THE SECOND BILL: PRICING ACTIVATION STEERING IN THE MONITOR'S CURRENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering is priced in one currency: how far an intervention moves the model's output distribution. Every linear activation monitor reading the same residual stream charges a second bill, poorly tracked by output-side quantities and computable in advance. The *monitor covector* $\boldsymbol{u}(x) = \boldsymbol{K}(x)^\top \boldsymbol{w}$ pulls a probe's readout back to the intervention site at one vector–Jacobian product per input: its second moment is a metric on interventions, which a Gaussian model of a threshold frozen at a fixed false-positive rate turns into two break conditions. A monitor can be *blinded* or made to *false-alarm*, and, at a given displacement, its margin protects only against the first: the second happens at a displacement of 0.970 within-class spreads whatever the margin. On 666 probes trained to graded margins at up to four depths above every block, the false-alarm rate well past that displacement is 79%–100% in every margin bin, while the blinding rate falls from 99% to 6%–7%; with block fixed effects the coefficient on $\log d'$ is -1.25 for blinding and -0.095 for false alarms. A standard difference-of-means vector, at the largest per-cell norm keeping generic KL and benchmark-accuracy drop within 0.05, makes 360 of 1,998 readings false-alarm, 22 of 618 on probes with $d' \geq 8$; the harmfulness probe crosses one norm further out, at $1.46\times$ that capability budget. A covector-only screen catches 66% of those breaks at precision 88%. The joint budget $(\boldsymbol{G} + \lambda \boldsymbol{N})^{-1}\boldsymbol{q}$ pays $0.367\times$ the output-optimal direction's monitor bill at $1.01\times$ its KL, and on an unpriced monitor $0.579\times$ Euclidean's, no better than a random-probe prior. Both break conditions and the in-budget false alarms replicate on three further models (1B–14B, pooled). All code and results will be released at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Two kinds of inference-time machinery now read the same residual stream of a deployed language model. One writes to it: activation steering adds a fixed vector $\boldsymbol{\delta}$ to the hidden state at some layer to make the model more truthful, less sycophantic, or more or less willing to refuse (Turner et al., 2023; Li et al., 2023; Zou et al., 2023a; Panickssery et al., 2024; Arditi et al., 2024). The other reads from it: linear activation monitors (probes for truthfulness, harmfulness, deception or knowledge awareness) score the hidden state at a later layer and flag when the score crosses a threshold calibrated on clean activations (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Orgad et al., 2025; Ferrando et al., 2025). Both are proposed for the same deployment stack (Duan, 2026b; Bailey et al., 2024); this paper asks what the first does to the second.

The steering literature prices its interventions in a single currency: the movement of the model's *output* distribution, as KL or perplexity on generic text, benchmark accuracy, or judged fluency (Wu et al., 2025; Tan et al., 2024; Stickland et al., 2024; Macocco et al., 2026). Wang et al. (2026) recently made that price explicit: to second order the output KL induced by $\boldsymbol{\delta}$ is $\frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{G}\boldsymbol{\delta}$, with $\boldsymbol{G}$ the softmax Fisher metric pulled back through the layers above the intervention, and the direction minimising it at a fixed first-order effect on the target behaviour follows. We take that framework as given. Our claim is that the same $\boldsymbol{\delta}$ is charged a *second* bill by every activation monitor above it, denominated in a currency the first cannot see, and computable in closed form before the intervention is applied.

A linear monitor with weight $\boldsymbol{w}$ at layer ℓ_j scores an input by $s(x) = \boldsymbol{w}^\top \boldsymbol{h}_{\ell_j}(x)$. Under an intervention at layer $\ell < \ell_j$ the score moves by $\boldsymbol{u}(x)^\top \boldsymbol{\delta}$ to first order, where the *monitor covector*

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.

$\boldsymbol{u}(x) = \boldsymbol{K}(x)^\top \boldsymbol{w}$ is the readout pulled back through the intermediate layers, at one vector–Jacobian product per input and no steered forward pass. Its second moment $\boldsymbol{N} = \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top]/\sigma^2$, in units of the monitor's clean score spread, is a metric on interventions in exactly the sense that $\boldsymbol{G}$ is, and splitting $\boldsymbol{u}$ into a common-mode part, which moves the operating point, and a differential part, which moves AUROC, explains why the metric the probing literature reports cannot see the damage.

We test this on three instruction-tuned models from two families (Qwen2.5-7B, Llama-3.2-3B, Qwen2.5-1.5B) at three depths, three behaviours, nine direction families, five strengths and five probes: 1,215 steered cells and 6,075 monitor readings (1,755/8,775 with the precision controls), each recording output KL and benchmark accuracy, and repeat the primary and graded-margin matrices on Llama-3.2-1B, Llama-3.1-8B and Qwen2.5-14B (Appendix O).

The second bill is predictable without evaluation, and neither output bills nor AUROC reveal it. Across all 6,075 primary readings the first-order prediction has rank correlation 0.972 with the measured bill (median 0.974 within model–depth–strength strata); as a detector of the event that matters operationally, the probe's operating point moving by 0.2 at its frozen threshold, it has precision 0.874 and recall 0.578 against a base rate of 4%. Output KL ranks the breaks at KL-reliable strengths at average precision 0.294 (0.121 within strata) against the covector's 0.705, and being one number per intervention cannot say *which* monitor broke; nor can AUROC, since 66% of the 71 readings whose true-positive rate fell by more than half keep at least 95% of their clean AUROC. Re-estimating the threshold on steered negatives repairs 78% of these breaks when the pipeline knows the vector is on, but trades false alarms for misses (Appendix Q).

A probe breaks on two sides, and its margin protects only one. A displacement that lowers the scores blinds the probe once it approaches its class separation d'; one that raises them makes it fire on negatives at $\Delta_F = 0.970$ within-class spreads, a constant that does not contain d'. We test this where margin and block are not confounded, on 666 probes trained to graded margins at up to four depths above every block (8 distinct layers): at displacements well past the break point ($\Delta_0/\sigma_w \geq 1.5$, 7,065 readings) the false-alarm rate is 79%–100% in every margin bin from $d' \in [1, 2.5)$ to $d' \geq 12$, while the blinding rate over the same bins falls from 99% to 6%–7%. Both arms show it: the matched families (5,442 readings) and a CAA vector at its own norms (1,623). Over the 21,120 readings whose displacement can break something, a logistic fit with block fixed effects puts the coefficient on $\log d'$ at -1.25 $[-2.26, -0.610]$ for blinding and -0.095 $[-0.629, 0.301]$ for false alarms.

A standard steering vector already does this inside a capability budget. We sweep a difference-of-means (CAA) vector over relative norms 0.020–0.400 and, in each of the 27 cells, take the largest norm at which that cell's generic KL and ARC-Easy drop are both at most 0.05. There 360 of 1,998 readings false-alarm, 22 of 618 on probes with $d' \geq 8$. The harmfulness probe does not break at that budget (largest rise 0.171) but crosses at the next norm up, at 0.073 nats of KL and a 0.016 drop on the 64-item benchmark.

Pricing both bills jointly wins on both axes, and pricing only the first is mostly optimism. The direction $(\boldsymbol{G} + \lambda \boldsymbol{N})^{-1}\boldsymbol{q}$ dominates Euclidean steering on both axes in 97% of matched pairs at $0.256\times$ the monitor bill and $0.440\times$ the KL; priced against a metric built from held-out probe inputs it still pays $0.272\times$, and against a monitor it never saw, $0.579\times$, which a random-probe prior matches. Its bill is under-predicted, so it needs a steered forward-pass check (§5.1). The output-optimal direction of the pooled Fisher metric, by contrast, cuts the monitor bill only to $0.806\times$, and most of its in-sample output saving is optimism (Appendix M).

What is ours. The pullback Fisher metric and output-optimal direction are Wang et al. (2026)'s, in pooled form (Appendix M). Monitor-evading directions exist under an attacker (Bailey et al., 2024; Gupta & Jenner, 2025), whereas ours is incidental to benign steering; the covector is attribution patching (Nanda, 2023; Syed et al., 2023) with a probe as readout. Ours are its use as a metric, four elementary propositions that organise the measurements, the predictor and screen, and the joint budget.

2 TWO BILLS FOR ONE INTERVENTION

Setting. A decoder-only language model maps a prompt x to a residual stream $\boldsymbol{h}_\ell(x) \in \mathbb{R}^d$ at layer ℓ, then through the remaining blocks to a final representation $\boldsymbol{f}(x)$, logits $\boldsymbol{W}_U \boldsymbol{f}(x)$ and a next-token

distribution $p(x)$. *Activation steering* adds a fixed $\boldsymbol{\delta} \in \mathbb{R}^d$ to $\boldsymbol{h}_\ell$ at every position (Turner et al., 2023; Panickssery et al., 2024). Write $\boldsymbol{J}(x) = \partial \boldsymbol{f}/\partial \boldsymbol{\delta}|_0$ for the Jacobian of the final representation with respect to the intervention (the total derivative through all positions) and $\boldsymbol{K}_j(x) = \partial \boldsymbol{h}_{\ell_j}/\partial \boldsymbol{\delta}|_0$ for the same derivative stopped at a later layer ℓ_j. A *linear activation monitor* $(\boldsymbol{w}_j, \ell_j)$ scores $s_j(x) = \boldsymbol{w}_j^\top \boldsymbol{h}_{\ell_j}(x) + b_j$ and is deployed at a threshold τ_j calibrated on clean activations, typically to a fixed false-positive rate; its clean scores have class means μ_0, μ_1, pooled spread σ_j, within-class spread σ_w, and margin $d'_j = (\mu_1 - \mu_0)/\sigma_w$.

The first bill. With $\boldsymbol{H}(x) = \mathrm{Cov}_{p(x)}(\boldsymbol{e})$ the softmax Fisher information (the covariance of the next token's unembedding row $\boldsymbol{e}$ under $p(x)$), the pooled form of Wang et al. (2026)'s result is

$$D_{\mathrm{out}}(\boldsymbol{\delta}) \;=\; \mathbb{E}_x[\mathrm{KL}(p(x) \,\|\, p_{\boldsymbol{\delta}}(x))] \;=\; \tfrac{1}{2}\,\boldsymbol{\delta}^\top \boldsymbol{G}\,\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3), \qquad \boldsymbol{G} = \mathbb{E}_x\big[\boldsymbol{J}(x)^\top \boldsymbol{H}(x) \boldsymbol{J}(x)\big], \tag{1}$$

so that $\boldsymbol{G}$, the softmax Fisher pulled back to the intervention site, is the metric in which the output-side cost is a quadratic form. At a fixed first-order effect $\boldsymbol{q}^\top \boldsymbol{\delta} = \rho$ on a target behaviour with covector $\boldsymbol{q}$, the direction minimising (1) is $\boldsymbol{\delta}_{\boldsymbol{G}} = \rho\, \boldsymbol{G}^+ \boldsymbol{q}/(\boldsymbol{q}^\top \boldsymbol{G}^+ \boldsymbol{q})$. We call D_{out} the *output bill*; our $\boldsymbol{q} = \mathbb{E}_x[\boldsymbol{J}(x)^\top (\boldsymbol{e}_+ - \boldsymbol{e}_-)]$ for two contrasting answer tokens is the forced-choice logit-gap gradient.

The second bill. Under the intervention the monitor's score becomes

$$s_j^{\boldsymbol{\delta}}(x) \;=\; s_j(x) + \boldsymbol{u}_j(x)^\top \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^2), \qquad \boldsymbol{u}_j(x) = \boldsymbol{K}_j(x)^\top \boldsymbol{w}_j. \tag{2}$$

We call $\boldsymbol{u}_j(x)$ the *monitor covector*: the readout pulled back to the intervention site, one vector–Jacobian product per input with $\boldsymbol{w}_j$ as cotangent, no steered forward pass and no labels (the screen of §3 also needs the calibration labels, to separate the class means). Over the monitor's own input distribution define the common-mode covector $\bar{\boldsymbol{u}}_j = \mathbb{E}[\boldsymbol{u}_j]$, the differential covector $\Delta\bar{\boldsymbol{u}}_j = \mathbb{E}[\boldsymbol{u}_j \mid y{=}1] - \mathbb{E}[\boldsymbol{u}_j \mid y{=}0]$, the dispersion $\boldsymbol{\Sigma}_j = \mathrm{Cov}(\boldsymbol{u}_j)$, and

$$\boldsymbol{N}_j = \frac{\mathbb{E}[\boldsymbol{u}_j \boldsymbol{u}_j^\top]}{\sigma_j^2} = \frac{\bar{\boldsymbol{u}}_j \bar{\boldsymbol{u}}_j^\top + \boldsymbol{\Sigma}_j}{\sigma_j^2}. \tag{3}$$

$\boldsymbol{N}_j$ is the *monitor metric*, and the *monitor bill* $B_{\mathrm{mon}}(\boldsymbol{\delta}) = \sum_j \omega_j\, \boldsymbol{\delta}^\top \boldsymbol{N}_j \boldsymbol{\delta} = \sum_j \omega_j\, \mathbb{E}[(\Delta s_j/\sigma_j)^2]$ is the mean squared standardised displacement of the score over deployed monitors with weights ω_j: a quadratic form for the same reason (1) is one, estimated from the held-out probe inputs on which the bill is later measured (§5.5 bounds the optimism this invites). Ratios of the bill between directions are reported as ratios of $B_{\mathrm{mon}}^{1/2}$, the RMS standardised displacement.

Common mode and differential mode. The three moments just defined price three kinds of damage. The common-mode covector $\bar{\boldsymbol{u}}_j$ shifts every score equally and moves the *operating point*: at a frozen threshold it changes the true- and false-positive rates without changing the ranking of inputs. The differential covector $\Delta\bar{\boldsymbol{u}}_j$ changes class separation and moves *discrimination*: AUROC, a ranking statistic, depends on $\Delta\bar{\boldsymbol{u}}_j^\top \boldsymbol{\delta}$ and on the within-class variance inflation $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_w \boldsymbol{\delta}$ but is exactly invariant to $\bar{\boldsymbol{u}}_j^\top \boldsymbol{\delta}$. Under a Gaussian score model the closed-form predictions are

$$d'_{\boldsymbol{\delta}} = \frac{\mu_1 - \mu_0 + \Delta\bar{\boldsymbol{u}}^\top \boldsymbol{\delta}}{\sqrt{\sigma_w^2 + \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_w \boldsymbol{\delta}}}, \qquad \mathrm{AUROC}_{\boldsymbol{\delta}} \approx \Phi\big(d'_{\boldsymbol{\delta}}/\sqrt{2}\big), \qquad \mathrm{TPR}_{\boldsymbol{\delta}}(\tau) \approx \Phi\Big(\frac{\mu_1 + \bar{\boldsymbol{u}}_{|1}^\top \boldsymbol{\delta} - \tau}{\sigma_{1,\boldsymbol{\delta}}}\Big), \tag{4}$$

with $\bar{\boldsymbol{u}}_{|y} = \mathbb{E}[\boldsymbol{u}_j \mid y]$ the class-mean covectors, $\boldsymbol{\Sigma}_w$ their within-class dispersion and $\sigma_{1,\boldsymbol{\delta}}$ the steered positive-class spread (Green & Swets, 1966). Prediction reads AUROC and rates off the per-input scores of (2); (4) serves §3. The measured $\|\bar{\boldsymbol{u}}_j\|/\|\Delta\bar{\boldsymbol{u}}_j\|$ ranges from 0.515 to 14.9 over 45 probe–block pairs: above one for most truthfulness pairs, below one for the harmfulness probe in eight of nine; where it exceeds one, AUROC misses most of what a generic intervention does to a monitor.

3 The exchange rate, the norm price, the frontier, and the trigger

Fix a target behaviour with covector $\boldsymbol{q}$ and the *matched-effect family* $V_\rho = \{\boldsymbol{\delta} : \boldsymbol{q}^\top \boldsymbol{\delta} = \rho\}$ of interventions with the same first-order on-target effect. Every comparison between direction families

is made inside V_ρ: the directions differ, the promised effect does not, and Appendix F verifies that the measured effect does not either. The regularised path

$$\boldsymbol{\delta}_\alpha = \rho\,\frac{(\boldsymbol{G}+\alpha\boldsymbol{I})^{-1}\boldsymbol{q}}{\boldsymbol{q}^\top(\boldsymbol{G}+\alpha\boldsymbol{I})^{-1}\boldsymbol{q}}, \qquad \alpha\in(0,\infty], \tag{5}$$

interpolates between Euclidean (gradient) steering, $\boldsymbol{\delta}_\infty \propto \boldsymbol{q}$, and the output-optimal direction $\boldsymbol{\delta}_{\boldsymbol{G}}$ of Wang et al. (2026) at $\alpha \to 0$. Write λ_i for the eigenvalues of $\boldsymbol{G}$ and q_i, u_i for the coordinates of $\boldsymbol{q}$ and of a monitor's common-mode covector $\bar{\boldsymbol{u}}$ in its eigenbasis; both bills along the path are then explicit rational functions of α ((9), Appendix A, where all proofs are given).

