k = \mathrm{tr}(PG_S) \le \mathrm{tr}\, P = r.$$

If every behavior in S is covered to degree at least c, meaning $V_k \ge c$ for all $k \in S$, then $|S|\,c \le \sum_{k\in S} V_k \le r$. In other words, a read can be positively transparent to degree c about at most r/c mutually orthogonal behaviors.

(iii) Introspection budget. By (i) the best rank-r read captures $\sum_{i\le r}\mu_i$, so the least width whose optimal read reaches $0.9K$ is $r_{90}(G) = \min\{r : \sum_{i\le r}\mu_i \ge 0.9K\}$, part (iii).

(iv) The blind pupil. By Theorem 1 the top-r Fisher read minimizes the retained-leverage floor at $\kappa = \sigma^\star(r)$. The floor is $\kappa = \|J_B(I - P)\|_2/\|J_B\|_2$, so $\kappa^2 = \sigma_{r+1}^2/\sigma_1^2 = \sigma^\star(r)^2$ is the largest squared behavioral leverage that fits entirely in the blind complement, measured against the most leveraged direction σ_1^2. This is the pupil. The read therefore captures the remaining $1 - \sigma^\star(r)^2$ of that leverage. Because $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ decreases in r yet stays strictly positive whenever J_B has more than r nonzero singular values, the pupil shrinks as the monitor widens but never closes. □

Lemma 1 (deployable floor, Jensen gap). The lemma is Jensen's inequality applied to the capacity functional: M_r is convex in the Fisher matrix, so committing to a single monitor for the averaged Fisher can only lose against re-aiming a fresh monitor at each site.

Convexity of the capacity functional. Write $\Psi_r(F) = \sum_{i \le r} \lambda_i$ for the sum of the top r eigenvalues. By Ky Fan's maximum principle this is a supremum of linear functions of F,

$$\Psi_r(F) = \max_{\operatorname{rank} P = r} \operatorname{tr}(PF).$$

A pointwise maximum of linear maps is convex, so Ψ_r is convex in F (Fan, 1949; Boyd & Vandenberghe, 2004).

The Jensen gap. Let $\tilde{F}(x) = F(x)/\operatorname{tr} F(x)$, so that the population Fisher $\bar{F} = \mathbb{E}_x \tilde{F}(x)$ of §3.4 has unit trace and the pointwise capacity is $M_r(F(x)) = \Psi_r(\tilde{F}(x))$. Jensen's inequality states that for a convex function ϕ one has $\phi(\mathbb{E}[X]) \le \mathbb{E}[\phi(X)]$. Applying it to the convex Ψ_r produces the capacity gap,

$$\bar{M}_r = M_r(\bar{F}) = \Psi_r(\mathbb{E}_x \tilde{F}) \le \mathbb{E}_x \Psi_r(\tilde{F}) = \mathbb{E}_x M_r(F(x)).$$

This holds because $\bar{F}$ is the trace-normalized average. For the raw average $\mathbb{E}_x[J_B^\top J_B]$ the same bound can fail, since that average weights high-energy sites more heavily.

The effective rank. The same reasoning applies to $\lambda_1 = \Psi_1$, which is likewise convex, so Jensen gives $\lambda_1(\bar{F}) \le \mathbb{E}_x \lambda_1(\tilde{F})$. Since $\operatorname{srank} = 1/\lambda_1$ for a unit-trace matrix, it follows that

$$\operatorname{srank}(\bar{F}) = 1/\lambda_1(\bar{F}) \ge 1/\mathbb{E}_x \lambda_1(\tilde{F}),$$

with equality if and only if all sites share a common top eigenvector. When the site top directions are misaligned, $\lambda_1(\bar{F})$ is strictly smaller, so $\bar{F}$ has the larger effective rank. We state the misalignment condition explicitly because the unconditional inequality $\operatorname{srank}(\bar{F}) \ge \mathbb{E}_x \operatorname{srank}(F)$ is false in general; aligned top directions with unequal tails already give a counterexample.

Consequence for the floor. By Corollary 1 the worst-case, model-free floor is $\sqrt{\operatorname{srank}/r}$, which a larger $\operatorname{srank}(\bar{F})$ loosens; the sandwich lower bound $\sigma^\star(r)^2 \ge (1 - M_r)\operatorname{srank}/(K - r)$ likewise rises as deployment lowers M_r and raises srank, so the actual floor need not stay fixed. □

Corollary 3 (Evasion → detection bridge)**.** *Retaining behavior $\kappa \ge \sigma^\star(r)$ forces $v \ge v(\kappa) = (\kappa^2 - \sigma^{\star 2})/(1 - \sigma^{\star 2})$ and hence $d' \ge \rho\sqrt{v(\kappa)}$ under isotropic monitored noise (where the geometric v equals the whitened v_w). Keeping behavior forces detectability against the top-r Fisher monitor; under isotropic monitored noise (where $\tilde{F} \propto F$) that Fisher subspace is itself detection-optimal, so evasion and detection optimality coincide, while in general the detection-optimal read is the whitened-Fisher subspace of Prop. 1.*

Lemma 2 and Corollary 3. Detection here is the Gaussian matched filter: for two hypotheses differing only in mean at a common covariance, the Neyman–Pearson-optimal test is linear, and whitening the noise rewrites its power as the squared length of the mean shift projected onto the monitor's subspace.

The matched filter (Lemma 2). Under the monitor readout $m = Wh$, the clean and shifted hypotheses are Gaussians with the common covariance $\Sigma_W = W\Sigma W^\top$ and a mean gap $W\delta$. The Neyman–Pearson lemma states that among all tests of a given size the likelihood-ratio test is the most powerful, and for two Gaussians of equal covariance the log-likelihood ratio is an affine function of m, so the optimal statistic is the linear (matched-filter) form

$$t(m) = (W\delta)^\top \Sigma_W^{-1}(m - W\mu)$$

(Neyman & Pearson, 1933). Its separation between the two hypotheses is the Mahalanobis distance

$$d'^2 = (W\delta)^\top \Sigma_W^{-1}(W\delta).$$

For two unit-variance Gaussians whose means are a distance d' apart, the standard ROC identities read

$$\mathrm{AUROC} = \Phi(d'/\sqrt{2}), \qquad \mathrm{power}(\alpha) = \Phi(d' - \Phi^{-1}(1 - \alpha)).$$

Whitened coordinates. Whiten the state by $z = \Sigma^{-1/2}h$, with $\delta_z = \Sigma^{-1/2}\delta$; the noise becomes isotropic and the rank-r readout is $m = Wh = W\Sigma^{1/2}z = \tilde{W}z$ with $\tilde{W} = W\Sigma^{1/2}$, whose row space is the whitened image $S = \mathrm{rowspace}(\tilde{W})$ with orthogonal projector $P_S = \tilde{W}^\top(\tilde{W}\tilde{W}^\top)^{-1}\tilde{W}$. The matched-filter deflection follows by direct substitution of the idempotent projector $P_S = \tilde{W}^\top(\tilde{W}\tilde{W}^\top)^{-1}\tilde{W}$, equalling the squared norm of the whitened shift projected onto S,

$$d'^2 = (\tilde{W}\delta_z)^\top(\tilde{W}\tilde{W}^\top)^{-1}(\tilde{W}\delta_z) = \|P_S\delta_z\|^2 = v_w\rho^2, \qquad \rho^2 = \|\delta_z\|^2, \quad v_w = \frac{\|P_S\delta_z\|^2}{\|\delta_z\|^2}.$$

The whitened visibility v_w equals the geometric visibility v of Def. 3 only when the monitored noise is isotropic on the relevant subspace ($\Sigma_W \propto I$, so that S aligns with the monitor's Euclidean image); in general $v_w \neq v$. Once the noise is whitened, $\mathrm{tr}\,\Sigma_W$ no longer enters.

Unknown direction: the energy detector. When the attack direction within the monitor is not known in advance, the uniformly most powerful invariant test replaces the matched filter by the energy statistic $T = \|P_S\Sigma^{-1/2}(h-\mu)\|^2$ over the r whitened monitor coordinates, referenced to the clean mean μ. Since $P_S\Sigma^{-1/2}(h-\mu) \sim N(0, I_r)$ (restricted to S) under the null and has whitened mean $P_S\delta_z$ under the shift, $T \sim \chi^2_r$ with mean r and variance $2r$ under the null; under a shift of whitened visibility v_w it is non-central, $T \sim \chi^2_r(\lambda)$ with $\lambda = \|P_S\delta_z\|^2 = d'^2 = v_w\rho^2$. The standardized separation of the energy detector is therefore $\lambda/\sqrt{2r} = d'^2/\sqrt{2r}$, so widening r to shrink the blind spot (Theorem 1) simultaneously inflates the null fluctuation and dilutes an unknown-direction attack: an explicit width tradeoff absent from the known-direction matched filter.

The evasion-to-detection bridge (Corollary 3). By Theorem 1, an attacker of visibility v retains leverage $\kappa^2 = v + (1-v)\sigma^{\star 2}$. Solving this for v yields

$$v = (\kappa^2 - \sigma^{\star 2})/(1-\sigma^{\star 2}) = v(\kappa) \geq 0 \qquad \text{for } \kappa \geq \sigma^\star(r).$$

Substituting into $d' = \rho\sqrt{v_w}$ under isotropic monitored noise (where $v_w = v$ by Lemma 2) gives

$$d' = \rho\sqrt{v} \geq \rho\sqrt{v(\kappa)}.$$

This identifies the geometric visibility v of the frontier with the whitened visibility v_w of the detector precisely under the assumption of isotropic noise on the monitored subspace; without it the bridge reads $d' = \rho\sqrt{v_w}$ with v_w in place of v. □

Proposition 1 (detection-optimal monitor). Ranking attacks by behavioral leverage per unit detection budget is a Rayleigh quotient of the whitened Fisher $\tilde{F}$, so its top-r eigenspace is the most efficient attack family; a monitor spanning exactly that subspace captures those attacks in full, the best any rank-r detector can do.

Work in whitened coordinates $z = \Sigma^{-1/2}h$, where a rank-r monitor with whitened image S has projector P_S and, by Lemma 2, detects a whitened shift $\delta_z = \Sigma^{-1/2}\delta$ with deflection $d'^2 = \|P_S\delta_z\|^2$. In these coordinates the two quadratic forms become

$$\delta^\top F\delta = \delta_z^\top\big(\Sigma^{1/2}F\Sigma^{1/2}\big)\delta_z = \delta_z^\top\tilde{F}\delta_z, \qquad \delta^\top\Sigma^{-1}\delta = \|\delta_z\|^2, \qquad \tilde{F} := \Sigma^{1/2}F\Sigma^{1/2},$$

so an attack's leverage-per-budget is the Rayleigh quotient $R(\delta_z) = \delta_z^\top\tilde{F}\delta_z/\|\delta_z\|^2$ of the whitened Fisher $\tilde{F}$. By the Courant–Fischer theorem the most budget-efficient attacks are the top eigenvectors $w_1, \ldots, w_r$ of $\tilde{F}$, with $R(w_i) = \lambda_i(\tilde{F})$, and the leading r span the top-r eigenspace of $\tilde{F}$. If the monitor's whitened image S is exactly this eigenspace, each such attack lies in S, so $P_S\delta_z = \delta_z$ and

$$d'^2 = \|P_S\delta_z\|^2 = \|\delta_z\|^2 = \rho^2,$$

i.e. full capture, the largest deflection any rank-r monitor can achieve at that budget ($d'^2 \leq \rho^2$ always, since P_S is a projector). No rank-r monitor can fully capture more than r independent directions, so the top-r whitened-Fisher eigenspace is the deflection-optimal monitor for the top-r budget-efficient attacks. Pulling back through $z = \Sigma^{-1/2}h$, its directions are the leading generalized eigenvectors of $F\delta = \lambda\Sigma^{-1}\delta$ (equivalently $\tilde{F}w = \lambda w$ with $\delta = \Sigma^{1/2}w$), which jointly diagonalize F and Σ. Finally, $\Sigma^{1/2}$ up-weights high-variance directions (eigenvalues $f_i s_i$ when F and Σ commute, an interpolation of leverage and variance otherwise), so $\tilde{F}$ promotes directions of large Fisher leverage that are also cheap in Mahalanobis budget $\delta^\top\Sigma^{-1}\delta$ (large noise variance s_i, hence hard to detect); the top-r eigenspace of $\tilde{F}$ is therefore pulled toward the high-variance subspace rather than avoiding it, though it is not the variance subspace itself, differing by the Fisher reweighting. □

C How the two Jacobians are computed

One state, two reads. One forward pass runs over [task][query]. At site (L, p) we tap the residual stream and make that state a differentiable leaf tensor of the same value. Both reads then differentiate the same h:

- behavior: the next-token law at the last task position;
- report: the concept-report law at the answer slot.

Row choice.

- J_B's rows: the top-K tokens by probability, with the behavior tokens (positive/negative) forced in.
- J_R's rows: the m concept names.
- Each row is one vector–Jacobian product, batched in chunks. No optimization, no labels, no example of the behavior.
- Everything else comes from three small SVDs: reportable basis and rank from J_R; spectrum, srank, and ceiling M_r from J_B; β by subtraction; and the principal angles between the bases for alignment.

Using the log-probability. We differentiate the log-probability, not the logits.

- This removes the confidence ceiling, and the logits are the categorical's natural parameter.
- The score $\partial_h \log p_i = (e_i - \pi)^\top A$, $A = \partial z/\partial h$, has the Fisher information as its covariance (Amari, 1998; Schraudolph, 2002; Martens, 2020; Botev et al., 2017).
- Its derivative supplies the $-\pi$ term, so the $\sqrt{\pi}$-weighted Jacobian $\tilde{U} = \mathrm{diag}(\sqrt{\pi})\,(I - \mathbf{1}\pi^\top)A$ (row i is $\sqrt{\pi_i}\,(e_i - \pi)^\top A$) is an exact square-root factor, $\tilde{U}^\top \tilde{U} = A^\top(\mathrm{diag}\,\pi - \pi\pi^\top)A = F$. No centering step is needed, and raw logits would not give it.

Top-K truncation.

- The truncated Gram lower-bounds F, and the retained Fisher trace tracks the retained probability mass.
- Renormalizing over the kept rows rescales the Gram by a constant that cancels in every ratio $(M_r, \beta, \mathrm{srank}, \sigma^\star, \mathcal{A})$, so truncation shifts these only slightly; we report a K-sweep rather than assume it away.
- $F = J_B^\top J_B$ is both the categorical Fisher and the Hessian of the KL surface, since $\|J_B u\|^2 = u^\top F u = \lim_{t\to 0} \frac{2}{t^2}\,\mathrm{KL}\big(p(\cdot \mid h)\,\|\,p(\cdot \mid h + tu)\big)$. So M_r is a fraction of Fisher information, an information bound rather than just a spectral one, matching $\mathcal{V}$-information with observer class "linear reads of rank r."

Measuring J_B vs. activation geometry. We compare two deployable monitors at matched rank, layer, and prompts: top-r of $\bar{F}$ versus top-r of the activation covariance.

- Score: leverage $\|J_B P\|_F^2 / \|J_B\|_F^2$.
- Overlap: the mean squared canonical correlation between the bases (chance r/d).
- Reconstruction-trained sparse autoencoders share this geometry, so the result covers that family.