Proposition 1 (The output bill can vanish while the monitor bill cannot)**.** *Let $R(\boldsymbol{\delta}) = B^{\bar{\boldsymbol{u}}}_{\mathrm{mon}}(\boldsymbol{\delta})/D_{\mathrm{out}}(\boldsymbol{\delta})$, with $\boldsymbol{G}$ taken over generic text and $\boldsymbol{q}$ over the behaviour's prompts. If $\boldsymbol{q}$ and $\bar{\boldsymbol{u}}$ both have non-zero coordinates q_m, u_m on a simple eigendirection of $\boldsymbol{G}$ with eigenvalue $\lambda_{\min} > 0$, then as $\alpha \to 0$ and then $\lambda_{\min} \to 0$ (rest of spectrum fixed), the quadratic form $D_{\mathrm{out}}(\boldsymbol{\delta}_\alpha) \to 0$ while $B^{\bar{\boldsymbol{u}}}_{\mathrm{mon}}(\boldsymbol{\delta}_\alpha) \to \rho^2 u_m^2/(\sigma^2 q_m^2) > 0$, so $R(\boldsymbol{\delta}_\alpha) = \Theta(1/\lambda_{\min})$. Along a near-null on-target mode of $\boldsymbol{G}$, output auditing on generic text certifies the intervention as free; the monitor still pays.*

Proposition 1 says where the danger lives (Appendix B bounds it) but not why, and not because monitors read the output-null directions of $\boldsymbol{G}$, which we measured and is false (Appendices D and Q). The mechanism is a price paid in norm.

Proposition 2 (The norm price of frugality)**.** *Among $\boldsymbol{\delta} \in V_\rho$ the minimum-norm member is $\boldsymbol{\delta}_{\boldsymbol{q}} = \rho\boldsymbol{q}/\|\boldsymbol{q}\|^2$ and the output-optimal member is $\boldsymbol{\delta}_{\boldsymbol{G}}$. Let μ be the probability measure on the spectrum of $\boldsymbol{G}$ with $\mu_i \propto q_i^2$. Then*

$$\frac{\|\boldsymbol{\delta}_{\boldsymbol{G}}\|^2}{\|\boldsymbol{\delta}_{\boldsymbol{q}}\|^2} = \frac{\mathbb{E}_\mu[\lambda^{-2}]}{\mathbb{E}_\mu[\lambda^{-1}]^2} = 1 + \mathrm{CV}_\mu(\lambda^{-1})^2 \;\geq\; 1, \tag{6}$$

with equality if and only if $\boldsymbol{q}$ lies in a single eigenspace of $\boldsymbol{G}$.

Output-frugality is bought with displacement norm, and the price is the dispersion of the inverse spectrum as seen by $\boldsymbol{q}$. The output-optimal direction lands where *both* metrics are small per unit norm (§5.5: at matched effect it carries $1.83\times$ the Euclidean norm), but a monitor has no reason to be cheap along the small eigenvalues of $\boldsymbol{G}$ and is charged for that norm, which is why the output bill collapses under the Fisher direction while the monitor bill barely moves.

Proposition 3 (The two-currency frontier)**.** *For $\lambda \geq 0$, the interventions in V_ρ that minimise $\frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{G}\boldsymbol{\delta} + \frac{\lambda}{2}\boldsymbol{\delta}^\top\boldsymbol{N}\boldsymbol{\delta}$ are $\boldsymbol{\delta}_\lambda \propto (\boldsymbol{G}+\lambda\boldsymbol{N})^+\boldsymbol{q}$, and the map $\lambda \mapsto (D_{\mathrm{out}}(\boldsymbol{\delta}_\lambda), B_{\mathrm{mon}}(\boldsymbol{\delta}_\lambda))$ traces the Pareto frontier of the two bills. For $\boldsymbol{G}, \boldsymbol{N}$ of full rank the frontier is non-degenerate if and only if $\boldsymbol{G}^{-1}\boldsymbol{q} \nparallel \boldsymbol{N}^{-1}\boldsymbol{q}$; the empirical $\boldsymbol{N}$ is rank-deficient and the solver adds a ridge $10^{-4}\bar{\lambda}\mathbf{I}$ (Appendix C).*

We call $\boldsymbol{\delta}_\lambda$ the *joint budget*; it needs only $\boldsymbol{G}$ from generic text and $\boldsymbol{N}$ from one VJP per monitor input, and, optimised against an estimated metric, it inherits the winner's curse of $\boldsymbol{\delta}_{\boldsymbol{G}}$ at large λ (§5).

When does a monitor actually break? The bills price displacement, but a deployed monitor fails discretely, and in one of two ways: the positive class stops crossing the frozen threshold (it is *blinded*), or the negative class starts to (it *false-alarms*). Under the equal-variance Gaussian model with the threshold at clean false-positive rate ε, $\tau = \mu_0 + z_{1-\varepsilon}\sigma_w$, the class-mean displacements $\Delta_1 = \bar{\boldsymbol{u}}_{|1}^\top\boldsymbol{\delta}$ and $\Delta_0 = \bar{\boldsymbol{u}}_{|0}^\top\boldsymbol{\delta}$ move the two rates:

$$\mathrm{TPR}_{\boldsymbol{\delta}} = \Phi\big(d' - z_{1-\varepsilon} + \Delta_1/\sigma_w\big), \qquad \mathrm{FPR}_{\boldsymbol{\delta}} = \Phi\big(-z_{1-\varepsilon} + \Delta_0/\sigma_w\big), \tag{7}$$

to first order in $\boldsymbol{\delta}$ at fixed within-class spread (dropping its first-order change $2\boldsymbol{\delta}^\top\mathrm{Cov}(\boldsymbol{u}, s)/\sigma_w^2$, which the per-input predictor of §5 does not need). Negative Δ_1 blinds; positive Δ_0 false-alarms.

Proposition 4 (The trigger condition)**.** (a) *The sensitivity of the true-positive rate to Δ_1/σ_w is $\varphi(d' - z_{1-\varepsilon})$, Gaussian-tail small in the margin; the sensitivity of the false-positive rate to Δ_0/σ_w is $\varphi(z_{1-\varepsilon})$,* independent of the margin *(0.103 at $\varepsilon = 0.05$). A wide margin protects a monitor from being blinded, not from being made to fire.* (b) *The displacement that costs a drop c in true-positive rate is $\Delta_T(d') = d'\,t^\star(d')$ with*

$$t^\star(d') = \frac{(d' - z_{1-\varepsilon}) - \Phi^{-1}\big(\Phi(d' - z_{1-\varepsilon}) - c\big)}{d'} \;> 0; \tag{8}$$

the displacement that raises the false-positive rate by c is $\Delta_F = z_{1-\varepsilon} + \Phi^{-1}(\varepsilon + c)$, a constant. For $c = 0.2$, $\varepsilon = 0.05$: $\Delta_F = 0.970\,\sigma_w$ for every monitor, while Δ_T bottoms at $0.507\,\sigma_w$ near $d' = 1.90$ and grows with the margin above it, so that $\Delta_T > \Delta_F$ (a probe can be made to fire while still catching its positives) for $d' > 3.29$; below that, down to $d' = 0.970$ (Appendix A), a probe goes blind first. On the blinding side $t = |\Delta_1|/(\sigma_w d')$ compresses our 14.9-fold range of margins into a 3.66-fold range of break points; on the false-alarm side there is nothing to compress.

Corollary 1 (An evaluation-free screen)**.** *Δ_1, Δ_0 are unavailable before deployment, but $\hat{\Delta}_y = \bar{\boldsymbol{u}}_{|y}^\top \boldsymbol{\delta}$ is available from the covectors of the probe's own calibration inputs, one class mean each. The rule "flag monitor j if $-\hat{\Delta}_1/\sigma_w \geq \Delta_T(d'_j)$ or $\hat{\Delta}_0/\sigma_w \geq \Delta_F$" is a per-monitor screen with no constant fitted to the steered data: both thresholds are the break displacements of* (b)*, set by ε and c. Its pooled form applies the same rule to $|\bar{\boldsymbol{u}}^\top \boldsymbol{\delta}|$ against $\Delta^\star = \min\{\Delta_T, \Delta_F\}$ (Table 10).*

4 Experimental setup

Models and depths. Qwen2.5-7B, Qwen2.5-1.5B (Yang et al., 2024) and Llama-3.2-3B (Grattafiori et al., 2024), instruction-tuned, 28 blocks each, $d = 3{,}584$, $1{,}536$ and $3{,}072$; we intervene at $\ell \in \{7, 14, 21\}$ and read every monitor at $\ell_j = \ell + 7$.

Behaviours. Three contrastive behaviours from the CAA suite (Panickssery et al., 2024; Perez et al., 2022): sycophancy, refusal and corrigibility (*corrigible-neutral-HHH*), each a set of A/B questions with a behaviour-matching and a non-matching answer. The on-target effect ψ is the mean logit gap between the two answer letters, the standard forced-choice steering metric; $\boldsymbol{q}$ is its gradient with respect to $\boldsymbol{\delta}$ over 64 items held out from the CAA vectors, the same items on which $\Delta\psi$ is measured (on 64 disjoint items the match holds, $0.995\times$; Appendix Q).

Monitors. Five linear probes trained by logistic regression on the clean residual stream at ℓ_j: four truthfulness probes on the geometry-of-truth statement sets (Marks & Tegmark, 2024), namely *cities*, *sp–en* translation, *larger* (numerical comparison) and *claims* (CommonClaim, Casper et al., 2023), and a harmfulness probe separating AdvBench requests (Zou et al., 2023b) from Alpaca instructions (Taori et al., 2023). Each is trained on 260 inputs and evaluated on 140 held-out ones, threshold frozen at a clean false-positive rate of 5%; clean margins span $d' = 0.991$ (*claims*) to 14.7 (harmfulness).

Metrics. $\boldsymbol{G}$ is estimated on 512 WikiText-2 passages (Merity et al., 2017) at $\ell = 14$ and 160 at the other depths, with an exact softmax Fisher per example truncated to its top 384 eigenpairs (95%–98% of its trace); snapshots at $n \in \{8, \ldots, 512\}$ are saved in the same pass, and an independent $\boldsymbol{G}_B$ on 256 disjoint passages serves the optimism analysis. Monitor covectors are one VJP per held-out probe input, and $\boldsymbol{N}$ is their second moment over those inputs.

Directions and strengths. Nine direction families, seven matched to $\boldsymbol{q}^\top \boldsymbol{\delta} = \rho$: Euclidean ($\boldsymbol{\delta} \propto \boldsymbol{q}$); the Fisher path (5) at $\alpha \in \{1, 10^{-2}, 10^{-4}\}$ in units of the mean eigenvalue of $\boldsymbol{G}$; the joint budget at $\lambda \in \{10^{-2}, 10^{-1}, 1\}$ with $\boldsymbol{N}$ the mean of the five monitor metrics; and two random controls with the norm of the Euclidean and of the $\alpha = 10^{-4}$ direction. Every comparison is between directions of equal measured on-target effect at strengths with little capability cost (Appendix F); $\rho \in \{0.25, 0.5, 1, 2, 4\}$ in logit units, relative norm 4.1×10^{-4}–0.641. Difference-of-means CAA vectors are nearly orthogonal to $\boldsymbol{q}$ (median $|\cos| = 0.020$) and cannot be matched without destroying the model (Appendix D); they are excluded from the matched family and swept at their own relative norms in §5.4. A second matrix (§5.3) adds 74 probes per block on average, trained on prefixes of each dataset's training split so that margins are graded *within* a block, at up to four depths ($\ell + 3$, $\ell + 7$, $\ell + 11$ and the final layer, where the model has them), 8 distinct layers in all. To reach the displacements a wide-margin probe needs, three of the families (Euclidean, $\alpha = 10^{-4}$, joint $\lambda = 10^{-2}$) are also applied at $\rho \in \{8, 16, 32\}$ (243 cells, 17,540 readings). These are not matched-effect comparisons: past $\rho = 4$ the effect saturates and capability moves (median KL 0.054, 90th-percentile ARC drop 0.375). They are used only in statements conditioned on the displacement a reading received, never to compare families.

Measurements. Every cell records the measured $\Delta\psi$, the output bill as mean next-token KL on 48 held-out generic passages, accuracy on 64 ARC-Easy questions (Clark et al., 2018), and per monitor

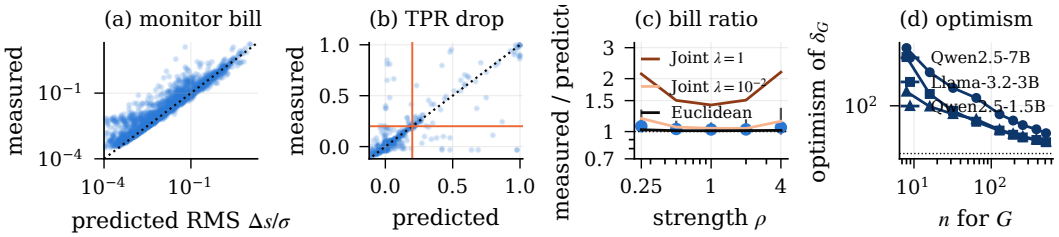


Figure 1: Evaluation-free prediction, all 6,075 primary readings. (a) Predicted against measured RMS standardised displacement (Spearman 0.972). (b) Predicted against measured drop in true-positive rate at the frozen threshold; lines mark the 0.2 break criterion. (c) Median measured/predicted bill by strength (points, interquartile bars) and by family: the direction optimised against $\boldsymbol{N}$ at $\lambda = 1$ is under-predicted at every strength. (d) The same curse for $\boldsymbol{\delta_G}$ (Appendix M): bill under independent $\boldsymbol{G}_B$ over bill under the choosing $\boldsymbol{G}_n$.

the measured and predicted AUROC, true-positive rate at the frozen threshold, mean standardised shift and RMS standardised displacement; predictions use only $\boldsymbol{u}_j(x)$ and the clean scores. The primary matrix is 1,215 cells (3 models $\times$ 3 depths $\times$ 3 behaviours $\times$ 9 directions $\times$ 5 strengths) in `float16`; precision controls on Qwen2.5-1.5B (`float32` at $\ell = 14$, `bfloat16` at all three depths) bring the total to 1,755 cells and 8,775 monitor readings. Four further matrices use the same blocks: transfer (2,430 cells), half-split (756), metric decomposition (1,215), and the graded-margin and deployed-norm sweep of §§5.3–5.4 (756 cells, 41,516 monitor readings from the matched families and 13,986 at deployed norms). A reading is *non-trivial* if it comes from a matched direction and $|\Delta\psi| \geq 0.2$, which keeps all 4,725 matched readings; the random controls are reported separately. A monitor *breaks* if its operating point moves by at least 0.2 at the frozen threshold: its true-positive rate falls (*blinded*) or its false-positive rate rises (*false-alarmed*). Forward passes and VJPs run on one 24 GB GPU (48 GB for Appendix O); nothing is generated outside Appendix P.

Numerical precision. Generic KL at half precision sits on a noise floor: against a `float32` control, `float16` KL disagrees by a median factor of 4.12$\times$ at $\rho = 0.25$ and 1.18$\times$ at $\rho = 1$, and `bfloat16` by 1.95$\times$ even at $\rho = 4$ (Appendix E). The monitor bill is unaffected (rank correlation 0.999 between `float16` and `float32`, 0.966 for `bfloat16`), being a displacement of a scalar in units of that scalar's spread. Every KL-based statistic below is computed on $\rho \geq 1$, where `float16` KL is within a median factor of 1.50 of `float32` on all three models and the KL ratios we report within 0.022 (Appendix Q); monitor-bill statistics use all ρ.

5 Results

5.1 The monitor bill is predictable before the intervention is applied

Figure 1 compares the first-order prediction of (2), from one VJP per monitor input and the clean scores, with the steered forward pass. Over all 6,075 readings the predicted and measured RMS displacement have Spearman correlation 0.972 $[0.960, 0.983]$ (cluster bootstrap over the 27 model–depth–behaviour cells) and a median measured-to-predicted ratio of 1.04. It does not depend on how far above the intervention the monitor sits: over 8 depths from $\ell + 3$ to the final layer it has Spearman 0.971–0.990 and ratio 1.01–1.06 (Table 20). Nor on the monitor, behaviour or model: Spearman is 0.995 on MLP probes, 0.996 on a deception probe, 0.971 on four more behaviours and 0.931–0.981 on the scale-check models (Appendix O). It survives removing the shared dependence on ρ and monitor scale: within each of the 45 model–depth–strength strata the median is 0.974 (minimum 0.808), and 0.989 within a single monitor. On rebuilt cells the Jacobian-free $\boldsymbol{w}^\top \boldsymbol{\delta}$ reaches only 0.718 against the covector's 0.965 (Appendix Q).

A deployer needs the binary answer. On the 4,725 non-trivial readings (180 breaks, base rate 4%), the rule "flag if either predicted rate moves by 0.2" flags 119 readings with precision 0.874 $[0.803, 0.911]$ and recall 0.578 $[0.311, 0.714]$. The two sides differ: on the 68 blindings recall is 0.721 at precision 0.790, on the 112 false alarms 0.455 at 0.879, because the measured displacement of the negative class exceeds the first-order prediction by 2.77$\times$ in the median, and at a 5% threshold the false-positive rate is convex in that displacement, so a shortfall in the prediction shrinks the predicted rise more than

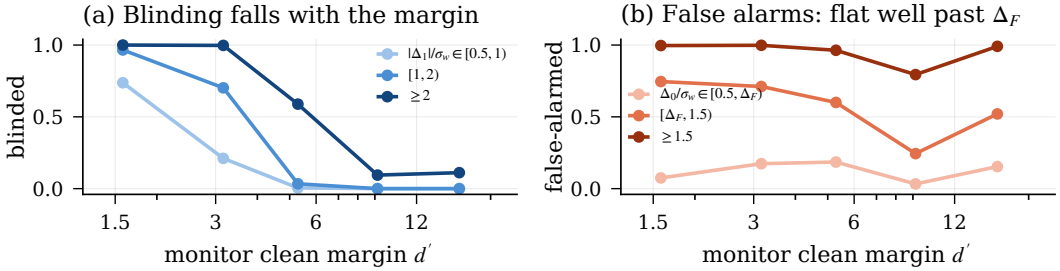


Figure 2: The two break conditions, measured. Graded-margin probes; each side is conditioned on the class displacement its own condition uses, in that probe's measured within-class spread σ_w. (a) Blinding, by the downward displacement of the positive class. (b) False alarms, by the upward displacement of the negative class; $\Delta_F = 0.970\sigma_w$ is Proposition 4's break point for every probe. Bins and population as in Table 19.

proportionally (a flag at 0.1 recovers 0.643). No steered forward pass is needed: once the covectors are computed per monitor, every candidate $\boldsymbol{\delta}$ costs a matrix–vector product.

The optimised direction is under-predicted, and the screen is blind on it. Figure 1(c) shows where first order fails: the measured bill exceeds the prediction by 1.02× at $\rho = 4$ for Euclidean and Fisher directions but by 2.18× for the joint budget at $\lambda = 1$, because minimising a first-order form selects the directions where the neglected second-order term is largest. Broken out by family (Table 14), the pre-deployment rule has recall 0.634 on Euclidean and 0.736 on Fisher directions but *zero* on all three joint budgets: a direction chosen by minimising the form the screen reads cannot be screened by it. Two facts bound the damage: the joint budget breaks monitors 3.34× less often, and the displacement rule of Corollary 1, fed the *measured* shift, recovers 90% of its breaks at $\lambda \leq 10^{-1}$. Screen an unoptimised direction; verify an optimised one.

5.2 Output bills do not proxy the monitor bill, and AUROC cannot see it

We ask the operational question: does output KL rank the 172 readings that broke above the 2,835 KL-reliable non-trivial readings? Pooled it does moderately (AUROC 0.845), because both grow with ρ; within strata that falls to 0.712, and its average precision is 0.294 pooled and 0.121 within strata (base rate 6%) against 0.705 for the covector. Benchmark accuracy loss is worse than chance (AUROC 0.228). And KL is one number per intervention: among the 90 interventions that broke some monitors and not others, the covector ranks the broken one higher in 79% of 428 pairs.

Nor does AUROC, the quantity the probing literature reports. Of the 71 readings whose true-positive rate fell by more than half, 66% keep at least 95% of their clean AUROC. §2 says why: a ranking statistic is exactly invariant to the common-mode component $\bar{\boldsymbol{u}}^\top \boldsymbol{\delta}$ and the threshold is not, and for the truthfulness probes that component is usually the larger one ($\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$ up to 14.9).

5.3 When a monitor breaks: the trigger condition

A probe's margin is mostly set by how much labelled data it saw, so we train probes on prefixes of each dataset's training split at up to four depths above the intervention. Every block then holds 74 probes on average, spanning $d' = 1.00$–21.6, making the margin effect a within-block comparison (666 probes, 756 cells, 41,516 matched readings; Appendix N). Matched and unmatched readings are pooled to populate the wide-margin bins. Equation (7), with no fitted constant, fits: the mean absolute error of the predicted operating point is 0.018 on the true-positive rate and 0.033 on the false-positive rate, and 0.025 and 0.067 on the 14,998 readings where a class mean moved by $\geq 0.5\sigma_w$.

The two sides then separate cleanly (Table 19, Figure 2). Well past its break displacement ($\Delta_0/\sigma_w \geq 1.5$, 7,065 readings) the false-alarm rate is 79%–100% *whatever the margin*, in all 5 bins from $d' \in [1, 2.5)$ to $d' \geq 12$: 100%, 100%, 96%, 79% and 99%. It is not a property of one kind of direction: the two arms of the sweep give 80%–100% over the matched families (5,442 readings) and 77%–100% at deployed norms (1,623), bin by bin within three points of each other (Table 19). The blinding side is not comparable at a fixed displacement, since its break point $\Delta_T(d') = d'\, t^\star(d')$ grows with the margin while Δ_F does not. Against each side's *own* break point both conditions are

decisive: at $1.5\times$ it, 78%–100% false-alarm and 98%–100% are blinded. What differs is reachability: Δ_F is exceeded 18% of the time and Δ_T only 4%. That is the asymmetry Proposition 4(b) predicts for a threshold set at a fixed false-positive rate; one set between the class means would move Δ_F with d'. The flatness is a property of the regime well past the break point, not of the crossing itself: in the band that merely crosses it, $\Delta_0/\sigma_w \in [\Delta_F, 1.5)$ (2,798 readings), the rate does move with the margin, from 75% at $d' < 2.5$ down to 24% at $d' \in [7, 12)$, a departure from the Gaussian model near the break point. Over the 21,120 readings whose displacement can break something, a logistic fit with block fixed effects and the displacement controlled gives a coefficient on $\log d'$ of -1.25 $[-2.26, -0.610]$ for blinding, excluding zero, and -0.095 $[-0.629, 0.301]$ for false alarms, containing it. At a given displacement, wide-margin probes show no detectable protection on this side; they are merely displaced past Δ_F less often, for reasons these data do not separate (the interventions, or how $\|\bar{\boldsymbol{u}}_{|0}\|$ scales with σ_w). Pooled over the scale-check models the conditions replicate: false alarms well past Δ_F at 84%–97% in every bin, and coefficients -1.60 $[-2.19, -1.20]$ for blinding and 0.350 $[-0.100, 0.758]$ for false alarms (Table 23). As a screen, the pooled form of Corollary 1 beats a pooled threshold at the same flag count, on measured displacements (recall 0.844 against 0.756) and on predicted ones (0.600 against 0.478; Table 10).

5.4 A STANDARD STEERING VECTOR, INSIDE A CAPABILITY BUDGET, MAKES MONITORS FIRE

Published vectors are not matched to a target effect but applied at a chosen norm. We sweep each behaviour's difference-of-means CAA vector (Panickssery et al., 2024) unmatched over relative norms 0.020–0.400 of $\mathbb{E}\|\boldsymbol{h}_\ell\|$ (13,986 readings; Table 21). The covector depends on the model and probe, not on how the direction was built, and the prediction holds out of family (Spearman 0.994).

The norm is selected per cell, before any monitor is read: for each of the 27 (model, depth, behaviour) cells we take the largest of the 7 swept norms at which *that cell's* generic KL and ARC-Easy drop are both at most 0.05 (chosen relative norms 0.080–0.240, median 0.160; Table 21 gives every norm; per-cell values are in the released results). At those norms 360 of 1,998 readings false-alarm (18%), 22 of the 618 on probes with $d' \geq 8$ among them, and 46 are blinded. The harmfulness probe does not false-alarm: its largest rise, 0.171, is short of the criterion. It crosses at $0.240\,\mathbb{E}\|\boldsymbol{h}_\ell\|$ (rise 0.229), and getting there costs 0.073 nats of KL, $1.46\times$ the bar and still invisible to the 64-item benchmark (0.016). A standard CAA vector therefore fires wide-margin monitors inside the capability budget, and, on these models, reaches the safety probe just outside it. The screen of §5.1 catches 66% of the 406 breaks here at precision 88%. On the scale-check models it false-alarms 157 of 855 readings with $d' \geq 8$ and the screen catches 70% of all breaks at precision 88%, but the safety probe crosses only at 2.32–$3.15\times$ the bar or not at all (Table 24).

5.5 THE TWO-CURRENCY FRONTIER AND THE JOINT BUDGET

The exchange rate and the joint budget. Table 7 and Figure 3 (Appendix F) quantify Propositions 1–3. Moving from Euclidean to the output-optimal direction buys a median $2.39\times$ reduction in output KL, but the monitor bill falls only to $0.806\times$ (lower than Euclidean in 70% of pairs) while the direction carries $1.83\times$ the norm: the output saving is real, the monitor saving marginal. A matched pair is one (block, behaviour, strength, monitor) with $|\Delta\psi| \geq 0.2$ and Euclidean displacement $\geq 0.05\sigma$ at KL-reliable strengths (223 pairs; intervals cluster-bootstrapped over blocks). At $\lambda = 10^{-2}$ the joint budget is lower than Euclidean steering on *both* axes in 97% of pairs $[93\%, 100\%]$, paying $0.256\times$ the monitor bill and $0.440\times$ the KL (79% with KL read at every position; Appendix J). Against the output-optimal direction itself it pays $1.01\times$ the KL $[1.000, 1.06]$ and $0.367\times$ the monitor bill, a trade rather than a domination, won in 92% of pairs $[85\%, 98\%]$ at KL within 25%. Larger λ lowers the bill further ($0.116\times$ at $\lambda = 1$) at a falling domination rate, and at $\ell = 14$ alone, where $\boldsymbol{G}$ is best estimated, it dominates in 100% of 65 pairs at $0.251\times$ the bill. The saving is not optimism from estimating $\boldsymbol{N}$ on the inputs the bill is read on: reading it on the half of each probe's inputs that did not build $\boldsymbol{N}$, the ratio moves only from 0.245 to 0.272 at $\lambda = 10^{-2}$ ($[0.204, 0.360]$, 219 pairs), with domination at 96% (756 cells, Appendix I). Projecting onto the monitors' null space does worse: the output-optimal direction with zero first-order monitor displacement pays $0.355\times$, and $0.650\times$ on an unconstrained monitor against $0.570\times$ for the joint budget in the same rerun (Appendix Q). The pricing cost grows with λ ($1.11\times$ to $1.57\times$) as the solver enters the null space of $\boldsymbol{N}$ (rank ≤ 700).

Transfer to monitors it was not priced against. Re-solving $\boldsymbol{\delta}_\lambda$ with $\boldsymbol{N}$ rebuilt from four monitors and reading the fifth (2,430 further cells), the held-out monitor's bill relative to Euclidean is 0.579 at $\lambda = 10^{-2}$ against 0.256 when it is in $\boldsymbol{N}$; that out-of-sample figure is the one to plan on. What transfers is not monitor-specific: an ignorant prior of 8 random unit probes at the monitor layer matches it (0.542) and beats it on both axes at $\lambda = 10^{-1}$, while the common-mode part $\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top$ of $\boldsymbol{N}$ alone (one direction per monitor) gives 0.436. The harmfulness probe gains nothing over the ignorant prior (1.000×, against 2.01–2.79× for the other four). Appendices G and L give the tables.

6 Related work

Pricing steering in the output currency. Steering by adding a vector to the residual stream (Turner et al., 2023; Li et al., 2023; Zou et al., 2023a; Panickssery et al., 2024) has its cost measured as perplexity, benchmark accuracy and judged fluency (Arditi et al., 2024; Wu et al., 2025; Tan et al., 2024; Macocco et al., 2026; Stolfo et al., 2025). A line of work reduces that cost by KL-regularised fine-tuning before steering (Stickland et al., 2024), steering fewer positions (Feng et al., 2026), conditioning on a trigger (Lee et al., 2025), closing the loop on strength (Nguyen et al., 2026a; Vogels et al., 2025), rotating (Pham & Nguyen, 2024), null-space projection (Sheng et al., 2026), as in editing (Fang et al., 2025), or sanitising against a reference behaviour (Aravindan et al., 2026; Nguyen et al., 2026b); Wang et al. (2026) make the output cost exact through the pullback Fisher metric, and Dalili et al. (2026) and Skifstad et al. (2026) treat steering as geometry or control on a locally linear model. All price the output side. The quadratic-form view of damage has precedents in concept erasure (Ravfogel et al., 2020; Belrose et al., 2023), natural gradients (Martens, 2020; Martens & Grosse, 2015) and influence functions (Grosse et al., 2023); the monitor covector is attribution patching (Nanda, 2023; Syed et al., 2023; Kramár et al., 2024) with a probe readout, to which we add its second moment as a metric, the common/differential split, and the consequences above.

Side effects, their prediction, and monitors under stress. Steering erodes refusal (Xiong et al., 2026; Li et al., 2026), moves untargeted behaviours (Ong et al., 2026), leaves the prompt-reachable manifold (Mishra et al., 2026) and disrupts attention (Luo et al., 2026; Venkatesh & Kurapath, 2026); these side effects are forecast from unsteered representations at the level of outputs or SAE features (Duan, 2026a; Jafari et al., 2026; Cyberey et al., 2026). Linear probes (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Orgad et al., 2025; Ferrando et al., 2025) are proposed as deployment-time monitors; their fragility is studied under model updates (Duan, 2026b) and under an adversary optimising inputs, activations or weights to evade them (Bailey et al., 2024; Gupta & Jenner, 2025). None of this prices what a benign intervention does to a reader of the residual stream.

7 Discussion, limitations and conclusion

What to do with the second bill. One VJP per monitor input prices any candidate vector against every monitor in the stack. We recommend the per-input screen of §5.1 before deployment, a steered forward-pass check for any direction optimised against $\boldsymbol{N}$, which the screen cannot see, the joint budget at moderate λ with the random-probe prior when the stack is unknown, and operating-point statistics rather than AUROC. If the pipeline knows the vector is on, re-estimating τ on steered negatives removes 78% of matched breaks but trades false alarms for misses (Appendix Q).

Limitations. The monitors are probes on six datasets; one-hidden-layer MLP probes are predicted about as well (Spearman 0.995, Appendix O); deeper architectures are untested. Every effect claim is scoped to the forced-choice logit gap, which the matched-effect family equalises only to first order; our free-generation check was uninformative at this scale (Appendix P). Matched-strength breaks concentrate on Qwen2.5-1.5B (159 of the primary 180) and Llama-3.2-1B (71 of 1,575 readings, against 1 and 0 on Llama-3.1-8B and Qwen2.5-14B), so their frequency depends on the model; pooled, the two break conditions hold at scale and the deployed-norm false alarms do not fade (Appendix O; Appendix J lists what else bounds the numbers). The claim is not that steering usually breaks monitors, but that the displacement behind a break is computable in advance.

Conclusion. A monitor reading the same residual stream is charged a second price, computable in advance; at a given displacement, its margin defends it against only one of two failures.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments are forward passes and vector–Jacobian products on publicly available models and datasets; §4 and Appendix C give every estimation setting. The supplementary material contains the code, the complete result files for every matrix reported here (the 1,755 cells and 8,775 readings of the main matrix, and the transfer (2,430), half-split (756), decomposition (1,215) and graded-margin and deployed-norm (756) matrices, the scale-check and extension matrices of Appendix O and the re-calibration and baseline matrices of Appendix Q), the metric spectra and optimism curves, and the script that regenerates every number and figure in this paper from those files. Proofs of all propositions are in Appendix A. Numerical precision is treated as an experimental variable and reported in §4 and Appendix E.

ETHICS STATEMENT

This work studies the interaction between two safety-relevant inference-time mechanisms. Its results could in principle inform an attempt to steer a model in a way that evades a monitor; we note that the existence of such directions was established by prior work under an explicit attacker model, that the screen we propose is defensive, and that a wide-margin safety probe is shown here to resist being blinded at practical strengths. The same analysis shows that, at a given displacement, no margin protects a probe from being made to fire, and that a standard CAA steering vector, at norms inside a capability budget, already makes wide-margin monitors fire, reaching the harmfulness probe itself just outside that budget on the primary models (§5.4) and only further out, or not at all, on the scale-check models (Appendix O); that failure is loud rather than silent, and the screen we propose flags 66% of those breaks. The joint budget, which by construction moves the priced monitors less, could in principle be pointed at a harmful target to keep a monitor quiet; we evaluate it only on the three benign behaviours, and deliberate evasion of that kind is already achievable by direct optimisation under the attacker model of the prior work above (Bailey et al., 2024). No human subjects or personal data are involved; the harmful-request dataset (AdvBench) is a standard public benchmark and the only generation from it is the small demonstration of Appendix P.

REFERENCES

Kavin Aravindan, Arihant Rastogi, Krishak Aneja, Aadi Prasad, Saiyam Jain, Vaishnavi Shivkumar, and Ponnurangam Kumaraguru. OPIUM: Mitigating steering externalities and over-refusal via dual objective latent optimization. *arXiv preprint arXiv:2607.19806*, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

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. *arXiv preprint arXiv:2412.09565*, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Hannah Cyberey, Yangfeng Ji, and David Evans. White-box sensitivity auditing with steering vectors. *Transactions on Machine Learning Research*, 2026.