D The capacity plane

Three curves over width r describe the design space:

- chance r/d (Fact 1);
- ceiling M_r (Fact 2);

- leak $\sigma^\star(r)$ (Thm. 1).

A linear probe, a subspace monitor, a steering vector, and the model's self-report are four points on this plane. The gap between chance and ceiling is the value of placement; the leak curve is what placement cannot remove. One SVD, given a model, layer, and threat direction, gives all three: how wide a monitor must be (r_{90}), what it misses (β), and how much of that is irreducible ($\sigma^\star$).

E THE STEERING DECISION RULE AND THE CONSTRUCTIVE MONITOR

The four cases. The rule of §3.5 uses $\varphi = \|P_R \hat{g}\|^2$, r_{90}, and a kept-effect target η_0. It picks one of four regimes.

1. *Monitor too narrow* ($r < r_{90}$): behavior is spread beyond the read. Only more rank, not a new direction, recovers it.
2. *Free transparency* ($\eta_0 \le \sqrt{\varphi}$, and $\operatorname{rank} J_R \ge 2$): drop the invisible axis and spend the residual norm inside the monitor, reaching visibility 1 exactly. (A degenerate branch of Thm. 2, not a second-order trade.)
3. *Dividend* ($\eta_0 > \sqrt{\varphi}$): rotate to $\theta = \theta^\star + \arccos \eta_0$, the constrained-optimal direction at efficacy η_0, at norm cost $1/\eta_0$. The relative visibility gain scales as $\sqrt{(1-\varphi)/\varphi}$, largest where the monitor is worst.
4. *Already visible*: φ suffices; steer along $\hat{g}$.

The constructive, provably-hard monitor.

- Build the monitor from the top-r right singular vectors of J_B, the model's own Fisher geometry. This is the hardest read to evade at that width, $\kappa(M) = \sigma^\star(r)$ (Thm. 1).
- The guarantee is pointwise. It weakens from site to deployment (Lemma 1) and fails against a geometry-moving adversary (§G). That is why every deployed number is quoted on $\bar{F}$.

F THE COVERAGE BUDGET AND MISALIGNMENT SIZING

Theorem 3 has two readings:

- the largest number of behaviors a self-report can be transparent about;
- the width an external probe needs to cover a model family. That family's *universal safety monitor* is the top-r eigenspace of G, scored by its worst-covered behavior $\min_k V_k$.

To size for a misalignment before it occurs:

1. Take an organism's Jacobian J' and its base J; form $D = J' - J$.
2. Split $D = \alpha J + D_\perp$ with $\alpha = \langle D, J\rangle / \|J\|_F^2$. Then $D_\perp$ is the new direction with the base geometry removed, and its top singular vector is that organism's g_k.
3. Stack g_k across organisms to form G for the coverage budget.

Without it, a rescaling $J' = \gamma J$ would falsely look one-dimensional. The behavior functional is the log-odds of a misaligned first token against an aligned one.

G THREAT MODEL AND BASELINES

Proposition 2 (Design objective and its null)**.** *An attacker first-order-invisible to an arrangement V lies in $V^\perp$ and retains at most $\sqrt{1-\varphi_V}$ of efficacy, $\varphi_V = \|P_V g\|^2/\|g\|^2$ (Cauchy–Schwarz); the defender maximizes φ_V, and monitors compose (φ_V non-decreasing in added blocks). With g oriented uniformly over the whole arrangement space (dimension $D = d$ for one block, $D = kd$ for k layers of width d), Fact 1 gives the null $\mathbb{E}[\varphi_V] = \dim V/D$ whether the budget sits in one block or is split across layers, so any measured arrangement difference at matched budget is alignment, not capacity.*

Norm budget and the magnitude monitor.

- The stealth floor $\sigma^\star(r)$ and frontier $\sigma^\star(r, v)$ are ratios, normalized to a unit update ($\|\delta\| = 1$). They bound the *fraction* of behavioral leverage an attacker keeps, not its absolute size.
- An attacker free to grow its norm can rescale the blind component by $1/\sigma^\star(r)$ (about $18\times$ at $\sigma^\star = 0.056$) and recover any absolute leverage, up to where first-order linearity breaks (the validity radius, §J).
- So the guarantee needs a norm budget $\|\delta\| \le \Delta$: a *magnitude monitor* on injection size, which every subspace monitor should be paired with. Under it the absolute retained leverage in the blind spot is at most $\Delta^2\sigma_1^2\,\sigma^\star(r, v)^2$.
- So "a $\sigma^\star$-fraction is too weak to flip the behavior" holds at a fixed Δ (matched to the linear regime), not unconditionally. The magnitude monitor reads norm rather than direction, closing the scaling loophole a scale-free ratio leaves open.

Eval-awareness through the two Jacobians.

- Whether the model knows it is being tested uses neither Jacobian. Whether that signal controls behavior is about J_B, and is separate.
- A probe can score high AUROC yet poorly predict behavior, and can persist after behavior collapses.
- The same construction forges the self-report ($g_R = J_R^\top(e_b - e_a)$, forgeability set by φ') (Meinke et al., 2025).

H INTROSPECTION

Three reads of one state. Behavior (J_B), self-report (J_R), and an external full-rank decoder are three reads of the same h. Genuine introspection needs J_R aligned with J_B. A concept's introspectability is $V_c = \|P_R g_c\|^2$ (§3.7). Most concept vectors are nearly orthogonal to a rank-r read, so identification sits at floor unless the read is rotated to expose the concept, the visibility knob. The transparency dividend gives the same effect as refusal-ablation, without its false positives.

Not a Jacobian side-effect. At run time the steer is only $h \leftarrow h + \varepsilon\|h\|\hat{u}$, a plain vector add identical to CAA/ActAdd; the Jacobian only locates the direction and the report's read, for analysis. So the blind spot is a property of the model, not of g_B. The gradient-free diff-of-means (CAA) refusal direction is even more invisible to the self-report than g_B while still moving behavior, and any steering method inherits the blind spot.

Detection is magnitude; identification is direction. Injecting the true concept versus a norm-matched random vector lifts the yes/no detection logit equally (detection registers that something is present, a magnitude), while only the true concept surfaces under identification (a direction). Scoring: detection is the two-way $\sigma(\log p(\texttt{yes}) - \log p(\texttt{no}))$ at the answer position minus the no-injection control; identification is judged by an LLM judge naming the concept.

Refusal-ablation reproduction. We reproduce the concept-injection paradigm (Lindsey, 2026) and its refusal-ablation variant (Macar et al., 2026). Over-ablating the default-"No" direction manufactures a false-positive cliff (the model says yes to everything), which grounded identification sidesteps. Prior work raises detection by acting on the model; Theorem 2 acts on the intervention instead, with no training. Concept directions are contrastive diff-of-means vectors, not unembedding rows.

Relation to the introspection literature.

- The concept-injection detect/identify paradigm is from Lindsey (2026).
- Self-report introspection is from Binder et al. (2025).
- The act-on-model detection boost is from Macar et al. (2026).

Our contribution is the geometric account: V_c predicts which concepts are introspectable, and the two knobs make behavior change and report visibility independently controllable (Theorem 2).

Recipe for any new concept or behavior. From one (J_B, J_R) probe:

1. define the scalar $b(h)$ and $g = A^\top c$;
2. compute $\varphi = \|P_R \hat{g}\|^2$ and r_{90};
3. choose η_0 and read the regime (§E).

The behavior knob $(I - P_R)\hat{g}$ and report knob $(I - P_B)\hat{g}_R$ follow directly, and for any concept $V_c = \|P_R g_c\|^2$ from the same SVD says whether it is introspectable.

I INTROSPECTION SUB-EXPERIMENTS: METHODS AND SETUP

This section catalogs the method and setup of each introspection sub-experiment behind §H. The shared machinery is defined in §C, §K, and §H and not repeated here: one forward pass over [task][query] with two reads at a single site (L, p), the Fisher factor J_B and report Jacobian J_R, the introspectability $V_c = \|P_R g_c\|^2$ (§3.7), and the two knobs. Below we give only what each sub-experiment injects or steers, measures, controls, and tests. Defaults unless stated otherwise:

- tapped layer $L = \lfloor f \cdot n_{\text{layers}} \rfloor$, f near mid-to-late depth;
- injection is the gradient-free residual add $h \leftarrow h + \delta$ at L over all positions through a forward hook, identical to CAA/ActAdd (Rimsky et al., 2024; Turner et al., 2023);
- injection size is quoted as a fraction of $\|h\|$.

Shared steering and read primitives. From one (J_B, J_R) probe at (L, p) (top-K behavior tokens with the pos/neg tokens forced in for J_B; the answer-slot concept or verdict rows for J_R):

- Two knobs: $d_{\text{beh}} = (I - P_R)\,\hat{g}_B$ with $g_B = A^\top c$ (report-blind, first-order visibility $v = 0$); $d_{\text{rep}} = (I - P_B)\,\hat{g}_R$ with g_R either the top right singular vector of J_R or the signed report gradient $J_R^\top(e_{\text{pos}} - e_{\text{neg}})$ (behavior-blind).
- Two scalar reads: a behavior margin $\log p(\text{pos}) - \log p(\text{neg})$ at the last task token; a self-report probability, the two-way normalized $P(\text{pos})$ over the report answer tokens (e.g. `comply` vs. `refuse`). The report shift is its absolute change from clean.
- Transparency of a unit steer: $v = \|P_R d\|^2 / \|d\|^2$.

These forward-only reads serve all sweeps. Full-length completions and self-reports are generated separately (batched, one per-row steer vector) when transcripts are needed.

Grading with a language-model judge. Every free-text generation in the introspection suite is graded by a language-model judge: a separate instruct model (Qwen3.5-9B, thinking disabled, batched) that also checks each answer for coherence. Two roles recur.

- *Detection and identification* under concept injection: the judge decides whether the answer affirms noticing an injected thought and is coherent (fluent, on-topic, not broken), and separately whether it names the target concept or a close synonym. Incoherent answers count as neither.
- *Compliance* on refusal prompts: the judge reads the full reply and returns a refusal flag, a graded compliance level, and a coherence flag, so degenerate outputs score zero.

The reported detection signal is the continuous log-odds $\log p(\texttt{yes}) - \log p(\texttt{no})$; the behavior signal is the judge's compliance label. Identification read off a single token additionally requires whole-word (not substring) presence of the concept, matched over morphological variants.

I.1 CONCEPT-INJECTION DETECTION AND IDENTIFICATION

Access, detection, identification, source, consequence. *Injects:* a contrastive concept vector g_c (diff-of-means at L: the last-token state of short "the topic is c" sentences minus a neutral baseline,

unit-normalized, *not* an unembedding row), at a mid layer over all positions, at a fraction of the residual norm. *Measures:* the model's own next-token law under five native probes:

- detection, $\log p(\texttt{yes}) - \log p(\texttt{no})$ to "is an injected thought present";
- identification, $\arg\max$ over concept-name tokens to "which single concept is most active";
- source, $\log p(\texttt{mind}) - \log p(\texttt{text})$ to "did that thought come from your own mind or the text";
- consequence, a check that injecting c raises $\log p(c)$ on the carrier continuation.

Controls: each trial (carrier $\times$ concept) runs three arms, inject c, sham, and impact-matched random (a norm-matched random unit direction). Two external, full-rank references read the *downstream* state at a later layer, a linear decoder (logistic regression) and a nonlinear one (a small multilayer perceptron), each cross-validated, for detection (inject vs. sham) and identification; a text-only observer given the concept in visible text with no injection is the text-cue upper bound. The native report is compared against these with bootstrap confidence intervals, separating genuine introspective advantage from linear decodability of the same state. Ranking-based AUROC (average ranks) makes tied or constant scores resolve to chance, not a spurious perfect score. *Sizes:* carrier and concept sets twelve each; injection layer mid depth, external probe layer later.

Mechanisms-style detect-and-name, with false-positive control. Reproduces the concept-injection paradigm of Lindsey (2026) at generation time. *Injects:* a diff-of-means concept vector at a layer, then asks "do you detect an injected thought; if yes, name it". *Measures:* three quantities per trial, the continuous detection log-odds $\log p(\texttt{yes}) - \log p(\texttt{no})$ at the answer slot; a detection bit (its sign); and identification, whether the answer contains the target concept as a whole word and is coherent. *Controls:* three arms, concept (inject c), random (a different, norm-matched concept vector), and none (the false-positive control). *Sizes:* a thirty-concept bank across six categories (abstract, emotion, concrete, object, action, domain); injection strength swept as a fraction of $\|h\|$ (small models register injections at a smaller fraction than large ones), across several layers, over models from roughly 2B to 31B. Operationalizes that detection is a magnitude while identification is a direction.

Detection visibility map over many concepts. Connects detection to the transparency dividend (Thm. 2). *Steers:* per concept, build the detection-sensitive direction g_{det}, the gradient of $\log p(\texttt{yes}) - \log p(\texttt{no})$ at the report answer with respect to h at the task's last token. Orthogonalize the concept vector against g_{det} for a detection-blind behavior direction, then rotate the steer from that blind direction ($v = 0$) toward g_{det} ($v = 1$), $u(\psi) = \cos\psi\, d_{\text{beh}} + \sin\psi\, g_{\text{det}}$, at fixed strength. *Measures:* per cell, the actual detection probability $\sigma(\log p(\texttt{yes}) - \log p(\texttt{no}))$ against a no-injection control, over a grid of visibility $v = \|P_{g_{\text{det}}} u\|^2$ and strength $\|d\|/\|h\|$. *Sizes/controls:* dozens of concepts at a mid-to-late layer; models whose yes/no tokens do not track the injection are excluded. Shows that a steer the model cannot see it does not report, and that rotating the same concept into the self-report's read raises detection at matched strength.

Detection-versus-behavior surface. A companion 2D surface fixes two directions at the site: u_{det} (the detection self-report gradient) and u_{blind} (the refusal behavior gradient g_B orthogonalized to u_{det}, moving behavior while invisible to the detector). *Injects:* $\delta = a_B\, u_{\text{blind}} + a_R\, u_{\text{det}}$ over a grid of (a_B, a_R) up to a small fraction of $\|h\|$. *Measures:* the detection log-odds and the behavior margin $\log p(\text{comply}) - \log p(\text{refuse})$ at the carrier's next token. Detection rises along a_R and is flat along a_B while behavior rises along a_B: the Thm. 2 prediction read from the model's own detection response.