Seyed Arshan Dalili, Ajay Narayanan Sridhar, Vijaykrishnan Narayanan, and Mehrdad Mahdavi. Conditional optimal bridge for Riemannian activation steering. *arXiv preprint arXiv:2607.10517*, 2026.

Evan Duan. Pre-intervention prediction of sparse autoencoder steering side effects. *arXiv preprint arXiv:2606.08365*, 2026a.

Evan Duan. Do activation monitors survive model updates? Benchmarking, predicting, and repairing activation-monitor staleness. *arXiv preprint arXiv:2606.15980*, 2026b.

Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, et al. Fine-grained activation steering: Steering less, achieving more. In *International Conference on Learning Representations (ICLR)*, 2026.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? Knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhishek 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.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

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

Mehdi Jafari, Hao Xue, and Flora Salim. Mechanistic indicators of steering effectiveness in large language models. *arXiv preprint arXiv:2602.01716*, 2026.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. *arXiv preprint arXiv:2603.24543*, 2026.

Haoyan Luo, Mateo Espinosa Zarlenga, and Mateja Jamnik. Don't lose focus: Activation steering via key-orthogonal projections. *arXiv preprint arXiv:2605.06342*, 2026.

Iuri Macocco, Pau Rodríguez, Arno Blaas, et al. On the effectiveness-fluency trade-off in LLM conditioning: A systematic study. *arXiv preprint arXiv:2606.12234*, 2026.

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 (COLM)*, 2024.

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

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM activations are non-surjective. *arXiv preprint arXiv:2604.09839*, 2026.

Neel Nanda. Attribution patching: Activation patching at industrial scale. `https://www.neelnanda.io/mechanistic-interpretability/attribution-patching`, 2023.

Dung V. Nguyen, Hieu M. Vu, et al. Activation steering with a feedback controller. In *International Conference on Learning Representations (ICLR)*, 2026a.

Tam Nguyen, Tu Anh Nguyen, Sina Alemohammad, and Richard G. Baraniuk. Minimizing collateral damage in activation steering. *arXiv preprint arXiv:2605.01167*, 2026b.

Chong Yong Ong, Alson Wei Jie Sim, Peixin Zhang, and Jun Sun. Forecasting side effects of activation steering. *arXiv preprint arXiv:2608.11227*, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Van-Cuong Pham and Thien Huu Nguyen. Householder pseudo-rotation: A novel approach to activation editing in LLMs with direction-magnitude perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations (ICLR)*, 2026.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. Steering without side effects: Improving post-deployment control of language models. *arXiv preprint arXiv:2406.15518*, 2024.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations (ICLR)*, 2025.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *NeurIPS 2023 Workshop on Attributing Model Behavior at Scale*, 2023.

Daniel Tan, David Chanin, Aengus Lynch, Adrià Garriga-Alonso, Dimitrios Kanoulas, Brooks Paige, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

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.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Arthur Vogels, Benjamin Wong, Yann Choho, Annabelle Blangero, and Milan Bhan. In-distribution steering: Balancing control and coherence in language model generation. *arXiv preprint arXiv:2510.13285*, 2025.

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.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Chen Xiong, Zhiyuan He, Pin-Yu Chen, Ching-Yun Ko, and Tsung-Yi Ho. Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models. *arXiv preprint arXiv:2602.04896*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A PROOFS

Throughout, $\boldsymbol{G} = \sum_i \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ is the eigendecomposition of the pullback Fisher metric, $q_i = \boldsymbol{v}_i^\top \boldsymbol{q}$ and $u_i = \boldsymbol{v}_i^\top \bar{\boldsymbol{u}}$. We work on the range of $\boldsymbol{G}$ and write $\boldsymbol{G}^+$ for the pseudo-inverse; the ridge $\alpha \boldsymbol{I}$ in (5) makes every quantity below finite for $\alpha > 0$.

The two bills along the regularised path. With $\boldsymbol{\delta}_\alpha = \rho\, \boldsymbol{c}/(\boldsymbol{q}^\top \boldsymbol{c})$ and $\boldsymbol{c} = (\boldsymbol{G} + \alpha \boldsymbol{I})^{-1} \boldsymbol{q} = \sum_i q_i \boldsymbol{v}_i/(\lambda_i + \alpha)$, we have $\boldsymbol{q}^\top \boldsymbol{c} = \sum_i q_i^2/(\lambda_i + \alpha)$, $\boldsymbol{c}^\top \boldsymbol{G} \boldsymbol{c} = \sum_i \lambda_i q_i^2/(\lambda_i + \alpha)^2$ and $\bar{\boldsymbol{u}}^\top \boldsymbol{c} = \sum_i u_i q_i/(\lambda_i + \alpha)$. Substituting into $D_{\mathrm{out}} = \frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{G} \boldsymbol{\delta}$ and into the common-mode part $B^{\bar{\boldsymbol{u}}}_{\mathrm{mon}} = (\bar{\boldsymbol{u}}^\top \boldsymbol{\delta})^2/\sigma^2$ of $\boldsymbol{\delta}^\top \boldsymbol{N} \boldsymbol{\delta}$ (the dispersion part $\boldsymbol{\delta}_\alpha^\top \boldsymbol{\Sigma} \boldsymbol{\delta}_\alpha/\sigma^2$ has a numerator of the same order in $\lambda_{\min}$ as the denominator, so it does not affect the limit ratio) gives

$$D_{\mathrm{out}}(\boldsymbol{\delta}_\alpha) = \frac{\rho^2}{2} \frac{\sum_i \lambda_i q_i^2/(\lambda_i + \alpha)^2}{\big(\sum_i q_i^2/(\lambda_i + \alpha)\big)^2}, \qquad B^{\bar{\boldsymbol{u}}}_{\mathrm{mon}}(\boldsymbol{\delta}_\alpha) = \frac{\rho^2}{\sigma^2} \frac{\big(\sum_i u_i q_i/(\lambda_i + \alpha)\big)^2}{\big(\sum_i q_i^2/(\lambda_i + \alpha)\big)^2}. \tag{9}$$

Proof of Proposition 1. Let m index the (simple) eigendirection with eigenvalue $\lambda_{\min}$ and suppose $q_m u_m \neq 0$; the limit $\alpha \to 0$ is taken first, with the rest of the spectrum and of $\boldsymbol{q}, \bar{\boldsymbol{u}}$ held fixed. As $\alpha \to 0$ every sum in (9) converges (they run over the range of $\boldsymbol{G}$). Isolating the m-th terms and writing A, B, C for the remaining sums, which do not depend on $\lambda_{\min}$: $D_{\mathrm{out}}(\boldsymbol{\delta}_\alpha) \to \frac{\rho^2}{2}\,(q_m^2/\lambda_{\min} + A)/(q_m^2/\lambda_{\min} + B)^2 \sim \frac{\rho^2}{2} \lambda_{\min}/q_m^2 \to 0$, while $B^{\bar{\boldsymbol{u}}}_{\mathrm{mon}}(\boldsymbol{\delta}_\alpha) \to \frac{\rho^2}{\sigma^2}(u_m q_m/\lambda_{\min} + C)^2/(q_m^2/\lambda_{\min} + B)^2 \to \rho^2 u_m^2/(\sigma^2 q_m^2)$, a positive constant. Their ratio R therefore behaves as $2u_m^2/(\sigma^2 \lambda_{\min}) = \Theta(1/\lambda_{\min})$: the divergence is entirely the output bill vanishing, not the monitor bill growing. The hypothesis needs $\boldsymbol{q}$ and $\boldsymbol{G}$ to be averaged over different inputs. Were both taken over the same x, Cauchy–Schwarz in the $\boldsymbol{H}(x)$ inner product would give $q_m^2 \leq \bar{\kappa}\, \lambda_m$ with $\bar{\kappa} = \mathbb{E}_x[(\boldsymbol{e}_+ - \boldsymbol{e}_-)^\top \boldsymbol{H}(x)^+ (\boldsymbol{e}_+ - \boldsymbol{e}_-)]$, hence $D_{\mathrm{out}} \geq \rho^2/(2\bar{\kappa})$ on V_ρ, and the output bill on the behaviour's own prompts could not vanish. In the paper $\boldsymbol{G}$ is estimated on generic text and $\boldsymbol{q}$ on the behaviour's prompts (§4), so the statement concerns an output audit on generic text, which is how the output bill is measured throughout. The proposition is about the quadratic form (1); the KL itself carries an $O(\|\boldsymbol{\delta}\|^3)$ remainder that does not vanish with $\lambda_{\min}$ along this path, since $\|\boldsymbol{\delta}_\alpha\| \to \rho/|q_m|$. A repeated $\lambda_{\min}$ must be treated with its eigenspace projector $\boldsymbol{P}$, giving the limit $\rho^2(\bar{\boldsymbol{u}}^\top \boldsymbol{P} \boldsymbol{q})^2/(\sigma^2 \|\boldsymbol{P} \boldsymbol{q}\|^4)$, which can be zero; hence the simplicity assumption. □

Proof of Proposition 2. Any $\boldsymbol{\delta} \in V_\rho$ satisfies $\rho = \boldsymbol{q}^\top \boldsymbol{\delta} \leq \|\boldsymbol{q}\| \|\boldsymbol{\delta}\|$ with equality iff $\boldsymbol{\delta} \propto \boldsymbol{q}$, so $\boldsymbol{\delta}_{\boldsymbol{q}} = \rho \boldsymbol{q}/\|\boldsymbol{q}\|^2$ is the minimum-norm member. For the output-optimal member, $\|\boldsymbol{\delta}_{\boldsymbol{G}}\|^2 = \rho^2 \|\boldsymbol{G}^+ \boldsymbol{q}\|^2/(\boldsymbol{q}^\top \boldsymbol{G}^+ \boldsymbol{q})^2 = \rho^2 \sum_i q_i^2 \lambda_i^{-2}/\big(\sum_i q_i^2 \lambda_i^{-1}\big)^2$. Dividing by $\|\boldsymbol{\delta}_{\boldsymbol{q}}\|^2 = \rho^2/\|\boldsymbol{q}\|^2$ and writing $\mu_i = q_i^2/\|\boldsymbol{q}\|^2$,

$$\frac{\|\boldsymbol{\delta}_{\boldsymbol{G}}\|^2}{\|\boldsymbol{\delta}_{\boldsymbol{q}}\|^2} = \frac{\sum_i \mu_i \lambda_i^{-2}}{\big(\sum_i \mu_i \lambda_i^{-1}\big)^2} = \frac{\mathbb{E}_\mu[\lambda^{-2}]}{\mathbb{E}_\mu[\lambda^{-1}]^2} = 1 + \frac{\mathrm{Var}_\mu(\lambda^{-1})}{\mathbb{E}_\mu[\lambda^{-1}]^2},$$

which is ≥ 1 by Jensen, with equality iff λ^{-1} is μ-almost surely constant, i.e. iff $\boldsymbol{q}$ is supported on a single eigenspace. □

Proof of Proposition 3. Minimising $\frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{G} \boldsymbol{\delta} + \frac{\lambda}{2} \boldsymbol{\delta}^\top \boldsymbol{N} \boldsymbol{\delta}$ subject to $\boldsymbol{q}^\top \boldsymbol{\delta} = \rho$ is a convex quadratic programme with one linear constraint; stationarity gives $(\boldsymbol{G} + \lambda \boldsymbol{N}) \boldsymbol{\delta} = \nu \boldsymbol{q}$, so $\boldsymbol{\delta} \propto (\boldsymbol{G} + \lambda \boldsymbol{N})^+ \boldsymbol{q}$ (with $\boldsymbol{\delta}$ in the range). Scalarisation of a two-objective convex problem with non-negative weights traces its Pareto frontier. For non-singular $\boldsymbol{G}, \boldsymbol{N}$ each objective has the unique constrained minimiser $\propto \boldsymbol{G}^{-1} \boldsymbol{q}$ or $\boldsymbol{N}^{-1} \boldsymbol{q}$; if these are parallel the frontier is a single point, and conversely a one-point frontier means a common minimiser exists, so they are parallel. When $\boldsymbol{N}$ is singular and $\boldsymbol{q}$ has a component in its null space, the monitor-bill minimiser is an affine set on which $B_{\mathrm{mon}} = 0$ and the frontier reaches the axis; the ridge in the solver keeps every $\boldsymbol{\delta}_\lambda$ finite, and the empirical consequences of that null space are measured in §5.5 and Appendix I. □

Proof of Proposition 4. Under $s \mid y{=}0 \sim \mathcal{N}(\mu_0, \sigma_w^2)$ and $s \mid y{=}1 \sim \mathcal{N}(\mu_1, \sigma_w^2)$, the threshold with clean false-positive rate ε is $\tau = \mu_0 + z_{1-\varepsilon} \sigma_w$, and the clean true-positive rate is $\Pr(s > \tau \mid y{=}1) = \Phi\big((\mu_1 - \tau)/\sigma_w\big) = \Phi(d' - z_{1-\varepsilon})$. To first order the intervention adds $\Delta_1 = \bar{\boldsymbol{u}}_{|1}^\top \boldsymbol{\delta}$

to the positive-class mean and $\Delta_0 = \bar{\boldsymbol{u}}_{|0}^\top \boldsymbol{\delta}$ to the negative-class mean, and leaves τ frozen, giving (7); the variance inflation $\boldsymbol{\delta}^\top \boldsymbol{\Sigma} \boldsymbol{\delta}$ is second order in $\boldsymbol{\delta}$. (a) $\partial\,\mathrm{TPR}_{\boldsymbol{\delta}}/\partial(\Delta_1/\sigma_w) = \varphi(d' - z_{1-\varepsilon})$ at $\Delta_1 = 0$, and $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$; a monitor with headroom $d' - z_{1-\varepsilon} = h > 0$ needs $|\Delta_1|/\sigma_w \geq h - \Phi^{-1}(\Phi(h) - c)$ before its true-positive rate falls by c, which for $h \geq 3$ equals $h - z_{1-c}$ to within 1%. (b) Setting $\Phi(d' - z_{1-\varepsilon} + \Delta_1/\sigma_w) = \Phi(d' - z_{1-\varepsilon}) - c$ and solving for $|\Delta_1|/\sigma_w$ gives the numerator of (8); dividing by d' gives $t^\star$. The stated range is obtained by evaluating (8) at $c = 0.2$, $\varepsilon = 0.05$ on the 45 measured per-block margins (Tables 2 and 12). The false-positive rate is $\Phi(-z_{1-\varepsilon} + \Delta_0/\sigma_w)$ with derivative $\varphi(z_{1-\varepsilon})$ at $\Delta_0 = 0$, which contains no d'; setting it to $\varepsilon + c$ gives Δ_F. $t^\star$ is *not* monotone in d': it falls to a minimum of 0.236 near $d' = 2.40$ and rises on either side, approaching $1 - (z_{1-\varepsilon} + z_{1-c})/d'$ from above as d' grows, and diverging as $d' \downarrow z_{1-\varepsilon} + \Phi^{-1}(c)$, where the clean true-positive rate itself falls to c and a drop of c becomes unreachable. Over the margins we measure it stays inside $[0.257, 0.940]$; the content of (b) is the compression, not a constant. The break displacement $\Delta_T = d'\, t^\star(d')$ itself is defined for $d' > 0.803$, diverges at that edge, falls to a minimum of 0.507 near $d' = 1.90$ and then grows; it exceeds Δ_F on $(0.803, 0.970)$, a sliver where the clean true-positive rate is barely above c, and on $(3.29, \infty)$, which is the range the ordering statement of (b) concerns. □

B A REMARK ON OUTPUT INVISIBILITY

Let $N_\tau(\boldsymbol{G})$ be the span of eigenvectors with $\lambda_i \leq \tau$. Any $\boldsymbol{\delta} \in N_\tau(\boldsymbol{G})$ has $D_{\mathrm{out}}(\boldsymbol{\delta}) \leq \frac{1}{2}\tau\|\boldsymbol{\delta}\|^2$, so an audit based on the output distribution can charge it no more than that to second order, while a monitor with common-mode mass $\|P_{N_\tau}\bar{\boldsymbol{u}}\|^2$ in that subspace can be displaced by up to $\|P_{N_\tau}\bar{\boldsymbol{u}}\|\,\|\boldsymbol{\delta}\|$ (attained when $\boldsymbol{\delta} \parallel P_{N_\tau}\bar{\boldsymbol{u}}$). This is consistent with the empirical finding that an attacker can find monitor-evading directions (Bailey et al., 2024); it is not a new existence claim. We add one measured caveat: $\boldsymbol{G}$ is strongly anisotropic (participation ratio 52.6–247 against d of 1,536–3,584) but not low-rank (the rank capturing 99% of its trace is 0.461–0.856 d), so the certificate concerns a genuinely small-eigenvalue subspace rather than a large null space.

C ESTIMATION AND IMPLEMENTATION DETAILS

Metric estimation. For each of 512 (primary depth) or 160 (other depths) WikiText-2 passages truncated to 96 tokens, we compute the exact softmax Fisher $\boldsymbol{H}(x) = \mathrm{Cov}_{p(x)}(\boldsymbol{e})$ at the final position, take its top 384 eigenpairs $(\eta_k, \boldsymbol{a}_k)$, and accumulate $\boldsymbol{G} \mathrel{+}= \sum_k \eta_k (\boldsymbol{J}^\top \boldsymbol{a}_k)(\boldsymbol{J}^\top \boldsymbol{a}_k)^\top$ using one VJP per eigenpair, chunked to fit memory. The fraction of $\mathrm{Tr}\,\boldsymbol{H}$ captured by the truncation is 95%–98% (Table 1); with $r = 128$ it was as low as 0.87 on Llama-3.2-3B, and because truncation flatters output-frugal directions we chose the conservative rank. Snapshots of the running $\boldsymbol{G}_n$ at $n \in \{8, 16, 32, 64, 128, 192, 256, 384, 512\}$, up to the number of passages used at that depth, are saved in the same pass; $\boldsymbol{G}_B$ is estimated identically on 256 disjoint passages. Eigendecompositions are done in `float32` on the GPU and checked against a `float64` host path in the unit tests.

Monitors. Each probe is a logistic regression on standardised residual-stream features at the last token of ℓ_j, trained on 260 examples and evaluated on 140; the threshold is the empirical 95th percentile of clean negative scores. Monitor covectors $\boldsymbol{u}_j(x) = \boldsymbol{K}_j(x)^\top \boldsymbol{w}_j$ are computed by one VJP per held-out example with the probe's raw-space weight as cotangent, through all positions; $\boldsymbol{N}_j$ is their second moment divided by the clean score variance, and the $\boldsymbol{N}$ of the joint budget is the mean of the five. Per-block margins and common-to-differential ratios are in Table 2.

Directions. The Fisher path solves $(\boldsymbol{G} + \alpha\bar{\lambda}\boldsymbol{I})\boldsymbol{\delta} = \boldsymbol{q}$ with $\bar{\lambda}$ the mean eigenvalue of $\boldsymbol{G}$; the joint budget solves $(\boldsymbol{G} + \lambda\boldsymbol{N} + 10^{-4}\bar{\lambda}\boldsymbol{I})\boldsymbol{\delta} = \boldsymbol{q}$. Every direction is then scaled so that $\boldsymbol{q}^\top\boldsymbol{\delta} = \rho$. Random controls draw a fixed-seed Gaussian direction and scale it to the norm of the reference direction.

Compute. All 1,755 cells, the optimism curves, the spectra and the controls ran on one shared 24 GB GPU in about 34 wall-clock hours; a cell of the Qwen2.5-7B matrix takes 2–6 s. The scale check and the extensions of Appendix O ran on one 48 GB GPU in about 11 wall-clock hours with several jobs sharing the card, rebuilding the primary blocks' metrics and monitors from scratch,

which reproduce the primary matrix to within `float16` noise. The only generation in the paper is in the two controls of Appendix P.

Table 1: Spectrum of $\boldsymbol{G}$ per block at the largest estimation sample.

model	ℓ	d	n	$\operatorname{Tr} \boldsymbol{G}$	participation ratio	$\text{rank}_{99\%}/d$	captured by $r = 384$
Qwen2.5-7B	7	3584	160	2.32	52.6	0.46	0.982
Qwen2.5-7B	14	3584	512	1.71	211.8	0.77	0.979
Qwen2.5-7B	21	3584	160	0.38	134.5	0.68	0.982
Llama-3.2-3B	7	3072	160	50.97	80.4	0.57	0.949
Llama-3.2-3B	14	3072	512	22.78	183.8	0.81	0.950
Llama-3.2-3B	21	3072	160	7.44	247.4	0.84	0.949
Qwen2.5-1.5B	7	1536	160	1.07	72.1	0.70	0.974
Qwen2.5-1.5B	14	1536	512	1.33	209.0	0.86	0.971
Qwen2.5-1.5B	21	1536	160	0.39	236.8	0.84	0.974

Table 2: Clean monitor margins d' and common-to-differential covector ratios $\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$ per block.

model	ℓ	harmfulness		truth (sp–en)		truth (cities)		truth (larger)		truth (claims)	
		d'	$\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$	d'	$\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$	d'	$\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$	d'	$\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$	d'	$\|\bar{\boldsymbol{u}}\|/\|\Delta\bar{\boldsymbol{u}}\|$
Qwen2.5-7B	7	6.85	0.94	5.40	3.05	4.31	2.44	5.60	2.06	1.02	6.09
Qwen2.5-7B	14	13.12	0.55	11.97	0.59	10.76	0.57	7.71	0.61	1.98	1.93
Qwen2.5-7B	21	14.41	0.61	11.46	0.76	9.28	0.67	9.40	0.75	1.82	2.26
Llama-3.2-3B	7	10.92	0.51	5.64	0.94	6.87	0.85	5.81	1.38	1.56	2.17
Llama-3.2-3B	14	11.93	0.54	5.69	2.00	6.78	1.41	4.67	2.17	1.45	2.85
Llama-3.2-3B	21	14.73	0.62	4.68	2.14	6.11	1.30	4.12	2.90	1.38	4.81
Qwen2.5-1.5B	7	6.25	1.08	3.54	6.14	3.27	4.31	3.70	14.88	0.99	7.46
Qwen2.5-1.5B	14	12.52	0.65	4.59	1.78	4.41	1.35	5.27	2.25	1.43	3.09
Qwen2.5-1.5B	21	12.59	0.74	4.36	3.55	3.99	2.27	4.65	4.20	1.26	4.56

D PILOT: THE FALSIFIED MECHANISM AND THE VERIFIED NORM PRICE

Before the main run we tested, on Qwen2.5-0.5B-Instruct at $\ell = 12$ with the monitor at $\ell_j = 18$, the working hypothesis, fixed before the main run, that monitor covectors sit in the near-null space of $\boldsymbol{G}$. It is false, with the sign reversed: the fraction of squared mass carried by the subspace of $\boldsymbol{G}$ holding $\leq 1\%$ of its trace is 0.53 for a random direction and 0.32–0.34 for the three monitor covectors tested, which look spectrally like on-target covectors (0.34–0.39). Monitors read the model's computationally active subspace, so the reason output-frugality fails to buy monitor-frugality is the norm price of Proposition 2, not kernel occupancy. As an implementation check of (6) on the same model, the predicted $\|\boldsymbol{\delta}_{\boldsymbol{G}}\|^2/\|\boldsymbol{\delta}_{\boldsymbol{q}}\|^2$ of 9.77 (sycophancy) and 6.62 (refusal) agree with the solver's own output to 9 significant figures; this checks the code, not the theory. Both sides are recomputed by the released script. The pilot also established that contrastive-activation-addition difference-of-means vectors have small cosine with $\boldsymbol{q}$, so that matching them to $\boldsymbol{q}^\top\boldsymbol{\delta} = \rho$ inflates their norm to 1.6–$12.5 \times \|\boldsymbol{h}_\ell\|$ and destroys the model; they are reported separately in the released results and excluded from the matched family. The pilot's own headline, a monitor's true-positive rate falling to zero at unchanged AUROC, is reproduced in weaker form at scale (true-positive rate halved at $\geq 95\%$ of clean AUROC, §5.2), but the pilot's strengths were outside the linear regime (ARC-Easy $0.65 \to 0.31$), and the base rate at practical strengths is the one reported in §5.1. Appendix Q repeats the spectral measurement on all nine primary blocks.

E NUMERICAL PRECISION

Table 3: Precision control on Qwen2.5-1.5B: identical cells in `float32` and `float16` at $\ell = 14$, and in `bfloat16` at all three depths (540 cells beyond the 1,215 of the primary matrix). `float16` KL is unreliable below $\rho = 1$ and `bfloat16` KL well beyond it; the monitor bill is neither.

ρ	median KL (`float32`)	`float16` / `float32`	`bfloat16` / `float32`	Spearman bill `float16` vs `float32`
0.25	4.80e-06	4.12×	206×	0.995
0.5	1.92e-05	1.91×	46.8×	0.999
1.0	7.70e-05	1.18×	14.6×	0.997
2.0	3.11e-04	1.03×	4.53×	0.999
4.0	1.28e-03	1.01×	1.95×	0.999

Table 3 pairs the `float16`, `bfloat16` and `float32` runs of the same cells. Generic KL at `bfloat16` sits on a floor near 10^{-3} nats where `float32` measures 10^{-5} or less; `float16`, with three more mantissa bits, agrees with `float32` within 18% in the median from $\rho = 1$ upward. The monitor bill agrees across all three precisions at every strength (rank correlation 0.999 and 0.966 with `float32`), because it is a displacement of a scalar score in units of that score's spread. This is why the primary matrix is `float16` and why KL-based statistics are restricted to $\rho \geq 1$.

F ADDITIONAL RESULTS

Table 4: Per-block prediction quality and breakage counts (`float16` primary matrix). "blinded": non-trivial readings whose true-positive rate fell by ≥ 0.2; "false-alarmed": whose false-positive rate rose by ≥ 0.2; "broke": either. Breakage concentrates in the smallest primary model at the shallowest depth, but every primary model contributes.

model	ℓ	readings	Spearman (bill)	med. ratio	MAE TPR	MAE AUROC	blinded	false-alarmed	broke
Qwen2.5-7B	7	675	0.947	1.06	0.0010	0.0003	9	6	15
Qwen2.5-7B	14	675	0.994	1.04	0.0002	0.0000	0	0	0
Qwen2.5-7B	21	675	0.994	1.04	0.0002	0.0000	0	0	0
Llama-3.2-3B	7	675	0.980	1.03	0.0003	0.0001	0	0	0
Llama-3.2-3B	14	675	0.993	1.04	0.0007	0.0001	0	6	6
Llama-3.2-3B	21	675	0.980	1.05	0.0034	0.0002	0	0	0
Qwen2.5-1.5B	7	675	0.922	1.07	0.0355	0.0065	51	40	91
Qwen2.5-1.5B	14	675	0.990	1.02	0.0074	0.0016	6	39	45
Qwen2.5-1.5B	21	675	0.972	1.03	0.0026	0.0006	2	21	23

Table 5: Matching check: measured on-target effect of each direction relative to Euclidean steering at the same ρ, over all model–depth–behaviour–strength cells.

direction	median $\Delta\psi/\Delta\psi_{\mathrm{Eucl}}$	IQR
Fisher $\alpha = 1$	0.999	$[0.991, 1.005]$
Fisher $\alpha = 10^{-2}$	0.996	$[0.986, 1.004]$
Fisher $\alpha = 10^{-4}$	0.997	$[0.987, 1.004]$
Joint $\lambda = 10^{-2}$	0.997	$[0.986, 1.004]$
Joint $\lambda = 10^{-1}$	0.996	$[0.985, 1.007]$
Joint $\lambda = 1$	0.999	$[0.984, 1.007]$

Table 6: Trigger bins over all 4,725 non-trivial readings of the primary matrix, by the pooled shift $|\Delta\bar{s}|$ in clean score spreads σ divided by the margin (a pooled analogue of t in Proposition 4). "move": the larger of the true-positive drop and the false-positive rise.

$\lvert\Delta\bar{s}\rvert/(\sigma d')$	readings	median move	p90 move	blinded	false-alarmed	broke	median AUROC ratio
$[0, 0.1)$	4274	0.000	0.014	0.0%	0.0%	0.1%	1.000
$[0.1, 0.25)$	227	0.057	0.207	0.9%	9.7%	10.6%	1.000
$[0.25, 0.5)$	117	0.279	0.557	16.2%	41.9%	58.1%	0.997
$[0.5, 1)$	56	0.664	0.936	30.4%	46.4%	76.8%	0.993
≥ 1	51	0.806	0.943	54.9%	27.5%	82.4%	0.921

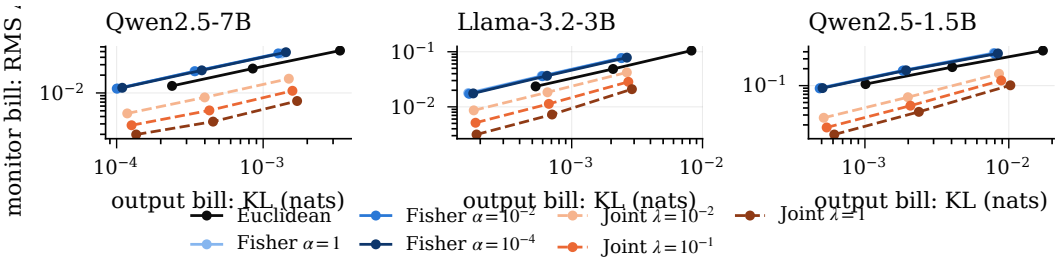


Figure 3: The two bills at matched effect, $\rho \in \{1, 2, 4\}$ (successive points), geometric means over behaviours, depths and monitors. At each ρ the output-frugal directions (blue) pay far less KL than Euclidean steering (black) but only a slightly smaller monitor bill (0.806× at matched effect), so as curves in the plane they lie at or above it; the joint budget (orange, dashed) lies below both.

Table 7: Exchange rate and joint budget at KL-reliable strengths ($\rho \geq 1$) over 223 matched pairs; ratios are medians relative to Euclidean steering. Left: the Fisher path cuts the output bill far more than the monitor bill, so $R = B_{\text{mon}}/D_{\text{out}}$ worsens. Right: the joint budget; "dominates": lower than Euclidean on both axes; "beats frugal": lower monitor bill than the $\alpha = 10^{-4}$ direction at KL within 25%. In-sample with respect to the inputs on which $\boldsymbol{N}$ is estimated.

α	KL saving	$B_{\text{mon}}^{1/2}$ ratio	R ratio	R worse	λ	dominates	beats frugal	$B_{\text{mon}}^{1/2}$ ratio	D_{out} ratio
1	2.56×	0.858×	1.86×	83%	10^{-2}	97%	92%	0.256	0.440
10^{-2}	2.58×	0.783×	1.71×	72%	10^{-1}	97%	78%	0.165	0.450
10^{-4}	2.39×	0.806×	1.65×	72%	1	86%	60%	0.116	0.455

Table 8: Breakage by the monitor's per-block clean margin, over the 4,725 non-trivial readings, split by side. Blinding (true-positive rate falls by 0.2) is confined to small margins; false alarms (false-positive rate rises by 0.2) are not, and 66 of the 112 occur at $d' \in [4, 6)$ on 3 models. The $[2, 4)$ margins occur only on Qwen2.5-1.5B (2 blocks), so in that bin margin and block cannot be separated; the $[4, 6)$ bin spans all three models. No monitor with $d' \geq 6$ broke on either side at these strengths.

per-block d'	readings	models	blinded	false-alarmed	broke	break rate
$[0, 2)$	945	3	27	1	28	3%
$[2, 4)$	420	1	39	45	84	20%
$[4, 6)$	1,470	3	2	66	68	5%
≥ 6	1,890	3	0	0	0	0%

Table 9: Exchange rate ($\alpha = 10^{-4}$ against Euclidean) and joint budget ($\lambda = 10^{-2}$) by strength, all five strengths. KL at $\rho < 1$ is at the `float16` noise floor and is shown for completeness only.

ρ	pairs	Fisher KL saving	Fisher $B_{\text{mon}}^{1/2}$ ratio	Fisher R ratio	joint dominates	joint $B_{\text{mon}}^{1/2}$ ratio	joint $D_{\text{out}}/D_{\text{out}}^{\text{Eucl}}$
0.25	17	1.43×	0.74×	0.79×	100%	0.173	0.842
0.5	34	2.04×	0.75×	1.16×	100%	0.175	0.523
1.0	52	2.23×	0.75×	1.24×	100%	0.187	0.450
2.0	70	2.47×	0.74×	1.57×	99%	0.222	0.438
4.0	101	2.39×	0.86×	1.77×	94%	0.353	0.440

Table 10: Pre-deployment screens for monitor breakage, all evaluated on the same 4,725 non-trivial readings (positive: the operating point moves by 0.2 on either side; 180 positives). Each theory rule is paired with a pooled empirical threshold tuned to the same flag count on the same input. "measured" uses the realised displacement Δ; "predicted" uses $\hat{\Delta} = \bar{\boldsymbol{u}}^{\top}\boldsymbol{\delta}$ or the per-input first-order scores, neither of which needs a steered forward pass. $\Delta^{\star}(d') = \min\{\Delta_T(d'), \Delta_F\}$ is the two-sided break displacement of Corollary 1, in the measured within-class spread and with no fitted constant; the last two rows split the per-input rule by side.

screen	input	flagged	precision	recall
theory, $\lvert\Delta\rvert/\sigma_w \geq \Delta^{\star}(d')$ (Cor. 1)	measured Δ	192	0.792	0.844
pooled threshold at equal flags ($t \geq 0.307$)	measured Δ	192	0.708	0.756
the same at half the break displacement	measured Δ	381	0.462	0.978
blinding-only rule, $\lvert\Delta\rvert/\sigma_w \geq \Delta_T(d')$	measured Δ	117	0.889	0.578
theory, $\lvert\hat{\Delta}\rvert/\sigma_w \geq \Delta^{\star}(d')$	predicted $\hat{\Delta}$	145	0.745	0.600
pooled threshold at equal flags ($t \geq 0.343$)	predicted $\hat{\Delta}$	145	0.593	0.478
per-input first-order, either rate moves ≥ 0.2 (§5.1)	predicted scores	119	0.874	0.578
blinding side only (TPR drop ≥ 0.2; 68 positives)	predicted scores	62	0.790	0.721
false-alarm side only (FPR rise ≥ 0.2; 112 positives)	predicted scores	58	0.879	0.455

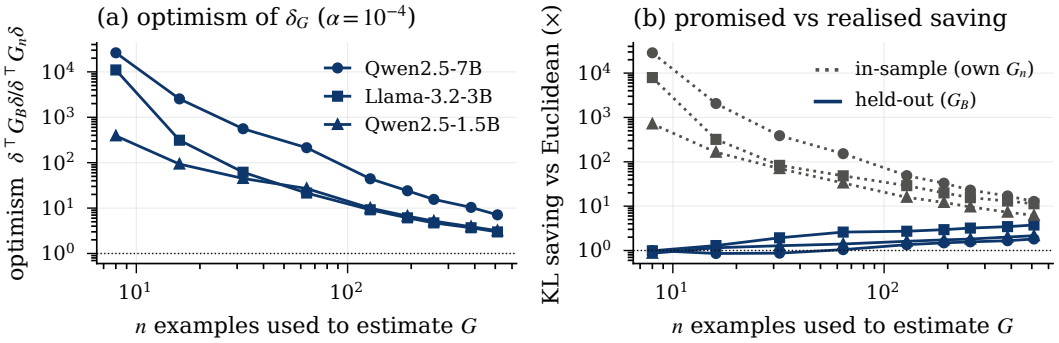


Figure 4: In-sample optimism of the output-optimal direction. (a) The ratio of the direction's output bill under an independently estimated metric $\boldsymbol{G}_B$ to its bill under the $\boldsymbol{G}_n$ that chose it, as a function of the n passages used to estimate $\boldsymbol{G}_n$; the direction is chosen with $\boldsymbol{G}_n$ at every n. (b) The KL saving over Euclidean steering that $\boldsymbol{G}_n$ promises in-sample (dotted) and that $\boldsymbol{G}_B$ awards held-out (solid). Medians over behaviours.