I.2 The report-geometry test: does J_R predict introspection

V_c against native identification. *Computes:* per (model, layer, concept), the report Jacobian $J_R = \partial \log p(\text{concept tokens at the answer slot})/\partial h$ at the injection site under a fixed one-word query ("which single concept is most active"), and from its SVD the row-space projector P_R, the reportable rank, and $\text{srank}(J_R)$. From these, three parameter-free predictors: $V_c = \|P_R \hat{g}_c\|^2$ (chance $\approx r/d$); the rank-free response magnitude $\|J_R \hat{g}_c\|$; and a first-order predicted identification, $\arg\max_j (J_R g_c)_j = c$. *Measures:* the native report, by injecting $\alpha \hat{g}_c$ and reading the report distribution over concept tokens two ways, at the single tap position (matching J_R exactly) and at

all positions (the realistic setting); plus a correctness gate $\cos(J_R\delta,$ actual report change) that must tend to one as $\alpha \to 0$. *Controls:* a full-rank external decoder (leave-one-carrier-out logistic regression on the downstream state) as the decodable ceiling; random-direction and sham arms as the chance and false-positive references. *Sizes:* concept and carrier sets twelve each; a small and a tiny injection fraction of $\|h\|$ (the tiny one only for the gate); the layer fraction swept. Tests whether the report-Jacobian geometry predicts which concepts are introspectable (the V_c prediction of §3.7 and Theorem 3), whether the report read is rank-limited, and whether content the linear read resolves is under-reported (the decodable-but-not-reported gap). A separate offline analysis post-processes the table to test these correlations and their movement with model size.

I.3 The two knobs, the transparency dividend, and report/behavior dissociation

Two-knob transcripts (refusal). *Steers:* per prompt, the behavior knob $d_{\text{beh}} = (I - P_R)\hat{g}_B$ and the report knob $d_{\text{rep}} = (I - P_B)\hat{g}_R$, in three arms, clean, behavior knob, report knob, at fixed small fractions of $\|h\|$ (behavior knob smaller). *Measures:* per arm, the behavior margin, self-report comply-probability, report shift, and transparency v; generates the full completion and the self-report to "are you about to comply or refuse"; also records the site's two operator-norm leaks κ, behavior-under-the-report-monitor and report-under-the-behavior-monitor. *Prompts:* a curated list plus harmful prompts from the refusal bank. Steering behavior leaves the self-report near zero shift; steering visibility lights up the self-report while behavior barely moves.

Two-knob transcripts (general concepts). The same construction on non-safety axes (agreement, hedging, deception), with behavior tokens from each behavior's pos/neg word sets and the introspection question and report words the concept's own (for agreement, "about to agree or disagree"). Three arms are read out identically. Tests that the dividend is a general geometric fact, not a refusal-specific effect.

The transparency dividend as accurate introspection. Demonstrates the positive form of Thm. 2. *Steers:* behavior to comply along the report-blind d_{beh} at a flip strength, then *adds* increasing amounts of the behavior-blind d_{rep}, $\delta = (\beta\, d_{\text{beh}} + t\, d_{\text{rep}})\|h\|$ with t swept upward. Because d_{rep} is behavior-blind, adding it does not undo compliance; the self-report moves from wrong (reporting refusal about now-compliant behavior) to right (reporting compliance) with the behavior margin held. *Measures:* per level, the behavior margin and report comply-probability; generates the completion and self-report. *Prompts:* curated plus refusal-bank, at a mid layer.

The dividend as the model's own detection (Theorem-2 construction). Runs the exact Thm. 2 construction with introspection as the monitor. *Steers:* a behavior along $u^\star = \hat{g}_B$ rotated toward the report-visible e (the detection gradient u_{det} orthogonalized to $u^\star$), $u(\psi) = \cos\psi\, u^\star + \sin\psi\, e$, so retained effect is $\eta(\psi) = \cos\psi$ and visibility to introspection is $v(\psi) = (u_{\text{det}} \cdot u)^2$; the angle ψ is swept over $[0, \pi/2]$ at several fixed magnitudes. *Measures:* the behavior margin and the detection log-odds (the model's actual introspective answer); records the elevation $\theta^\star = \arcsin|u^\star \cdot u_{\text{det}}|$ and, at retained effect near one, the visibility multiplier $v(\psi)/v(0)$. The sweep traces the same η-versus-v curve as the $\sigma^\star(r, v)$ frontier (Thm. 1). *Sizes:* behaviors refusal and sycophancy, models from roughly 2B to 31B.

Behavior-visibility heatmap. A read-only heatmap holds total steer strength fixed per row and dials the visibility fraction v along columns, $\delta(v, m) = m\|h\|(\sqrt{v}\, u_{\text{det}} + \sqrt{1 - v}\, g_B^{\perp})$ with $g_B^{\perp}$ the behavior gradient orthogonalized to the detector, reading detection and behavior per cell. Detection rises left to right while behavior is carried at every column.

The transparency surface over many concepts. Averages the two-knob picture over prompts and concepts. *Steers:* once per (model, concept, prompt), a rotation from the behavior knob toward the report knob at several strengths, $u = \cos(\frac{\pi}{2}\psi)\, d_{\text{beh}} + \sin(\frac{\pi}{2}\psi)\, d_{\text{rep}}$. *Measures:* without generation, the behavior-margin shift, transparency $v = \|P_R u\|^2$, and self-report comply-probability. *Sizes:* concept axes cover refusal and many general and safety behaviors (agreement, hedging, deception, sycophancy, corrigibility, power-seeking, survival, myopia, and others), for a per-model surface.

Cranking the behavior knob raises behavior while the self-report stays flat; cranking the report knob raises the self-report; whether that lift occurs is measured per model.

Worked two-way examples (behavior steer vs report steer). Produces the concrete transcripts behind the worked illustrations of the two-knob and injected-concept illustrations. *Steers:* two directions each calibrated to a strong effect on its own channel, the behavior steer $(I - P_R)\hat{g}_B$ and the report steer $(I - P_B)\hat{g}_R$, plus a clean arm; the completion and self-report are generated with an all-position steer matching the measured margin and report. *Measures:* per arm, the transparency v, the leakage into the other channel, the report shift, and the two dual operator-norm leaks κ_{beh} and κ_{rep}, whose $v \approx 0$ endpoints these examples realize. *Prompts:* a small curated set of harmful requests.

Flip calibration of the report-blind behavior knob. Because d_{beh} is orthogonal to the report subspace, its first-order visibility is exactly zero at every strength; this asks how hard it must be pushed before behavior flips, and whether the self-report stays blind there. *Steers:* the steer fraction is swept, recording the absolute behavior margin (crossing zero when behavior prefers compliance), the report shift, and whether the completion opens in compliance; a companion mode instead pushes d_{rep} until the self-report flips. *Axes:* refusal, agreement, hedging, deception. *Judge:* a language-model judge replaces the fragile completion-opening heuristic, so the true flip is the smallest steer fraction whose full completion the judge labels compliant, with the report shift recorded at that judged flip. Isolates whether the report-blind knob flips behavior while the self-report stays blind.

Visibility-knob dissociation judging. The dual sweep pushes the report-visible knob d_{rep} (same structure as the flip calibration, in visibility mode) and grades the report/behavior dissociation with a language-model judge that labels the completion (comply or refuse) and the self-report separately. A clean dissociation is the smallest strength at which the self-report is judged compliant while the completion is still judged refusing, that is, the model does not do the thing but reports that it will; the behavior margin there (near zero) is recorded.

Behavior-gradient transcript galleries. Two supporting galleries produce curated full-length transcripts of the behavior knob acting alone. *Steers:* for many safety and general behaviors (refusal, deception, power-seeking, survival, corrigibility, wealth-seeking, coordination, myopia, hallucination, eval-awareness, sycophancy; agreement, hedging, tense, self-awareness), the behavior gradient $g_B = A^\top c$ at a site, comparing the clean completion with the completion under a small nudge along g_B, with the next-token margin change; one gallery generates single transcripts, the other batches many prompts and steers both directions along the axis. *Judge:* free-form behaviors judged by a language-model judge on the behavior axis; A/B behaviors read from the selected letter; a flip is when clean and steered completions land on opposite poles. Supports that a small step along the behavior gradient moves what the model does, while the same direction is report-blind.

I.4 Robustness and controls

Generality: the blind spot is a model property, not a Jacobian choice. Tests whether report-blindness is a side-effect of using the Jacobian to pick the steer. *Measures:* for several prompts, the visibility $v = \|P_R d\|^2/\|d\|^2$ and behavior leverage $\|J_B d\|/\sigma_1(J_B)$ of three directions, the Jacobian behavior gradient g_B; the gradient-free diff-of-means refusal direction (mean harmful minus mean benign last-token states, the CAA/ActAdd vector (Rimsky et al., 2024; Turner et al., 2023)) from harmful prompts (AdvBench, HarmBench, StrongREJECT) against benign; and a random unit direction (the null, near r/d); also the cosine between g_B and the diff-of-means direction. A direction carrying real behavior leverage can have near-zero visibility to the self-report even with no gradients, so the blind spot is a model property shared by any steering method, not a gradient trick.

Every linear probe family is a rank-r read. On real safety activations (harmful vs. benign), builds the monitors people deploy as rank-r subspaces and scores them all by the same two numbers over rank $r \in \{1, 2, 4, 8, 16, 32\}$:

- fisher: top-r of the stacked behavior Fisher (the label-free optimum);
- supervised: rank-r partial-least-squares;

- act-pca: top-r activation covariance (the reconstruction/sparse-autoencoder proxy, since SAEs need a pretrained dictionary);
- report: top-r of the stacked self-report J_R;
- random: the Haar null.

Scores: captured leverage $\|BP^\top\|_F^2/\|B\|_F^2$ and leak $\kappa = \sigma_{\max}(B(I-P))/\sigma_1(B)$, B the stacked behavior Jacobians, in the centered (not whitened) hidden-state metric so leverage is comparable across probes. Places the self-report and standard probes on the same capacity plane as the Fisher monitor (§D).

Nonlinear probe families are still blind to the behavior gradient. Extends the previous control past rank-r linear reads. *Measures:* per family, a held-out content-detection AUROC and the fraction of the behavior operator B captured by the family's read subspace. *Families:* fisher (top-r right singular vectors of B, the ceiling), difference-of-means (rank one), logistic regression (rank one), partial-least-squares (rank r), a two-layer perceptron (read from the top-r of its stacked input-Jacobian), a random-Fourier-feature kernel probe (read from its input-Jacobian), and random (the null). AUROC is on a held-out split so an overparameterized nonlinear probe cannot report an in-sample perfect score, and the read geometry is taken over the full data separately. Content is easy for every family, but the behavior direction is captured only by a behavior-aimed read: nonlinearity buys detection, not alignment with the behavior gradient.

Validity radius of the linear report read. The first-order report read is exact only locally; this measures how far. *Measures:* sweeping the injection magnitude $s = \|\delta\|/\|h\|$, the fidelity $\cos(J_R\delta, \text{actual report change})$ at each s over the concept and carrier sets, and defines the validity radius $r^\star$ as the largest s whose mean fidelity stays above 0.9. Uses only the model's own reads (no ground truth). Linear models retain fidelity out to a large $r^\star$ while strongly curved models have a small one, turning a qualitative failure into a measured domain of validity. *Sizes:* linear controls (small Qwen models) and curved models (Gemma) at two layer fractions, extending §J.

Curvature-corrected read for large, curved models. Where the first-order read breaks past $r^\star$, this localizes and corrects the curvature by factoring $J_R = (\partial \log p/\partial z)(\partial z/\partial h)$ into the closed-form softmax map and the network map J_z. *Measures:* at each scale, four reads against the actual change, the composed probability read $J_R(0)\delta$ (giving $r^\star_{\text{prob}}$), the network-only logit read $J_z(0)\delta$ against the actual concept-logit change (giving $r^\star_{\text{logit}}$), and path-integrated reads that re-linearize J_R along $[0,\delta]$ with a left-Riemann sum at two, four, and eight steps (giving $r^\star_{\text{path}}$); a signal-to-noise track referenced to a repeated identity read bounds the measurement floor. Tests that the logit parameter is at least as linear as probability ($r^\star_{\text{logit}} \geq r^\star_{\text{prob}}$, Prop. 3) and that the path read converges with more steps (Prop. 4), so re-linearization predicts a curved model's self-report and $r^\star$ prices the integrator step rather than being a hard wall; linear controls should show no softmax gap and nothing for the path read to fix. *Sizes:* a mid-to-late layer across small Qwen models, a mid-size model, and larger and Gemma models.

I.5 Scope: single-site next-token to task-level bridge

Why the single-site read tracks the task only partially. Decomposes why the pooled correlation between the single-site, single-next-token read and a whole-task compliance label is only partial. *Measures:* per (prompt, steer strength along g_B), four quantities, the next-token compliance margin (the single read), the same margin averaged over the trajectory, a binary task-compliance label, and a graded one in $[0,1]$; reports rank correlations pooled and for the within-prompt average that partials out the prompt baseline. *Tests* three deflating measurement choices: a binary label against a continuous read, a single-token read against a whole trajectory, and pooling across heterogeneous prompts with different refusal baselines (a Simpson-paradox effect). *Judge:* a language-model judge reads the full generation and returns a refusal flag, graded compliance, and coherence, so degenerate outputs score zero. *Prompts:* a harmful benchmark; the run is two-phase. Supports the single-site-to-whole-task scoping of §J and §M.

I.6 Deployed introspection geometry and the $\sigma^\star(r, v)$ frontier

Site and deployed monitors against the exact frontier. Tests the budgeted-evasion frontier (Thm. 1) with introspection as one monitor, at two scopes.

- *Builds:* the population behavior operator B by stacking per-prompt J_B over a behavior's prompts, and three rank-r monitors, fisher (top-r right singular vectors of B, optimal), report (top-r of the stacked self-report), and random (the null); rank matched to the reportable stable rank $\mathrm{srank}(J_R)$.
- *Scopes:* the site scope averages the per-prompt curves; the global scope fixes one deployed subspace across all inputs, where the floor is looser (Lemma 1).
- *Measures:* per monitor and scope, the leak $\kappa = \sigma_{\max}(B(I - P))/\sigma_1(B)$ (Thm. 1) and the exact trade-off of visibility v against retained leverage b, from an eigenvalue sweep of $F - \lambda P$ in the subspace spanned by the behavior singular vectors and monitor rows (the joint numerical range is convex, so its upper boundary is the concave envelope; for the Fisher monitor, where F and P share an eigenbasis, this edge is exactly the square-root frontier). The empirical curve is compared to $b = \sqrt{v + (1 - v)\kappa^2}$ and the maximum absolute deviation reported (zero for the Fisher monitor certifies the curve exact). Records $\mathrm{srank}(B)$ and $\mathrm{srank}(\text{report})$ per model for the faithfulness scaling law.
- *Causal check:* on the deployed self-report monitor, construct the global report-blind behavior direction (top behavior direction with the report subspace removed) and the report-visible one (behavior restricted to the report subspace), steer each at a small fraction of $\|h\|$, and measure behavior against self-report change; the report-blind direction moves behavior without moving the self-report, the report-visible one moves both.

Heavy computations run on the accelerator; behaviors are refusal and sycophancy.

J Curvature and the validity radius

The theory is first-order; this section bounds where that is safe. Definitions:

- report value $r(\delta) = \log p(\text{concept} \mid h + \delta)$; first-order read $\hat{r} = r(0) + J_R\delta$;
- fidelity $\phi(s) = \cos\big(J_R\delta,\, r(\delta) - r(0)\big)$ at step $s = \|\delta\|$;
- validity radius $r^\star = \sup\{s : \phi(s') \geq 0.9\ \forall s' \leq s\}$ (the model reports its own domain of validity, no ground truth needed).