Table 11: Optimism of the Fisher path at $n = 512$ ($\ell = 14$): bill under $\boldsymbol{G}_B$ divided by bill under $\boldsymbol{G}_n$; KL saving over Euclidean steering promised in-sample, awarded by $\boldsymbol{G}_B$, and measured on held-out text. Medians over behaviours.

model	α	optimism	in-sample saving	held-out saving ($\boldsymbol{G}_B$)	measured saving
Qwen2.5-7B	1	2.65	5.4×	2.20×	2.10×
Qwen2.5-7B	10^{-2}	6.66	12.7×	1.92×	2.18×
Qwen2.5-7B	10^{-4}	7.10	12.8×	1.82×	2.05×
Llama-3.2-3B	1	1.65	5.6×	3.29×	2.61×
Llama-3.2-3B	10^{-2}	2.89	11.2×	3.85×	2.69×
Llama-3.2-3B	10^{-4}	2.99	11.3×	3.75×	2.58×
Qwen2.5-1.5B	1	1.78	3.8×	2.11×	2.26×
Qwen2.5-1.5B	10^{-2}	3.09	6.2×	2.21×	2.58×
Qwen2.5-1.5B	10^{-4}	3.18	6.2×	2.16×	2.63×

Table 12: Margin protects one side only. Per probe: clean margin d' over the nine blocks, Gaussian sensitivity of the true-positive rate $\varphi(d' - z_{0.95})$ (Prop. 4a) in the weakest block (the false-positive sensitivity is 0.103 for every probe), break point $t^\star(d')$ (8) at both extremes, the worst true-positive rate and largest false-positive rise measured, and the fraction of readings broken on each side.

probe	d' (min–max)	$\varphi(d'_{\min} - z)$	$t^\star(d'_{\min})$–$t^\star(d'_{\max})$	worst TPR	max FPR rise	blinded	false-alarmed
harmfulness	6.25–14.7	1.0×10^{-5}	0.602–0.831	0.957	0.171	0%	0%
truth (sp–en)	3.54–12.0	0.067	0.325–0.792	0	0.871	2%	2%
truth (cities)	3.27–10.8	0.107	0.292–0.769	0	0.912	2%	2%
truth (larger)	3.70–9.40	0.049	0.346–0.735	0.314	0.943	0%	8%
truth (claims)	0.991–1.98	0.322	0.940–0.257	0	0.211	3%	0%

The measured on-target effect of each non-Euclidean direction is 0.996–0.999 times Euclidean steering's at the same ρ for all six families (Table 5); and the random controls move ψ by at most 0.441. ARC-Easy accuracy is a coarse instrument (64 items, so one question is 0.016), and it does not move: the 90th-percentile drop is 0.016, one question, with 1% of cells above 0.05 and a worst case of 0.188 at $\rho = 4$. Generic KL is the capability statistic we rely on. The bills are directional rather than norm effects: at equal $\|\boldsymbol{\delta}\|$ a structured direction displaces the monitors 1.76× as much as a norm-matched random one. Table 8 stratifies breakage by the per-block margin and by side: blinding is confined to $d' < 6$ and concentrated below 4 (where the CommonClaim probe's clean true-positive rate is itself low, so a 0.2 drop is a large relative loss), while false alarms peak at $d' \in [4, 6)$. Table 4 gives both sides per block: 91 of the 180 breaks are on Qwen2.5-1.5B at $\ell = 7$, and 5 of the nine blocks, on all three models, contribute. The prediction quality is uniform across blocks (Spearman 0.922–0.994). Table 9 shows the joint budget's domination and monitor-bill ratio at every strength, including those where KL is unreliable; the monitor-side numbers are stable across ρ. On the optimism of Appendix M: the fitted decay puts optimism below 1.2 only beyond 3 thousand passages, and ambient dimension does not drive it: Llama-3.2-3B ($d = 3{,}072$) and Qwen2.5-1.5B ($d = 1{,}536$) are equally optimistic while Qwen2.5-7B ($d = 3{,}584$) is more than twice as much, so the spectrum of $\boldsymbol{G}$ is what matters.

G TRANSFER OF THE JOINT BUDGET TO HELD-OUT MONITORS

For every block, behaviour, strength and λ we rebuild $\boldsymbol{N}$ from four of the five monitors, re-solve $\boldsymbol{\delta}_\lambda$, and evaluate all five, so that the held-out reading and the four in-$\boldsymbol{N}$ readings come from the same forward passes; separately we build a monitor-agnostic $\boldsymbol{N}_{\mathrm{rand}}$ from 8 random unit probes at the monitor layer over 160 inputs. This is 2,430 additional cells. The directions remain matched (median $\Delta\psi$ relative to Euclidean steering 0.996) and capability is untouched (90th-percentile ARC-Easy drop 0.016).

Table 13 reports, for each monitor j, its measured bill under the joint budget divided by its bill under Euclidean steering at the same ρ, under the three pricings: j in $\boldsymbol{N}$; $\boldsymbol{N}$ from the four monitors other

than j; and $\boldsymbol{N}$ from random probes. All are restricted to KL-reliable strengths and to pairs where the Euclidean displacement on j is at least 0.05σ, and all three are evaluated on the identical index set, so the columns are paired.

Table 13: Transfer of the joint budget. Each entry is the held-out monitor's measured $B^{1/2}_{\mathrm{mon}}$ relative to its value under Euclidean steering at matched effect (lower is better; 1.0 = no protection); the last two rows give each pricing's KL relative to Euclidean and its rate of dominating Euclidean on both axes, over the same pairs. "in $\boldsymbol{N}$": the monitor was priced. "four others": $\boldsymbol{N}$ was built from the other four monitors. "random probes": $\boldsymbol{N}$ was built from 8 random unit probes at the monitor layer, with no knowledge of any monitor.

		$\lambda = 10^{-2}$			$\lambda = 10^{-1}$		
monitor j	pairs	in $\boldsymbol{N}$	four others	random probes	in $\boldsymbol{N}$	four others	random probes
harmfulness	20	0.610	0.669	0.610	0.511	0.594	0.449
truth (sp–en)	48	0.215	0.515	0.503	0.151	0.483	0.330
truth (cities)	45	0.244	0.482	0.493	0.157	0.464	0.444
truth (larger)	53	0.204	0.557	0.568	0.153	0.418	0.378
truth (claims)	57	0.299	0.665	0.603	0.162	0.554	0.479
all	223	0.256	0.579	0.542	0.165	0.497	0.404
KL / Euclidean		0.440	0.442	0.392	0.450	0.448	0.422
dominates Euclidean		97%	81%	88%	97%	83%	91%

The last two rows compare the priors on the output axis as well, because the priors have different traces and the same nominal λ is not the same trade-off: the random-probe prior pays *less* KL than the four-sibling prior at both λ (0.392 against 0.442 and 0.422 against 0.448) and dominates Euclidean steering more often, so its lower monitor bill is not bought with output. Pricing the monitor is worth roughly twice what not pricing it is for four of the five, and the two ignorant pricings are close to each other, with the random-probe prior ahead at $\lambda = 10^{-1}$ for all five monitors (3 of five at 10^{-2}). The harmfulness probe is the exception at every λ: pricing it explicitly ties the ignorant prior at $\lambda = 10^{-2}$ (0.610 against 0.610) and *loses* to it at $\lambda = 10^{-1}$ (0.511 against 0.449). Its covector is close to the layer's generic sensitivity, so there is nothing monitor-specific left for $\boldsymbol{N}$ to exploit, and at the larger λ the extra structure only costs the solver rank. Cluster-bootstrapped over the nine model–depth blocks, the four-sibling prior minus the random prior is 0.036 $[-0.024, 0.106]$ at $\lambda = 10^{-2}$, 0.092 $[0.034, 0.125]$ at $\lambda = 10^{-1}$ and 0.057 $[0.017, 0.141]$ at $\lambda = 1$: knowing four sibling monitors is not better than knowing none, and at the larger λ it is measurably worse. The random prior is not an approximation to any monitor: for $\boldsymbol{w}$ uniform on the unit sphere, $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}\boldsymbol{w}^\top] = \boldsymbol{I}/d$, so in expectation over the probes and up to the per-probe variance weights $\boldsymbol{N}_{\mathrm{rand}} = \mathbb{E}_x[\boldsymbol{K}_j^\top \boldsymbol{K}_j]/d$, with 8 probes a noisy estimate of the layer's own generic sensitivity, a prior that says only "do not move the residual stream at ℓ_j". That this ignorant prior transfers as well as four sibling monitors says the transferable part of $\boldsymbol{N}$ *is* that generic term; by (3) what remains is the monitor-specific $\bar{\boldsymbol{u}}_j \bar{\boldsymbol{u}}_j^\top + \boldsymbol{\Sigma}_j$, which is the natural place to locate the remaining factor of two, though we have not isolated it experimentally. A practical corollary is that the random-probe prior is the right default when the monitor stack is unknown or changing, since it costs one pass of 8 VJPs and needs no access to any probe's weights.

H THE SCREEN BY DIRECTION FAMILY

Table 14: Operating characteristics of the breakage screen, separately for each direction family, over the 4,725 non-trivial readings. "predicted": the pre-deployment rule of §5.1, which uses only the covectors and the clean scores. "measured": the pooled displacement rule of Corollary 1, applied to the displacement measured in one steered forward pass over the probe's inputs. The joint budget is chosen by minimising the first-order form the predicted rule reads, so the predicted rule is blind on it.

			predicted $\hat{\Delta}$ (pre-deployment)		measured Δ	
direction family	readings	breaks	flags	recall	flags	recall
Euclidean	675	41	30	0.634	43	0.805
Fisher path (3 α)	2025	106	89	0.736	113	0.849
joint, $\lambda = 10^{-2}$	675	11	0	0.000	12	0.909
joint, $\lambda = 10^{-1}$	675	10	0	0.000	11	0.900
joint, $\lambda = 1$	675	12	0	0.000	13	0.833

The headline precision and recall are robust to what supplies the breaks: dropping Qwen2.5-1.5B leaves 21 breaks at precision 0.917 and recall 0.524, dropping Qwen2.5-7B 165 at 0.870/0.570, dropping the weakest probe 152 at 0.891/0.539, and dropping the single block that supplies 91 of them 89 at 0.860/0.483. What is not robust is the direction family. The failure is a Goodhart effect, confined to the family that optimises against the screen. On Euclidean and Fisher directions the pre-deployment rule has recall 0.634 and 0.736 at precision 0.867 and 0.876 (the measured rule: 0.805 and 0.849 at 0.767 and 0.796, flagging roughly twice as many readings). On the joint budgets it flags nothing at all: their predicted displacement is small by construction, while the measured one exceeds it by 1.14–2.18×. Two things keep this from undermining the recommendation. The joint budget's own break rate is 3.34× lower than the rest of the matrix (2% against 5%), so the events the screen misses are correspondingly rarer; and the measured-displacement rule recovers most of them (recall 0.909 at $\lambda = 10^{-2}$ and 0.900 at 10^{-1}, 0.905 over the 21 breaks at $\lambda \leq 10^{-1}$; at $\lambda = 1$, where the under-prediction is largest, 0.833). The practical rule we draw from this is in §5.1: screen an unoptimised direction before deploying it, and verify an optimised one with a forward pass.

I OUT-OF-SAMPLE MONITOR METRIC

Half-split. Each monitor's 140 held-out probe inputs are split by parity. $\boldsymbol{N}_A$ is the mean over the five monitors of the second moment of their covectors on the even inputs, divided by the clean score variance on those inputs; the joint budget is solved against $\boldsymbol{N}_A$, and every monitor's displacement is then read separately on the even half (in sample for the metric) and on the odd half (inputs the metric never saw). Euclidean steering and the Fisher direction are read on the same halves, so each ratio is formed within one half and the A-versus-B gap isolates the optimism of the pricing rather than any difference between the halves. The random-probe prior is likewise built from even-half texts only. This is 756 cells over all 9 blocks; pairs are restricted as in Table 7 (KL-reliable strengths, non-trivial effect, Euclidean displacement at least 0.05σ on both halves), leaving 219 monitor readings per row.

Table 15: Pricing against a monitor metric built from inputs the bill is not measured on. Each entry is the measured $B_{\text{mon}}^{1/2}$ relative to Euclidean steering at matched effect, on the half of the probe's inputs that built $\boldsymbol{N}_A$ (in sample) and on the disjoint half (out of sample). "cost" is the ratio of the two. $\boldsymbol{N}_A$ uses 70 inputs per monitor, half of what Table 7 uses.

direction	λ	pairs	in sample (A)	out of sample (B)	95% interval	cost
Fisher $\alpha = 10^{-4}$	—	219	0.801	0.785	—	0.98×
joint, $\boldsymbol{N}_A$	10^{-2}	219	0.245	0.272	$[0.204, 0.360]$	1.11×
joint, $\boldsymbol{N}_A$	10^{-1}	219	0.156	0.208	$[0.173, 0.291]$	1.34×
joint, $\boldsymbol{N}_A$	1	219	0.121	0.190	$[0.151, 0.277]$	1.57×
joint, random prior	10^{-2}	219	0.539	0.528	$[0.418, 0.581]$	0.98×
joint, random prior	10^{-1}	219	0.413	0.425	$[0.329, 0.497]$	1.03×
joint, random prior	1	219	0.343	0.358	$[0.280, 0.476]$	1.05×

The Fisher direction and the random-probe prior are the controls: neither is fitted to a monitor's inputs, and neither pays a cost. Read on the out-of-sample half, the joint budget still dominates Euclidean steering on both axes in 96% of pairs at $\lambda = 10^{-2}$ and 95% at $\lambda = 1$. The joint budget pays 1.11× at $\lambda = 10^{-2}$ and 1.57× at $\lambda = 1$, the same direction of effect as the optimism of Appendix M (a ratio of quadratic forms, so these ratios of $B_{\text{mon}}^{1/2}$ must be squared before comparing), and at $\lambda = 10^{-2}$ far too small to explain the factor of two, on the same $B_{\text{mon}}^{1/2}$ scale, between pricing the deployed monitor and pricing a generic prior (Table 13).

J WHAT BOUNDS THE NUMBERS

Six choices set the resolution and the reach of everything reported here.

The output bill is measured at one position. D_{out} is a next-token KL at the final position of each generic passage, while $\boldsymbol{\delta}$ is added at every position. A last-position KL understates the output cost of an all-position intervention, so every exchange rate we quote is in an understated output currency; the monitor bill, by contrast, is read on the probe's actual readout. Re-measuring every primary cell's output bill as the KL averaged over all positions changes the ranking little (Spearman 0.992 with the last-position bill, median ratio 1.04) but does cost the joint budget: at $\lambda = 10^{-2}$ it dominates Euclidean steering on both axes in 79% of pairs instead of 97%, as expected of a direction priced by a $\boldsymbol{G}$ estimated at the final position only, while its win over the output-optimal direction (a lower monitor bill at KL within 25%) goes from 92% to 96%.

λ is nominal, not normalised across blocks. The Fisher path's α is in units of the mean eigenvalue of $\boldsymbol{G}$, but the joint budget's λ is not rescaled, and $\operatorname{Tr} \boldsymbol{G}$ spans two orders of magnitude across the nine blocks (Table 1). A median at fixed λ therefore pools different points on each block's frontier. The per-strength table (Table 9) and the $\ell = 14$-only figures are the ones least affected. Normalising per block, $\lambda = c \operatorname{Tr} \boldsymbol{G} / \operatorname{Tr} \boldsymbol{N}$ ($\operatorname{Tr} \boldsymbol{G} / \operatorname{Tr} \boldsymbol{N}$ spans 0.013–10.2 across the blocks), the joint budget dominates Euclidean steering in 94%, 97% and 97% of pairs at $c = 10^{-2}, 10^{-1}, 1$, at 0.413, 0.256 and 0.171× the output-optimal direction's monitor bill.

Operating points are read on 140 *inputs per probe.* Every true- and false-positive rate is estimated from about 70 items per class, and the frozen threshold is the 95th percentile of 70 clean negative scores. That sets the resolution of every rate here, and it is why the paper reports break *rates* over thousands of readings rather than individual rate changes. Appendix K shows the two break conditions are qualitatively unchanged when the break criterion moves between 0.1 and 0.3.

The matched-effect constraint is first order. $\boldsymbol{q}^\top \boldsymbol{\delta} = \rho$ equalises the first-order on-target effect; a direction could in principle buy a lower monitor bill through second-order on-target effects the constraint does not see. The measured effects agree to 0.996–0.999× of Euclidean steering's, and to 0.995× on disjoint items (Appendix Q), which bounds this but does not exclude it, and the CAA vectors are outside the family entirely because they cannot be matched without destroying the model.

$\boldsymbol{N}$ comes from the probes' own inputs. It survives a half-split of those inputs (Appendix I) and transfers to monitors it never saw (Appendix G), but the text distribution is the probes' own throughout.

Intervals rest on few clusters. Every interval is a percentile bootstrap over clusters: the 27 model–depth–behaviour cells in §5.1 and the nine model–depth blocks elsewhere. With nine clusters such intervals under-cover, and the clusters are imbalanced: 91 of the 180 matched-strength breaks come from one block. The intervals are therefore descriptive; the leave-one-out checks of Appendix H, which drop the dominant model, the dominant block and the weakest probe in turn, are the stronger evidence that no single cluster drives the headline detector.

K Robustness of the two break conditions

The break criterion. Every break in this paper is an operating-point move of at least $c = 0.2$. Table 16 repeats the two break conditions at $c \in \{0.1, 0.2, 0.3\}$ on the same readings. The false-alarm rate well past Δ_F stays between 71% and 100% across all three criteria and all five margin bins, and the blinding rate collapses with the margin under all three. The conditions are statements about where each break point lies, not about the particular height of the step.

side	criterion	$d' \in [1, 2.5)$	$[2.5, 4)$	$[4, 7)$	$[7, 12)$	$d' \geq 12$
false-alarmed	$c = 0.1$	100%	100%	98%	90%	100%
	$c = 0.2$	100%	100%	96%	79%	99%
	$c = 0.3$	99%	99%	94%	71%	94%
blinded	$c = 0.1$	100%	92%	37%	6%	8%
	$c = 0.2$	99%	85%	31%	6%	7%
	$c = 0.3$	76%	78%	27%	5%	7%

Table 16: The two break conditions at three break criteria. Upper block: false-alarm rate among readings with $\Delta_0/\sigma_w \geq 1.5$. Lower block: blinding rate among readings with $|\Delta_1|/\sigma_w \geq 1$. Columns are the per-block margin bins of Table 19.

Probes trained on the full split. Grading the margin by the size of the training prefix also grades the estimation noise in $\boldsymbol{w}$, so within a block a low-d' probe is also a noisily-estimated one. Restricting to the 90 probes trained on the full 260 examples, where the margin varies with the dataset and the read depth alone, leaves $d' = 1.64$–16.2 and 7,560 readings, and both conditions survive: the false-alarm rate well past Δ_F is 69%–100% across the five bins while the blinding rate falls from 99% to 6% (Table 17). The training-size grading widens the margin range; it does not produce the asymmetry.

side	displacement	$d' \in [1, 2.5)$	$[2.5, 4)$	$[4, 7)$	$[7, 12)$	$d' \geq 12$
false-alarmed	$\Delta_0/\sigma_w \geq 1.5$	100%	100%	99%	69%	99%
	readings	81	156	822	93	97
blinded	$\lvert\Delta_1\rvert/\sigma_w \geq 1$	99%	95%	29%	11%	6%
	readings	103	62	382	122	440

Table 17: The two break conditions restricted to probes trained on the full 260-example split, where margin varies with the dataset and the read depth but not with the amount of training data.

L Which part of the monitor metric does the work?

The metric splits exactly, $\boldsymbol{N}_j = (\bar{\boldsymbol{u}}_j \bar{\boldsymbol{u}}_j^\top + \Sigma_j)/\sigma_j^2$ with $\bar{\boldsymbol{u}}_j = \mathbb{E}_x[\boldsymbol{u}_j]$ and $\Sigma_j = \mathrm{Cov}_x(\boldsymbol{u}_j)$: a rank-one part per monitor (rank at most five in the mean the joint budget prices) that moves every input's score together (the part a ranking statistic cannot see, §2) and a dispersion part that reorders inputs, which carries the differential covector. Which one is worth pricing? We re-solve the joint budget against each part alone, and against $\mathrm{diag}(\boldsymbol{N})$, a control that keeps every coordinate's scale but discards the geometry (1,215 cells over 9 blocks, evaluated on the index set of Table 7).

Table 18: Measured $B^{1/2}_{\mathrm{mon}}$ relative to Euclidean steering at matched effect, when the joint budget is priced against a part of $\boldsymbol{N}$ rather than all of it (medians over the same 223 readings as Table 13; lower is better). The lower block repeats the two transfer priors of Appendix G for comparison.

metric priced against	$\lambda = 10^{-2}$	$\lambda = 10^{-1}$	$\lambda = 1$
full $\boldsymbol{N}$ (Table 7)	0.256	0.165	0.116
common-mode part $\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top/\sigma^2$	0.436	0.416	0.413
dispersion part Σ/σ^2	0.516	0.359	0.289
diagonal of $\boldsymbol{N}$ (scales only)	0.785	0.761	0.727
four sibling monitors	0.579	0.497	0.460
random probes	0.542	0.404	0.403

Three things follow. First, the parts are complementary: neither alone reaches the full metric (0.436 and 0.516 against 0.256 at $\lambda = 10^{-2}$), even though the common-mode part carries most of the trace (75% of $\mathrm{Tr}\,\boldsymbol{N}$ in the median block, 32%–94% across blocks). Second, the geometry is what matters, not the scales: the diagonal control leaves the bill at 0.785, about what the output-optimal direction achieves (0.806). Third, this locates the monitor-specific information that Appendix G could not isolate. Pricing against the common-mode part $\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top$ of the deployed probes, a single direction per monitor and the one AUROC is blind to, already beats both transfer priors (0.436 against 0.579 and 0.542). What a generic prior at the monitor layer cannot supply is each probe's own mean covector.

M THE POOLED OUTPUT-OPTIMAL DIRECTION'S IN-SAMPLE SAVING IS MOSTLY OPTIMISM

A fixed steering vector applied to every input must be priced under the *pooled* metric $\boldsymbol{G} = \mathbb{E}_x[\boldsymbol{J}^\top \boldsymbol{H} \boldsymbol{J}]$, which is estimated (the method of Wang et al., 2026 re-solves per input at the current activation, a different estimator whose optimism we do not measure). Minimising an estimated quadratic form selects the coordinates where the estimate is smallest, which are the noisiest (Efron, 2004); Figure 4 and Table 11 measure the consequence with snapshots of $\boldsymbol{G}_n$ and an independent $\boldsymbol{G}_B$ from disjoint text. At $n = 512$ the optimism of $\boldsymbol{\delta}_{\boldsymbol{G}}$ at $\alpha = 10^{-4}$ (its bill under $\boldsymbol{G}_B$ over its bill under the $\boldsymbol{G}_n$ that chose it) is 2.99 on Llama-3.2-3B, 3.18 on Qwen2.5-1.5B and 7.10 on Qwen2.5-7B. At the $n = 160$ of our secondary depths it is 6.10–24.1, so $\boldsymbol{\delta}_{\boldsymbol{G}}$ is a noisier comparison benchmark at $\ell \in \{7, 21\}$ than at $\ell = 14$; Table 7 pools all nine blocks, and the $\ell = 14$-only figures of §5.5 are the ones least affected. It is driven by the spectrum of $\boldsymbol{G}$, not the ambient dimension (Appendix F). Its realisable saving is about $2\times$, which the joint budget also delivers at a quarter of Euclidean steering's monitor bill.

N GRADED MARGINS, MONITOR DEPTHS, DEPLOYED NORMS

A probe's margin is mostly decided by how much labelled data it was trained on, so for each of the 5 monitor datasets we train probes on class-balanced prefixes of the training split of size 12–260, with the *same* held-out inputs, at every depth in $\{\ell + 3, \ell + 7, \ell + 11\}$ that the model has and at the final layer; each probe is read at the depth it was trained on. All probes of a dataset share one text batch and one covector pass, so the marginal cost over the primary matrix is a few suffix runs per cell: 666 probes over 9 blocks, 756 cells, 41,516 matched readings. A probe with $d' < 1$ is not a monitor and is excluded (the primary matrix keeps its five fixed probes, including the one claims block at $d' = 0.991$).

Table 19: The two break conditions at matched per-class displacement. Each side is conditioned on the displacement its own condition uses, in the probe's measured within-class spread: blinding on $|\Delta_1|$ (the positive class moving down), false alarms on Δ_0 (the negative class moving up). Entries are the fraction of readings broken on that side. The last row gives the fraction of all readings in which Δ_0 reaches Δ_F at all.

side	displacement	readings	$d' \in [1, 2.5)$	$[2.5, 4)$	$[4, 7)$	$[7, 12)$	$d' \geq 12$
blinded	$\lvert\Delta_1\rvert/\sigma_w$ $[0.5, 1)$	3288	74%	21%	1%	0%	0%
	$\lvert\Delta_1\rvert/\sigma_w$ $[1, 2)$	2654	97%	70%	3%	0%	0%
	$\lvert\Delta_1\rvert/\sigma_w \geq 2$	3446	100%	100%	59%	9%	11%
false-alarmed	Δ_0/σ_w $[0.5, \Delta_F)$	4429	7%	17%	19%	3%	15%
	Δ_0/σ_w $[\Delta_F, 1.5)$	2798	75%	71%	60%	24%	52%
	$\Delta_0/\sigma_w \geq 1.5$	7065	100%	100%	96%	79%	99%
matched families	$\Delta_0/\sigma_w \geq 1.5$	5442	100%	100%	97%	80%	99%
deployed norms	$\Delta_0/\sigma_w \geq 1.5$	1623	99%	100%	95%	77%	100%
reached	frac. of readings with $\Delta_0 \geq \Delta_F$		13.5%	28.2%	23.1%	11.8%	7.7%

Table 20: The covector prediction against monitor depth, over the same readings.

monitor depth	readings	Spearman	median measured/predicted
$\ell_j = 10$	2307	0.982	1.013
$\ell_j = 14$	3915	0.981	1.042
$\ell_j = 17$	4388	0.986	1.016
$\ell_j = 18$	4344	0.985	1.009
$\ell_j = 21$	4575	0.990	1.017
$\ell_j = 24$	4341	0.971	1.063
$\ell_j = 25$	4388	0.990	1.006
$\ell_j = 28$	13258	0.985	1.007

Table 21: Difference-of-means CAA vectors swept over relative norm, unmatched, on all 9 blocks and three behaviours. $|\Delta\psi|$, KL and the ARC-Easy drop are medians over cells; the last three columns count monitor readings blinded, false-alarmed, and false-alarmed on a probe with $d' \geq 8$.

$\|\boldsymbol{\delta}\|/\mathbb{E}\|\boldsymbol{h}_\ell\|$	readings	$\lvert\Delta\psi\rvert$	KL	ARC drop	blinded	false-alarmed	of which $d' \geq 8$
0.02	1998	0.03	3.5e-04	0.000	4	3	0
0.04	1998	0.06	1.4e-03	0.000	9	16	0
0.08	1998	0.11	5.3e-03	0.016	19	54	3
0.16	1998	0.24	2.1e-02	0.016	40	255	19
0.24	1998	0.44	4.8e-02	0.016	62	485	41
0.32	1998	0.51	8.4e-02	0.016	85	654	83
0.4	1998	0.51	1.3e-01	0.016	118	773	108

Equation (7) is tested directly here, with the two class means and the measured within-class spread and no fitted constant: the mean absolute error of the predicted operating point is 0.018 (true-positive rate) and 0.033 (false-positive rate) over 41,516 readings. The two class displacements are correlated but not equal (Spearman 0.440, median gap 0.100 and 90th percentile 2.59 in σ_w), which is why Corollary 1 states the screen with them separately.

O SCALE CHECK, FURTHER BEHAVIOURS AND UNPRICED MONITORS

Scale check. The primary matrix and the graded-margin and deployed-norm sweep of §5.3–5.4 are repeated on three further instruction-tuned models: Llama-3.2-1B (16 blocks, $d = 2048$), Llama-3.1-8B (32 blocks, $d = 4096$) and Qwen2.5-14B (48 blocks, $d = 5120$). The intervention depths are $\ell \in \{L/4, L/2, 3L/4\}$ and the primary monitors are read at $\ell_j = \ell + L/4$, the same fractions as $\{7, 14, 21\}$ and $\ell + 7$ on the 28-block models; the graded-margin read depths are the primary offsets

scaled by $L/28$. $\boldsymbol{G}$ uses 512 passages at the middle depth and 160 at the others, at rank 384, in `float16`. Behaviours, probe datasets and sizes, direction families, strengths and the break criterion are unchanged, and the break conditions and the deployed-norm rule are computed by the function that produces the primary numbers, with each model's three blocks as fixed effects.

Table 22: The primary matrix per model: non-trivial readings, breaks at the frozen threshold, the covector prediction (Spearman, predicted against measured RMS displacement, all readings), and the share of matched pairs in which the joint budget at $\lambda = 10^{-2}$ is below Euclidean steering on both axes (number of pairs). The first three rows are the primary models.

model	blocks	readings	breaks	blinded	false-alarmed	per 1,000	Spearman (pred.)	joint dominates
Qwen2.5-1.5B	3	1,575	159	59	100	101.0	0.957	95% (114)
Llama-3.2-3B	3	1,575	6	0	6	3.8	0.985	99% (67)
Qwen2.5-7B	3	1,575	15	9	6	9.5	0.938	100% (42)
Llama-3.2-1B	3	1,575	71	52	19	45.1	0.970	99% (98)
Llama-3.1-8B	3	1,575	1	1	0	0.6	0.981	100% (50)
Qwen2.5-14B	3	1,575	0	0	0	0.0	0.931	71% (21)

The prediction is as good as on the primary models (Spearman 0.931–0.981, against 0.938–0.985; Table 22). Breaks at matched strengths become rare with scale: 1 on Llama-3.1-8B and 0 on Qwen2.5-14B of 1,575 non-trivial readings each, against 71 on Llama-3.2-1B and 159 on Qwen2.5-1.5B. The joint budget dominates Euclidean steering in 99%, 100% and 71% of pairs, the last on only 21 pairs, because on Qwen2.5-14B few monitors move by the 0.05σ a pair requires at matched strengths.

Table 23: The two break conditions on the scale check. False alarms at $\Delta_0/\sigma_w \geq 1.5$ and blinding at $|\Delta_1|/\sigma_w \geq 1$, as the fraction of readings broken in each margin bin ("—": fewer than 20 readings), and the logistic coefficient on $\log d'$ with block fixed effects and the displacement controlled. The first two row groups pool nine blocks each (the primary models' as in §5.3, and the scale check's) and carry a 95% interval from a bootstrap over blocks; a single model has three blocks, too few for that bootstrap, so its rows give the point estimate.

model	side	$d' \in [1, 2.5)$	$[2.5, 4)$	$[4, 7)$	$[7, 12)$	$d' \geq 12$	coef. on $\log d'$
primary models, pooled	false-alarmed	100%	100%	96%	79%	99%	−0.095 [−0.629, 0.301]
	blinded	99%	85%	31%	6%	7%	−1.25 [−2.26, −0.610]
scale check, pooled	false-alarmed	96%	96%	87%	84%	97%	0.350 [−0.100, 0.758]
	blinded	93%	88%	14%	3%	0%	−1.60 [−2.19, −1.20]
Llama-3.2-1B	false-alarmed	96%	95%	37%	55%	—	−0.261
	blinded	92%	87%	12%	9%	—	−1.30
Llama-3.1-8B	false-alarmed	95%	100%	91%	85%	76%	−0.022
	blinded	—	—	11%	1%	1%	−2.26
Qwen2.5-14B	false-alarmed	100%	97%	95%	92%	100%	0.838
	blinded	98%	92%	21%	0%	0%	−2.02

Pooled, both break conditions reproduce (Table 23): false alarms well past Δ_F at 84%–97% in every bin, blinding collapsing with the margin, and a coefficient on $\log d'$ whose interval excludes zero for blinding and contains it for false alarms. Model by model the blinding coefficient is negative everywhere (−1.30, −2.26 and −2.02), while the false-alarm coefficient is −0.261, −0.022 and 0.838: at matched displacement the margin gives no protection against false alarms on Llama-3.1-8B and Qwen2.5-14B, and some on Llama-3.2-1B, a fifth of what it gives there against blinding. The uncontrolled rates show the same: on Llama-3.2-1B the false-alarm rate well past Δ_F falls to 37% and 55% in its two widest populated bins (38 and 87 readings), and on Llama-3.1-8B it drifts from 95% to 76% across the bins.