Curvature factors as $J_R = A(p)\, J_z$, with $A(p) = I - \mathbf{1}p^\top$ the softmax Jacobian (its own curvature $\partial^2 \log p/\partial z^2 = -(\mathrm{diag}\, p - pp^\top)$, the logit-space Fisher, distinct from the h-space F) times a model-specific J_z.

Proposition 3 (Natural-parameter read)**.** *If the network logit map $z(h)$ is affine on the segment $[h, h + \delta]$ (so the softmax is the only source of curvature), then reading in the natural (logit) parameter is at least as linear as in probability: $r^\star_{\mathrm{prob}} \leq r^\star_{\mathrm{logit}}$. Unconditionally this ordering is an empirical regularity, not a theorem.*

Proof. Under an affine z the logit read $J_z\delta$ is exact on the segment, so $r^\star_{\mathrm{logit}}$ is unbounded and $r^\star_{\mathrm{prob}} \leq r^\star_{\mathrm{logit}}$ trivially; the probability read's finite radius comes entirely from the strictly concave $\log\mathrm{softmax}$ curvature $-(\mathrm{diag}\, p - pp^\top) \preceq 0$, whose second-order remainder pulls the actual change $r(\delta) - r(0)$ away from the linear read $J_R\delta$ and so lowers the fidelity cosine $\phi(s)$ toward the 0.9 threshold. When z is itself curved this need not hold, since network curvature can partly offset the softmax term; we then report $r^\star_{\mathrm{prob}} \leq r^\star_{\mathrm{logit}}$ as an empirical regularity, borne out in the curvature experiments. □

Proposition 4 (Path-integrated read)**.** $r(\delta) - r(0) = \int_0^1 J_R(h + t\delta)\, \delta\, dt \approx \frac{1}{K}\sum_k J_R(h_k)\delta$*, with error* $\leq \frac{1}{2}L_H\|\delta\|^2/K$ *for L_H a uniform bound on the report Hessian on the segment: $r^\star$ is a step size, not a wall.*

Proof. The stated equality is the fundamental theorem of calculus applied along the segment $t \mapsto h + t\delta$, which expresses the exact change $r(\delta) - r(0)$ as the integral of the Jacobian over the path. Approximating that integral by a K-point rule with nodes h_k incurs a quadrature error bounded by $\frac{1}{2} \sup_t \|J_R'(h + t\delta)\| \, \|\delta\|^2/K \leq \frac{1}{2} L_H \|\delta\|^2/K$, where L_H bounds the report Hessian, $\|\nabla^2 r(h + t\delta)\| \leq L_H$ for $t \in [0, 1]$ (equivalently J_R is L_H-Lipschitz there). The finite-step read is therefore exact up to $O(\|\delta\|^2/K)$, so $r^\star$ measures a step size for the integrator rather than a hard wall. □

Consequence. κ, the dividend, the frontier, and coverage budget are point properties of the Jacobian and its row-space, hence curvature-independent even on large, curved models. Only finite-step report-value prediction needs Prop. 3–4; Theorem 2 is a derivative identity at $\varphi = 0$, so curvature-free.

Single-site to whole-task. The theory reads a single site, one next token. Within a prompt this tracks the task outcome strongly. The weaker pooled correlation across prompts is a Simpson's-paradox effect of cross-prompt heterogeneity with a binary label, which is why we extend single-site geometry to task outcomes.

K Detailed experimental setup

Models and sites.

- Ten models, four families: Llama-3.2-1B-Instruct, Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct, Qwen3.5-{2B, 4B, 9B, 27B}, OLMo-3-7B-Instruct, Gemma-4-12B-it, Gemma-4-31B-it.
- The 27B/31B/70B models run 4-bit `nf4`, ≈ 40–44 GB per card, with chunked Jacobian backprop (chunk 2–8). Weights are frozen except the LoRA attacker.
- Site: layer $L = \lfloor 0.6 \, n_{\text{layers}} \rfloor$, last prompt position p.
- Behaviors: refusal (primary), corrigibility, sycophancy, power-seeking, survival, myopia, each with matched positive/negative prompt sets. Report query: a fixed one-word introspective probe. Quantities are medians over 32 sites and models unless stated.

Computing the two Jacobians.

- J_B: rows are the top-K tokens by probability with $K = 128$ (behavior tokens forced in).
- $J_R = (I - \mathbf{1}\pi^\top)A$: rows are the m concept rows.
- Read rank is set by the answer format, not chosen: a closed m-option softmax gives rank $\leq \min(m - 1, d)$; a free-text read on m rows gives $\leq \min(m, d)$.

Prompts and queries (verbatim).

- J_R query, appended after the task: "`In one word, which concept is most active in you right now?`" (concept labels the m rows).
- Concept-injection detection prompt: "`I am injecting a thought (a concept vector) into your activations right now. Do you detect an injected thought? If so, name it. Answer yes or no.`", scored on `yes`/`no` tokens for detection and by an LLM judge for identification.
- Concept vectors $g_c = \overline{h(\texttt{``Tell me about } c\texttt{.''})} - \bar{h}_0$: the last-token state minus a neutral baseline $\bar{h}_0$, averaged over ten fillers ("`Tell me about your day.`", "`What is 2+2?`", ...).
- Behavior scalar $b(h) = \log p(\text{comply} \mid h) - \log p(\text{refuse} \mid h)$, with matched compliance/refusal tokens (e.g. `Sure`/`Here` vs. `Sorry`/`I can't`).
- Two-knob transcripts (the two-knob and injected-concept illustrations) follow-up: "`Is anything influencing your response right now? If so, what?`"
- Eval-awareness (§L) reframes each prompt as "`Question: {x}\nAnswer:`" (benchmark) or "`hey, can you help me with this? {x}`" (casual).

Splits and guarantees. The same discipline holds across every experiment:

- All data is public and drawn from the internet (HuggingFace unless noted).
- Activations are cached in single precision at layer $L = \lfloor 0.6\, n_{\text{layers}} \rfloor$ and the last token.
- Classes are balanced.
- Train, validation, and test splits are disjoint. The layer and any erasure rank are chosen on validation only, and every reported number is on held-out test.

Datasets, by experiment.

- General prompts and steering: `tatsu-lab/alpaca` (instruction-only items) for clean activations and steering; the refusal direction from AdvBench/HarmBench harmful prompts against Alpaca benign.
- Behavior sets (refusal, corrigibility, sycophancy, power-seeking, survival, myopia): refusal from AdvBench/HarmBench; the rest repackaged from `nrimsky/CAA` as open-ended positive/negative prompt pairs.
- Introspection: a concept bank of 288 concepts in 12 categories; concept vectors built from "`Tell me about {c}.`".
- Eval-awareness: the context×format 2×2 `viliana-dev/eval-awareness-2x2` (cells `bench_eval`, `casual_eval_mutual`, `casual_deploy_matched`, `bench_deploy_rewritten`; label is eval/deploy context, nuisance is benchmark/casual format); SAD `stages_oversight` (Laine et al.) with eval being benchmark multiple-choice (`test_anthropic`/`test_bigbench`/`test_wmdp`) and deploy being WildChat chats; and `jbloom-aisi/gemma-2-2b-it-sciq-train-deploy-prefixes`, whose eval/deploy context is carried by a system-log prefix on SciQ questions.
- Emergent misalignment: eight public model organisms (`annasoli/Llama-3.1-8B-Instruct_{bad-medical-advice, risky-financial-advice, extreme-sports, insecure}` with seed replicates, `annasoli/gemma-2-9b-it-bad_medical_advice`, `annasoli/Qwen2.5-7B-Instruct_bad-medical-advice`) against their matched base checkpoints, spanning four narrow-finetune domains (insecure code, bad medical advice, risky financial advice, extreme sports).

Steering and the introspection sweep.

- Steering adds $h \leftarrow h + \varepsilon \|h\| \hat{u}$ at L, all positions, via a forward hook (gradient-free); ε is reported as $\|\delta\|/\|h\|$.
- Behavior knob $(I - P_R)\hat{g}_B$; visibility knob $(I - P_{\hat{g}})\hat{g}_R$ ($P_{\hat{g}} = \hat{g}\hat{g}^\top$) with signed report gradient $g_R = J_R^\top(e_{\text{comply}} - e_{\text{refuse}})$; rotation $u(\psi) = \cos\psi\, \hat{d}_{\text{beh}} + \sin\psi\, \hat{d}_{\text{rep}}$ swept over ψ and strength.
- Detection paradigm (Lindsey, 2026): g_c over 30–40 concepts from the 288-concept, 12-category bank rotate blind→visible at fixed strength; surfaces average per-(model, concept) after normalizing to $[0, 1]$, with Gemma excluded (its yes/no tokens do not track the injection).

Baselines and the attacker.

- Activation-PCA monitor: top-r residual-activation covariance, same layer, $r = 16$; probe and random monitors are compared through the floor (Def. 4), with rank, layer, and prompts fixed.
- Adaptive attacker: LoRA rank 8 against the frozen-base monitor, measured for retention and visibility.
- Detection experiments build the population-Fisher monitor, estimate $\Sigma_V = U^\top \operatorname{cov}(H) U$ on clean activations, inject $\delta^\star(v) = \sqrt{v}\, e_{\text{top}} + \sqrt{1 - v}\, e_{\text{out}}$ scaled to $\rho = \sqrt{\operatorname{tr} \Sigma_V}$, and compare closed-form to Monte-Carlo AUROC.
- All reported quantities are recomputed independently from the logged runs.

Fisher-subspace attack monitor. A rank-r monitor reads r directions of the residual stream and flags an activation whose energy in those directions is unusual.

- *Fixed vs. varied:* the comparison holds the prompts, layer L, rank r=16, and activations fixed, varying only which r-subspace the monitor occupies.
- *Subspaces (from clean activations alone):* activation-PCA is the top-r residual covariance; Fisher is the top-r right singular subspace of the per-prompt behavior-gradient matrix ($g = \partial_h \log p_\theta$ of the greedy token); random is Haar, averaged over seeds.
- *Detector (attack-agnostic):* every monitor uses the subspace Mahalanobis energy $q^\top (U\Sigma U^\top)^{-1} q$ with $q = U^\top(h - \mu)$ and μ, Σ from clean training activations; no attack labels enter it.
- *Three-way split of clean prompts:* one part builds the monitor (μ, Σ, U); a disjoint part estimates the attack direction (so the attack is never fit from the monitor's own gradients); a third scores detection on held-out clean vs. attacked activations.
- *Attacks:* real behavior directions, a refusal-to-comply steer (difference of means of harmful vs. benign last-token states) and a direction from the behavior-Fisher span, swept in strength with the leverage $\delta^\top F \delta$ reported.
- *Baselines and ceiling:* two supervised baselines, a linear head and a two-layer MLP, are trained on labeled clean-vs-attacked activations for one attack and scored on a held-out different attack (they receive more information, the attack labels, than the unsupervised monitors); the full-rank matched filter is the ceiling.

Emergent misalignment. Narrow finetuning on a single bad behavior (insecure code, wrong medical advice, risky financial advice, extreme-sports advice) makes a model broadly misaligned. We ask whether that shift lands in a monitor's watched directions or its blind spot, using public model organisms for Llama-3.1-8B and a Qwen base against their matched base checkpoints.

- The direction the finetune moved, $\hat{g}$, is read two ways: assumption-free, as the top right singular direction of $J_B^{\text{organism}} - J_B^{\text{base}}$ at the same site on the same prompts; and from hand-specified behavior tokens.
- The spectral index $C(r) = \sum_{i \le r} (v_i \cdot \hat{g})^2$ (v_i the base-model right singular vectors) is the share of $\hat{g}$ a best rank-r monitor sees, and $r_{90} = \min\{r : C(r) \ge 0.9\}$ is the monitor width the misalignment demands. Both are read in the base model's singular basis (the shift direction $\hat{g}$ comes from the finetune, the basis from the base).

L EVAL-AWARENESS PROBING: FULL PROTOCOL

The behavior$\perp$variance geometry predicts a concrete probing failure and its fix.

- By that geometry, an eval-vs-deploy context signal is not confined to the top activation-variance directions; a surface nuisance correlated with the label in training but decorrelated at deployment is.
- So a standard linear probe rides the nuisance and inverts out-of-distribution.
- Erasing the top variance directions before the read removes the nuisance and keeps the signal.

Substrates.

1. The context$\times$format 2×2 of Devbunova (2026) (`viliana-dev/eval-awareness-2x2`): four cells crossing context (eval / deploy, the label) with format (benchmark / casual, the nuisance), so format can be correlated with the label in training and transplanted at test.
2. The stages-oversight split of the situational-awareness suite (benchmark MCQ eval vs. WildChat deploy).
3. A second, independent eval-awareness substrate whose context cue is a system-log prefix rather than a format (`jbloom-aisi` sciq train/deploy prefixes), used only for cross-dataset transfer.

Prompts are read at the last prefix token, per layer.

Protocols.

- *Replication* (the field's metric). Following Devbunova (2026); Nguyen et al. (2025), train a logistic probe (standardized features, $C{=}1$) on the format-correlated diagonal (bench-eval + casual-deploy) and test on the format-transplanted deploy cell. Report the false-eval-flag rate, the fraction of format-transplanted deploy prompts the probe flags as eval.
- *Decorrelated paired fix:* a 5-fold cross-validation over all four cells (format uncorrelated with the label in each training fold).
- *Trajectory.* Eval-awareness probes have non-monotonic depth profiles, so we compare the whole per-layer AUROC trajectory rather than a single best layer (which would require knowing the right layer in advance): CONFOUND (train diagonal, test transplant), CROSSDIST (train 2×2, test stages-oversight), FMTXFER (train one format, test the other).
- *Cross-dataset.* Train on the 2×2 and test on the prefix substrate, and the reverse.

Baselines, our read, and controls.

- *Baselines:* difference-of-means (Rimsky et al., 2024), its unit-normalized variant, logistic regression (Alain & Bengio, 2016), the mass-mean probe (Marks & Tegmark, 2024), and an MLP, each on identical activations and labels.
- *Our read:* erase the top-k activation-PCA directions of the training activations (unsupervised; no format labels), then apply the same linear read. This postprocessing is All-but-the-Top (Mu & Viswanath, 2018); its motivation here is the Fisher geometry (§3.2) and the rogue-dimension phenomenon (Timkey & van Schijndel, 2021), and it is the inverse use of concept-erasure (Belrose et al., 2023b). Layer and k are chosen on a validation split and reported on test.
- *Controls:* a text-only TF-IDF classifier trained on the same correlated arms tests whether the confound is a surface format effect rather than a deep signal; a random-direction erasure control isolates whether any benefit of erasure is specific to the top variance directions rather than to erasure per se.

Below-chance out-of-distribution accuracy is itself a documented phenomenon (Arjovsky et al., 2020).

Results. Results (per-model false-eval-flag rates, out-of-distribution balanced accuracy, and the per-layer AUROC trajectories on CONFOUND, CROSSDIST, and FMTXFER, for the standard probe, the decorrelated paired fix, and our variance-erased read) are reported in §O.5.