Table 24: The per-cell deployed-norm rule of §5.4 on the scale check: for each (model, depth, behaviour) cell, the largest swept CAA norm (in $\mathbb{E}\|\boldsymbol{h}_\ell\|$; median shown) at which that cell's generic KL and ARC-Easy drop are both at most 0.05. Readings false-alarmed, false-alarmed among probes with $d' \geq 8$, and blinded there; the harmfulness probe's largest false-positive rise there; the smallest norm at which it crosses 0.2, with that cell's KL as a multiple of the 0.05 bar; and the screen's precision and recall on the breaks. No harmfulness-probe reading is blinded at the selected norms on any scale-check model (0 of the pooled readings).

model	norm (med.)	readings	false-alarmed	on $d' \geq 8$	blinded	harm. rise	harm. crosses	screen P / R
primary models, pooled	0.16	1,998	360 (18%)	22 / 618	46	0.171	0.24 (1.5×)	0.88 / 0.66
scale check, pooled	0.16	1,953	305 (16%)	157 / 855	66	0.129	0.32 (3.1×)	0.88 / 0.70
Llama-3.2-1B	0.16	537	40 (7%)	0 / 72	48	0.043	0.40 (2.3×)	0.76 / 0.70
Llama-3.1-8B	0.24	693	165 (24%)	82 / 300	4	0.129	none ≤ 0.40	0.95 / 0.62
Qwen2.5-14B	0.16	723	100 (14%)	75 / 483	14	0.114	0.32 (3.1×)	0.91 / 0.81

At deployed norms the scale check false-alarms about as often as the primary models (16% of readings against 18%) but far more often on wide-margin probes: 157 of 855 readings with $d' \geq 8$, against 22 of 618, and 82 of 300 on Llama-3.1-8B alone (Table 24). The harmfulness probe is neither blinded nor made to fire at any in-budget norm; it crosses only outside the budget, at $0.320\,\mathbb{E}\|\boldsymbol{h}_\ell\|$ on Qwen2.5-14B and $0.400\,\mathbb{E}\|\boldsymbol{h}_\ell\|$ on Llama-3.2-1B, and not at all on Llama-3.1-8B up to $0.400\,\mathbb{E}\|\boldsymbol{h}_\ell\|$.

Four further behaviours. The primary matrix is repeated on the nine primary blocks for the four remaining behaviours of the CAA suite (Panickssery et al., 2024): hallucination, myopic reward, survival instinct and coordination with other AIs (1,620 cells, 8,100 readings), with $\boldsymbol{G}$, $\boldsymbol{N}$ and the monitors unchanged and $\boldsymbol{q}$ computed as in §4. The prediction (Spearman 0.971), the break detector (precision 89%, recall 76% on 164 breaks in 6,300 non-trivial readings) and the joint budget (dominating Euclidean steering in 93% of pairs, at 0.332× the output-optimal direction's monitor bill) behave as on the three primary behaviours (Table 25).

Unpriced monitors. Six monitors that no direction was priced against are read on every primary cell. One is a linear deception probe in the style of Zou et al. (2023a): each input is a true statement voiced after an instruction to act as an honest or as a dishonest person, the label is the instruction, and the probe is trained like the primary probes on 260 inputs and evaluated on 140. Its clean AUROC of 1.00 says it can read the instruction; it is a monitor, not a validated deception detector. The others are one-hidden-layer MLP probes (128 ReLU units), one per primary dataset, with the primary split and threshold convention. For an MLP probe the covector of an input is the vector–Jacobian product with the probe's input gradient at that input's clean activation as cotangent, i.e. its local linearisation. The prediction holds for both kinds (Spearman 0.996 and 0.995). The joint budget, priced only against the five linear probes, dominates Euclidean steering in 82% and 96% of pairs. Against the output-optimal direction it lowers the MLP probes' bill to 0.542× but leaves the deception probe's unchanged (0.990×), the uneven transfer of Appendix G.

Table 25: Extensions on the nine primary blocks, against the primary matrix: the covector prediction on all readings, the break detector of §5.1 on non-trivial readings (not computed for the unpriced monitors), and the joint budget at $\lambda = 10^{-2}$ against Euclidean steering (share of matched pairs dominated on both axes) and against the output-optimal direction (median monitor-bill ratio).

readings	n	Spearman (pred.)	detector P / R	joint dominates	joint/Fisher bill
primary matrix (5 probes, 3 behaviours)	6,075	0.972	0.87 / 0.58	97%	0.37×
four further behaviours	8,100	0.971	0.89 / 0.76	93%	0.33×
deception probe (unpriced)	1,215	0.996	—	82%	0.99×
MLP probes (unpriced)	6,075	0.995	—	96%	0.54×

P NEGATIVE RESULTS

Benign steering does not blind a well-separated probe. We steered Qwen2.5-7B and Qwen2.5-1.5B at $\ell = 14$ with the corrigibility vector at ρ up to 8 under all three direction families and asked whether any probe degrades. On Qwen2.5-7B the harmfulness probe holds AUROC 1.00 and true-positive rate 1.00 at every strength (maximum displacement 0.0064σ); the cities probe is identical

(0.0088σ); the CommonClaim probe moves from AUROC 0.913 to 0.911 at worst. On Qwen2.5-1.5B the CommonClaim probe is the only one that moves: AUROC $0.838 \to 0.812$, true-positive rate to 0.319, at a displacement of 0.438σ, below the blinding displacement $\Delta_T(d')$ that Proposition 4 gives for that block's margin, and the drop stays below 0.2. The threat model in which a benign vector silently blinds a strong safety probe is not supported at these strengths.

Free-generation judging is uninformative at this scale. A generation-based check on Qwen2.5-7B (48 harmful requests, four steering configurations, an LLM judge whose model and prompt are given in `code/scripts/21_safety_demo.py` of the supplement) returned refusal rates of 1.00 for the clean, Fisher and joint configurations and 0.98 for Euclidean steering, with no harmful content judged in any of the 188 completions the judge returned a label for (of 192 generated; 4 calls failed) and the harmfulness probe flagging every request in every configuration. A single complying completion supports no conclusion either way; the paper's safety argument rests on the measured trigger condition and margin dependence of §5.3 rather than on a demonstration. A judged free-generation check of the steered behaviours on the pilot model was likewise uninformative: the 0.5B model does not produce opinionated free-form answers on the contrastive prompts, so the same judge cannot resolve differences between configurations that the forced-choice logit gap resolves clearly.

Q Re-calibration, simpler baselines, disjoint items and precision

The checks below re-measure the nine primary blocks after rebuilding every prep artefact ($\boldsymbol{G}$, monitors, covectors, $\boldsymbol{q}$) from the same data and seeds. The rebuild reproduces the primary matrix: 4,725 non-trivial readings with 179 breaks (4,725 and 180 in §5.1), and at the per-cell deployed norms 356 of 1,998 readings false-alarm (360 of 1,998 in §5.4).

Re-calibrating the threshold. A deployer who knows that a vector is on can move the threshold instead of leaving it frozen. Table 26 applies four policies to the same steered scores: re-estimating τ as the 95th percentile of the *steered* negative scores, which costs one steered forward pass over the calibration inputs; shifting τ by the covector's predicted displacement of the negative class, $\bar{\boldsymbol{u}}_{|0}^{\top}\boldsymbol{\delta}$, or of all inputs, $\bar{\boldsymbol{u}}^{\top}\boldsymbol{\delta}$, which costs none; and shifting it by the Jacobian-free $\boldsymbol{w}^{\top}\boldsymbol{\delta}$. Re-estimation removes 78% of the matched breaks and every false alarm at the deployed norms, but it is not free and it is not neutral. It trades false alarms for misses: a threshold raised to restore the false-positive rate also cuts into a positive class whose separation has shrunk, so at the deployed norms the blinded readings rise from 47 to 59. The 40 matched breaks that survive, all on Qwen2.5-1.5B, are losses of discrimination that no threshold can repair. Estimated on one half of the calibration inputs and read on the other, re-estimation leaves 73 and 65 breaks, since 35 negatives per half fix a 95th percentile only coarsely. The covector's shift removes 31% of the matched breaks and 53% at the deployed norms without any forward pass; it stops short because the measured displacement of the negative class exceeds the first-order one (§5.1). Re-calibration therefore complements the screen where the pipeline knows when the vector is on and can afford a steered pass over its calibration set; conditional or per-request steering, where the monitor does not know, leaves the frozen-threshold analysis as the operative one.

Table 26: Breaks under five threshold policies, on the same steered scores. Left: the non-trivial readings of the primary matrix. Right: the graded-margin probes at the per-cell deployed norm of §5.4. "broke" counts readings whose operating point moved by at least 0.2 relative to the clean scores at the frozen threshold; a reading can break on both sides. The last row re-estimates τ on the even-indexed inputs and reads the odd-indexed ones.

	matched directions (4,725 non-trivial)			deployed CAA norms (1,998)		
threshold after steering	broke	blinded	false-alarmed	broke	blinded	false-alarmed
frozen τ	179	69	111	403	47	356
re-estimated on steered negatives	40	40	0	59	59	0
$\tau + \bar{\boldsymbol{u}}_{\|0}^{\top}\boldsymbol{\delta}$ (covector, no forward pass)	123	27	97	191	37	154
$\tau + \bar{\boldsymbol{u}}^{\top}\boldsymbol{\delta}$ (covector, pooled)	107	25	83	185	28	157
$\tau + \boldsymbol{w}^{\top}\boldsymbol{\delta}$ (Jacobian-free)	222	89	133	—	—	—
re-estimated, held-out half	73			65		

The Jacobian is needed. Replacing each covector by the monitor's own weight ($\boldsymbol{K} = \boldsymbol{I}$, so every input's predicted shift is $\boldsymbol{w}^\top \boldsymbol{\delta}$) is the zero-cost alternative to one VJP per input. It ranks the measured RMS displacement with Spearman 0.718 against the covector's 0.965 over the matched readings, and 0.402 against 0.965 within model–depth–strength strata. As a break detector it has precision 0.553 and recall 0.291 against 0.907 and 0.598, and average precision 0.414 against 0.742. Pricing against $\sum_j \boldsymbol{w}_j \boldsymbol{w}_j^\top / \sigma_j^2$ in place of $\boldsymbol{N}$ buys nothing: 0.797× Euclidean steering's bill, the same as the output-optimal direction's 0.788× in this rerun (0.806× in Table 7; Table 27). What an intervention does to a monitor several layers up is not what it does at the monitor's own layer.

Hard projection baselines. Two directions in V_ρ leave the first-order common-mode and differential displacement of all five monitors at zero, $\boldsymbol{U}^\top \boldsymbol{\delta} = 0$ for $\boldsymbol{U} = [\bar{\boldsymbol{u}}_j, \Delta\bar{\boldsymbol{u}}_j]_j$: $\boldsymbol{q}$ projected off the span of $\boldsymbol{U}$ (the minimum-norm member, in the spirit of nullspace projection; Ravfogel et al., 2020; Sheng et al., 2026) and the output-optimal member under the same constraint. Both lower the bill, neither as far as the joint budget (Table 27): 0.424× and 0.355× Euclidean's, against 0.258× at nearly the same KL, the first with no KL saving at all (0.971×). On a monitor left out of $\boldsymbol{U}$ they keep 0.849× and 0.650× of Euclidean's bill against the joint budget's 0.570× when that monitor is left out of $\boldsymbol{N}$. Zeroing the first-order mean displacement leaves the dispersion term and the second-order terms, which the joint budget prices and a projection does not.

Table 27: Pricing baselines at KL-reliable strengths over 222 matched pairs of the rebuilt blocks, defined as in Table 7 (223 pairs there): medians of the measured $B_{\text{mon}}^{1/2}$ and D_{out} relative to Euclidean steering, the rate of dominating it on both axes, the bill relative to the output-optimal direction, the on-target effect on the disjoint items relative to Euclidean steering's, and the bill on a monitor left out of the pricing (or of the projection).

direction	$B_{\text{mon}}^{1/2}$ ratio	D_{out} ratio	dominates	vs. output-optimal	effect, disjoint items	unpriced monitor
output-optimal ($\alpha = 10^{-4}$)	0.788	0.422	69%	1.00	0.981	—
Jacobian-free pricing $\sum_j \boldsymbol{w}_j \boldsymbol{w}_j^\top / \sigma_j^2$	0.797	0.431	70%	0.996	0.980	—
$\boldsymbol{q}$ projected off the monitors	0.424	0.971	67%	0.573	0.995	0.849
output-optimal, monitor-null	0.355	0.435	91%	0.531	0.979	0.650
joint budget ($\lambda = 10^{-2}$)	0.258	0.449	96%	0.391	0.980	0.570

The matched effect on disjoint items. In the main matrix $\boldsymbol{q}$ and $\Delta\psi$ come from the same 64 items. On 64 further CAA items disjoint from them and from the CAA vectors, the non-Euclidean directions reach 0.995× Euclidean steering's effect in the median (0.998× on the original items; 0.991–1.00× by family), so the matching is not an in-sample artefact. Every direction's absolute effect on the new items is 0.707× its in-sample effect in the median, with the sign kept in 97% of cells: part of each effect is specific to the 64 items, equally for all families. $\boldsymbol{q}$ itself is stable for refusal and corrigibility (cosine 0.632–0.971 with $\boldsymbol{q}$ rebuilt from the disjoint items) but not for sycophancy on the smaller models (down to 0.018; Table 28), whose forced-choice gradient over 64 items is dominated by item-specific directions. The median $|\cos|$ between $\boldsymbol{q}$ and the CAA vector, 0.020, is at the level of two random directions in these dimensions.

Where the covectors live. The pilot of Appendix D measured one small model. On all nine primary blocks, the subspace of $\boldsymbol{G}$ that holds 1% of its trace (0.144–0.489 of the dimensions) carries 0.386–0.716 times a random direction's squared mass for the per-input covectors and 0.324–0.649 for $\boldsymbol{q}$, while the monitors' own weights sit at 0.812–1.04 (Table 28). The pilot's finding holds at scale: monitors read the model's computationally active subspace, and the reason output-frugality does not buy monitor-frugality is the norm price, not kernel occupancy.

Table 28: Left: fraction of squared mass in the subspace of $\boldsymbol{G}$ holding 1% of its trace, for a random direction (in expectation, the subspace's share of dimensions), the per-input covectors and their mean, and $\boldsymbol{q}$ (averages over monitors and behaviours). Right: cosine between $\boldsymbol{q}$ and $\boldsymbol{q}$ rebuilt from 64 disjoint items.

		mass in the 1%-of-trace subspace of $\boldsymbol{G}$				$\cos(\boldsymbol{q}, \boldsymbol{q}_{\text{disjoint}})$		
model	ℓ	random	$\boldsymbol{u}(x)$	$\bar{\boldsymbol{u}}$	$\boldsymbol{q}$	sycophancy	refusal	corrigible
Qwen2.5-7B	7	0.488	0.349	0.354	0.317	0.540	0.848	0.786
Qwen2.5-7B	14	0.228	0.122	0.124	0.091	0.707	0.878	0.960
Qwen2.5-7B	21	0.282	0.174	0.173	0.164	0.858	0.947	0.632
Llama-3.2-3B	7	0.386	0.226	0.219	0.198	0.755	0.894	0.878
Llama-3.2-3B	14	0.186	0.074	0.071	0.083	0.678	0.836	0.950
Llama-3.2-3B	21	0.153	0.079	0.074	0.094	0.459	0.836	0.945
Qwen2.5-1.5B	7	0.269	0.167	0.167	0.150	0.279	0.810	0.912
Qwen2.5-1.5B	14	0.143	0.062	0.061	0.046	0.345	0.892	0.971
Qwen2.5-1.5B	21	0.153	0.059	0.060	0.098	0.018	0.854	0.888

Precision on all three models. Evaluating the identical matched $\boldsymbol{\delta}$ at $\ell = 14$ with the model in `float32` and in `float16`, the `float16` KL differs by a median factor of 1.12–1.50 at $\rho = 1$ and at most 1.11 at $\rho = 2$. The ratios the paper reports are stable: the joint budget's KL relative to Euclidean steering moves by at most 0.022 between precisions (for example 0.500 and 0.479 on Qwen2.5-7B), relative to the output-optimal direction by at most 0.042, its domination rate at $\ell = 14$ is 100% in both, and the monitor bill agrees with Spearman at least 0.995.