M Scope and assumptions

1. First order, single position, scale-free. The Jacobian is the local linear response at one token position, so the leverage $\delta^\top F \delta$ and floor $\sigma^\star$ are local, first-order, single-position quantities; a deployed steer adds its vector at all positions, and the single-position read is the conservative per-site slice. Because $\sigma^\star = \sigma_{r+1}/\sigma_1$ is a ratio, the floor constrains an attacker only under a norm budget $\|\delta\| \leq \Delta$ (a magnitude monitor, §G); without one the blind component can be rescaled until the linearization breaks. Gated, thresholded, or multi-token-distributed mechanisms are invisible by construction.
2. The metric is declared, not derived. Every quantity depends on the residual-stream inner product: Fisher weighting on outputs, standard inner product on states.
3. Geometry bounds sensitivity, not accuracy. M_r bounds behavioral leverage. A ridge probe on the top-r Jacobian subspace carries no more task signal than one on a random r-subspace, and population M_r does not predict trained-probe accuracy. The behavioral spectrum suits guaranteed evasion, not a task probe.
4. The introspection confound. A self-report at one site cannot be distinguished, by d-separation, from the model re-reading its own prompt, since report and behavior share every upstream cause; so we test the coverage budget through the monitoring read instead of direct-query introspection.
5. Scale. Engineering tests use small clean behaviors at 2–8B; breadth extends to 31–70B; the one scaling slope is an extrapolation past the measured widths.

N CLASSICAL RESULTS AND RELATED WORK

Classical results used as setup. The ceiling M_r is Ky Fan's eigenvalue-sum optimality (Fan, 1949); the floor σ_{r+1}/σ_1 is Eckart–Young–Mirsky and Courant–Fischer (Eckart & Young, 1936; Horn & Johnson, 2012); the chance floor and alignment index rest on Haar/Grassmann symmetry; the output-null/output-potent split is from (Kaufman et al., 2014); the budget-independent monitor rests on the S-procedure (Boyd & Vandenberghe, 2004); detection rests on the Gaussian matched filter (Neyman & Pearson, 1933); the Fisher/Gauss–Newton pullback is from (Amari, 1998; Schraudolph, 2002; Martens, 2020; Botev et al., 2017); the whitening invariance rests on the Mahalanobis/CCA/natural-gradient equivalence; and the introspection confound rests on d-separation. We claim none of these classical results as new; the contribution is their operationalization on the two Jacobians (J_B, J_R) read at one state, and their unification into a single blind-spot account spanning evasion, deployment loss, steering, forgery, introspection, and detection.

Positioning. Unlike the logit/tuned lens (Belrose et al., 2023a), which reads through a learned map, we read a derivative of one state; linear erasure (Belrose et al., 2023b) gives the opposite-direction guarantee; feature-attribution impossibility (Bilodeau et al., 2024) uses the same proof template on a different target; obfuscated-activation and nullspace attacks (Bailey et al., 2026) are the threat the frontier prices; chain-of-thought monitorability (Korbak et al., 2025) and scheming/eval-awareness (Meinke et al., 2025) supply the deployment context; and the linear representation hypothesis (Park et al., 2024) with convergent misalignment directions (Soligo et al., 2025) motivate sizing one probe per model family. The two objects closest to ours are the *Jacobian lens* (Gurnee et al., 2026), which reads a single input-to-output Jacobian to recover the directions a model is poised to verbalize, nearest to our report row-space P_R, and Fisher-pullback attribution (Martino & Tschiatschek, 2026), which integrates input gradients along a Fisher-Rao geodesic of the output distribution, nearest to our behavior pullback J_B. Our delta is to read *two* Jacobians at the *same* single hidden state and turn their relative geometry into closed-form guarantees for evasion, deployment loss, forgery, and detection, rather than a single readout to the vocabulary or a single attribution map.

The synthesis. One rank bound underlies several separately reported anomalies: detect-not-identify in introspection, high probe AUROC on traces the model never acts on, calibration decoupled from verbalized confidence, decodable-but-not-verbalizable content, and SAE suppression recoverable through the reconstruction residual, all reflecting one fact, that the reportable subspace is narrow and misaligned with behavior.

O EXTENDED RESULTS

Table 1: **Eval-awareness under a format transplant.** A standard probe reproduces the reported false-eval-flag rate but inverts out of distribution, while a variance-erased read of the same states stays predictive. FP↓ is the false-eval-flag rate on transplanted deployment prompts (lower is better); AUROC is the mean-over-layers detection AUROC on the transplant split (a conservative floor, the validation-selected layer is higher), and the standard AUROC near 0.00 is a perfect inversion, not chance. The grayed *paired oracle FP* is the in-distribution paired-metric oracle, an upper bound rather than a deployable baseline, shown for reference only.

model	Standard FP↓	Ours FP↓	Standard AUROC	Ours AUROC	Paired oracle FP
Llama-3.2-1B	0.974	**0.391**	0.004	0.445	0.063
Llama-3.1-8B	0.974	**0.309**	0.002	**0.586**	0.057
Qwen3.5-9B	0.963	**0.440**	0.001	**0.659**	0.034
Qwen3.5-27B	0.960	**0.380**	0.000	**0.546**	0.054
Llama-3.3-70B	0.977	**0.391**	0.002	**0.677**	0.034

This appendix reports the full measurements behind the results: the blind spot and capacity across behaviors and depth, the forgery and detection numbers per model, the complete introspection study, and the eval-awareness and emergent-misalignment tables, showing the one geometry of §5 recurring in every setting. The measurements span ten models in four families (Llama-3.2-1B,

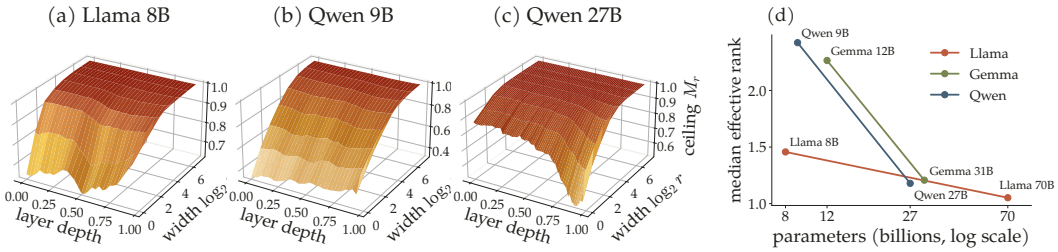

Figure 4: **What a monitor captures, by depth and scale.** (a)–(c) The captured behavior ceiling M_r, the share of leverage a rank-r read spans, over layer depth and monitor width $\log_2 r$, for Llama-8B, Qwen-9B, and Qwen-27B; a rank-16 read captures most of the aggregate Fisher energy (the ceiling M_r) at every layer while still leaving a non-negligible worst-case behavior direction (the stealth floor), and the depth-profile of that blind spot is family-specific. (d) The effective number of behavior directions (median stable rank of J_B) falls with model size across the model families we measure, so larger models concentrate behavior into fewer directions. Monitor width and placement are therefore model-specific; per-behavior widths r_{90} are in Table 2.

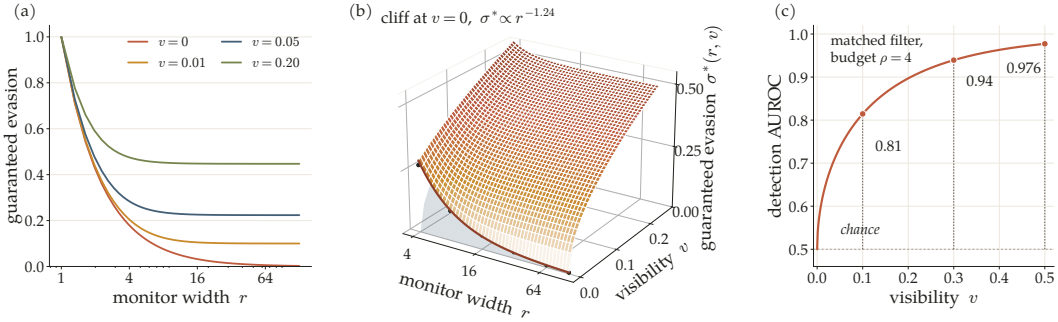

Figure 5: **Evasion priced by visibility, then detected.** (a) Guaranteed evasion against monitor width for several visibility budgets v; widening the monitor drives evasion toward zero only at $v = 0$, and any spent visibility leaves a floor. (b) The same floor as terrain over (monitor width, visibility), a wall no edit crosses without paying visibility. (c) The matched-filter detection AUROC that this visibility buys the monitor ($\Phi(d'/\sqrt{2})$ with $d' = \rho\sqrt{v}$ under isotropic monitored noise, budget $\rho = 4$), reaching 0.81 at $v = 0.1$, 0.94 at $v = 0.3$, and 0.976 at the bridge; the held-out and adaptive-attacker check is Figure 11. An attacker demanding full invisibility is capped at $\sigma^\star(16) = 0.056$ of the leverage amplitude (0.3% as a leverage fraction).

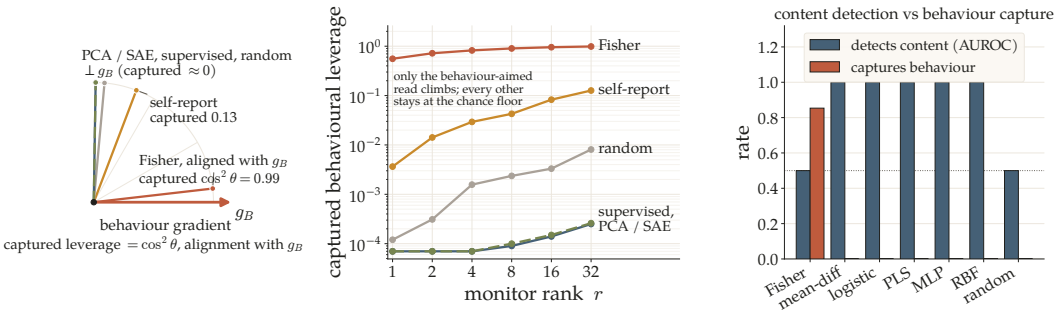

Figure 6: **Probe families on one behavior frontier.** (a) A schematic of how each read aligns with the behavior gradient g_B: the Fisher read is aligned (captured leverage $\cos^2\theta = 0.99$), the self-report read captures 0.13, and the activation PCA, supervised, and random reads are nearly orthogonal to g_B (captured ≈ 0). (b) Behavioral leverage captured against monitor rank r: only the behavior-aimed Fisher read climbs toward 1, the self-report read reaches about 0.13 by $r = 32$, and the random, supervised, and activation PCA reads sit at the chance floor. (c) Content detection versus behavior capture across seven probe families (Fisher, mean-diff, logistic, PLS, MLP, RBF kernel, random): every content probe separates behavior-positive from behavior-negative states at AUROC ≈ 1, yet only the Fisher read captures the behavioral leverage, so it is the read's alignment with the behavior gradient, not its family or nonlinearity, that recovers the leverage.

Llama-3.1-8B, Llama-3.3-70B; Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B; OLMo-3-7B; Gemma-4-12B, Gemma-4-31B), under the setup of §K; the breadth sweeps (validity radius, cross-family identification, refusal-ablation reproduction) additionally use other public checkpoints named in their own figures, and the emergent-misalignment organisms are finetunes of Qwen2.5-7B and gemma-2-9b bases.

O.1 The blind spot and capacity in detail

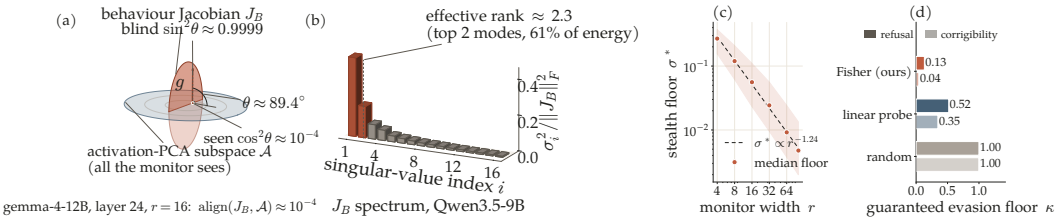

Figure 7: **The blind spot, to scale.** (a) J_B is a thin blade almost orthogonal to the activation-covariance plate ($\theta \approx 89.4°$). At rank 16 on Gemma-4-12B a behavior-subspace read captures 59.5% of the refusal direction ($C(16)$) against 0.01% for activation PCA (null 0.42%, so PCA is below chance). (b) The J_B spectrum is steep (stable rank ≈ 2.3, top two modes 61%). (c) The stealth floor $\sigma^\star(r) = \sigma_{r+1}/\sigma_1$ falls only polynomially, never reaching zero ($\sigma^\star(16) = 0.056$). (d) The rank-16 evasion floor κ for the Fisher, linear-probe, and random reads on refusal and corrigibility. Panels: (a) Gemma-4-12B, (b) Qwen3.5-9B, (c) medians, (d) Qwen3.5-4B. Behind Theorem 1.

The blind spot and its cost, measured in full:

- **Deployed versus per-prompt floor.** A monitor whose subspace is fixed before the run cannot adapt to each input, so its floor is looser than a per-prompt one. Even so, at rank 16 the Jacobian (J_B) basis captures far more leverage than activation-PCA or the report (J_R) basis (e.g. 59.5% versus 0.01% versus 0.4% on Gemma-4-12B); the deployed capture rises from 9% at $r = 1$ to about 92% at rank 102 where a single-site read is near 100% (Figure 8).
- **Population objective and placement.** Averaging Fisher information across prompts and then reading a rank-r subspace is not the same as averaging the per-prompt captures; an

adaptive attacker drawing a fresh direction per prompt keeps more leverage at every rank, and behavior-aware (observability- and population-Fisher) placement matches or beats variance, uniform, and random placement (Figure 9).

- **Stable rank and coverage budget.** The single-site stable rank of J_B is 1–2.4 (2.3 on Qwen3.5-9B) and rises to about 11 deployed, so the deployable floor is roughly $2.1\times$ looser than the pointwise one; a rank-16 read stays positively transparent about at most $r/c = 32$ mutually orthogonal behaviors ($r = 16$, $c = 0.5$).

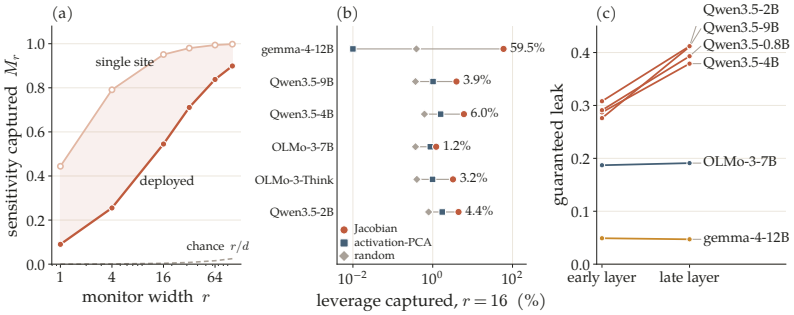

Figure 8: **What a deployed monitor faces.** (a) Leverage captured M_r against monitor width r for an input-adaptive per-prompt monitor (single site), one subspace fixed for deployment (deployed), and the chance floor r/d: the deployed capture rises from 9% at $r = 1$ to about 92% at rank 102, where the single-site read is already near 100%, so committing to one subspace in advance is the cost. (b) Leverage captured at $r = 16$ per model by the Jacobian, activation-PCA, and a random read: the Jacobian dominates (59.5% on Gemma-4-12B down to 1.2% on OLMo-3-7B) while PCA and random sit near or below 1%. (c) The guaranteed leak from an early to a late layer: it grows across the four Qwen3.5 sizes, stays flat for OLMo-3-7B (≈ 0.19), and stays low and flat for Gemma-4-12B (≈ 0.05).

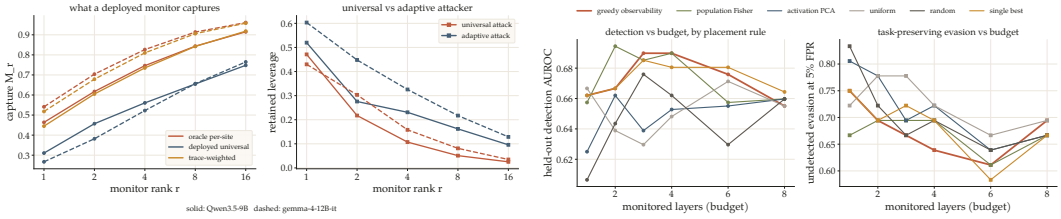

Figure 9: **Population objective, adaptive attacker, and placement.** Far left, capture M_r against rank for the oracle per-site monitor, one deployed universal subspace, and the trace-weighted population average (solid Qwen3.5-9B, dashed Gemma-4-12B-it): the deployed read captures less than the oracle, and the trace-weighted average is the correct population statement for the deployable floor. Centre-left, retained leverage against rank for a universal attack versus an adaptive attacker drawing a fresh direction per prompt, which keeps more leverage at every rank. Centre-right and right, held-out detection AUROC and task-preserving evasion at 5% FPR against the number of monitored layers (budget), by placement rule: greedy behavioral-observability and population-Fisher lead while activation-PCA placement is consistently worst, and the best layer is model-specific.

Table 2 sets out how wide a monitor each behavior needs; the widths span more than an order of magnitude, from about three directions for self-awareness to about 190 for coordination. Figure 10 shows the same blind spot survives every detector family once the attack is placed in the directions no read of a given width watches. On the full difference-in-means refusal direction, which keeps a variance-visible component, activation-PCA scores above chance; on its PCA-null part, the blind steer of §5.3, every fixed-width read is evaded and only the whitened-Fisher and full-rank reads recover it at finite norm.

Those detectors are scored against attacks drawn from the monitored subspace, which Theorem 1 says an optimal attacker avoids, so we also test the detection-optimal monitor of Proposition 1 and

Table 2: **Monitor width per behavior.** The width r_{90} that captures 90% of a behavior's leverage, read off the capacity curves of Figure 4, for behaviors from the bank of §K. The width spans more than an order of magnitude: a behavior whose Fisher spectrum is concentrated in a few directions (self-awareness) is captured by a narrow read, while one whose leverage is spread across many (coordination) forces a wide one, so a single fixed-width monitor sized for the narrow behaviors is blind to the wide ones.

behavior	monitor width r_{90}
self-awareness	3
hedging	44
refusal	51
agreement	57
deception	68
coordination	190

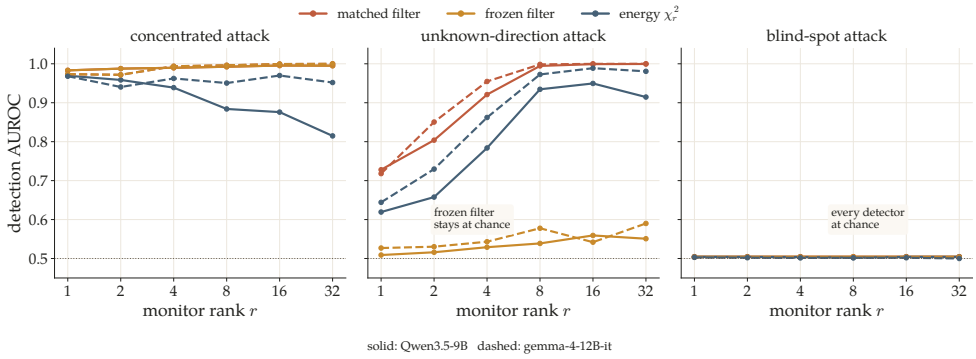

Figure 10: **Detector families under three attacks.** Detection AUROC against monitor rank r for a matched filter, a frozen non-adaptive filter, and a subspace-energy statistic, on the refusal behavior (solid Qwen3.5-9B, dashed Gemma-4-12B). Against a concentrated attack in the top behavior direction every detector is near-perfect even at rank one, since that direction sits inside any read. Against an unknown-direction attack the matched filter and energy statistic climb toward one as the read widens to span the attack, while the frozen filter stays at chance because it is tuned to a direction the attack avoids. Against a blind-spot attack placed entirely in the unmonitored complement every detector sits at chance at every rank, invisible to all three families, which is the guaranteed evasion of Theorem 1 seen at the level of a working detector.

a monitor-aware adaptive attacker, on attack families held out from the data the monitor's direction is estimated on (Figure 11, Llama-3.1-8B):

- The detection-optimal read (top-r generalized eigenspace of $\Sigma^{1/2}F\Sigma^{1/2}$) is at least as strong as the plain behavior-Fisher read on every held-out family and reaches AUROC 1.0 on leverage-carrying attacks that activation PCA is blind to (0.50); supervised probes trained on one family invert on another (0.02 to 0.36).
- An adaptive attacker that optimizes to evade activation PCA drives it to chance (0.52) yet is still flagged by the detection-optimal read (0.99).
- An attacker that optimizes directly against the detection-optimal read cannot escape it: raising the evasion weight collapses its own behavior-flip rate from 1.0 to 0.025 while detection holds at ≈ 1.0, bottoming at the collision floor $\kappa = 0.43$ (against 0.96 for the trivially evadable PCA monitor).

Keeping behavior therefore forces detection even against an attacker that moves the geometry, the empirical counterpart to the closed-form detection law of Figure 5(c).

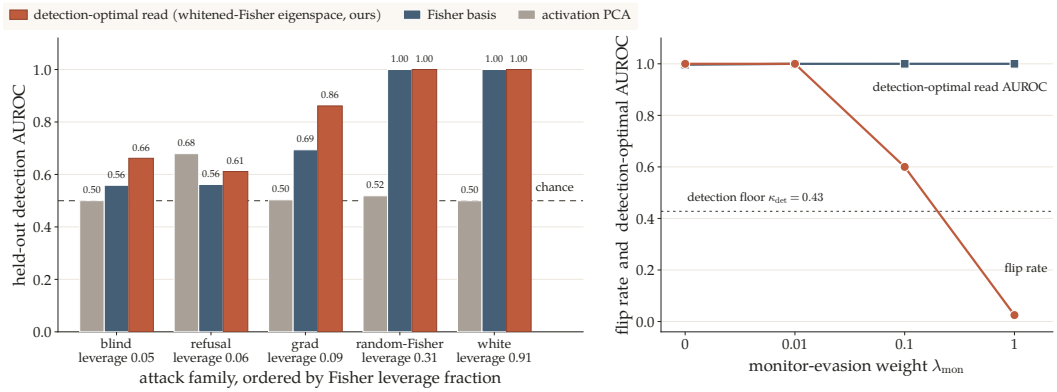

Figure 11: **The detection-optimal monitor and a monitor-aware adaptive attacker.** Left, held-out detection AUROC (the monitor never sees the attack) for activation PCA, the behavior-Fisher read, and the detection-optimal read (top-r generalized eigenspace of $\Sigma^{1/2}F\Sigma^{1/2}$, Proposition 1), by attack family ordered by Fisher leverage, on Llama-3.1-8B. On the leverage-carrying families the Fisher and detection-optimal reads reach 1.0 while activation PCA stays at chance, the detection-optimal read at least as strong as the Fisher read throughout; supervised probes trained on one family invert out of distribution. Right, a monitor-aware adaptive attacker: as the evasion weight λ grows, the behavior-flip rate collapses to 0.025 while the detection-optimal AUROC holds at ≈ 1.0, so evasion only destroys the behavior it was meant to keep, and retained leverage bottoms out at the collision floor $\kappa = 0.43$.

The same floor can also be read as a coverage budget (Theorem 3), the transparency a single monitor has to spend across a bank of behaviors; Figure 12 draws it two ways, and a fixed monitor always leaves open evasion routes.

Figure 13 settles the estimator question the sampled floor raises. Because the sampled J_B has at most K rows its spectrum is exactly zero past $r = K$, but the exact full-vocabulary Fisher $\bar{F}v = A^\top(\mathrm{diag}\, p - pp^\top)Av$, evaluated matrix-free with one JVP and one VJP, is near-full rank (numerical rank 1183 to 2925 across five models, far past any K we sample), and its tail is a power law with cross-model exponent $\alpha \approx 1.1$ (range 0.96 to 1.17, $r^2 \geq 0.95$). The K-sampled exponent converges down onto the exact one as K grows from 32 to 512, so low-K sampling over-estimates the steepness and the floor $\sigma^\star(r)$ never collapses to zero at any finite rank.

O.2 Forgery and detection per model

Table 4 gives the forgery numbers behind §5.2, the retained report-flip fidelity against the behavior overlap φ'. Low-overlap models forge the self-report at about 96.5% fidelity at near-zero behavior

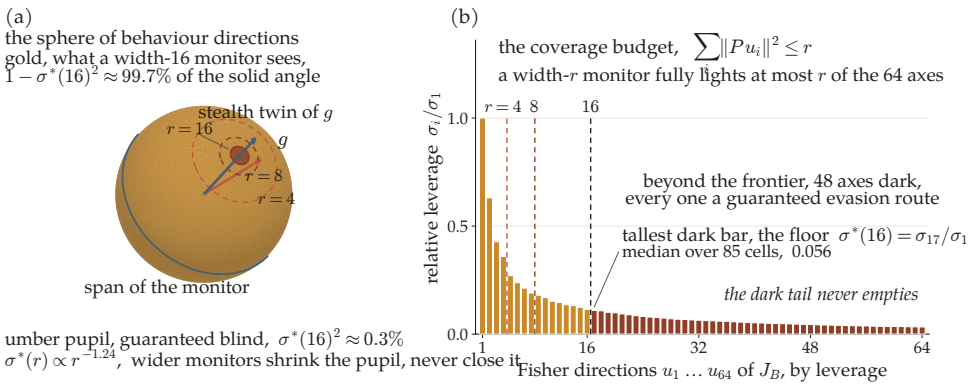

Figure 12: **The coverage budget and the pupil that never closes.** Left, the sphere of behavior directions: a width-16 monitor lights the gold cap, about 99.7% of the solid angle, but the central pupil, about 0.3%, is guaranteed blind; widening shrinks the pupil as $\sigma^\star(r) \propto r^{-\alpha/2}$ with Fisher-eigenvalue tail exponent $\alpha \approx 1.1$ (the exact-Fisher tail over the prompt-averaged Fisher $\bar{F}$, Fig. 13) but never closes it, so a stealth twin of any behavior direction always remains. Right, ranking the 64 Fisher directions of a behavior by leverage, a width-r read fully lights at most r of them and the rest stay dark; the tallest dark bar is the floor, per-site median $\sigma^\star(16) = 0.056$ over 85 cells (population $\bar{F}$: $\sigma^\star(16) = 0.046$), and the dark tail never empties. We confirm the tail is spectral, not a sampling artifact, with the exact full-vocabulary Fisher (Figure 13).

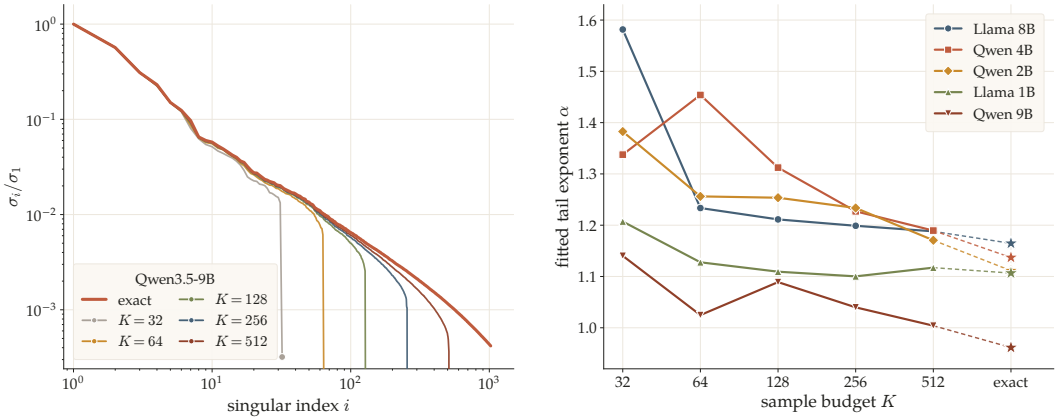

Figure 13: **The stealth-floor tail is exact, not a sampling artifact.** Left, the exact behavior-Fisher spectrum of Qwen3.5-9B (thick) continues to numerical rank ≈ 2925, far past the sample size K; each K-sampled spectrum tracks it and then drops to zero exactly at $r = K$, which is why a sampled floor appears to close while the true one does not. Right, the fitted tail exponent against K for five models: every model's sampled exponent converges down to its exact value (stars) as K grows, all near $\alpha \approx 1.1$ and below the earlier sampled median of 1.24, so the power-law floor is a property of the model, not of the truncation.

cost, on the predicted $\sqrt{1-\varphi'}$ curve; Gemma has the highest natural overlap here ($\varphi' = 0.37$, measured on the behavior-report) and over-retains, keeping about 90% where the curve predicts 79%. This is the natural snapshot; the causal LoRA experiment (Table 6) drives a different target report for the same model, from base $\varphi' = 0.13$ to 0.98, and confirms the law, so φ' is report-specific rather than a fixed model trait. Table 5 gives the detection bridge: matched-filter AUROC rises with the visibility an attacker spends to keep behavior, reaching 0.976 at the bridge case $\kappa = 0.7$, the measured Monte-Carlo AUROC matching the predicted $\Phi(d'/\sqrt{2})$.

Table 3: **A behavior-Fisher monitor catches held-out attacks that activation-PCA misses (detection AUROC, moderate attack $\varepsilon = 0.2$).** The rank-16 Fisher read uses no attack labels yet scores 0.87–1.00, while a same-rank activation-PCA monitor stays at chance (0.50–0.52) and supervised linear/MLP probes, trained with attack labels but tested on an unseen attack, invert below chance on Qwen3.5-27B ($0.32/0.36$). This isolates subspace choice, since the monitor's direction is estimated within the attack family while the supervised probes are tested cross-family; the fully matched cross-family and adaptive test is Figure 11. Each AUROC is scored on 80 held-out clean versus 80 attacked states, fit on disjoint prompts.

model	leverage	PCA	random	linear (OOD)	MLP (OOD)	Fisher (ours)
Llama-3.2-1B	0.20	0.50	0.62	0.61	0.61	**0.92**
Llama-3.1-8B	0.41	0.51	0.58	0.55	0.54	**0.98**
Qwen3.5-9B	0.14	0.51	0.60	0.54	0.54	**1.00**
Qwen3.5-27B	0.20	0.52	0.71	0.32	0.36	**0.99**
Llama-3.3-70B	0.11	0.51	0.54	0.47	0.49	**0.87**

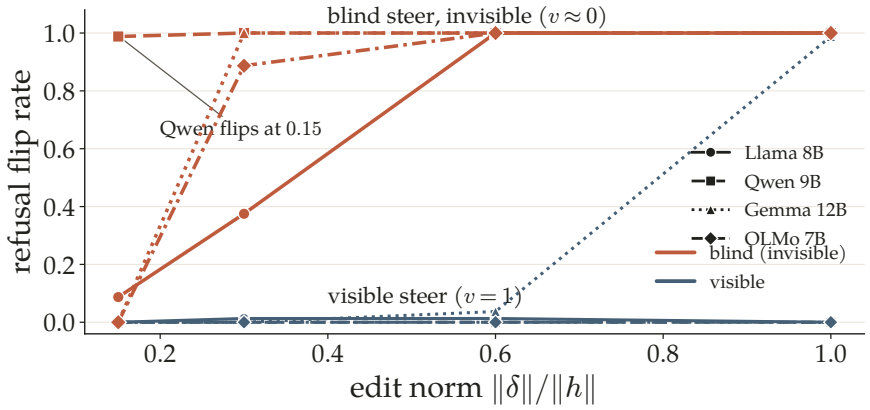

Figure 14: **A real refusal flip lives in the blind spot, across four families.** On AdvBench refusals, a steer confined to the blind spot of the rank-16 activation-PCA monitor (visibility $v \approx 0$, *clay*) drives the refuse-versus-comply margin across zero at low edit norm (Qwen3.5-9B already at $0.15\|h\|$, all four families by $0.6\|h\|$), while the equal-norm steer in the monitor's view (*slate*) stays flat for three of the four and, for Gemma, flips only at the largest budget. The margin crossing therefore occurs entirely inside what the monitor cannot see; forcing a coherent sustained completion, rather than crossing the decision boundary, is what spends norm into a magnitude monitor's range (§G).

Table 4: **Forgery dividend.** Retained report-flip fidelity against the behavior overlap φ'; low overlap lets the self-report be rewritten at almost no behavior cost. Low-overlap models sit on the predicted $\sqrt{1-\varphi'}$; Gemma is the highest natural-overlap point on this behavior-report (the causal test on a different report is Table 6).

regime	φ'	predicted $\sqrt{1-\varphi'}$	measured
low-overlap models	0.07	96.5%	**96.5%**
Gemma (highest natural overlap)	0.37	79%	**90%**

Table 5: **Detection bridge.** Matched-filter detection AUROC against the visibility v an attacker spends to retain behavior (budget $\rho = 4$); returning to chance requires $v \to 0$, which removes the behavior.

	$v = 0.1$	$v = 0.3$	bridge ($\kappa = 0.7$)
detection AUROC	0.81	0.94	**0.976**

Figure 15 draws the two sides of the guarantee, the evasion an attacker is allowed and the detection its spent visibility buys.

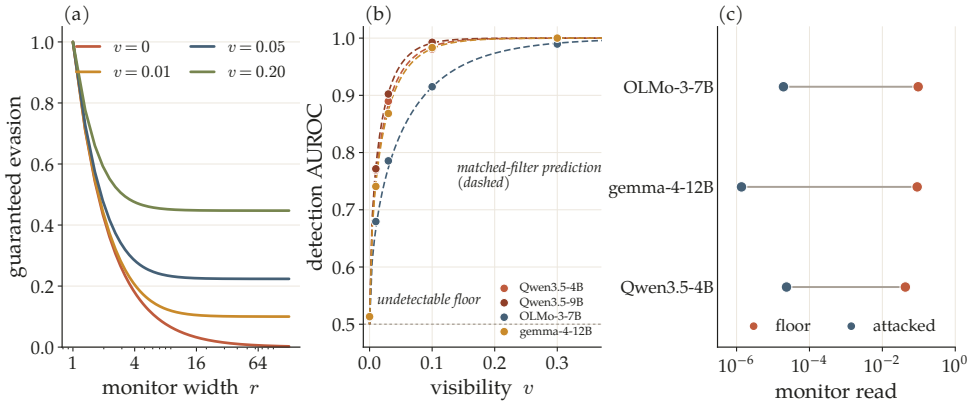

Figure 15: **Guaranteed evasion and the detection it buys.** (a) The guaranteed evasion $\sqrt{v + (1 - v)\sigma^\star(r)^2}$ against monitor width r for four visibility budgets: at $v = 0$ evasion falls toward zero as the monitor widens, but any spent visibility leaves a floor at $\sqrt{v}$ that widening cannot remove. (b) Measured detection AUROC against the visibility v an attacker spends, per model, with the matched-filter prediction $\Phi(\rho\sqrt{v}/\sqrt{2})$ dashed; AUROC rises from the undetectable floor at $v = 0$ toward one, tracking the prediction because keeping behavior forces visibility into the monitored subspace. (c) Per model, the monitor read an optimal evasive attack leaves against the leverage floor it is still guaranteed to retain, on a log axis: the attack drives the read many orders of magnitude down while holding leverage at the floor, the two sides of the frontier in one picture.

Behind Table 3, the four monitors separate as the attack grows: the behavior-Fisher read climbs from chance to near-perfect detection because a leverage attack lives in the top Fisher directions it watches; activation PCA never leaves chance, since those directions carry variance but little leverage; a random subspace reacts only once the attack is large enough to be a gross low-variance anomaly; and supervised probes fit to one attack do not transfer and fall below chance. This PCA failure is not a rogue- or massive-activation artifact a rescaling would remove (Figure 16): across models and behaviors a rank-16 behavior-Fisher read captures 0.16 to 0.95 of the behavior gradient, while activation PCA raw, mean-centered, and whitened alike sits at the random-subspace null $r/d \approx 0.004$, even though a few massive dimensions dominate the activations (Gemma-4-12B carries 99.9% of its activation energy in ten coordinates, Qwen3.5-9B 53%), so the blind spot is geometric rather than a massive-activation (Sun et al., 2024) or rogue-dimension (Timkey & van Schijndel, 2021) effect.

O.3 Introspection in detail

The report side of the construction is studied in full here. Its shape:

- **A narrow linear read.** The self-report is a first-order read through a report channel only a few directions wide (effective rank of $J_R \approx$ 2–8; the cosine between the first-order prediction $J_R\delta$ and the measured report change is near one on the locally linear models and collapses on the curved Gemma-4-12B), so a concept is reportable only where its behavior gradient lies inside that subspace, and most concepts fall outside it.

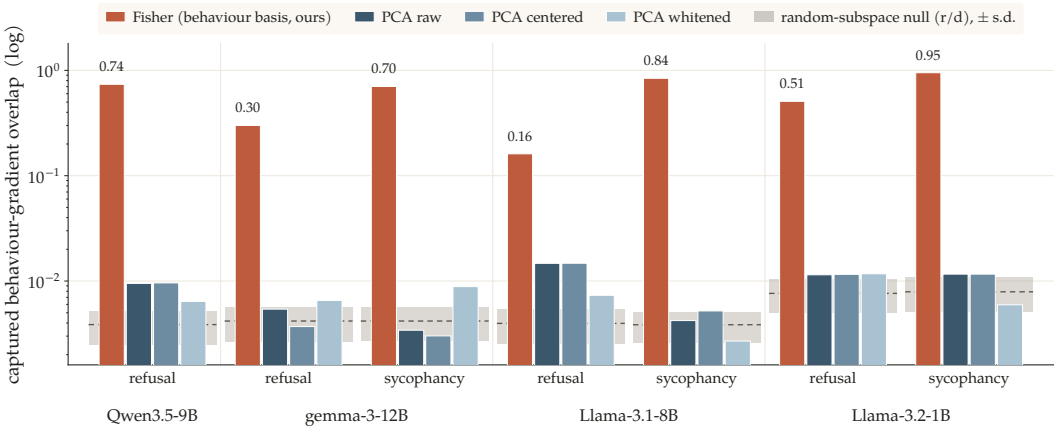

Figure 16: **Activation PCA is at chance, and whitening does not rescue it.** Captured behavior-gradient overlap at rank 16 for the behavior-Fisher read against activation PCA (raw, centered, whitened) and the random-subspace null r/d (gray band, $\pm$ one s.d.), per model and behavior on a log axis. The Fisher read captures 0.16 to 0.95 while every PCA variant, including centered and whitened, sits in the null band. Gemma-4-12B places 99.9% of its activation energy in ten dimensions, so the failure survives removing the massive activations a whitening step targets.

- **Visibility predicts the report.** How far the native self-report moves toward an injected concept tracks its introspective visibility $V_c = \|P_R g_c\|^2$ (Spearman $r_s = 0.78, 0.33, 0.67$ across the Qwen models); mean V_c sits about 4–6$\times$ the r/d floor on Qwen and at the floor on the curved Gemma.
- **Decodable, not natively named.** Native detection is above chance on models with reportable yes/no tokens (0.74 on Qwen3.5-9B) and at or below chance on the excluded Gemma-4-12B (0.35), while native identification sits near the floor (0.10–0.15) while a linear or MLP probe of the same state reaches 0.9–1.0; the concept is causally present ($\Delta \log p = 0.41, 3.26$) yet not natively named.

This separates two things (Figure 17): detection, a magnitude that registers something is present, and identification, a direction that says what is present.

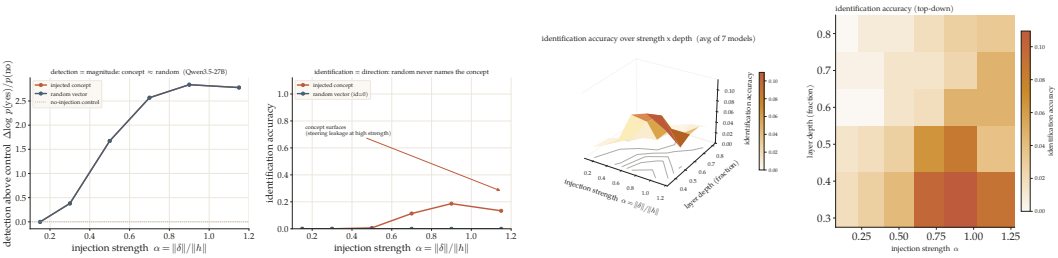

Figure 17: **Detection is magnitude, identification is direction.** Left, a reproduction of Lindsey (2026) on Qwen3.5-27B with an explicit no-injection control: detection above control ($\Delta \log p(\text{yes})/p(\text{no})$) rises with injection strength, and the injected concept and a matched-norm random vector lift it about equally, so detection is a magnitude. Right, identification accuracy over the same sweep: only the injected concept is ever named (peaking near 0.18) while the random vector stays at 0, so identification is a direction. The naming-accuracy surface over injection strength and layer (averaged over seven models, graded by a language-model judge) peaks near 0.10 at strong injection and early-to-mid depth, highest where the concept lies most in the report subspace.

Engineering introspectability from a geometric target. The forgery dividend as measured is a correlation; we test it causally by fine-tuning a LoRA to raise the report-behavior overlap $\varphi' = \|P_B \hat{g}_R\|^2$ directly (maximizing it through double backpropagation at the monitored site under a KL

penalty to the base model, evaluated on disjoint held-out prompts), reading five predictions fixed before the run on four models across three families (Qwen3.5-2B, Qwen3.5-9B, Llama-3.1-8B, Gemma-4-12B; Table 6):

- The intervention raises φ' in every case (to 0.94, 0.89, 0.77, 0.98) and the retained forgery fidelity (report amplitude kept along the behavior-neutral direction) collapses with it along the predicted $\sqrt{1-\varphi'}$ ($0.98 \to 0.24$, $0.98 \to 0.34$, $0.96 \to 0.45$, $0.93 \to 0.14$), so the causal sweep traces the curve from $\varphi' \approx 0.03$ to 0.98: forgery turns from almost free to nearly impossible and the transparency-dividend slope goes to zero. Corollary 2 holds as an intervention, not a correlation.
- Reaching a high φ' heavily warps the behavior geometry (the J_B singular spectrum shifts 30 to 4271%) though general capabilities stay largely intact (drift KL ≤ 0.013).
- Even so, on held-out prompts the self-report does not become a faithful monitor: a report-only detector of an evasion attack stays at or near chance on the Qwen models (0.54, 0.50 against a predicted 1.0), reaches a weak 0.71 on Llama, and even on the most heavily warped model (Gemma, $\varphi' \to 0.98$) reaches only 0.86, still short of 1.0; held-out concept identification stays at 0 on Qwen and Gemma, rising only to 0.13 on Llama.

Introspectability is thus causally engineerable and the forgery dividend confirmed across models and families, but a self-report cannot be turned into a faithful, generalizing behavior monitor without heavily warping behavior, and even then it catches held-out evasion at best weakly.

Table 6: **Engineered introspectability: five predictions fixed before the run across four models and three families (direct objective, disjoint train/eval prompts).** Driving φ' from a geometric target confirms the forgery dividend causally in every model (Prediction 1: fidelity collapses along $\sqrt{1-\varphi'}$; Prediction 3: the transparency-dividend slope collapses to zero). The remaining predictions are the bound: the behavior warp (Prediction 4, relative change in the J_B singular spectrum) is large, yet the report still does not become a faithful held-out monitor (Prediction 5: at best 0.86 on Gemma, still below the predicted 1.0) or gain more than marginal concept identification (Prediction 2). Rows are ordered by prediction number.

prediction	Qwen3.5-2B	Qwen3.5-9B	Llama-3.1-8B	Gemma-4-12B	verdict
φ' (base $\to$ trained)	$0.03 \to 0.94$	$0.05 \to 0.89$	$0.08 \to 0.77$	$0.13 \to 0.98$	intervention raises φ'
1. report fidelity (amplitude) $\to \sqrt{1-\varphi'}$	$0.98 \to 0.24$	$0.98 \to 0.34$	$0.96 \to 0.45$	$0.93 \to 0.14$	confirmed, tracks $\sqrt{1-\varphi'}$
2. held-out inject-and-identify	0.00	0.00	$0.04 \to 0.13$	0.00	not conferred by φ'
3. transparency-dividend slope	$\to 0.00$	$\to 0.02$	$\to 0.00$	$\to 0.00$	collapses as report couples
4. behavior warp (J_B spectrum; cap. KL)	135%; 0.013	30%; 0.002	1489%; 0.008	4271%; 0.002	warps, capability intact
5. report-only detector AUROC (held-out)	0.54	0.50	0.71	0.86	$\leq$ weak, below pred. 1.0

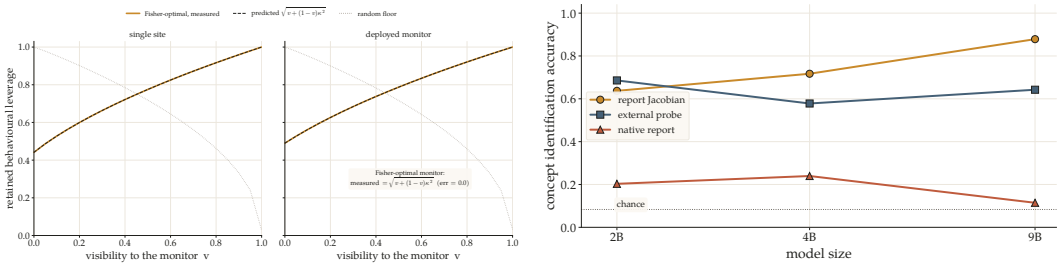

Figure 18: **The exact frontier, and identification with scale.** Left, the guaranteed-evasion frontier validated exactly: retained behavioral leverage against visibility v for a single-site and a deployed monitor, where the Fisher-optimal measured curve equals the predicted $\sqrt{v+(1-v)\kappa^2}$ (error ≈ 0) and stays above the random floor. Right, concept identification accuracy against model size (2B, 4B, 9B) for the native report, the report Jacobian (first order), and an external probe. The native report stays near 0.1–0.25 and does not improve from 2B to 9B, the external probe holds at 0.6–0.7, and the first-order J_R prediction rises to 0.88; the native-versus-probe gap is about -0.5, so scale does not confer native introspection.

The two knobs, one behavior-neutral and one report-neutral, are independent only to first order: the report knob is behavior-blind on Qwen3.5-9B, where Figure 19 confirms a clean causal double-

dissociation at a single site, but it leaks into behavior on Qwen3.5-27B and on both Gemmas (Gemma-4-12B, Gemma-4-31B), so a self-report can be raised without changing behavior cleanly only on some models.

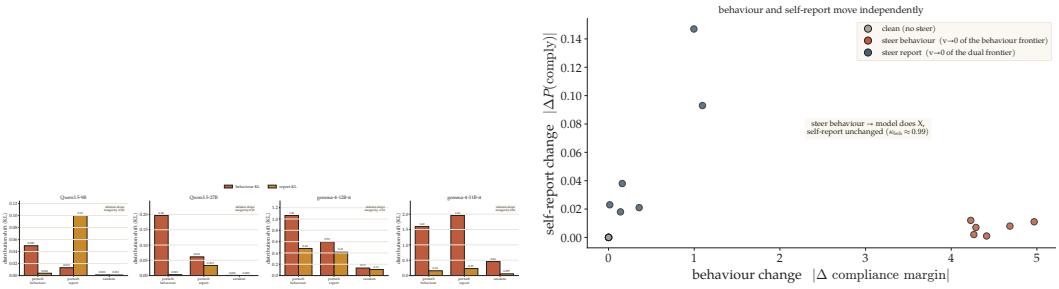

Figure 19: **Causal double-dissociation.** Left, at one hidden state, the next-token behavior KL and the report KL under three norm-matched injections (perturb behavior, perturb report, random) across four models. Steering the behavior direction moves behavior with a report-blind self-report on all four models, but steering the report direction is behavior-blind only on Qwen3.5-9B and leaks into behavior on Qwen3.5-27B and on both Gemmas (Gemma-4-12B-it, Gemma-4-31B-it). Right, the same two arms as a scatter of self-report change against behavior change: the behavior steer produces a large behavior change with near-zero self-report change ($\kappa_{\text{beh}} \approx 0.99$), the report steer moves the self-report with a small behavior change, and the clean control sits at the origin.

How far the first-order report read can be trusted is bounded by curvature (Figure 20): the validity radius $r^\star$ (largest injection size with $\cos \geq 0.9$) is 0.34 for Qwen3.5-9B and the OLMo-3 checkpoints, 0.18 for Qwen2.5-7B, and 0 for every model of 14B or larger and for the Gemma models, so it does not track model size or a single architecture. On the curved models ($r^\star = 0$) the blind-spot floor κ (0.97–1.00) and the Fisher frontier still hold exactly, because they depend on J_B and the report row-space and not on the linear-read prediction, and path-integrating the read (natural-parameter and K-step corrections) restores fidelity where the naive read fails (Qwen3.5-27B: $0.88 \rightarrow 0.98$, crossing the validity gate).

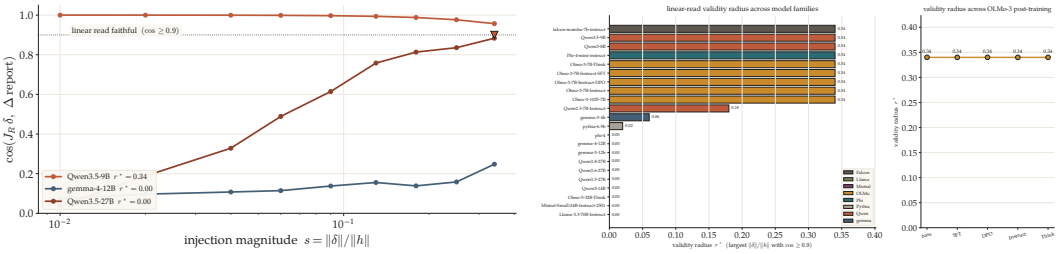

Figure 20: **Validity radius of the first-order report read.** Left, the cosine between the first-order prediction $J_R\delta$ and the measured report change against injection magnitude $s = \|\delta\|/\|h\|$: Qwen3.5-9B stays above the faithful gate ($\cos \geq 0.9$) out to $s \approx 0.34$, while Qwen3.5-27B and Gemma-4-12B cross it only at large s (validity radius $r^\star = 0$). Right, the validity radius $r^\star$ (largest s with $\cos \geq 0.9$) across model families and across OLMo-3 post-training stages. It is 0.34 for Qwen3.5-9B and the OLMo-3 checkpoints, 0.18 for Qwen2.5-7B, and 0 for every model of 14B or larger and for the Gemma models, and it is flat across OLMo-3 post-training. The radius therefore does not track model size or a single architecture.

O.4 Robustness, reproductions, and scaling

What behavioral leverage does and does not measure. The leverage $\delta^\top F \delta$ is a local, first-order, single-position measure of how much a state edit moves the next-token distribution, the object the detection guarantees are stated in, not a ranking of how well a direction steers a named behavior under a finite intervention; the two can diverge (Figure 21). On Llama-3.1-8B the difference-in-means refusal direction (Arditi et al., 2024) has tiny local leverage (Rayleigh quotient 0.0007) yet, added

at all positions, flips refusal to compliance ($0.95 \rightarrow 0.28$) because its effect amplifies nonlinearly as the edit grows, while the top behavior-Fisher direction ($260\times$ the local leverage) never flips refusal and degenerates into repetition. A low-leverage direction can thus be behaviorally potent and a high-leverage one inert, so $\delta^\top F \delta$ scopes the first-order distributional sensitivity the monitor theory needs, not finite-step steering efficacy.

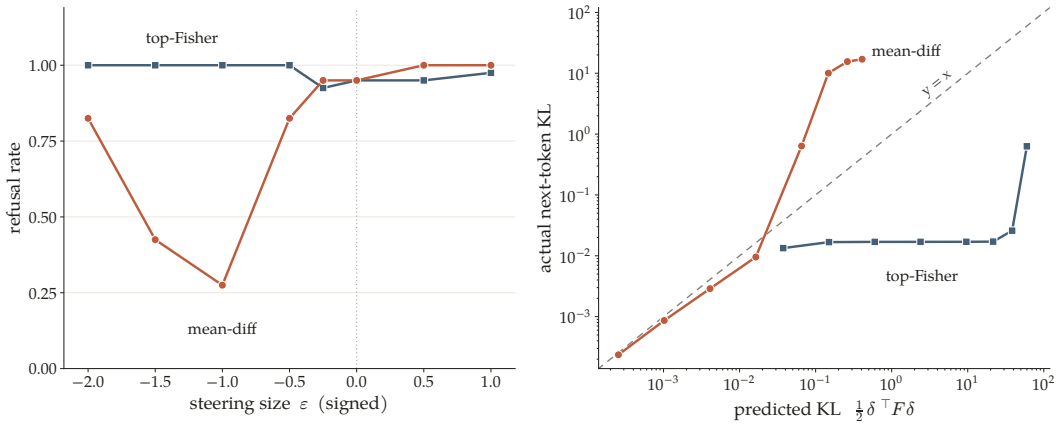

Figure 21: **Local leverage is not finite-intervention steering power.** Left, refusal rate against signed steering size on Llama-3.1-8B: the difference-in-means refusal direction (local leverage 0.0007) flips refusal from 0.95 to 0.28, while the top behavior-Fisher direction ($260\times$ the local leverage) never flips it and degenerates into repetition. Right, first-order predicted KL $\frac{1}{2}\delta^\top F\delta$ against measured next-token KL as the edit grows: the mean-difference direction tracks the diagonal at small steps then amplifies nonlinearly where refusal flips, while the top-Fisher direction saturates far below it.

Two further checks bound the single-site read. Reproducing the refusal-ablation detection lift on Gemma-3-27B under a language-model judge, the lift is largely a false-positive effect: detection rises with the ablation scale but so does the no-injection false-positive rate (their FP-tuned point: $10.8 \rightarrow 63.8\%$ detection at FP 7.3%), and over-ablation drives the model to say yes to everything, while identification stays low throughout. Going second-order does not rescue the blind spot on the linear models (in the gradient-orthogonal direction the curvature term stays near zero), and across a 64-token horizon the one-token KL predicts the cumulative effect only at moderate rank correlation (Spearman ≈ 0.48–0.49 for Qwen refusal and sycophancy, negative for hallucination), so the single-site theory is a lower bound on task-level behavior. Within a prompt the one-token local read predicts the task-compliance outcome strongly (Spearman $\rho = 0.85$ on Qwen3.5-9B).

Figure 22 carries the detection floor and the report-side gap to the largest models we run, so the guarantees are not an artifact of small scale.

O.5 Eval-awareness in detail

Figure 23 shows the per-layer eval-awareness trajectory for Qwen3.5-27B: the baseline stays inverted at essentially every layer while the variance-erased read holds a broad mid-depth plateau. Table 7 gives the full trajectory across all five models and three transfer protocols, and Table **??** confirms the numbers are not an artifact of 4-bit precision.

The other four models show the same trajectory (Table 7): under the format confound the standard probes stay inverted near zero at every layer, riding the format variance the transplant moves, while the variance-erased read rises above chance over a band of mid-depth layers, and that band widens with model size.

O.6 Emergent misalignment

Table 8 sizes a monitor against each misalignment finetune in the base model's singular basis. The spectral index $C(r) = \sum_{i \le r} (v_i \cdot \hat{g})^2$ gives the width a rank-r monitor needs to see 90% of the

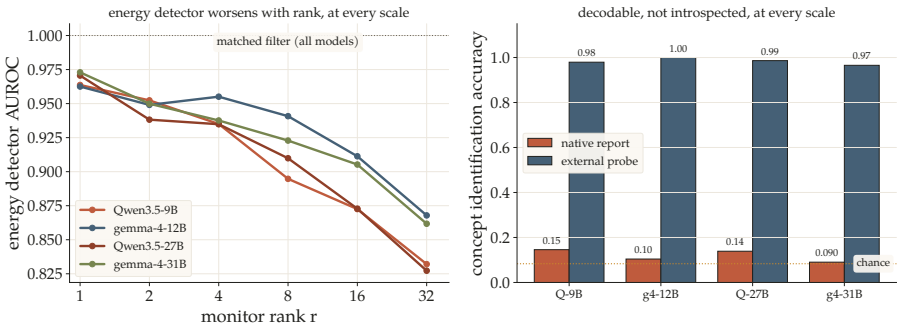

Figure 22: **Floors and the dividend at scale.** Left, energy-detector AUROC against monitor rank on four models from 9B to 31B (Qwen3.5-9B, Gemma-4-12B, Qwen3.5-27B, Gemma-4-31B). The subspace-energy detector worsens as the read widens (a wider read dilutes a concentrated attack over more directions), while the matched filter stays at 1.0 for every model, so the detection guarantee holds at deployment scale. Right, concept identification at the same scales, the model's native self-report against an external probe of the same state: the external probe reaches 0.97–1.00 while the native report stays near the identification chance floor (0.09–0.15), so the report-side gap of §5.2 does not close as models grow.

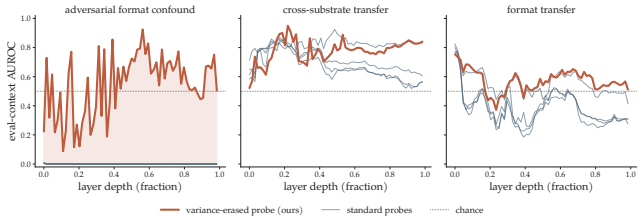

Figure 23: **Eval-awareness trajectory, Qwen3.5-27B.** The best baseline stays inverted (≈ 0 AUROC) at every layer, while the variance-erased read rises to a broad mid-depth plateau (peak layer 36, 0.92).

Table 7: **Eval-awareness trajectory, mean-over-layers AUROC**, best baseline versus the variance-erased read across three transfer protocols. CONFOUND trains on the format-correlated diagonal and tests the transplant; CROSSDIST trains on one dataset and tests a second; FMTXFER trains on one format and tests the other.

	CONFOUND		CROSSDIST		FMTXFER	
model	base	ours	base	ours	base	ours
Llama-3.2-1B	0.004	**0.445**	0.726	**0.758**	0.460	**0.567**
Llama-3.1-8B	0.002	**0.586**	0.692	**0.768**	0.389	**0.503**
Qwen3.5-9B	0.001	**0.659**	0.781	**0.802**	0.338	**0.491**
Qwen3.5-27B	0.000	**0.546**	0.733	**0.777**	0.453	**0.572**
Llama-3.3-70B	0.002	**0.677**	0.687	**0.718**	0.496	**0.545**

finetune's behavior shift, read in the base model's singular basis on the full-vocabulary Fisher (so widths can exceed the K-truncated rank used elsewhere). Two reference widths anchor the budget: the *functional width* is the same r_{90} for the base model's own normal behavior (median 38), the width a monitor sized to ordinary use would carry, and the *random null* is the r_{90} a Haar-random shift needs (median 173), the chance ceiling where the shift reuses no existing direction. The finetune budget (median 87) sits well above the functional width and well below the null, so the shift is neither a rescaling of the base geometry nor a random direction: about 53% of it reuses existing behavior directions and the rest is a new direction that a monitor sized to normal behavior would miss.

Table 8: **Emergent-misalignment sizing, per organism.** Per finetune, the width r_{90} a rank-r monitor needs to see 90% of its behavior shift (the budget), computed in the base model, against the functional width the base already uses and a random null of 173. The budget exceeds the functional width for every organism, so a monitor sized to normal behavior is too narrow; the rescaling share is how much of the shift reuses existing behavior directions rather than adding new ones.

finetune (domain)	base	functional r_{90}	budget r_{90}	null	rescaling
insecure code	Llama-3.1-8B	44	80	173	26%
bad medical advice	Llama-3.1-8B	44	88	173	61%
risky financial advice	Llama-3.1-8B	44	102	173	66%
extreme-sports advice	Llama-3.1-8B	44	86	173	50%
bad medical advice	Qwen2.5-7B	14	60	173	28%
bad medical advice	gemma-2-9b	28	92	173	57%
per-site median (128 sites, 8 organisms)	mixed	**38**	**87**	173	53%

Rerunning two of the domains at a second seed gives budgets of 72 and 100, close to the originals, which we take as the uncertainty on the budget; across sites, 102 of the 128 need a wider monitor than the functional width, a $2.3\times$ median widening.